

Anom3W: A Unified Benchmark for Video Anomaly Grounding and Understanding

Liang Liang^{1,2}, Yun Zhu², Juekai Lin¹, Haoyang Wang¹, Jinhui Li¹,
Weilin Zhou³, Yuehan Zhu⁴, Lijun Wu², Wenqiao Zhang^{1*}, Jun Xiao¹

¹Zhejiang University ²Shanghai Artificial Intelligence Laboratory

³Xinjiang University ⁴Sun Yat-sen University

liangxliang@zju.edu.cn, zhuyun_dcd@zju.edu.cn, linjuekai@zju.edu.cn

22651112@zju.edu.cn, 22651289@zju.edu.cn, 20232501631@stu.xju.edu.cn

zhuyh68@mail2.sysu.edu.cn, lijun_wu@outlook.com

wenqiaozhang@zju.edu.cn, junx@cs.zju.edu.cn

Abstract

Understanding anomalies in real-world videos requires reasoning about when events occur, where they happen, and what actions unfold. Existing benchmarks, however, largely ignore one or more of these dimensions, making it difficult to evaluate whether multimodal large models (MLLMs) truly understand abnormal events beyond simple detection. To address this gap, we introduce **Anom3W**, a benchmark designed to evaluate MLLMs across **temporal, spatial, and semantic dimensions**. SemEviGround, our semantics-driven evidence grounding framework, consolidates heterogeneous anomaly resources into verified spatio-temporal annotations, enabling rigorous, reasoning-oriented evaluation. By treating spatial boxes and tubes as evidence cues, Anom3W challenges models to reason about entities, locations, and behaviors, rather than stopping at coarse abnormality detection. Experiments demonstrate that even state-of-the-art MLLMs struggle, highlighting the critical importance of spatial grounding and semantic understanding. Anom3W establishes a new standard for interpretable, context-aware video anomaly reasoning.

1 Introduction

Real-world video anomalies, such as traffic accidents, public safety incidents, or suspicious behaviors (Nayak et al., 2021; Liu et al., 2024; Sultani et al., 2018; Wu et al., 2020), rarely occur in isolation. They involve specific entities, take place at particular locations, and unfold through distinct actions. Accurately understanding these events is critical for enhancing public safety, enabling timely interventions, and producing interpretable

insights that support decision-making. Achieving this requires reasoning not only about **when** anomalies occur, but also **where** they happen and **what** actions unfold, which we refer to as **3W** understanding.

Existing benchmarks cover these dimensions in a fragmented way. Conventional video anomaly detection (VAD) benchmarks primarily measure whether and when anomalies occur (Sultani et al., 2018; Wu et al., 2020; Cao et al., 2023). Video anomaly understanding (VAU) benchmarks emphasize what happened, and sometimes why, through textual descriptions or question answering (Yuan et al., 2024; Du et al., 2024, 2026; Zhu et al., 2025). A few anomaly benchmarks provide masks, boxes, or tracks (Accintoae et al., 2022; Yao et al., 2023; Landi et al., 2019), yet these annotations are primarily used as localization targets rather than as evidence for reasoning. Figure 1 illustrates this paradigm evolution, and Table 1 further compares representative benchmarks along the same dimensions. Together, they show that **unified evaluation covering what, where, and when remains largely underexplored**.

To address this gap, we introduce Anom3W, a unified benchmark that evaluates multimodal large models (MLLMs) along three key dimensions: **temporal (when), spatial (where), and semantic (what)**. We develop SemEviGround, a semantics-driven evidence grounding framework that starts from anomaly semantics and turns event descriptions into verified visual evidence. Specifically, semantic cues are used to identify anomaly-relevant subjects and interactions, localize them as boxes, and track them into tubes, yielding aligned *what*, *where*, and *when* annotations at benchmark scale. Under this unified representation, we define a protocol for temporal grounding, spatio-temporal

*Corresponding author.

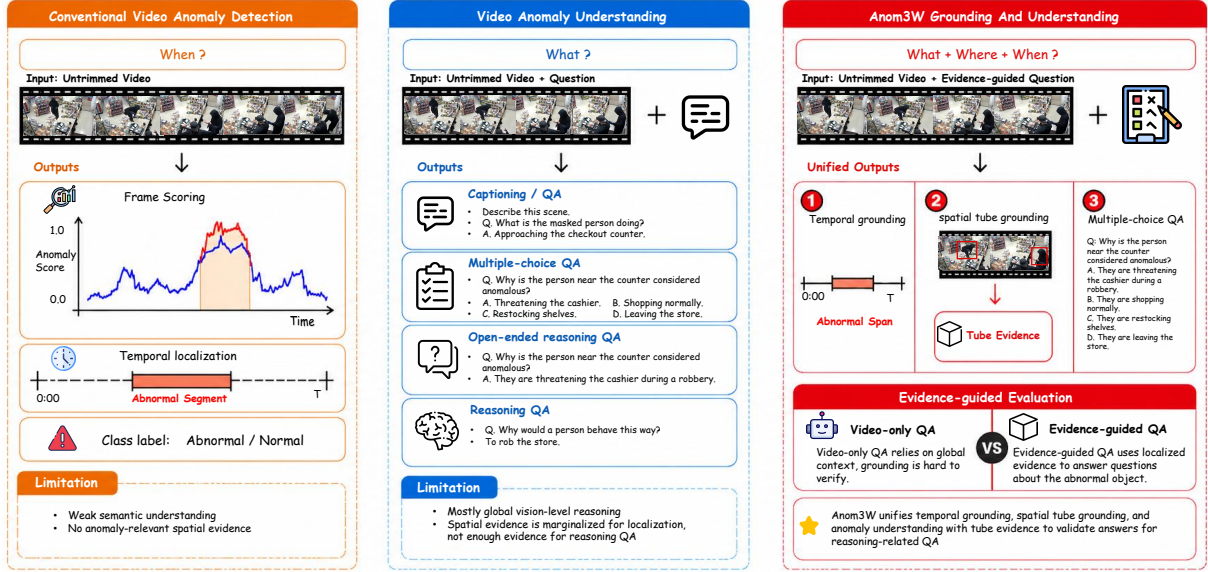


Figure 1: Paradigm evolution of video anomaly benchmarks. Existing anomaly benchmarks typically emphasize **when** or **what** separately, while Anom3W jointly evaluates **what**, **where**, and **when** with anomaly-relevant tube evidence.

grounding, and anomaly question answering. This protocol assesses models not only on detection and description, but also on whether their reasoning is grounded in responsible visual evidence.

By treating spatial boxes and tubes as visual cues, Anom3W challenges models to reason about entities, locations, and behaviors, rather than stopping at coarse abnormality detection. Experiments show that even state-of-the-art MLLMs struggle under this unified 3W evaluation, highlighting the critical importance of spatial grounding and semantic understanding. Anom3W establishes a new standard for interpretable, context-aware video anomaly reasoning.

Our contributions are summarized as follows:

- We introduce Anom3W, a benchmark for video anomaly grounding and understanding across what, where, and when.
- We define a 3W evaluation protocol for anomaly question answering, temporal grounding, and spatio-temporal grounding, using anomaly-relevant boxes and tubes as visual evidence under both video-only and evidence-assisted QA settings.
- We develop SemEviGround, a semantics-driven evidence grounding framework that converts anomaly semantics into verified anomaly-relevant boxes and tubes at benchmark scale.

2 Related Work

2.1 Video Anomaly Benchmarks

Video anomaly benchmarks have evolved from detection-centered evaluation to language-level understanding, as summarized in Table 1. Conventional VAD benchmarks evaluate anomaly scoring, video-level classification, frame-level detection, or temporal localization (Wang and Miao, 2010; Lu et al., 2013; Luo et al., 2017; Ramachandra and Jones, 2020; Sultani et al., 2018; Wu et al., 2020; Cheng et al., 2021; Cao et al., 2023; Zhu et al., 2024; Jin et al., 2022), but provide limited support for grounding responsible subjects or explaining event semantics.

Spatially annotated VAD benchmarks add masks, boxes, or tracks (Acsintoae et al., 2022; Yao et al., 2023; Landi et al., 2019). They make localization measurable, but mainly use spatial labels as detection targets rather than evidence for semantic reasoning.

Recent VAU benchmarks emphasize the *what* dimension through descriptions, causation, conversational QA, reasoning, joint grounding, and fine-grained assessment (Yuan et al., 2024; Du et al., 2024, 2026; Zhang et al., 2025b; Gani et al., 2025; Liu et al., 2025; Zhu et al., 2025; Huang et al., 2025; Gao et al., 2026; Pereira et al., 2026), while spatio-temporal video grounding benchmarks provide mature localization protocols (Chen et al., 2019; Zhang et al., 2020; Tang et al., 2022; Yuan

Table 1: Comparison with representative video anomaly and grounding benchmarks along the **when**, **where**, and **what** dimensions.

Dataset Statistics				Grounding		Understanding	
Benchmark	Domain	#Eval. Videos	#Anom. Cat.	Temporal	Spatial	MCQ	Evidence-grounded QA
<i>Conventional video anomaly detection benchmarks</i>							
UCF-Crime (Sultani et al., 2018)	Crime	290	13	Frame-level	None	✗	✗
XD-Violence (Wu et al., 2020)	Violence	800	6	Frame-level	None	✗	✗
<i>Spatially annotated VAD benchmarks</i>							
UBnormal (Acsintoae et al., 2022)	Synthetic	211	22	Frame-level	Masks	✗	✗
DoTA (Yao et al., 2023)	Traffic	1,402	18	Temporal segments	Boxes	✗	✗
UCFCrime2Local (Landi et al., 2019)	Crime	90	6	Temporal segments	Boxes	✗	✗
<i>Video anomaly understanding benchmarks</i>							
UCA (Yuan et al., 2024)	Crime	310	13	Temporal segments	None	✗	✗
CUVA (Du et al., 2024)	Multiple	1,000	42	Temporal segments	None	✗	✗
ECVA (Du et al., 2026)	Multiple	2,240	100	Temporal segments	None	✗	✗
VAGU (Gao et al., 2026)	Multiple	7,567	21	Temporal segments	None	✓	✗
Anom3W	Multiple	674	81	Temporal segments	Tubes	✓	✓

et al., 2025a,b). Yet anomaly-relevant boxes or tubes are rarely used as visual evidence for QA, leaving *when*, *where*, and *what* loosely coupled. To address this gap, Anom3W uses boxes and tubes not only as localization labels, but also as evidence for grounded anomaly understanding.

2.2 Evolution of Video Anomaly Models

This benchmark gap becomes more important as anomaly models move from abnormality scoring toward language-conditioned and multimodal reasoning. Classical and deep VAD methods rely on normality modeling, reconstruction or prediction, one-class classification, or weakly supervised multiple-instance learning (Nayak et al., 2021; Liu et al., 2024), typically producing scores, labels, or temporal segments.

Recent work extends VAD to open-set, open-vocabulary, and vision-language settings (Cao et al., 2023; Zhu et al., 2024; Wu et al., 2024; Zanella et al., 2024; Xu et al., 2025). Anomaly-oriented models further describe events, answer questions, or organize abnormal activities into structured semantics (Zhang et al., 2025b; Shao et al., 2025; Li et al., 2025; Tang et al., 2024; Wang et al., 2026). Such outputs can be fluent but weakly grounded, e.g., naming an anomaly while attending to the wrong subject or missing the key interaction. Anom3W targets this gap by evaluating whether anomaly understanding is supported by responsible visual evidence.

3 Anom3W Benchmark

3.1 Overview

Anom3W is an evidence-grounded benchmark for video anomaly understanding. Each instance is organized around three core questions: **when** the anomalous event occurs, **where** the anomaly-relevant visual evidence appears, and **what** happens. This 3W design connects temporal anomaly evaluation in VAD with language-level interpretation in VAU, while introducing spatial evidence as an explicit grounding signal for reasoning.

Formally, Anom3W represents each anomalous event with a unified instance schema:

$$x = (V, T, E_{\text{evt}}, E_{\text{ent}}, G, Q), \quad (1)$$

where V denotes the video clip and T denotes the temporal span of the anomalous event. E_{evt} captures the event semantics, E_{ent} specifies the anomaly-relevant subject or region, G contains verified boxes or tubes associated with the responsible visual evidence, and Q consists of anomaly-oriented questions. This schema supports 3W evaluation in an event-conditioned manner. T evaluates **when** the anomaly occurs. E_{ent} and G jointly evaluate **where** the responsible visual evidence appears, since spatial evidence is meaningful only with respect to the anomalous event and its relevant subject or region. E_{evt} together with Q evaluates **what** happens. We retain only instances with a well-defined temporal extent, visible and verifiable spatial evidence when grounding is required,

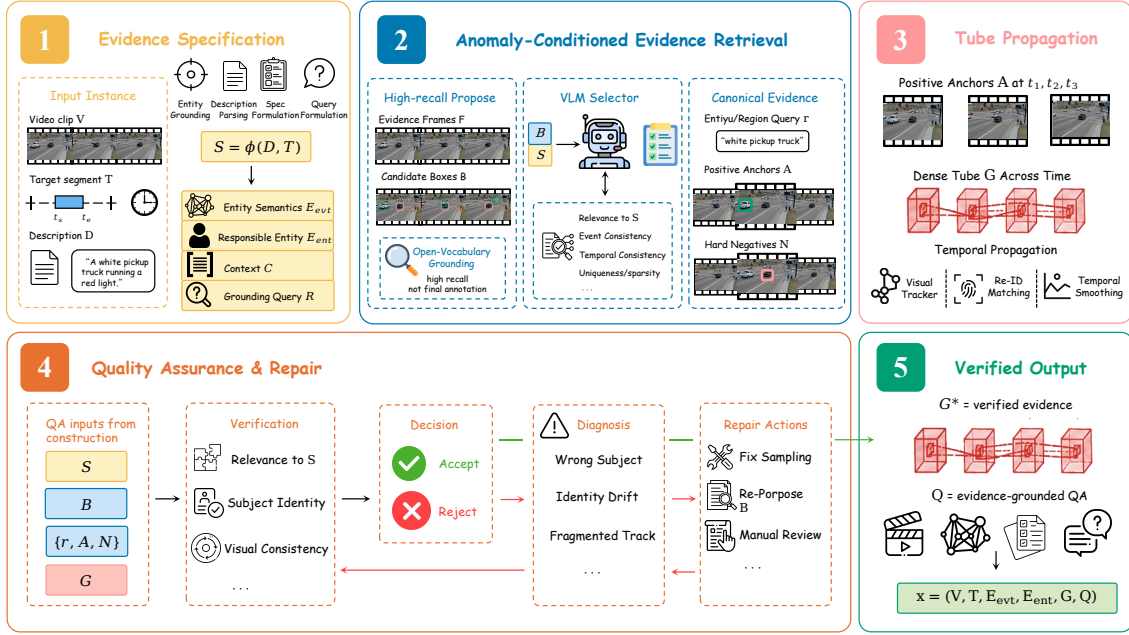


Figure 2: SemEviGround construction pipeline. The framework converts event-level anomaly semantics into verified spatial evidence by specifying anomaly-relevant entities, retrieving anomaly-conditioned evidence, propagating it into tubes, and applying quality-controlled repair before benchmark release.

and semantics that support reliable question answering.

To construct reliable spatial evidence at benchmark scale, we develop **SemEviGround**, a semantics-driven evidence grounding framework. Given the event semantics and anomaly-relevant subject or region, SemEviGround converts heterogeneous anomaly resources into verified boxes and tubes through four stages: evidence specification, anomaly-conditioned evidence retrieval, tube propagation, and quality-controlled repair. To mitigate circular evaluation in model-assisted annotation, LLM/VLM outputs are used only as proposals or preliminary reviews, while human annotators make final acceptance decisions after repeated filtering, visual inspection, and repair. The generated evidence is checked through automatic diagnostics, VLM-based review, and human verification by six annotators over approximately 340 annotator-hours. Appendix A details the source normalization and annotation process, and Appendix B provides implementation details. The resulting benchmark statistics are summarized in Figure 3.

3.2 SemEviGround: Semantics-driven Evidence Grounding

Evidence specification. Many anomaly resources provide event descriptions or temporal

annotations, but do not specify reliable spatial evidence for the anomalous event. To address this limitation, SemEviGround first converts heterogeneous annotations into a segment-level evidence specification:

$$S = \phi(D, T) = (E_{\text{evt}}, E_{\text{ent}}, C, R), \quad (2)$$

where D denotes the source annotation or description, and $T = [t_s, t_e]$ denotes the abnormal temporal segment. We derive T from the source temporal annotation and convert it into a unified interval format when necessary. Given (D, T) , ϕ is implemented as an LLM-based parser that extracts four structured fields: E_{evt} captures the abnormal action and its context, E_{ent} identifies the anomaly-relevant subject or region, C specifies contextual constraints, and R denotes grounding queries derived from the event semantics. Appendix B.2 provides example cases for these fields. The resulting specification transforms open-ended anomaly descriptions into event-conditioned visual evidence queries, while preserving the rationale for why the target evidence is relevant to the anomaly.

Anomaly-conditioned evidence retrieval.

Given the evidence specification S , SemEviGround retrieves candidate spatial evidence in two steps. It first samples evidence frames $F = \{f_k\}_{k=1}^K$ within T using uniform coverage,

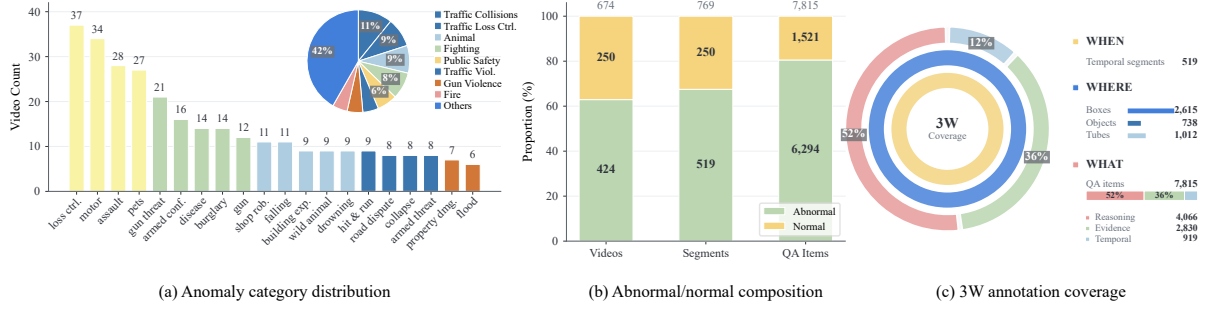


Figure 3: Benchmark statistics of Anom3W. (a) Top anomaly categories and coarse category groups. (b) Normal/abnormal composition across videos, temporal segments, and QA items. (c) Coverage of the 3W annotations used for temporal grounding, spatial grounding, and anomaly QA.

then applies GroundingDINO, an open-vocabulary object grounding model, to each grounding query $r \in R$ on each sampled frame. For a sampled frame f_k and query r , the grounding model returns a set of candidate boxes with confidence scores:

$$\gamma(f_k, r) = \{(\hat{q}_{k\ell}, \hat{s}_{k\ell})\}_{\ell=1}^{L_k}, \quad (3)$$

where $\hat{q}_{k\ell}$ is the normalized candidate box and $\hat{s}_{k\ell}$ is the proposal confidence. After normalization, filtering, and merging, the resulting proposal set is denoted as

$$\mathcal{B} = \{(f_k, r, \hat{q}_{k\ell}, \hat{s}_{k\ell})\}. \quad (4)$$

Since $\mathcal{B}$ is optimized for recall, it may contain visually plausible but semantically irrelevant regions, and is therefore not directly accepted as benchmark evidence.

To obtain event-aligned evidence, SemEviGround performs event-evidence consistency selection. Given S and $\mathcal{B}$, a vision-language selector evaluates whether each proposal supports the specified anomalous event and groups consistent proposals into canonical anomaly-relevant targets:

$$\mathcal{O} = \psi(S, \mathcal{B}) = \{(r_j, A_j, N_j)\}_{j=1}^M, \quad (5)$$

where r_j is the entity or region query, A_j is a sparse set of positive anchor boxes, and N_j contains negative samples from visually similar distractors. This stage filters out objects that are visually present but not responsible for the anomaly, ensuring that the selected anchors are grounded in both event semantics and visual evidence.

Tube propagation. The selected anchors are then propagated into dense spatio-temporal tubes using promptable video segmentation or video

object tracking models, such as SAM 3 and XMem (Carion et al., 2026; Cheng and Schwing, 2022). For each canonical target, SemEviGround constructs a tube

$$G_j = \{g_{j,t}\}_{t \in T}, \quad (6)$$

conditioned on the positive anchors A_j . Different targets are tracked independently to reduce identity interference. Anchor-frame corrections are used to mitigate identity drift and spatial deviation, while localized windowed tracking is applied to repair long or unstable intervals. The final spatial evidence is represented as $G = \{G_j\}_{j=1}^M$.

Quality assurance and repair. Before accepting G as benchmark evidence, SemEviGround applies a closed-loop quality assurance protocol. Automatic diagnostics check anchor alignment, abrupt area or center changes, empty gaps, temporal coverage, and tube continuity. VLM-based review further evaluates subject correctness, localization quality, identity stability, temporal consistency, and evidence sufficiency. Ambiguous or low-quality cases are routed to targeted repair, including frame resampling, proposal regeneration, object reselection, and local retracking. Final evidence is accepted only when the temporal extent is clear, the responsible entity or region is visible, the tube remains stable, and the QA semantics are visually supported.

3.3 3W Evaluation Protocol

Anom3W evaluates models along the same three dimensions used for construction: temporal anomaly grounding for **when**, spatio-temporal anomaly grounding for **where**, and anomaly question answering for **what**. Models are evaluated only on compatible subsets; for example, methods that cannot output boxes or tubes are excluded

from spatio-temporal grounding. The end-to-end protocol is gated by binary normal/abnormal discrimination: only videos predicted as abnormal are allowed to contribute temporal or spatio-temporal grounding scores, while abnormal videos predicted as normal are counted as detection failures with no valid grounding prediction. We additionally report binary F1 and balanced accuracy for this first-stage discrimination.

In the end-to-end grounding setting, temporal and spatio-temporal predictions are considered valid only for videos predicted as abnormal. An abnormal video predicted as normal is recorded as a detection failure and treated as having no valid temporal or spatio-temporal grounding prediction.

Normality discrimination. We report binary F1 and balanced accuracy for normal-vs-abnormal classification, which serves as the gating step before temporal and spatio-temporal grounding.

Temporal anomaly grounding. Temporal grounding evaluates whether a model localizes the abnormal event in time. Given a predicted span $p = [s_p, e_p]$ and a ground-truth span $g = [s_g, e_g]$, we compute:

$$\text{tIoU}(p, g) = \frac{\max(0, \min(e_p, e_g) - \max(s_p, s_g))}{\max(e_p, e_g) - \min(s_p, s_g)}. \quad (7)$$

Following temporal action localization protocols (Heilbron et al., 2015), we report mAP at tIoU thresholds $\{0.3, 0.5, 0.7\}$.

Spatio-temporal anomaly grounding. Spatio-temporal grounding evaluates whether a model localizes the anomaly-relevant subject or region over time. Given a predicted tube $P = \{p_t\}_{t=s_p}^{e_p}$ and a ground-truth tube $G = \{g_t\}_{t=s_g}^{e_g}$, we compute:

$$\text{vIoU}(P, G) = \frac{1}{|T_\cup|} \sum_{t \in T_\cap} \text{IoU}(p_t, g_t), \quad (8)$$

where $T_\cap = [s_p, e_p] \cap [s_g, e_g]$ and $T_\cup = [s_p, e_p] \cup [s_g, e_g]$. This jointly penalizes temporal misalignment and spatial localization errors. We report mean video IoU and success rates at vIoU thresholds $\{0.3, 0.5\}$ (Chen et al., 2019; Zhang et al., 2020; Tang et al., 2022).

We also report a diagnostic spatial-localization setting with the ground-truth temporal interval provided. This oracle setting removes temporal localization errors and evaluates whether a model can localize the anomaly-relevant subject or region within the correct event interval. We com-

pute mean spatial IoU over the ground-truth interval and success rates at spatial IoU thresholds $\{0.3, 0.5\}$, and report these results separately from the end-to-end benchmark.

Evidence-assisted anomaly QA. The QA task evaluates semantic anomaly understanding. Questions cover anomaly presence, category, responsible subject, interactions, temporal order, and intervention or prevention. For multiple-choice questions, we report accuracy. For open-ended questions, we use a reference-guided LLM-as-judge protocol with 1–5 rubrics (Zheng et al., 2023; Chiang et al., 2024). The semantic rubric emphasizes event correctness, anomaly/normality judgment, rationale quality, and false-alarm control, while the evidence rubric emphasizes subject grounding, interaction correctness, temporal consistency, and hallucination control.

These QA items provide complementary semantic probes alongside the grounding tasks. We separately analyze *video-only QA* versus *evidence-assisted QA*, where the latter uses human-verified box or tube overlays as controlled evidence cues.

3.4 Benchmark Statistics

At the standardized evaluation-set level, Anom3W contains **674** videos, **769** temporal segments, **7,815** QA items, and **81** anomaly categories. Fig. 3(a) shows the category distribution, with the top-20 anomaly subtypes summarized in the main panel and coarse anomaly groups shown in the inset. Fig. 3(b) summarizes the normal/abnormal composition across videos, temporal segments, and QA items. Fig. 3(c) presents the 3W annotation coverage: **519** temporal segments for **when**, **2,615** boxes, **738** objects, and **1,012** tubes for **where**, and **7,815** QA items for **what**, including **4,066** reasoning, **2,830** evidence, and **919** temporal questions.

4 Experiments

4.1 Experiment Setup

We evaluate general-purpose MLLMs from the Gemini, Qwen (Bai et al., 2025), GLM, and Kimi families; video-centric MLLMs including VideoL-LaMA3, LLaVA-Video, LLaVA-OneVision, and Video-XL variants (Zhang et al., 2025a, 2024; Li et al., 2024; Shu et al., 2025); and task-specific anomaly models including Holmes-VAU, Event-VAD, VADTree, and AnyAnomaly (Zhang et al., 2025b; Shao et al., 2025; Li et al., 2025; Ahn

Table 2: Main end-to-end evaluation results on Anom3W. Percentage-scale metrics are marked with (%); QA Judge uses the original 1–5 rubric scale.

Model	Binary F1 (%)	Bal. Acc. (%)	mAP@0.3 (%)	mAP@0.5 (%)	mAP@0.7 (%)	m_vIoU (%)	vIoU@0.3 (%)	vIoU@0.5 (%)	QA Acc. (%)	QA Judge (1–5)
<i>General-purpose MLLMs</i>										
Gemini 3.1 Pro	98.04	97.87	38.76	23.74	11.72	22.61	33.27	14.06	99.46	4.73
Gemini 3.1 Flash	95.78	92.93	42.45	26.49	13.23	25.97	38.32	20.20	99.46	4.79
Gemini 3.1 Flash Lite	98.60	97.60	40.49	25.47	11.84	13.90	19.21	7.62	97.86	4.64
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	82.20	83.00	28.81	17.89	9.30	3.46	4.46	1.19	93.29	4.38
Qwen3.6-Plus	87.01	85.88	39.39	23.88	11.10	10.57	13.96	4.46	94.53	4.31
Qwen3.6-35B-A3B	85.64	85.43	34.20	18.39	7.63	6.28	5.64	1.98	94.44	4.25
Qwen3.6-27B	85.38	84.67	35.85	20.16	9.71	8.27	10.79	3.27	93.79	4.22
Qwen3.5-397B-A17B	72.99	77.28	2.99	0.87	0.10	1.22	0.50	0.10	99.04	4.24
Qwen3.5-122B-A10B	84.22	85.35	36.08	20.80	11.20	7.07	8.02	3.07	91.87	4.37
Qwen3.5-9B	82.29	84.01	30.36	19.75	8.31	3.37	3.66	1.29	98.31	4.09
GLM-5V-Turbo	27.39	51.41	1.43	0.33	0.08	0.56	0.40	0.00	98.90	4.12
Kimi K2.5	87.06	85.27	37.77	20.97	9.94	7.92	9.90	2.67	97.89	4.40
Kimi K2.6	85.39	83.78	34.05	18.40	8.97	7.36	8.32	2.18	99.01	4.29
<i>Video-centric MLLMs</i>										
VideoLLaMA3-7B (Zhang et al., 2025a)	78.36	81.47	23.07	12.75	6.25	0.13	0.20	0.00	93.04	2.66
LLaVA-Video-7B-Qwen2 (Zhang et al., 2024)	72.67	77.61	9.98	6.59	3.69	0.28	0.40	0.20	90.88	2.67
LLaVA-OneVision-1.5-4B-Instruct	80.43	80.98	0.16	0.13	0.00	0.24	0.50	0.10	92.23	3.11
LLaVA-OneVision-1.5-8B-Instruct	86.63	84.03	9.11	2.44	0.41	0.67	0.20	0.00	93.33	2.98
LLaVA-OneVision-Qwen2-7B (Li et al., 2024)	53.52	67.80	0.42	0.15	0.03	0.12	0.00	0.00	89.24	2.85
Video-XL-Pro-3B (Shu et al., 2025)	66.87	74.55	3.56	1.98	0.09	1.91	1.98	0.30	90.96	2.99
Video-XL-2 (Shu et al., 2025)	55.78	69.34	7.48	3.36	0.26	0.00	0.00	0.00	92.17	3.05
<i>Task-specific anomaly models</i>										
Holmes-VAU (Zhang et al., 2025b)	77.08	49.96	1.20	0.06	0.02	0.00	0.00	0.00	80.09	1.95
EventVAD (Shao et al., 2025)	77.08	75.31	0.52	0.06	0.00	0.48	0.30	0.00	90.74	3.26
VADTree (Li et al., 2025)	72.67	77.61	9.98	6.59	3.69	0.37	0.40	0.30	90.88	2.70
AnyAnomaly (Ahn et al., 2026)	32.86	52.05	0.00	0.00	0.00	0.00	0.00	0.00	79.13	2.59

et al., 2026). MLLMs are evaluated zero-shot with default inference configurations, while task-specific models use released checkpoints or official pipelines. Temporal and spatio-temporal grounding are scored only after a model predicts the video as abnormal. Appendix C provides model-scale, compute, and execution details.

4.2 Main Results

Table 2 summarizes performance under the unified end-to-end 3W protocol, covering normality discrimination, temporal grounding, spatio-temporal grounding, and anomaly QA. Grounding metrics are the central stress test and are computed only on samples first predicted as abnormal.

General-purpose MLLMs are strongest overall, but no model dominates all dimensions. Gemini 3.1 Flash Lite gives the best abnormal F1, Gemini 3.1 Pro gives the best balanced accuracy, Gemini 3.1 Flash obtains the highest open-ended QA judge score and the strongest grounding scores among the Gemini models, and Qwen3.6-Plus remains the best non-Gemini model on temporal and spatio-temporal grounding. Video-centric MLLMs are more reliable on normality discrimination than on grounding: VideoLLaMA3-7B gives the strongest temporal localization in this group, but spatial scores remain weak after temporal and spatial errors compound. Task-specific anomaly models remain competitive detectors, but their QA and spatial grounding scores lag behind

stronger MLLMs. Overall, the main bottleneck is not only recognizing anomalous content, but binding that recognition to the responsible visual evidence. The high multiple-choice scores of several models should therefore be read as evidence of coarse anomaly recognition, rather than as evidence that the model has solved the grounding problem itself. At the same time, the large gaps on temporal and spatio-temporal grounding show that precise evidence localization remains the more discriminative challenge in this benchmark. This makes the benchmark useful not only for comparing final scores, but also for separating coarse semantic recognition from evidence-grounded reasoning.

4.3 Evidence-Assisted Anomaly Understanding

We compare *video-only QA* and *evidence-assisted QA* to test whether localized, human-verified evidence improves semantic anomaly understanding. Fig. 4 shows that evidence-assisted QA improves MCQ accuracy for 23 of 24 models, with an average gain of 4.49 points. Open-ended QA shows a weaker but positive trend: 18 of 24 models improve, with an average LLM-as-judge gain of 0.21. Gains are larger for several video-centric and task-specific models, while the strongest general-purpose MLLMs sometimes decrease slightly, suggesting that explicit evidence is most useful when video-only grounding is weak.

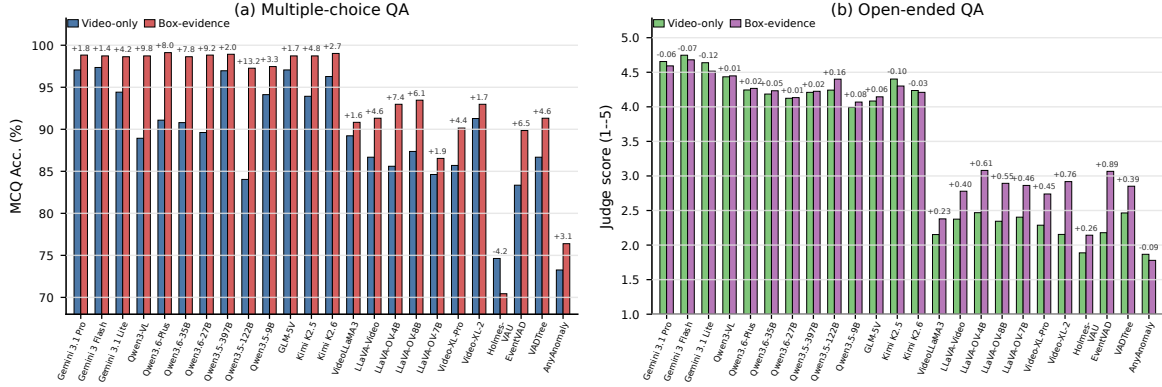


Figure 4: Effect of human-verified evidence on anomaly QA. Values above paired bars denote the evidence-input gain for multiple-choice accuracy and open-ended LLM-as-judge score.

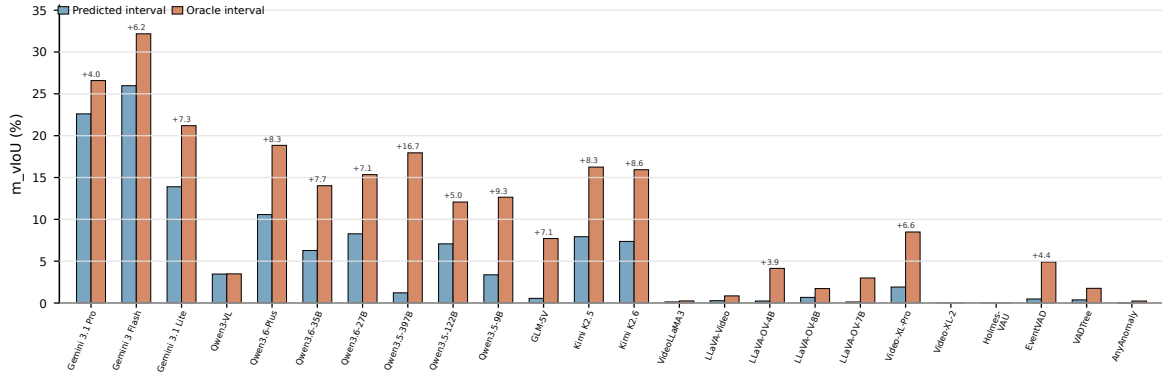


Figure 5: Effect of temporal localization on tube grounding. The oracle setting provides the ground-truth abnormal event interval; numbers above bars show m_vIoU gains of at least 3 points.

4.4 Temporal Effects on Tube Grounding

We disentangle temporal and spatial errors by comparing tube grounding under predicted temporal spans with an oracle setting that provides the ground-truth abnormal event interval. Fig. 5 shows that the oracle setting improves m_vIoU for 22 of 24 models, raising the average score from 5.12% to 9.98%. This confirms that temporal errors compound spatial localization errors. However, the remaining low oracle scores indicate that even when the event interval is known, localizing the anomaly-relevant subject remains a difficult spatial grounding problem.

4.5 Human Validation of LLM-as-Judge

To validate the reference-guided LLM-as-judge protocol for open-ended QA, we conduct a human validation study on 200 open-ended responses covering 24 evaluated models and both normal and abnormal videos. The validation set is stratified by LLM-judge score, with 40 responses sampled from each score level on the 1–5 rubric. As shown in Fig. 6, human and LLM scores show strong

agreement, with Spearman’s $\rho = 0.817$, Pearson’s $r = 0.819$, 71.0% exact agreement, 93.0% agreement within one point, and 0.40 mean absolute error. The mean signed difference (Human – LLM) is +0.19. These results support the benchmark-level use of the judge, while we do not treat it as interchangeable with human evaluation.

5 Conclusion

We presented Anom3W, a unified benchmark for video anomaly grounding and understanding organized around **what**, **where**, and **when**. Motivated by the separation between conventional VAD benchmarks and recent VAU benchmarks, Anom3W jointly evaluates temporal grounding, spatio-temporal grounding, and anomaly question answering under a shared 3W protocol. We further introduced SemEviGround, a semantics-driven evidence grounding framework that converts anomaly semantics into verified boxes and tubes at benchmark scale. By treating anomaly-relevant boxes and tubes as visual evidence, Anom3W evaluates whether models can ground

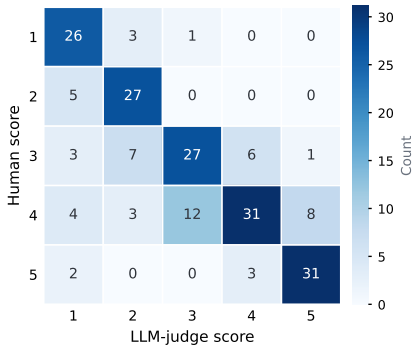


Figure 6: Human validation of LLM-as-judge scores. Rows denote human scores and columns denote LLM-judge scores; darker cells indicate more samples.

anomaly understanding in the responsible subjects and regions.

This unified design enables comparison under a common evaluation target while preserving each component task’s distinct difficulty: temporal grounding captures when an anomaly occurs, spatial evidence localization captures where the responsible entity or region appears, and anomaly QA captures what a model can explain about the event. Together, these views test whether models merely recognize anomalous content or align their understanding with supporting visual evidence, offering a useful perspective for future evidence-grounded video understanding beyond anomaly benchmarks.

Limitations

Anom3W is designed as an evaluation benchmark rather than a training dataset, so its coverage is bounded by the available source videos and verifiable annotations. This conservative construction favors reliable temporal, spatial, and semantic evaluation, but it also means that rarer or highly domain-specific anomalies may be under-represented. In addition, although SemEviGround provides verified visual evidence, tube construction remains challenging under occlusion, fast motion, and visually ambiguous cases. Open-ended QA also relies on a reference-guided LLM-as-judge protocol, whose scores may be affected by prompt design, judge-model behavior, or incomplete references.

Ethics Statement

Scope and Responsible Use. Anom3W is designed for offline research evaluation of video anomaly grounding and understanding, rather than

autonomous decision-making or real-time intervention. As with other benchmarks involving public-safety scenarios, its results should be interpreted as performance on a curated research benchmark and not as a guarantee of real-world robustness. Applications in high-stakes settings should include domain-specific validation and human oversight.

Artifact Licensing. Anom3W is constructed from source resources whose use and distribution are governed by the terms stated in their original releases or accompanying project pages. Where datasets specify explicit licenses or access conditions, we follow those terms in constructing and releasing the benchmark. Where redistribution terms are not clearly stated, we do not claim any additional rights beyond the original releases, and we distribute only benchmark-created annotations, metadata, and access mappings to the extent permitted by the source conditions.

Intended Use and Access Conditions. Our use of source resources is limited to research-oriented benchmark construction and evaluation, consistent with their role as research data resources where such intended use is specified. Correspondingly, Anom3W and its derived annotations are intended for research evaluation of video anomaly understanding and remain subject to applicable source access conditions. Any distribution of Anom3W will provide benchmark-created annotations, metadata, and access mappings only to the extent permitted by the original conditions, without expanding redistribution rights for the source media.

AI Assistance. AI assistants were used only for language polishing, grammar checking, and improving clarity during writing. They were not used to replace benchmark annotation, experimental execution, result computation, or the authors’ responsibility for the paper’s claims and conclusions.

Acknowledgments

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM103), the Zhejiang NSF (LZ25F020004), the Key R&D Program in Zhejiang Province (2025C02156).

References

- Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2022. [Ubnormal: New benchmark for supervised open-set video anomaly detection](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2011120121. IEEE.
- Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. 2026. [Anyanomaly: Zero-shot customizable video anomaly detection with lvlm](#). In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 30263035. IEEE.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-sheng Li, and 45 others. 2025. Qwen3-vl technical report.
- Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. 2023. [A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation](#). In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2039220401. IEEE.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Bais-han Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, and 19 others. 2026. Sam 3: Segment anything with concepts.
- Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. [Localizing natural language in videos](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):81758182.
- Ho Kei Cheng and Alexander G. Schwing. 2022. [XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model](#), page 640658. Springer Nature Switzerland.
- Ming Cheng, Kunjing Cai, and Ming Li. 2021. [Rwf-2000: An open large scale video database for violence detection](#). In *2020 25th International Conference on Pattern Recognition (ICPR)*, page 41834190. IEEE.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.
- Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Ji Zhang, Min Lei, Pengxuan Mao, Jinli Ou, Xiaofeng Tao, and Jun Liu. 2026. [Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly](#). *International Journal of Computer Vision*, 134(8).
- Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 2024. [Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1879318803. IEEE.
- Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. 2025. [Vane-bench: Video anomaly evaluation benchmark for conversational llms](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 31233140. Association for Computational Linguistics.
- Shibo Gao, Peipei Yang, Yangyang Liu, Yi Chen, Han Zhu, Xu-Yao Zhang, and Linlin Huang. 2026. [Vagu & gts: Llm-based benchmark and framework for joint video anomaly grounding and understanding](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):41674175.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. [Activitynet: A large-scale video benchmark for human activity understanding](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 961970. IEEE.
- Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. 2025. [Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought](#). In *Advances in Neural Information Processing Systems* 38, NeurIPS 2025, page 131353131385. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Pu Jin, Lichao Mou, Gui-Song Xia, and Xiao Xiang Zhu. 2022. [Anomaly detection in aerial videos with transformers](#). *IEEE Transactions on Geoscience and Remote Sensing*, 60:113.
- Federico Landi, Cees GM Snoek, and Rita Cucchiara. 2019. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer.
- Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. 2025. [Vadtrees: Explainable training-free video anomaly detection via hierarchical granularity-aware tree](#). In *Advances in Neural*

- Information Processing Systems* 38, NeurIPS 2025, page 164634164666. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. 2025. [Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models](#). *arXiv preprint*.
- Yang Liu, Dingkan Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. 2024. [Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models](#). *ACM Computing Surveys*, 56(7):138.
- Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. [Abnormal event detection at 150 fps in matlab](#). In *2013 IEEE International Conference on Computer Vision*, page 27202727. IEEE.
- Weixin Luo, Wen Liu, and Shenghua Gao. 2017. [A revisit of sparse coding based anomaly detection in stacked rnn framework](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 341349. IEEE.
- Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. 2021. [A comprehensive review on deep learning-based methods for video anomaly detection](#). *Image and Vision Computing*, 106:104078.
- Chirag Parikh, Deepti Rawat, Rakshitha R. T., Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. 2025. [Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1900219011. IEEE.
- Joao Alexandre Cardeira Pereira, Vasco Lopes, João C. Neves, and David Semedo. 2026. [Finevau: A novel human-aligned benchmark for fine-grained video anomaly understanding](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):84038411.
- Bharathkumar Ramachandra and Michael J. Jones. 2020. [Street scene: A new dataset and evaluation protocol for video anomaly detection](#). In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, page 25582567. IEEE.
- Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Xiaochen Wang, Hao Tang, Yan Wang, and Shuyan Li. 2025. [Eventvad: Training-free event-aware video anomaly detection](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM 25, page 25862595. ACM.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. [Video-xl: Extra-long vision language model for hour-scale video understanding](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2616026169. IEEE.
- Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. [Real-world anomaly detection in surveillance videos](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 64796488. IEEE.
- Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. 2024. [Hawk: Learning to understand open-world video anomalies](#). In *Advances in Neural Information Processing Systems* 37, NeurIPS 2024, page 139751139785. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2022. [Human-centric spatio-temporal video grounding with visual transformers](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):82388249.
- Shu Wang and Zhenjiang Miao. 2010. [Anomaly detection in crowd scene](#). In *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, page 12201223. IEEE.
- Zhuonan Wang, Zhenxuan Fan, Siwen Tan, Yu Zhong, Yuqian Yuan, Haoyuan Li, Hao Jiang, Wenqiao Zhang, Feifei Shao, Hongwei Wang, and Jun Xiao. 2026. [Mau-gpt: Enhancing multi-type industrial anomaly understanding via anomaly-aware and generalist experts adaptation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(31):2678726795.
- Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. [Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision](#), page 322339. Springer International Publishing.
- Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. 2024. [Open-vocabulary video anomaly detection](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1829718307. IEEE.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. [Next-qa: Next phase of question-answering to explaining temporal actions](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 97729781. IEEE.
- Chenting Xu, Ke Xu, Xinghao Jiang, and Tanfeng Sun. 2025. [Plovad: Prompting vision-language models for open vocabulary video anomaly detection](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):59255938.
- Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. 2023.

[Dota: Unsupervised detection of traffic anomaly in driving videos](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):444459.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. 2024. [Towards surveillance video-and-language understanding: New dataset, baselines, and challenges](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2205222061. IEEE.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. 2025a. [Viderefer suite: Advancing spatial-temporal object understanding with video llm](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1897018980. IEEE.

Yuqian Yuan, Wenqiao Zhang, Xin Li, Shihao Wang, Kehan Li, Wentong Li, Jun Xiao, Lei Zhang, and Beng Chin Ooi. 2025b. [Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity](#). *arXiv preprint*.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. [Harnessing large language models for training-free video anomaly detection](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1852718536. IEEE.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. [Videollama 3: Frontier multimodal foundation models for image and video understanding](#). *arXiv preprint*.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. 2025b. [Holmes-vau: Towards long-term video anomaly understanding at any granularity](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1384313853. IEEE.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. [Llava-video: Video instruction tuning with synthetic data](#). *arXiv preprint arXiv:2410.02713*.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. [Where does it exist: Spatio-temporal video grounding for multi-form sentences](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1066510674. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haoteng Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *ArXiv*, abs/2306.05685.

Table 3: Video-level statistics of the released evaluation set.

Statistic	Value
Total duration	13.54 hours
Duration, min / mean / max	4.99 / 72.34 / 1,342.12 s
Resolution range	320×240 – 2560×1440
Mean pixel count per frame	0.49 MP
FPS, min / mean / max	15.00 / 27.49 / 60.13

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

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. 2024. [Advancing video anomaly detection: A concise review and a new dataset](#). In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, page 8994389977. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

A Dataset Construction and Annotation Details

Anom3W is constructed by materializing heterogeneous anomaly and normal-video resources into two unified benchmark splits. The abnormal split combines ECVA, HC-STVG, Vid-STG, HIVAUXD-Violence, RoadSocial, Highlights, and Vad-R1, while the normal split combines normal videos from XD-Violence and NExT-QA. Because these sources provide different native supervision, including labels, temporal moments, descriptions, rule-derived intervals, QA annotations, and spatio-temporal tubes, construction converts each retained example into a common video–segment–event–entity–evidence–QA representation.

The final evaluation set contains 674 videos and 769 temporal segments: 424 abnormal videos with 519 abnormal segments, and 250 normal videos represented as full-video segments. The released benchmark contains 7,815 QA items in total, including 6,294 abnormal QA items and 1,521 normal QA items. These QA items further break down into 4,066 reasoning questions, 2,830 evidence questions, and 919 temporal questions. Table 3 reports video-level statistics, and Table 4 summarizes the released source composition.

Source normalization. All sources are normalized into a segment-centered schema. For abnormal videos, each segment corresponds to a visually grounded event with a start time, end time,

Table 4: Released benchmark composition and native supervision before normalization.

Source	Split	Videos	Native supervision used in construction
ECVA (Du et al., 2026)	Abnormal	142	Video ID, anomaly classification, annotated moment, description, reason/result, and key sentences.
HC-STVG (Tang et al., 2022)	Abnormal	88	Original temporal spans, frame ranges, and spatio-temporal grounding tubes, with anomaly-category assignment.
RoadSocial (Parikh et al., 2025)	Abnormal	80	Rule-derived decision, category hints, social/video tags, high-value QA signals, and grounding intervals.
VidSTG (Zhang et al., 2020)	Abnormal	39	Original temporal spans, frame ranges, and spatio-temporal grounding tubes, with anomaly-category assignment.
Highlights	Abnormal	34	Curated clips extracted from compilation-style videos in XD-Violence and ECVA, with major/minor category labels and descriptions.
Vad-R1 (Huang et al., 2025)	Abnormal	33	Normalized anomaly categories, event-level time spans, reasoning fields, descriptions, and key sentences.
HIVAU/XD-Violence (Wu et al., 2020)	Abnormal	8	HIVAU annotations over XD-Violence videos, including normalized anomaly categories, summaries, descriptions, and event-level temporal spans.
XD-Violence normal (Wu et al., 2020)	Normal	150	Normal-video summaries, split captions, raw labels, and source split metadata.
NExT-QA (Xiao et al., 2021)	Normal	100	Normal-video QA pairs, relation objects, relation annotations, and frame metadata.

Table 5: Temporal segment normalization by source family.

Source family	Segment definition
ECVA	Annotated abnormal moment; fallback to full video only when the record is otherwise valid.
HIVAU / Vad-R1	Event-level event_sec span from normalized segment records.
RoadSocial	Rule-derived grounding interval; invalid or rejected rule records are removed.
Highlights	Curated clip-level interval, treated as one abnormal segment.
VidSTG / HC-STVG	Original time_range_sec and frame_range from spatio-temporal annotations.
Normal videos	Full-video segment from 0 to video duration.

category label, and short event description. Table 5 lists the temporal source used for each family. Highlights is treated as a curated segment-level source, with each retained clip corresponding to one abnormal event.

Normal videos are represented as full-video normal segments, with QA or caption information retained only as context for normal-video QA construction. For compilation-style videos, annotators re-segment each file into event-level clips using scene cuts, transitions, changes of location or subject, and switches to different incidents; consecutive views of the same event may be merged. After temporal normalization, each abnormal segment is converted into the unified instance schema used in the main paper, while source traceability is preserved in the released files.

Category taxonomy. We use a two-level taxonomy with `major_category` and `minor_category`, together with a textual classification field. Category signals are parsed or normalized from source labels when available, inferred from rule-based records for RoadSocial, and assigned

through anomaly-candidate classification for VidSTG and HC-STVG. Normal videos are assigned `major_category=Normal`, with the minor category indicating the normal source dataset.

During normalization, annotators merge synonymous labels, separate visually different subtypes when a source label is too broad, and assign “Other” only when the event is clearly abnormal but cannot be reliably disambiguated further. Each category is checked against the retained segment so that it reflects the visible event rather than source metadata alone.

The documentation of Anom3W focuses on event domains, anomaly categories, temporal and spatial evidence, and QA supervision. Event descriptions are manually written in English. We do not construct demographic identity labels or claim demographic subgroup coverage, and demographic subgroup evaluation is outside the scope of the benchmark.

Segment filtering. Filtering is part of the multi-round review process. It considers not only whether a video is abnormal, but also whether each retained segment can support the 3W benchmark protocol. Table 6 summarizes the acceptance checks. Records with explicit filter flags, rejected rule decisions, failed status fields, missing media, or unusable temporal intervals are skipped before manual review; annotators then judge each segment within its specified time range.

Segments are rejected when they are too short, visually degraded, dominated by transitions or text overlays, repeated without new information, inconsistent with the visible event, or unsuitable for stable localization and tracking. Global scene

Table 6: Acceptance checks used during construction and review.

Check	Retained instance must satisfy
Media validity	Video path is available and the retained segment is decodable.
Temporal clarity	The interval contains the visible abnormal event or the full normal video.
Category consistency	Major/minor category matches the visible event rather than source metadata alone.
Entity visibility	The abnormal subject or localized region is visible enough for evidence construction when required.
Spatial reliability	Anchors or tubes can be localized, tracked, and verified without identity drift or severe ambiguity.
QA answerability	Questions can be answered from the retained video evidence.

states and crowd-level conditions are excluded unless a localized responsible subject or region can be identified. When needed, annotators adjust temporal boundaries to better cover the visible and trackable portion of the abnormal event.

Human annotation and verification. Human annotation and verification instantiate these checks in several rounds. Annotators first remove segments whose spatial evidence is absent, not visually meaningful, or unsuitable for the grounding–selection–tracking pipeline. For compilation-style anomaly videos in ECVA and XD-Violence, annotators re-annotate retained incidents as event-level segments with start/end times, major/minor categories, and concise English descriptions naming the visible abnormal subject and action. In practice, SemEviGround serves as an annotation aid for efficiency and consistency, while the final temporal boundaries, category assignments, and acceptance decisions are made by human annotators after inspection.

For spatial evidence, annotators manually verify pipeline-generated boxes against the original frame, event semantics, entity query, canonical object identity, and neighboring anchors from the same object. Decisions rely on visual evidence rather than confidence scores or geometric heuristics alone. Visible targets are redrawn rather than discarded; invisible, unverifiable, or distractor-confused anchors are removed. After tube generation, Gemini-based review flags potential identity drift, poor localization, fragmented tracks, incorrect re-acquisition, and visibility errors, which annotators inspect before acceptance. Verified anchors are treated as identity references, so tube repair modifies propagated frame-level boxes or visibility intervals without changing the reviewed anchors. Multiple-choice and open-ended QA items

are finally reviewed for answerability and consistency with the retained video evidence.

Quality control. Before acceptance, each instance is checked for temporal clarity, event/category consistency, visible evidence, entity correctness, cross-frame identity stability, and QA answerability. Automatic diagnostics and VLM review help identify suspicious boxes or tubes, but final acceptance prioritizes visual evidence and human verification. Ambiguous or unverifiable samples are removed rather than retained with unreliable temporal, spatial, or semantic annotations.

B SemEviGround Implementation Details

B.1 Pipeline Artifacts

SemEviGround is organized around inspectable intermediate artifacts. For each temporally localized segment, the semantic generation step uses GLM-5.1 by default and produces two views: an event view describing the abnormal action, interaction, causal cue, and scene context, and an entity view specifying the subjects, grounding queries, and contextual constraints needed for localization. These fields are stored with the segment metadata and reused by evidence retrieval, canonical selection, evidence review, and QA construction.

The construction protocol follows a human-in-the-loop, model-assisted annotation workflow. LLM/VLM outputs are used to propose event fields, candidate evidence, QA items, and preliminary review signals, but they do not directly define the released annotations. Annotators perform multiple rounds of filtering, verification, editing, and repair before any sample is accepted. This reduces the risk of circular evaluation, where model-generated labels are later treated as ground truth for evaluating similar models. Appendix A and the algorithm below summarize the full review-and-repair flow.

B.2 Evidence Specification Examples

Figure 7 shows three representative cases in a compact card layout. Each card pairs a visual strip with the corresponding event view, entity view, contextual constraints, and selection takeaway.

Table 7 lists the main model components used in construction, separating evidence construction from QA generation and validation.

Evidence frames are sampled uniformly from each segment. We use at least 3 frames, at most

A000089. Brown-dog attack in a home interior.



Event: a brown dog bites a brown-and-white dog.

Entity: the two interacting dogs.

Constraint: keep the box on the contact pair, not the furniture.

Takeaway: follow the responsible subjects.

A000143. Vehicle fire with spread to roadside grass.



Event: a Ram EcoDiesel burns on a highway.

Entity: the burning vehicle, flames, smoke, and nearby grass fire.

Constraint: passing vehicles and bystanders are not anomaly targets.

Takeaway: a region-style fire zone is acceptable.

A000137. Intersection violation and rollover crash.



Event: a car fails to yield and clips a Ute.

Entity: the violating car, the rolling Ute, and the give-way context.

Constraint: select involved vehicles, not unrelated traffic.

Takeaway: context disambiguates the responsible vehicle.

Figure 7: Representative evidence specification cases. The examples illustrate how event semantics and contextual constraints determine anomaly-relevant evidence for close-contact interaction, region-centric fire, and traffic collision.

10 frames, and a default interval of 2.0 seconds, covering the beginning, middle, and end of most abnormal intervals. Algorithm 1 summarizes the construction and repair flow.

B.3 Evidence Retrieval and Selection

The pipeline uses GroundingDINO-base as the first-pass proposer. When the proposal set misses the target or contains poor boxes, expanded grounding introduces additional open-vocabulary proposers, including OWLv2, YOLO-World, and Grounded-SAM-2. Grounded-SAM-

2 uses grounding boxes as prompts, refines them with SAM2 masks, and converts the masks back to boxes. Each proposal is stored with a proposal ID, frame index, timestamp, query text, score, and normalized box.

Proposal generation is recall-oriented; semantic precision is handled by canonical selection. The Qwen selector receives the event/entity semantics, sampled-frame overlays with candidate boxes, and per-frame proposal metadata. For each canonical object, it records an object ID, entity query, object summary, selected anchors, and hard nega-

Table 7: Model configuration used in benchmark construction.

Stage	Backend
Event/entity specification	GLM-5.1
First-pass evidence retrieval	GroundingDINO
Expanded evidence retrieval	OWLv2; YOLO-World; Grounded-SAM-2
Canonical object selection	Qwen selector
Tube propagation	SAM3
Abnormal QA generation	Qwen3.6-Plus
Normal QA generation	Gemini-3.1 Flash Lite
QA validation	Gemini-3-Flash-Preview
Evidence review	Gemini-3.1-Pro

Table 8: Automatic tube diagnostics.

Diagnostic	Default criterion
Area jump	Adjacent visible box area change > 0.65
Center jump	Center displacement > 1.15 previous-box diagonals
Empty gap	Internal invisible interval ≥ 3 frames
Discontinuity	Breaks in tracking frame indices
Anchor alignment	Anchor-frame IoU < 0.30 counted as low alignment
Tube coverage	Visible ratio, span coverage, holes, visible intervals

tives. The selected anchors serve as sparse positive evidence for tracking, while the negatives record clear distractors that should not be followed. The selector also returns a segment-level decision with coverage confidence, evidence sufficiency, failure type, and a suggested next action.

B.4 Tracking, Review, and Repair

For spatial subsets that require newly constructed tubes, verified anchors are propagated with SAM3. Source datasets that already provide tube annotations, such as HC-STVG or VidSTG-style sources, are converted into the compatible tube format instead of being re-tracked. Each canonical object is tracked independently, and SAM3 masks are converted to boxes when box-level tubes are required. If tracking drifts after an anchor has been manually verified, repair modifies the propagated tube or marks affected frames as invisible; it does not overwrite the reviewed anchor identity.

After tracking, automatic diagnostics in Table 8 identify common tube failures before VLM or human review. They are not final acceptance rules: the reviewer prompt treats them as supporting signals and prioritizes visual evidence.

The Gemini-3.1-Pro reviewer receives the semantic requirement view, selected anchors and negatives, selector confidence, SAM3 masks or converted boxes, automatic diagnostics, and review frames or overlay video. It returns an accept-

Algorithm 1 SemEviGround evidence construction and repair

Require: Segment $x = (V, T, D)$ with video V , temporal span T , and source description D

Ensure: Verified evidence G^* or a manual-review target

```

1:  $S \leftarrow \phi(D, T)$   $\triangleright$  event/entity specification
2:  $F \leftarrow \text{UNIFORMSAMPLE}(V, T)$ 
3:  $B \leftarrow \text{GROUND}(F, S; \text{GroundingDINO})$ 
4:  $O \leftarrow \text{SELECT}(F, B, S)$   $\triangleright$  anchors and hard negatives
5: for  $r = 1$  to  $R_{\text{qwen}}$  do
6:   if  $O$  is erroneous or insufficient then
7:      $(F, B, O) \leftarrow \text{QWENREPAIR}(x, S, F, B, O)$ 
8:   else
9:     break
10:  end if
11: end for
12: if  $x$  has source tubes then
13:    $G \leftarrow \text{CONVERTTUBE}(x)$ 
14: else
15:    $G \leftarrow \text{TRACKSAM3}(V, T, O)$ 
16: end if
17:  $E \leftarrow \text{EVALUATE}(G, O, S)$   $\triangleright$  diagnostics + Gemini review
18: for  $r = 1$  to  $R_{\text{eval}}$  do
19:   if  $\text{ACCEPTABLE}(E)$  then
20:     break
21:   end if
22:    $c \leftarrow \text{PLANCALLBACK}(E)$ 
23:   if  $c = \text{manual\_review}$  then
24:     return  $\text{MANUALTARGET}(x, O, G, E)$ 
25:   end if
26:    $\text{ROLLBACKANDRERUN}$  from callback stage  $c$ 
27:   update  $F, B, O, G, E$  with merged artifacts
28: end for
29: return  $\text{MERGEVERIFIED}(x, O, G, E)$ 

```

ability label, a root-cause category, and a callback action. The root causes cover wrong-subject selection, identity drift, poor localization, fragmented tracking, insufficient evidence, and semantic ambiguity. The compressed prompt templates above show the main decision criteria used for selection and review.

B.5 Repair Controller

Repair addresses failures detected during canonical selection or tube review. Selection-stage repair handles insufficient or erroneous evidence before tracking. It may resample evidence frames, rerun grounding with expanded proposers, or rerun canonical selection. If the target remains missing or semantically ambiguous after these attempts, the segment is routed to manual review rather than being forced into the benchmark.

Evaluation-stage repair handles failures after tube propagation. The repair decision is based on the root cause returned by the reviewer and the automatic diagnostics. Tracking-related failures, such as identity drift, poor localization,

Canonical Evidence Selection Prompt

System: Select anomaly-relevant visual evidence. Output VALID JSON only.

Input:

- Event summary and event context
- Entity subjects and grounding queries
- Sampled frame overlays with candidate boxes
- Proposal metadata: proposal_id, frame_idx, timestamp, query_text, score, bbox

Instruction:

Select only boxes that directly support the abnormal event semantics.
Merge boxes into a canonical object only when they refer to the same real-world subject. Keep sparse high-quality positive anchors for tracking.
Keep hard negatives only when they are plausible distractors but clearly not the target. If evidence is insufficient, report the dominant failure type.

Return JSON schema:

```
{
  "canonical_objects": [
    {
      "canonical_object_id": "...",
      "entity_query": "...",
      "object_summary": "...",
      "selected_anchors": [{"proposal_id": "...", "frame_idx": 0}],
      "negative_items": [
        {"proposal_id": "...", "frame_idx": 0,
         "negative_quality": "...", "discard_reason": "..."}
      ]
    }
  ],
  "segment_level_decision": {
    "coverage_confidence": "high|medium|low",
    "coverage_confidence_score": 0.0,
    "evidence_sufficiency": "sufficient|insufficient",
    "evidence_failure_type": "none|sample_missed_event|grounding_missing_target|...",
    "suggested_next_action": "keep_current|resample_evidence|rerun_grounding|manual_review"
  }
}
```

Evidence Review and Repair Routing Prompt

System: Review tracked visual evidence. Output VALID JSON only.

Input:

- Event/entity semantics
- Selected anchors and hard negatives
- Tube overlays or sampled review frames
- Programmatic diagnostics: area jump, center jump, empty gaps, discontinuities, anchor IoU, tube coverage
- Optional repair history

Instruction:

Judge whether the tracked object still matches the intended abnormal subject.
Use visual evidence as primary evidence and diagnostics as supporting evidence.
Do not treat gaps as automatic failures when explained by occlusion, scene cuts, or leaving the view. Diagnose the dominant failure mode and recommend the minimal repair callback.

Return JSON schema:

```
{
  "subject_correctness": "...",
  "spatial_localization_quality": "...",
  "identity_stability": "...",
  "temporal_continuity_stability": "...",
  "acceptability": "acceptable|borderline|bad",
  "root_cause": "...",
  "root_cause_category": "wrong_subject_selected|identity_drift|poor_localization|...",
  "recommended_callback": "fix_tracking|fix_qwen_selection|fix_sampling|..."
}
```

fragmented tracks, and incorrect re-acquisition, are treated as tube errors rather than evidence-selection errors when the verified anchors remain consistent with the event semantics. In these cases,

repair modifies only the affected propagated boxes or visibility intervals, preserving the manually verified anchors as identity references.

The repair process is conservative. It stops

when the evidence is acceptable, when the repair budget is exhausted, or when the case requires manual review. Ambiguous cases are not automatically repaired by changing semantics or switching to a different subject, since doing so could introduce reliable-looking but semantically incorrect spatial evidence.

C Evaluation Execution Protocol

This section summarizes the experimental protocol used to evaluate heterogeneous API models, local video MLLMs, and task-specific anomaly models. It complements the metric definitions in the main paper by specifying model scale reporting, input settings, task prompts, and evaluation flow.

C.1 Model Scale and Compute Reporting

Table 9 summarizes the parameter-scale information available for the evaluated model families. For open or locally executed models, we report the scale indicated by the released checkpoint or public model identifier. For hosted API models, we report parameter counts only when they are part of the public model name or documentation; otherwise, provider-side parameter counts are not disclosed to users of the API.

For compute accounting, we separate benchmark construction, local evaluation, and hosted API calls. All local GPU jobs use NVIDIA A100-SXM4-80GB GPUs. The construction pipeline consumed 169.90 GPU-hours when counting all jobs that occupied GPU time, including failed, cancelled, timed-out, and retried jobs; restricting to successfully completed construction jobs gives 100.51 GPU-hours. For benchmark evaluation, local video MLLMs consumed 343.35 GPU-hours in total, or 148.72 GPU-hours for successfully completed jobs only. Task-specific anomaly/VAD models consumed 145.34 GPU-hours in total, or 136.70 GPU-hours for successfully completed jobs only. Hosted API models are executed through provider services; provider-side serving hardware and GPU-hours are not observable from our side and are therefore not estimated.

C.2 Evaluation Workflow

We use a two-stage evaluation workflow. First, each model is run with a unified task interface to produce structured predictions. Second, the saved predictions are scored against benchmark annotations to compute binary normal/abnormal dis-

crimination, temporal grounding, spatio-temporal grounding, and QA metrics. This separation keeps model inference fixed while allowing deterministic offline scoring.

The prediction schema is task-specific. Detection outputs contain video-level normal/abnormal judgments and predicted abnormal spans. QA outputs contain one answer per question, with multiple-choice answers constrained to option letters and open-ended answers kept concise. Tube outputs contain segment-level anomaly-relevant subjects and normalized boxes or tubes.

C.3 Normal and Abnormal Flows

Normal and abnormal videos follow different task flows. For normal videos, models answer the detection prompt and the normal-video QA set; tube prediction is skipped because no abnormal subject tube is annotated. For abnormal videos, models first answer the detection prompt, which requires both a normal/abnormal decision and predicted abnormal time spans. When a video is predicted as abnormal, the tube prompt is applied to the model-predicted temporal span to localize the anomaly-relevant subject or region. Segment-level anomaly QA is evaluated separately using the corresponding benchmark question set.

For the main end-to-end grounding protocol, abnormal videos classified as normal are recorded as detection failures and are treated as having no valid temporal or spatio-temporal grounding prediction. Spatio-temporal grounding is therefore evaluated on tubes produced from model-predicted abnormal spans, rather than from ground-truth temporal intervals. For controlled diagnostic studies, we also run tube localization with benchmark-provided temporal intervals when appropriate, so that detection and temporal-localization errors can be separated from spatial-grounding ability.

C.4 Video Inputs and Prompt Tasks

API models are evaluated with native video input when supported. Local video MLLMs use their released video-loading conventions under zero-shot settings. Detection prompts use the video context required to decide normality and predict abnormal time spans. Tube prompts use clips cropped to the predicted abnormal segment in the main end-to-end setting, and to benchmark-provided segments only in controlled diagnostic settings.

The inference interface uses three task prompts.

Table 9: Parameter-scale information for evaluated model families. MoE denotes mixture-of-experts models, where the first number is the total saved parameter scale and the “A” number denotes the approximate activated parameter scale per token.

Model group	Evaluated models	Parameter-scale reporting
Task-specific anomaly models	Holmes-VAU; EventVAD; VADTree; AnyAnomaly	Evaluated with released checkpoints or official inference pipelines. Parameter counts are not consistently disclosed across these task-specific releases.
Video-centric MLLMs	VideoLLaMA3-7B; LLaVA-Video-7B-Qwen2; LLaVA-OneVision-1.5-4B/8B-Instruct; LLaVA-OneVision-Qwen2-7B; Video-XL-Pro-3B; Video-XL-2	Public model identifiers provide the reported scale when available: 3B, 4B, 7B, and 8B variants. LLaVA-Video and LLaVA-OneVision-Qwen2 use Qwen2-7B-scale backbones. Video-XL-2 is reported by model identifier rather than assigned an inferred parameter count.
API MLLMs with public scale in name	Qwen3-VL-235B-A22B-Instruct; Qwen3.6-35B-A3B; Qwen3.6-27B; Qwen3.5-397B-A17B; Qwen3.5-122B-A10B; Qwen3.5-9B	Dense models use the stated parameter scale, such as 9B and 27B. MoE models use total/activated notation, such as 235B-A22B, 397B-A17B, 122B-A10B, and 35B-A3B.
API MLLMs without disclosed scale	Gemini 3.1 Pro; Gemini 3 Flash; Gemini 3.1 Flash Lite; Qwen3.6-Plus; GLM-5V-Turbo; Kimi K2.5; Kimi K2.6	Provider-side parameter counts and serving hardware are not disclosed through the hosted APIs, so we report the public model names used for evaluation.

The detection prompt asks for a normal/abnormal decision and abnormal time spans. The QA prompt batches questions from the same video and requires a JSON answer for every question ID; multiple-choice questions are constrained to option letters, while open-ended answers are kept concise. The tube prompt asks the model to output anomaly-relevant subjects with boxes in a normalized coordinate system for the temporal segment under evaluation.

Evidence-assisted QA is handled separately from video-only QA. Questions that refer to marked subjects or spatial evidence are answered with annotated videos containing boxes or tube overlays, while video-only QA uses the original video input. Failed or incomplete requests are re-run before scoring; these reruns replace missing outputs and are not treated as repeated experimental trials.

C.5 Execution Across Model Families

API models are evaluated through their hosted inference interfaces with the model names listed in Section 4. We use native video input when available and do not estimate undisclosed provider-side serving details.

Local video MLLMs are run on GPU compute nodes using released checkpoints or official model releases, with the same task definitions and output schemas as API models. They are evaluated zero-shot under default inference settings unless a model-specific release requires a fixed video-loading or frame-sampling convention. We do not perform task-specific fine-tuning or hyperparameter search for these models.

Task-specific anomaly models are also run on GPU compute nodes using released checkpoints or official inference pipelines. Their native outputs are converted into the compatible subset of benchmark predictions, such as temporal spans, anomaly labels, explanations, or QA-style outputs when supported. Models without language or spatial-output capability are not forced into unsupported QA or tube tasks. All reported scores are computed from completed predictions under the task interfaces described above.