OPENVA-BENCH: EVALUATING OPEN-WORLD VIDEO ANOMALY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world video anomaly understanding requires models to recognize anomalies from known categories and make evidence-based inferences about anomalies outside the predefined category set. When available labels fail to describe an event, forcing it into a known category can produce an incorrect yet plausible interpretation. We introduce OpenVA-Bench to examine this risk through prediction accuracy, confidence reliability, and explanation quality. The benchmark integrates 1,330 videos from three datasets and constructs explicit Known–Unknown category splits, with Normal included in the Known set and Unknown Anomaly available as a rejection option. Across 25 Video-LLMs, stronger known-class recognition does not consistently coincide with reliable unknown rejection. Failures concentrate on semantically related category pairs, incorrect predictions can receive high confidence, and coherent explanations can conflict with reference event descriptions. Neither a larger model nor more input frames consistently yields better performance in the evaluated configurations. These findings expose a gap between recognizing known anomalies and providing reliable judgments in open-world applications. OpenVA-Bench supports systematic evaluation of this gap through prediction, confidence, and reasoning.

1 INTRODUCTION

Real-world videos contain anomalous events that extend beyond the categories anticipated when a system is designed. Models deployed in these settings should recognize events from known categories and use visual evidence to reject events outside the predefined category set. Such interpretations should reflect the observed actions and interactions that define an event. Recent advances in video large language models (Video-LLMs) have expanded their capabilities for event description and anomaly reasoning (Huang et al., 2025; Zhang et al., 2025b). Reliable open-world use requires determining when none of the available categories adequately describes an observed event.

Figure 1 illustrates this requirement with Theft and Robbery. Both involve taking another person's property, but the confrontation between people changes the nature of the event. When Theft is available and Robbery is absent, choosing Theft preserves a superficial similarity while missing this distinction. The model may then build a plausible explanation around an inapplicable label. A reliable model should recognize the mismatch and leave room for an interpretation beyond the available categories. Ideally, models should use visual evidence to reject anomalies beyond the predefined category set. A basic requirement is to recognize when none of the available categories adequately describes the observed event, rather than force it into a related known category. We evaluate this ability through unknown-anomaly rejection under a controlled candidate set.

Motivated by this perspective, we introduce OpenVA-Bench, a benchmark for evaluating open-world video anomaly understanding in Video-LLMs. We use an open-set evaluation protocol to make category applicability testable under controlled Known–Unknown splits. The benchmark integrates 1,330 videos from UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), and MSAD (Zhu et al., 2024), and harmonizes their category labels. Models select a Known category, including Normal, or output Unknown Anomaly for anomalous events outside the predefined category set. The default split retains semantically similar categories on opposite sides of the Known–Unknown boundary, requiring models to attend to the event details that distinguish them.

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.

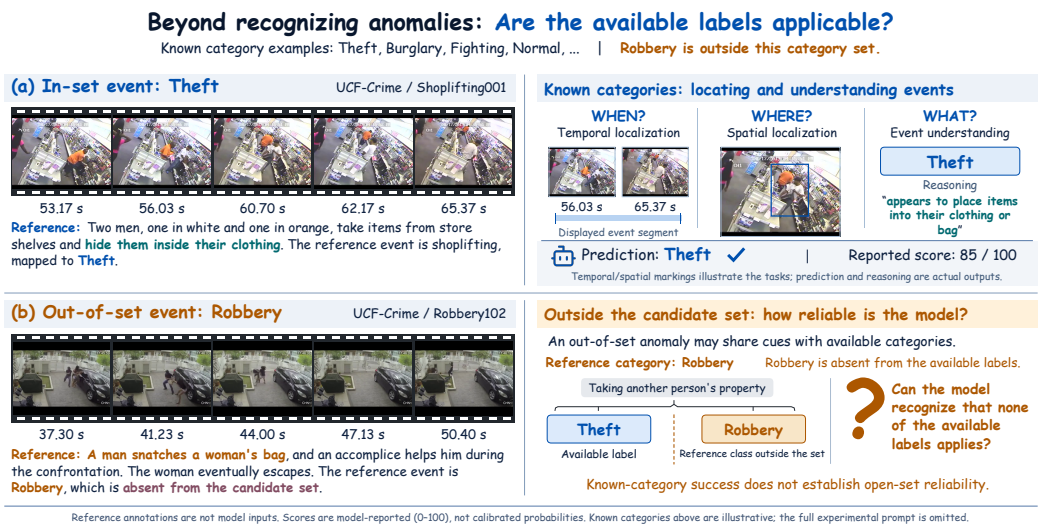

Figure 1: Known and unknown anomalies with similar visual cues.

The reliability of an event judgment also depends on how the model supports it. A final label alone does not reveal whether uncertainty is expressed or whether an explanation describes the underlying event. We therefore evaluate prediction, confidence, and reasoning: known-class recognition and unknown rejection; agreement between reported confidence and actual performance; and explanation consistency with reference descriptions and internal logical coherence. We further vary input frame count, candidate category configurations, and prompt definitions to examine how reliably models handle changes in the evaluation setting.

Across 25 Video-LLMs, stronger known-class recognition does not consistently coincide with reliable unknown rejection. Models repeatedly assign events outside the predefined category set to related known categories, report high confidence for incorrect predictions, and produce coherent explanations that conflict with reference descriptions. Choosing a larger variant or providing more frames does not consistently improve prediction performance, while detailed prompt definitions leave substantial rejection errors. These findings reveal a gap between known-class recognition and the reliability required for open-world applications.

Our main contributions are as follows:

- We introduce OpenVA-Bench, with a unified video collection, category mappings, and Known–Unknown splits for evaluating how Video-LLMs handle events beyond predefined categories.
- We establish an evaluation framework spanning prediction, confidence, and reasoning to examine category applicability and explanation reliability in open-world video anomaly understanding.
- We evaluate 25 Video-LLMs across model families and parameter scales, revealing recurring category confusions and examining sensitivity to candidate categories and input settings.

2 Related Work

2.1 Video Anomaly Datasets and Open-World Evaluation

Video anomaly datasets provide a foundation for video anomaly understanding. UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020) provide video-level labels for weakly supervised video anomaly detection (VAD). MSAD (Zhu et al., 2024) covers diverse scenes, viewpoints, and environmental variations to assess anomaly detection under complex surveillance conditions. These datasets broaden the coverage of anomalous events and scenes.

Existing studies have also explored open-set settings for video anomaly analysis. OpenVAD (Zhu et al., 2022) considers settings in which known and novel anomalies coexist at test time and uses uncertainty modeling to improve the handling of unknown anomalies. Wu et al. (2024) introduce open-vocabulary video anomaly detection (OVVAD), extending the task to the detection and category recognition of both seen and unseen anomalies. UBnormal (Acsintoae et al., 2022) provides a synthetic video dataset with pixel-level annotations and disjoint anomaly types across training and test sets, supporting supervised open-set video anomaly detection. These studies advance video anomaly analysis through open-set detection and open-vocabulary category recognition.

2.2 Video Anomaly Understanding with MLLMs

Advances in multimodal large language models enable video anomaly analysis to combine visual content with natural-language reasoning. Vad-R1 (Huang et al., 2025) enhances anomaly reasoning through a perception-to-cognition chain of thought and reinforcement learning. Holmes-VAU (Zhang et al., 2025b) combines hierarchical instruction data with anomaly-focused temporal sampling to support long-video anomaly understanding at multiple temporal granularities. Meanwhile, general-purpose Video-LLMs such as Qwen3-VL (Bai et al., 2025) have advanced long-video understanding, temporal grounding, and multimodal reasoning.

VALU (He et al., 2026) includes an Anomaly Description Choice (ADC) task that evaluates whether models can identify an appropriate anomaly description or reject the provided descriptions. VirtueBench (Yu et al., 2026) evaluates abstention in long-video question answering when sampled frames omit the evidence needed to answer.

While these benchmarks examine description selection and abstention when visual evidence is insufficient, OpenVA-Bench evaluates whether models recognize when the available categories do not apply to an observed anomaly, and assesses the confidence and explanations accompanying their judgments.

3 The Proposed OpenVA-Bench Benchmark

OpenVA-Bench examines category applicability in open-world video anomaly understanding through an open-set evaluation protocol. It uses explicit Known–Unknown category splits and evaluates prediction, confidence, and reasoning. Figure 2 summarizes the task and evaluation framework.

3.1 Known–Unknown Evaluation Protocol

Let C_K denote the Known categories provided at test time. Models predict from $C_K \cup \{\text{Unknown Anomaly}\}$, with Known and Unknown defined relative to this candidate set.

Known samples. Known samples have ground-truth labels in C_K, comprising Abuse, Arrest, Burglary, Explosion, People Falling, Theft, Fighting, and Normal by default.

Unknown samples. Under the default split, these samples have ground-truth categories outside C_K: Assault, Traffic Accident, Robbery, Riot, Water Incident, Object Falling, Fire, Vandalism, and Shooting. Their original labels support benchmark construction and category-wise evaluation but are not selectable outputs. The prompt includes some unknown-type examples to explain rejection without expanding the selectable labels. All Unknown samples should receive Unknown Anomaly; any Known prediction, including Normal, is a rejection failure.

3.2 Benchmark Construction

Video collection. UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), and MSAD (Zhu et al., 2024) contribute 290, 800, and 240 videos, respectively, totaling 1,330 videos: 760 anomalous and 570 normal. Table A1 in the appendix reports source and category distributions.

Category harmonization. To address differences in category naming and granularity across datasets, we harmonize their labels into a unified category set. We merge Stealing and Shoplifting into Theft, standardize Traffic Accident, Car Accident, and Road Accident as Traffic Accident, and merge Arson and Fire into Fire without distinguishing intent. We retain Abuse/Assault,

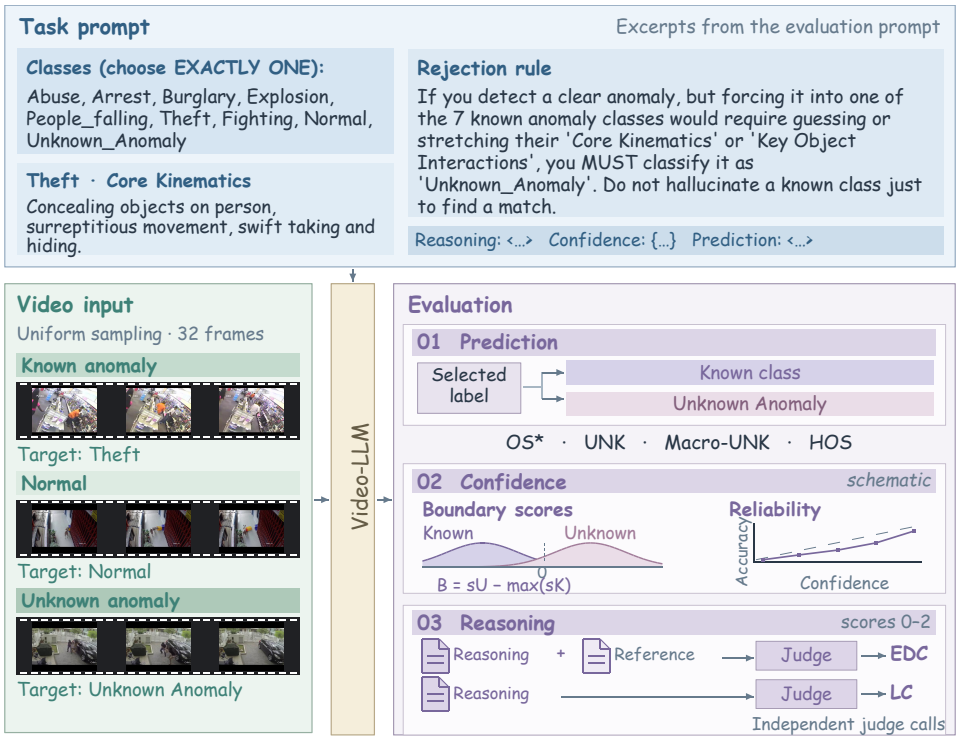

Figure 2: OpenVA-Bench: prediction, confidence, and reasoning evaluation.

Theft/Robbery, and Explosion/Fire as separate categories. Complete mappings appear in Appendix A.1.

Split construction. The default split contains 897 Known videos (327 known anomalies and 570 normal videos) and 433 Unknown videos. Section 4.3 varies the number and composition of known anomaly categories to assess sensitivity to the candidate set.

Reference annotations. Reasoning evaluation uses reference event descriptions from Holmes-VAU (Zhang et al., 2025b) for XD-Violence and Alert-CLIP (Zhu et al., 2026) for UCF-Crime and MSAD. These reference descriptions are compared with model-generated explanations to assess semantic agreement on the core event and key details. Reference matching, coverage, and examples are detailed in Appendix A.2.

3.3 Task Definition and Evaluation Metrics

Given video frames, Known categories and their definitions, and the Unknown Anomaly option, models return a final label (Prediction), scores for all candidate labels (Confidence), and a supporting explanation (Reasoning). The task prompt, output format, and input protocol appear in Appendix B.

3.3.1 Open-set Prediction

Known-class accuracy (OS^*). We macro-average per-class accuracy over C_K, including Normal, so larger categories do not dominate.

Unknown rejection (UNK / Macro-UNK). UNK measures the fraction of Unknown videos correctly rejected. Macro-UNK averages rejection rates equally across unknown categories, providing a class-balanced complement to UNK. The default split has eight Known and nine Unknown categories.

Harmonic Open-set Score (HOS). Following Bucci et al. (2020), we summarize recognition and rejection using

$$\mathrm{HOS} = \frac{2 \times OS^* \times \mathrm{UNK}}{OS^* + \mathrm{UNK}}. \tag{1}$$

We report all component metrics and examine the mappings from true Unknown categories to incorrectly predicted Known categories to identify concentrated rejection failures. Further metric details appear in Appendix C.1.

3.3.2 CONFIDENCE EVALUATION

Models report a score from 0 to 100 for each candidate label; scores need not sum to 100. Let s_U denote the Unknown Anomaly score and $s_K^{\max} = \max_{c \in C_K} s_c$ the highest Known score.

Boundary awareness. We compare Known and Unknown distributions of

$$B(x) = s_U(x) - s_K^{\max}(x). \tag{2}$$

Positive values favor Unknown Anomaly over every Known category; negative values favor at least one Known category. Boundary scores use unscaled model scores.

Confidence reliability. Confidence is the score assigned to the final Prediction, divided by 100. We bin predictions by confidence and compare each bin's mean confidence with its per-sample accuracy, exposing discrepancies such as overconfidence (Guo et al., 2017). Binning and scaling details appear in Appendix C.2.

3.3.3 REASONING EVALUATION

Event Description Consistency (EDC). The judge compares the model explanation with the reference annotation, assessing semantic agreement on the core event and key details.

Logical Coherence (LC). Given only the explanation, the judge assesses whether its stated evidence supports its interpretation or conclusion without critical gaps or contradictions.

DeepSeek-V4-Flash (DeepSeek-AI et al., 2026) scores EDC and LC independently on a 0–2 scale: 2 indicates full satisfaction, 1 partial satisfaction, and 0 a critical error or insufficient evaluable content. We normalize each dimension's mean as $100 \times \text{mean}/2$. Scoring criteria, judge prompts, and aggregation procedures appear in Appendices D, F, and C.3, respectively.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluated models. We evaluate 25 locally deployed and API-accessed Video-LLM variants across model families and parameter scales. Table A4 in Appendix B.1 lists the evaluated variants, recorded API identifiers, and access methods.

Evaluation protocol. We uniformly sample 32 frames across each full video and present them chronologically. All models share the Known categories, definitions, instructions, and Prediction/Confidence/Reasoning requirements, with interface-specific input formatting. Appendix B provides the task prompt and input protocol; Section 4.3 and Appendix E specify changes for each ablation.

4.2 EVALUATION RESULTS AND FINDINGS

Table 1 summarizes prediction and reasoning results, and Figure 4(a) examines confidence. Prediction metrics use all 1,330 videos, counting missing or invalid predictions as errors. Confidence analysis requires valid predictions and complete category scores.

Known-class Recognition and Unknown Rejection. Model rankings differ between recognition and rejection. Kimi-K2.6 exceeds Doubao-Seed-2.1-Turbo in OS^* (71.89% vs. 64.73%) but trails in UNK (58.66% vs. 75.06%) and HOS (64.60% vs. 69.51%). Doubao-Seed-2.1-Turbo achieves the highest HOS among evaluated models. Higher known-class accuracy does not necessarily translate into better unknown-anomaly rejection.

Aggregate rejection also conceals category-level weaknesses. Qwen3.5-397B-A17B scores 63.97% UNK but 46.09% Macro-UNK: it rejects 116/139 Traffic Accident videos (83.45%), compared with 3/19 Robbery and 0/11 Assault videos.

Table 1: Open-set prediction and reasoning results.

Model	Open-set prediction				Reasoning	
	OS^* ↑	UNK ↑	Macro-UNK ↑	HOS ↑	EDC ↑	LC ↑
API-accessed models						
Doubao-Seed-2.0-Lite (ByteDance Seed, 2026a)	67.46	70.90	65.64	69.14	79.95	98.76
Doubao-Seed-2.1-Turbo (ByteDance Seed, 2026b)	64.73	75.06	64.45	69.51	84.27	99.38
Gemini-2.5-Flash (Comanici et al., 2025)	65.59	58.66	45.52	61.93	79.20	97.33
Kimi-K2.6 (Moonshot AI, 2026)	71.89	58.66	49.23	64.60	83.94	98.91
Qwen3.5-397B-A17B (Qwen Team, 2026)	64.88	63.97	46.09	64.42	81.94	99.36
Locally deployed models						
GLM-4.6V-Flash (Z.ai, 2025)	53.83	39.95	41.87	45.87	74.13	95.92
InternVL3.5-1B (Wang et al., 2025)	35.91	32.79	25.54	34.28	64.33	90.18
InternVL3.5-2B	40.90	27.71	22.70	33.04	70.73	92.70
InternVL3.5-4B	56.43	55.66	43.81	56.04	76.45	90.07
InternVL3.5-8B	54.72	28.87	22.85	37.80	71.71	90.71
InternVL3.5-14B	56.03	43.65	40.61	49.07	72.25	95.29
InternVL3.5-38B	63.36	46.65	38.92	53.74	78.71	93.64
LLaVA-OneVision-2-8B-Instruct (An et al., 2026)	53.17	26.79	20.70	35.63	74.76	95.75
MiMo-VL-7B-no-thinking (Xiaomi MiMo, 2025)	38.81	66.28	64.56	48.95	65.62	93.01
MiMo-VL-7B-thinking	55.30	69.98	64.07	61.78	69.55	92.10
MiniCPM-V-4.5 (Yu et al., 2025)	48.55	43.19	39.74	45.71	74.29	95.01
Molmo2-8B (Clark et al., 2026)	45.36	51.50	44.11	48.23	74.46	95.27
Ovis2.5-9B (Lu et al., 2025)	49.20	36.26	31.06	41.75	72.84	94.06
Qwen3-VL-2B (Bai et al., 2025)	41.03	4.39	5.42	7.93	67.38	88.56
Qwen3-VL-4B	52.60	40.88	32.80	46.00	73.21	95.52
Qwen3-VL-8B	44.78	39.72	37.81	42.10	73.51	97.10
Qwen3-VL-30B-A3B	59.87	49.42	48.02	54.15	78.67	97.82
Qwen3-VL-32B	53.78	55.89	43.71	54.82	79.27	99.02
VideoLLaMA3-2B (Zhang et al., 2025a)	22.52	63.74	62.87	33.28	55.53	87.74
VideoLLaMA3-7B	45.29	32.79	26.59	38.04	68.27	87.49

All scores are percentages (higher is better).

Misclassification Preferences of Unknown Anomalies. Rejection failures concentrate on particular category pairs (Figure 3). Doubao-Seed-2.1-Turbo maps 7/12 Robbery failures to Theft and 12/28 Riot failures to Fighting; Qwen3-VL-8B maps 19/45 Riot failures to Fighting. These errors illustrate the failure highlighted in Figure 1: an event outside the predefined category set is assigned to a related known category.

Unknown-to-Normal errors form a distinct failure mode: Qwen3-VL-8B assigns 66/79 Traffic Accident rejection failures (83.54%) to Normal, indicating missed anomaly recognition rather than selection of another known anomaly category.

Confidence Behavior. Doubao-Seed-2.1-Turbo has mean boundary scores of -91.09/37.64 on Known/Unknown videos, compared with -90.16/-19.88 for Qwen3-VL-8B (Figure 4(a)). The latter favors Known options on average even for Unknown samples.

Reported confidence can exceed accuracy. On 1,329 valid records, Qwen3.5-397B-A17B reports 96.22% mean confidence but achieves 81.49% per-sample accuracy. For Qwen3-VL-8B's 1,277 predictions with confidence at least 90%, the corresponding values are 98.56% and 71.10%. Here, accuracy includes correct Known classification and Unknown rejection within the valid subset; it differs from OS^*, which macro-averages Known-category accuracy. Self-reported certainty is therefore insufficient as a standalone signal of a correct classification.

Effect of Model Scale. HOS varies non-monotonically with scale: Qwen3-VL's 4B/8B/32B variants score 46.00%/42.10%/54.82%, and InternVL3.5's 4B/8B/14B/38B variants score

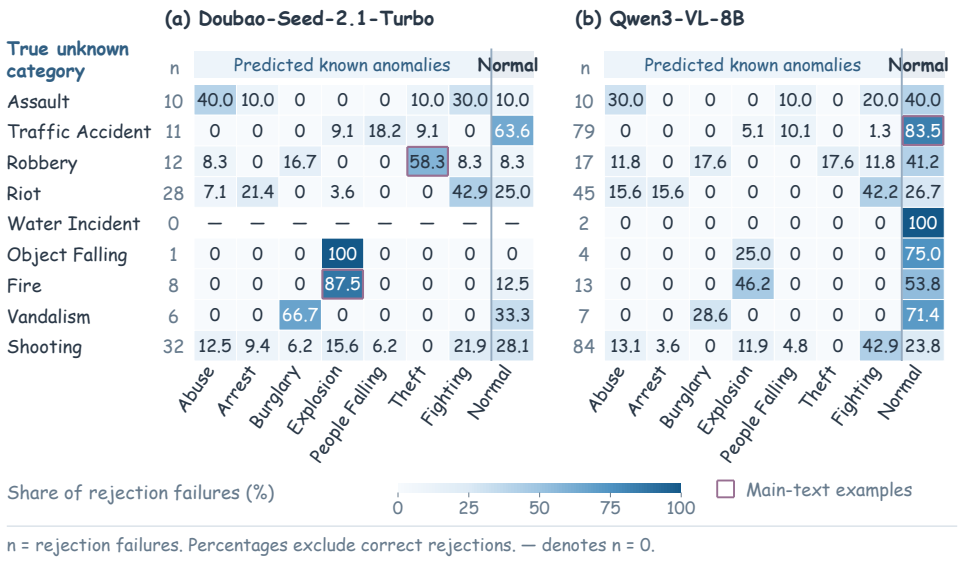

Figure 3: Unknown-to-known errors as percentages of each row's rejection failures.

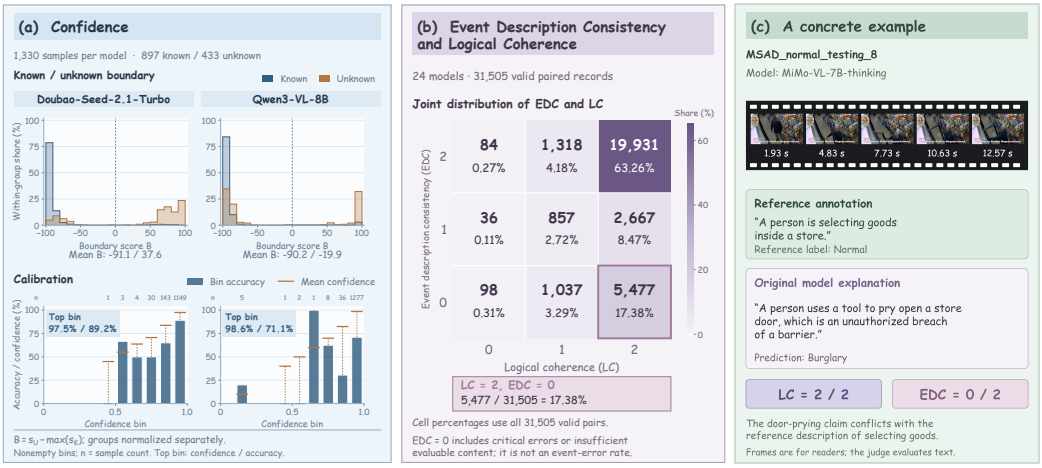

Figure 4: Confidence reliability, joint EDC–LC scores, and an illustrative example.

56.04%/37.80%/49.07%/53.74%. InternVL3.5-38B improves OS^* over 4B (63.36% vs. 56.43%) but reduces UNK (46.65% vs. 55.66%), yielding lower HOS. Training varies across these models; the practical finding is that choosing a larger variant does not consistently improve rejection.

Event Description Consistency and Logical Coherence. Across 24 models, 31,505 model–video records have valid EDC and LC scores, with coverage varying by model. Table 1 uses the same valid pairs for each model's two means. As shown in Figure 4(b), 5,477 of these records (17.38%) receive LC$=2$ but EDC$=0$, showing that coherent explanations can be inconsistent with reference event descriptions.

For MSAD_normal_testing_8, the reference describes a customer selecting goods. MiMo-VL-7B-thinking instead describes "a person uses a tool to pry open a store door" and interprets this as unauthorized entry (Figure 4(c)). The judge assigns LC$=2$ because the stated action supports that interpretation, but EDC$=0$ because it conflicts with the reference. This case shows how an explanation can justify a wrong interpretation.

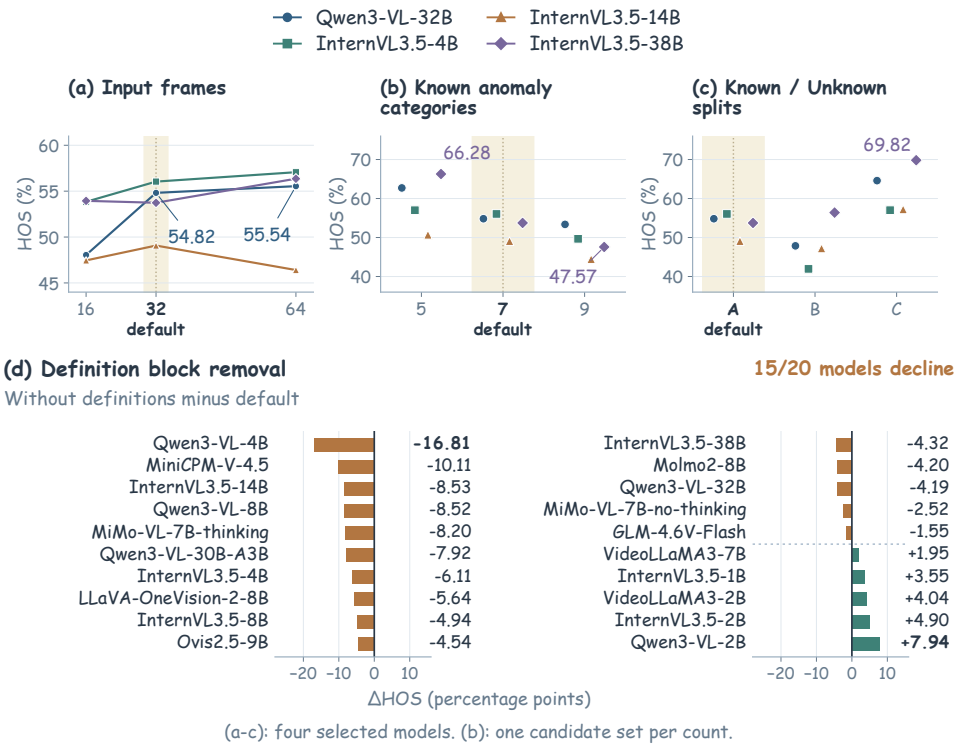

Figure 5: Sensitivity to frame count, candidate categories, splits, and definitions.

4.3 Further Analysis and Discussion

We examine four evaluation settings within the 20-model ablation cohort using the same 1,330 videos (Figure 5; Table 2). The default uses 32 frames and seven known anomaly categories, plus Normal and Unknown Anomaly. When candidate categories change, we update each sample's Known–Unknown status and recompute metrics. We report HOS alongside OS^*, UNK, and Macro-UNK.

Input Frame Count. With the category split and prompt fixed, uniformly sampling 16, 32, and 64 frames over each video's full duration yields Qwen3-VL-32B HOS values of 48.04%, 54.82%, and 55.54% (Figure 5(a)). From 16 to 64 frames, UNK rises from 45.27% to 58.66% and Macro-UNK from 35.37% to 48.26%, while OS^* changes from 51.19% to 52.74%, indicating gains mainly in rejection. InternVL3.5-4B also improves, with HOS values of 53.84%, 56.04%, and 57.07%. However, InternVL3.5-14B declines from 49.07% at 32 frames to 46.41% at 64. Additional frames primarily improve unknown-anomaly rejection for Qwen3-VL-32B, with diminishing gains beyond 32 frames. The contrasting responses show that increasing the input budget does not reliably improve performance under the same candidate set.

Number of Known Anomaly Categories. We compare five, seven, and nine known anomaly categories while retaining Normal and Unknown Anomaly (Figure 5(b)). Relative to the default, the five-category setting removes Abuse and Burglary, while the nine-category setting adds Traffic Accident and Vandalism. Known/Unknown sample counts are 874/456, 897/433, and 1,047/283, respectively, with Normal included in Known. Qwen3-VL-32B's HOS decreases from 62.71% to 54.82% and 53.36%. From five to nine categories, its OS^* falls from 69.67% to 53.02%, whereas UNK changes from 57.02% to 53.71%. InternVL3.5-38B's HOS also declines from 66.28% to 47.57%, with weaker recognition and rejection. Across the tested configurations, 19 of 20 models achieve lower HOS with nine known anomaly categories than with five, highlighting their difficulty in maintaining open-set classification performance under an expanded candidate set.

Table 2: Candidate category configurations and video counts.

Candidate category set			Videos	
Configuration	No.	Known anomaly categories	Known	Unknown
Default / Split A	7	Abuse, Arrest, Burglary, Explosion, People Falling, Theft, Fighting	897	433
Varying category count				
Five-category setting	5	Arrest, Explosion, People Falling, Theft, Fighting	874	456
Nine-category setting	9	Abuse, Arrest, Burglary, Explosion, People Falling, Theft, Fighting, Traffic Accident, Vandalism	1,047	283
Varying category composition (seven categories)				
Split B	7	Assault, Traffic Accident, Robbery, Water Incident, Object Falling, Fire, Vandalism	786	544
Split C	7	Arrest, Burglary, Explosion, People Falling, Riot, Shooting, Traffic Accident	1,086	244

Category counts exclude Normal; Known video counts include 570 normal videos.

Known–Unknown Category Splits. Three splits with seven known anomaly categories produce substantial variation (Figure 5(c)). InternVL3.5-38B achieves HOS values of 53.74%, 56.39%, and 69.82% on Splits A, B, and C; Qwen3-VL-32B achieves 54.82%, 47.86%, and 64.58%. InternVL3.5-38B's corresponding UNK values are 46.65%, 47.98%, and 65.16%. For Fighting, unknown in both B and C, Qwen3-VL-32B rejects 14 of 131 samples under B; 105 of its 117 failures are classified as Assault. Under C, which excludes Assault, it rejects 102 samples (77.86%). The concentration of Fighting errors in Assault reveals systematic confusion between semantically similar known and unknown anomalies.

Visual Feature Definitions. Using the default frames and split, we remove the prompt definition block while retaining candidate labels, general evidence rules, rejection requirements, and output format (Figure 5(d)). HOS decreases for 15 of 20 models, including Qwen3-VL-32B (54.82% to 50.62%) and InternVL3.5-4B (56.04% to 49.93%). For Qwen3-VL-32B, both OS^* and UNK decline, from 53.78% to 49.99% and from 55.89% to 51.27%, respectively. Conversely, InternVL3.5-2B's HOS improves from 33.04% to 37.95% as UNK rises from 27.71% to 63.51%, despite OS^* falling from 40.90% to 27.05%. Definitions improve HOS for most models, but substantial rejection errors remain even with the full prompt. Explicit category guidance helps without resolving the difficulty of recognizing when known labels do not apply.

5 Conclusion

OpenVA-Bench exposes a gap between recognizing known anomalies and handling events beyond predefined categories. The tested increases in model scale and frame count do not yield consistent gains, and substantial rejection errors persist even with detailed prompts. This gap is especially consequential when incorrect judgments remain confident or are accompanied by coherent explanations. The evaluated Video-LLMs therefore fall short of the reliability needed for standalone open-world anomaly understanding. Progress requires models that recognize when known categories are inadequate, communicate uncertainty, and explain events without forcing them into an available label. OpenVA-Bench provides a basis for assessing these requirements.

AI USE STATEMENT

We used OpenAI Codex to aid or polish writing, for retrieval and discovery (e.g., finding related work), for research ideation or execution, to draft sections of the paper. AI assistance also supported reviewing the evaluation methodology and interpreting experimental results. LLM-based reasoning evaluation is described in Section 3.3 and Appendices D and F. The authors take responsibility for the scientific claims, reported results, and final content of this work.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20143–20153, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Acsintoae_UBnormal_New_Benchmark_for_Supervised_Open-Set_Video_Anomaly_Detection_CVPR_2022_paper.html.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence. *arXiv preprint arXiv:2605.25979*, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the Effectiveness of Image Rotation for Open Set Domain Adaptation. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123610409.pdf.

ByteDance Seed. Seed2.0, 2026a. URL https://seed.bytedance.com/en/seed2. Official model documentation. Accessed September 26, 2026.

ByteDance Seed. Seed2.1, 2026b. URL https://seed.bytedance.com/en/seed2_1. Official model documentation. Accessed September 26, 2026.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

DeepSeek-AI et al. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Yixiao He, Menghao Zhang, Haifeng Sun, Jing Wang, Kangheng Lin, Jinghan Wang, Chenye Xu, Pengfei Ren, Qi Qi, and Jingyu Wang. VALU: A Benchmark for Video Anomaly Temporal Localization and Understanding at Multiple Semantic Levels. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1252–1296, 2026. doi: 10.18653/v1/2026.acl-long.56. URL https://aclanthology.org/2026.acl-long.56/.

Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought. *arXiv preprint arXiv:2505.19877*, 2025. URL https://arxiv.org/abs/2505.19877.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 Technical Report. *arXiv preprint arXiv:2508.11737*, 2025. URL https://arxiv.org/abs/2508.11737.

Moonshot AI. Kimi-K2.6, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6. Official model documentation. Accessed September 26, 2026.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026. URL https://qwen.ai/blog?id=qwen3.5. Official model documentation. Accessed September 26, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123750324.pdf.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-Vocabulary Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18297–18307, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Open-Vocabulary_Video_Anomaly_Detection_CVPR_2024_paper.html.

Xiaomi MiMo. MiMo-VL-7B-RL-2508, 2025. URL https://github.com/XiaomiMiMo/MiMo-VL. Official model documentation. Accessed September 26, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe. *arXiv preprint arXiv:2509.18154*, 2025. URL https://arxiv.org/abs/2509.18154.

Xueqing Yu, Bohan Li, Yan Li, and Zhenheng Yang. VirtueBench: Evaluating Trustworthiness under Uncertainty in Long Video Understanding. *arXiv preprint arXiv:2603.07071*, 2026. URL https://arxiv.org/abs/2603.07071.

Z.ai. GLM-4.6V-Flash, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash. Official model documentation. Accessed September 26, 2026.

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. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.pdf.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://msad-dataset.github.io/.

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-CLIP: Abnormality-aware Latent-Enhanced Representation Tuning of CLIP for Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Zhu_Alert-CLIP_Abnormality-aware_Latent-Enhanced_Representation_Tuning_of_CLIP_for_Video_Anomaly_CVPR_2026_paper.pdf.

Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards Open Set Video Anomaly Detection. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2208.11113.

A Data Composition and Reference Annotations

A.1 Dataset composition and category harmonization

The benchmark contains 1,330 videos from UCF-Crime, XD-Violence, and MSAD. Table A1 gives the complete source-by-category distribution. The default candidate set includes seven anomaly categories and Normal, yielding 897 Known videos and 433 Unknown videos. Unknown status is defined by exclusion from this candidate set, rather than by whether an event occurred in a model's pretraining data.

Table A1: Video counts by source and category under the default split.

Category and status		Source dataset			Total
Category	Status	UCF-Crime	XD-Violence	MSAD	
Known anomalies					
Abuse	Known	2	8	0	10
Arrest	Known	5	0	0	5
Burglary	Known	13	0	0	13
Explosion	Known	21	91	8	120
People Falling	Known	0	0	22	22
Theft	Known	26	0	0	26
Fighting	Known	5	120	6	131
Known anomalies subtotal		72	219	36	327
Normal	Known	150	300	120	570
Unknown anomalies					
Assault	Unknown	3	0	8	11
Traffic Accident	Unknown	23	96	20	139
Robbery	Unknown	5	0	14	19
Riot	Unknown	0	101	0	101
Water Incident	Unknown	0	0	5	5
Object Falling	Unknown	0	0	10	10
Fire	Unknown	9	0	12	21
Vandalism	Unknown	5	0	6	11
Shooting	Unknown	23	84	9	116
Unknown anomalies subtotal		68	281	84	433
Total		290	800	240	1,330

Table A2 summarizes the category normalization used in the manuscript. Different spellings and separators are presentation variants, whereas combining Stealing and Shoplifting into Theft, or Arson and Fire into Fire, changes category granularity. The Fire category does not distinguish deliberate ignition from other fire events. Related categories retained separately include Abuse/Assault, Theft/Robbery, and Explosion/Fire. Retaining these labels does not imply that every video has mutually exclusive visual characteristics.

Table A2: Source labels and harmonized categories.

Source label or stored variant	Harmonized category
Stealing; Shoplifting; Theft	Theft
Traffic Accident; Car Accident; Road Accident; Traffic_accident	Traffic Accident
Arson; Fire; Fire_Arson	Fire
People_falling	People Falling
Water_incident	Water Incident
Object_falling	Object Falling
Abuse; Arrest; Burglary; Explosion; Fighting	Corresponding label retained
Assault; Robbery; Riot; Vandalism; Shooting	Corresponding label retained
Normal	Normal

Category frequencies are uneven. Normal accounts for 570 videos, whereas the smallest anomaly categories contain five videos each. Consequently, we report both sample-weighted and class-balanced rejection measures; the benchmark is not rebalanced by duplicating minority-category videos when computing the reported metrics. The exact denominators for each category remain those in Table A1.

A.2 Reference text matching, coverage, and provenance

Prediction evaluation uses harmonized category labels. EDC instead compares each generated explanation with a reference event description. These two annotation objects serve different purposes and are not interchangeable. Table A3 reports the stored reference files and the number of distinct videos represented in the final EDC records. These are file-level provenance and coverage statistics, not evidence of who originally wrote or verified the descriptions.

Table A3: Reference files and matched video coverage.

Video source	Reference file	Videos
MSAD	`Final_ucf_fgclip_translated_qwen_test.json`	240
UCF-Crime	`Final_ucf_fgclip_translated_qwen_test.json`	254
XD-Violence	`xd_database_test.json`	800
UCF-Crime	`ucf_database_test.json`	35

Counts are distinct matched videos; total coverage is 1,329 of 1,330 videos.

The reference pipeline matches normalized video identifiers within each dataset. For UCF-Crime, a fallback file is consulted only when the primary reference is absent; matching uses exact identifiers rather than fuzzy names. The stored annotation contains event descriptions or captions, together with available video and label information. The final records preserve the assembled text in `annotation` and its source file in `annotation_source`. Across the archived EDC records, the same dataset–video identifier has a consistent reference text. This consistency check concerns the stored inputs, not their factual accuracy.

One UCF-Crime video, `Normal_Videos_100_x264`, lacks a matched reference in the available files. It remains in prediction evaluation but cannot contribute an EDC score. Thus, the existence of 1,330 prediction samples does not imply 1,330 valid reference-based reasoning records for every model.

For example, the stored reference for `MSAD_normal_testing_8` describes a normal store scene and a person selecting goods. The reference for `Assault_2` describes a man in a blue T-shirt knocking down a person wearing a black hoodie. These examples illustrate the contents supplied to the judge; they are not additional video-level verification performed for this appendix.

B Inference Protocol and Task Prompt

B.1 Models and input configuration

Table A4 lists the 25 evaluated model variants and their access modes. Provider identifiers are preserved where they are recorded in the API requests. Names containing parameter scales identify the evaluated variants; they should not be interpreted as a consistent measure of active parameters across dense and mixture-of-experts architectures.

The main experiment samples 32 frames across the full video and presents them in temporal order. Frame indexing, resizing, image tokenization, and input packing follow the model adapter. For example, the available Qwen3-VL adapter uses endpoint-spaced frame indices, whereas the MiniCPM and InternVL adapters use segment-based sampling. Therefore, equal frame counts do not imply identical pixel tensors or identical visual token budgets. The common task prompt is supplied together with the resulting visual inputs.

The inspected API request records specify 32 uniformly sampled frames, an output token limit of 2,048, temperature 0, and top-p of 1. Thinking is disabled for the two Doubao variants, Kimi, and Qwen3.5; the Gemini request specifies a zero reasoning-token budget. These are request settings,

Table A4: Evaluated models and access identifiers.

Model / family	API identifier or local variants	Access
Doubao-Seed-2.1-Turbo (ByteDance Seed, 2026b)	`doubao-seed-2-1-turbo-260628`	Volcengine Ark
Doubao-Seed-2.0-Lite (ByteDance Seed, 2026a)	`doubao-seed-2-0-lite-260215`	Volcengine Ark
Kimi-K2.6 (Moonshot AI, 2026)	`Pro/moonshotai/Kimi-K2.6`	SiliconFlow
Qwen3.5-397B-A17B (Qwen Team, 2026)	`Qwen/Qwen3.5-397B-A17B`	SiliconFlow
Gemini-2.5-Flash (Comanici et al., 2025)	`google/gemini-2.5-flash`	OpenRouter
Qwen3-VL (Bai et al., 2025)	2B, 4B, 8B, 30B-A3B, 32B; Instruct	Local
InternVL3.5 (Wang et al., 2025)	1B, 2B, 4B, 8B, 14B, 38B	Local
MiMo-VL-7B (Xiaomi MiMo, 2025)	RL-2508; thinking and no-thinking	Local
VideoLLaMA3 (Zhang et al., 2025a)	2B, 7B	Local
LLaVA-OneVision-2 (An et al., 2026)	8B-Instruct	Local
GLM (Z.ai, 2025)	4.6V-Flash	Local
MiniCPM-V (Yu et al., 2025)	4.5	Local
Molmo2 (Clark et al., 2026)	8B	Local
Ovis2.5 (Lu et al., 2025)	9B	Local

rather than guarantees about a provider's internal execution. Local generation settings are adapter-dependent: for example, the saved Qwen3-VL configuration disables sampling with a 2,048-token limit, whereas MiniCPM-V-4.5 enables sampling with temperature 0.7. Uniform prompts should not be interpreted as uniform decoding settings.

B.2 Default prompt and response fields

The following block reproduces the shared task prompt extracted from the available main-experiment adapters. The prompt text is identical across the 19 inspected adapter files. Model-specific frame tokens and input wrappers are outside this block. Machine-readable labels use underscores, such as `People_falling` and `Unknown_Anomaly`; the main text uses spaces for readability. The examples in the Unknown definition illustrate rejection and do not add selectable output labels.

```
You are an expert video anomaly detection analyst operating in an
OPEN-WORLD setting.  Your task is to classify the video into EXACTLY ONE
of the following 9 categories based on the provided strict visual
evidence rules.
Classes (choose EXACTLY ONE):
1) Abuse
2) Arrest
3) Burglary
4) Explosion
5) People_falling
6) Theft
7) Fighting
8) Normal
9) Unknown_Anomaly
```

```
Visual Feature Definitions (IMPORTANT - STRICTLY adhere to these
boundaries):
**Abuse** - Core Kinematics:  Aggressive striking, grappling, choking,
forceful restraint, or violent pushing.  - Key Object Interactions:
Weapon (stick/knife) used against a person/animal; person/animal being
struck, held down, or overpowered.  - Spatiotemporal Dynamics:  Sudden
escalation to sustained, forceful physical contact with intent to harm.
**Arrest** - Core Kinematics:  Sudden forceful approach, physical
restraint, and controlled subduing.  - Key Object Interactions:  Law
enforcement personnel applying handcuffs or weapons to detain a resisting
individual.  - Spatiotemporal Dynamics:  Abrupt transition from
individual autonomy to enforced physical control by authorities.
**Burglary** - Core Kinematics:  Forcibly prying, kicking, or climbing to
breach barriers.  - Key Object Interactions:  Unauthorized manipulation
of doors, windows, cash registers, or safes.  - Spatiotemporal Dynamics:
Sudden forceful action on access points, followed by unauthorized
interior rummaging.
**Explosion** - Core Kinematics:  Sudden, violent expansion; rapid energy
release; forceful ejection of materials.  - Key Object Interactions:
Objects fragmenting, deforming, or being violently displaced; intense
flames and/or smoke plumes appear.  - Spatiotemporal Dynamics:  Abrupt
transition from a stable state to chaotic, high-energy dispersal;
shockwaves may cause secondary shaking.
**People_falling** - Core Kinematics:  Sudden, uncontrolled descent; loss
of upright posture.  - Key Object Interactions:  Person separates from
support surface (ground, vehicle, stool).  - Spatiotemporal Dynamics:
Rapid transition from stable locomotion or stance to prone/impact state.
**Theft** - Core Kinematics:  Concealing objects on person, surreptitious
movement, swift taking and hiding.  - Key Object Interactions:
Merchandise or valuables moved from public display to hidden on body or
in bag.  - Spatiotemporal Dynamics:  Covert interaction with item
followed by departure without payment or authorization.
**Fighting** - Core Kinematics:  Sudden, forceful strikes (punches,
kicks, pushes) and aggressive grappling (wrestling, restraining).  - Key
Object Interactions:  Direct, harmful physical contact between human
bodies; potential use of handheld weapons/objects as striking tools.  -
Spatiotemporal Dynamics:  Rapid escalation from non-contact to sustained,
reciprocal violent physical engagement.
**Normal** - Core Kinematics:  People walking, standing, sitting, or
driving vehicles at steady, predictable speeds.  - Key Object
Interactions:  Vehicles moving along roads, people interacting with
static objects like benches or store shelves.  - Spatiotemporal Dynamics:
Continuous, smooth motion within designated pathways (sidewalks, roads,
lanes).
**Unknown_Anomaly** - Core Kinematics:  A visually obvious, dangerous,
disruptive, or illegal event is occurring, but its specific actions DO
NOT strictly match the Core Kinematics of the 7 specific anomalies listed
above (e.g., Arson, Vandalism, Shooting, Traffic Accidents).  - Rule:
Use this to cleanly reject out-of-distribution (OOD) novel anomalies.
Evidence rule (Open-World Setting):  - Base your decision on explicit
visual evidence.  If you observe any distinct actions that match the Core
Kinematics of an anomaly class, prioritize classifying it as that
anomaly.  - Only classify as 'Normal' if the ENTIRE video strictly
consists of routine, harmless activities without any suspicious events.
- CRITICAL ANTI-HALLUCINATION RULE: If you detect a clear anomaly, but
forcing it into one of the 7 known anomaly classes would require guessing
or stretching their 'Core Kinematics' or 'Key Object Interactions', you
MUST classify it as 'Unknown_Anomaly'.  Do not hallucinate a known class
just to find a match.
Task:
1) Reasoning:  Describe ONLY what is visually observable.  If selecting
```

```
'Unknown_Anomaly', briefly state what the novel anomaly actually is and
why it doesn't fit the 7 known classes.
2) Confidence:  Output integer scores for ALL 9 classes.
- All scores must be integers in [0,100].
3) Prediction:  Output the class name that has the HIGHEST score in the
Confidence list above.  (This MUST be consistent with your scores).
Response Format (STRICT): ---------------------------------- Reasoning:
<1-3 sentences, explicit observations only> Confidence:  {Abuse:  X,
Arrest:  X, Burglary:  X, Explosion:  X, People_falling:  X, Theft:  X,
Fighting:  X, Normal:  X, Unknown_Anomaly:  X} Prediction:  <one of:
Abuse / Arrest / Burglary / Explosion / People_falling / Theft / Fighting
/ Normal / Unknown_Anomaly> ----------------------------------
```

Scores are independently reported integers from 0 to 100 and are not required to sum to 100. The prompt requests that Prediction agree with the largest reported score. Evaluation nevertheless uses the parsed final Prediction as the decision and its assigned score as confidence; the boundary score is a separate analysis quantity and is not used to silently replace a model's decision.

C Metric Definitions and Aggregation

C.1 Prediction metrics and rejection failures

Let $\mathcal{D}$ be the fixed set of 1,330 videos, y_i a harmonized source category, C_K the current Known set including Normal, and C_U the remaining source categories. Let u denote the output Unknown Anomaly. The evaluation target is $t_i = y_i$ if $y_i \in C_K$, and $t_i = u$ otherwise. Write $n_c = |\{i : y_i = c\}|$ and let $\hat{y}_i$ be the final prediction. Missing, failed, or invalid predictions are treated as incorrect rather than removed from prediction denominators.

$$\mathrm{Acc}_c = \frac{1}{n_c} \sum_{i:y_i=c} \mathbf{1}[\hat{y}_i = c], \tag{A1}$$

$$OS^* = \frac{1}{|C_K|} \sum_{c \in C_K} \mathrm{Acc}_c, \tag{A2}$$

$$\mathrm{UNK} = \frac{\sum_{i:y_i \in C_U} \mathbf{1}[\hat{y}_i = u]}{\sum_{c \in C_U} n_c}, \tag{A3}$$

$$\text{Macro-UNK} = \frac{1}{|C_U|} \sum_{c \in C_U} \frac{1}{n_c} \sum_{i:y_i=c} \mathbf{1}[\hat{y}_i = u], \tag{A4}$$

$$\mathrm{HOS} = \frac{2OS^*\mathrm{UNK}}{OS^* + \mathrm{UNK}}. \tag{A5}$$

All metrics above are defined on a $[0, 1]$ scale and displayed as percentages. HOS is set to zero if both components are zero. Macro-UNK weights each Unknown category equally, whereas UNK weights categories in proportion to their sample counts. Neither metric alone describes every category's behavior.

For the displayed unknown-to-known confusion analysis, let $F_c = \{i : y_i = c,\ \hat{y}_i \in C_K\}$ be the valid Known-label errors for an Unknown category c. The entry for Known label k is

$$P(k \mid F_c) = 100\, \frac{\sum_{i \in F_c} \mathbf{1}[\hat{y}_i = k]}{|F_c|}. \tag{A6}$$

Correct rejections are excluded. The row denominator is reported explicitly, and an empty row is undefined rather than zero. The two models shown in the manuscript's confusion analysis have no missing or invalid predictions on Unknown videos, so these row denominators equal their rejection-failure counts. An assignment to Normal is a rejection failure just as an assignment to a Known anomaly category is.

C.2 CONFIDENCE VALIDITY, BOUNDARY SCORES, AND RELIABILITY

Confidence analysis uses records with a valid final prediction and complete scores for all current candidate labels. This subset can differ by model. Missing confidence records are not assigned an artificial confidence of zero. With $s_{i,u}$ the Unknown score and $s_{i,k}$ the score for Known label k, define

$$B_i = s_{i,u} - \max_{k \in C_K} s_{i,k}, \qquad c_i = s_{i,\hat{y}_i}/100. \tag{A7}$$

The boundary score uses the original 0–100 score scale and lies in $[-100, 100]$. Positive values favor Unknown; negative values favor at least one Known label; zero is a tie. Histograms use bins of width 10 and are normalized separately within the ground-truth Known and Unknown subsets. Means are descriptive summaries and do not determine the overlap of the two distributions.

For reliability analysis, predictions are grouped into ten bins $[0, 0.1), \ldots, [0.8, 0.9), [0.9, 1]$. If I_b denotes the records in a nonempty bin,

$$\mathrm{conf}_b = \frac{1}{|I_b|} \sum_{i \in I_b} c_i, \qquad \mathrm{acc}_b = \frac{1}{|I_b|} \sum_{i \in I_b} \mathbf{1}[\hat{y}_i = t_i]. \tag{A8}$$

Empty bins are omitted. Confidence above accuracy indicates overconfidence within that bin. The sample accuracy here includes correct Known classification and Unknown rejection; it is not the Known-only macro-average OS^*. The high-confidence analysis uses $c_i \geq 0.9$ and reports its own sample count.

C.3 REASONING COVERAGE AND PAIRED AGGREGATION

Reasoning records are matched by model, dataset, and normalized video identifier. Let P_m contain the records for model m with valid integer scores in $\{0, 1, 2\}$ for both EDC and LC and matching explanation text. For $d \in \{\mathrm{EDC}, \mathrm{LC}\}$, the reported mean is

$$S_{m,d} = \frac{100}{2|P_m|} \sum_{i \in P_m} r_{m,i,d}. \tag{A9}$$

The two scores for a model therefore share a denominator. Missing scores are excluded from this paired reasoning analysis, unlike the treatment of missing predictions in classification metrics. A valid zero is retained: zero can express an actual scoring judgment, including insufficient evaluable content. An absent score and a score of zero are not interchangeable.

Table A5 reports coverage explicitly. The pooled analysis contains 31,505 valid pairs from 24 models; Doubao-Seed-2.1-Turbo has no reasoning evaluation in the reported table. The LC$= 2$, EDC$= 0$ cell contains 5,477 pairs, or 17.38% of this valid paired set. This is a record-weighted joint-score proportion, not a per-video prevalence or a verified event-description error rate.

C.4 FIGURE AND TABLE DISPLAY CONVENTIONS

In Figure 1, UCF-Crime frames show a Known Theft event and an Unknown Robbery event. Panel (a) contains a prediction, self-reported score, and reasoning excerpt from one model response; temporal and spatial markings are manual illustrations. Panel (b) poses the category-applicability question. Reference summaries, abbreviated candidate lists, and omitted prompt details serve the illustration; reference summaries are not model inputs. Scores are not calibrated probabilities. In Figure 2, video labels are shown for readers; the two reasoning judges receive text only.

Figure 3 uses true Unknown categories as rows and predicted Known categories as columns. Each row is normalized over its rejection failures (n), excluding correct rejections. The two models share a 0–100% scale; outlines mark examples discussed in the text, and a dash denotes $n = 0$.

Figure 4(a) shows boundary distributions and confidence calibration for Doubao-Seed-2.1-Turbo and Qwen3-VL-8B, with 1,330 valid records each; binning follows Appendix C.2. Panel (b) pools 31,505 EDC–LC pairs from 24 models. Panel (c) presents five frames and the model's verbatim explanation for readers; frames are not judge inputs. The interpretation of EDC$= 0$ is given in Appendix C.3.

Table A5: Valid EDC, LC, and paired records by model.

Model	EDC valid	LC valid	Paired valid
Doubao-Seed-2.1-Turbo	0	0	0
Doubao-Seed-2.0-Lite	1,329	1,330	1,329
Kimi-K2.6	1,329	1,330	1,329
Qwen3.5-397B-A17B	1,329	1,330	1,329
Gemini-2.5-Flash	1,310	1,311	1,310
MiMo-VL-7B-thinking	1,322	1,323	1,322
InternVL3.5-4B	1,329	1,330	1,329
Qwen3-VL-32B	1,329	1,330	1,329
Qwen3-VL-30B-A3B	1,329	1,330	1,329
InternVL3.5-38B	1,329	1,330	1,329
InternVL3.5-14B	1,328	1,329	1,328
MiMo-VL-7B-no-thinking	1,280	1,281	1,280
Molmo2-8B	1,204	1,204	1,204
Qwen3-VL-4B	1,329	1,330	1,329
GLM-4.6V-Flash	1,237	1,238	1,237
MiniCPM-V-4.5	1,233	1,234	1,233
Qwen3-VL-8B	1,329	1,330	1,329
Ovis2.5-9B	1,329	1,330	1,329
VideoLLaMA3-7B	1,327	1,328	1,327
InternVL3.5-8B	1,329	1,330	1,329
LLaVA-OneVision-2-8B-Instruct	1,329	1,330	1,329
InternVL3.5-1B	1,329	1,330	1,329
VideoLLaMA3-2B	1,329	1,330	1,329
InternVL3.5-2B	1,329	1,330	1,329
Qwen3-VL-2B	1,329	1,330	1,329

EDC and LC use paired records; zero coverage denotes unavailable evaluation.

Figure 5(a–c) shows four models, with shading marking default settings and one category combination per category count. Panel (d) shows all 20 ablation models, read down the left block and then the right. ΔHOS is HOS without definitions minus default HOS, in percentage points.

Table 1 groups models by access method, then alphabetically by family; specified total parameter counts increase within each family, and same-size variants are adjacent. Table A4 lists five API variants and 20 local variants; API aliases are recorded request identifiers. Complete run metadata is addressed in Appendix B.1.

D Reasoning Judge and Scoring Criteria

The archived evaluation records identify the judge as `deepseek-v4-flash`. The stored scoring timestamps span September 6–10, 2026 (UTC). EDC and LC are obtained through separate calls. EDC receives the assembled reference annotation and the generated explanation; LC receives only the explanation. Neither judge receives video frames or the separately reported final prediction. A reference annotation may itself contain a source label, which is part of the EDC reference rather than a model prediction input.

EDC uses the output field `event_recognition_score`. LC uses `logical_coherence_score`, not the auxiliary `reasoning_quality_score`. Auxiliary fields are retained in the logs but are not substituted for the two paper metrics. Table A6 summarizes the scoring criteria; Appendix F reproduces the complete prompts from the scoring scripts.

EDC permits different wording and missing nonessential details. For a Normal reference, recognizing the absence of abnormal activity can receive full credit. Recognizing only a consequence, a broad event family, or one of several important annotated events can receive partial credit when the core event is incompletely recognized. The intermediate diagnostic flags do not mechanically determine the score. LC concerns the support offered by the evidence stated within the explanation,

Table A6: EDC and LC scoring criteria.

Score	EDC: event description consistency	LC: internal logical coherence
2	Main event and key evidence recognized; minor omissions allowed. Correct recognition of normality can suffice for a Normal reference.	The main reasoning chain is clear and consistent; its conclusion reasonably follows from its stated evidence.
1	Partial event recognition, secondary clues, a related broad event family, or only one of multiple important events.	A logical connection is present but incomplete, weak, ambiguous, or affected by a noticeable gap.
0	Different or contradictory event, false normalization, invented anomaly in a Normal reference, or non-evaluable content.	Contradictory, irrelevant, logically invalid, or non-evaluable reasoning.

including whether the central conclusion follows without a major contradiction or reasoning gap. A coherent explanation can therefore conflict with the reference event and receive high LC but low EDC.

The saved repair pipeline recovers final explanations from supported text, Markdown, and JSON formats and merges repeated model–dataset–video tasks. It does not invent explanations from category labels, coordinates, or unfinished internal reasoning. Unresolved extraction failures remain missing. Existing scores are reused only when the relevant stored inputs and judge identifier match. These parsing and matching checks establish input consistency; they do not establish human agreement with the judge.

The scoring scripts default to temperature 0 and disabled thinking. The historical records preserve model identifiers, inputs, scores, and reuse metadata, but not prompt hashes for every original call. The complete prompts provided here are the archived script templates associated with the repaired pipeline; a cryptographically verified prompt history is unavailable. No human–judge agreement statistic or multi-judge agreement result is asserted in this appendix.

E Ablation Configurations

All ablations reuse the same 1,330 videos. Changing the candidate set changes the target labels and metric denominators, while Normal remains Known and Unknown Anomaly remains a selectable rejection option. The default candidate set and Split A are identical.

The frame-count ablation compares 16, 32, and 64 frames over each full video while holding the category split and task prompt fixed. The category-count ablation compares five, seven, and nine Known anomaly categories, excluding Normal from the category count. The five-category setting removes Abuse and Burglary from the default set; the nine-category setting adds Traffic Accident and Vandalism. Their Known/Unknown sample counts are 874/456, 897/433, and 1,047/283.

Split B contains Assault, Traffic Accident, Robbery, Water Incident, Object Falling, Fire, and Vandalism as Known anomalies, giving 786 Known and 544 Unknown videos. Split C contains Arrest, Burglary, Explosion, People Falling, Riot, Shooting, and Traffic Accident, giving 1,086 Known and 244 Unknown videos. Each split has seven Known anomaly categories plus Normal. These settings change the category composition as well as the membership of the Known and Unknown subsets; they are not controlled interventions on a single semantic boundary.

The prompt ablation removes the entire Visual Feature Definitions block while retaining the candidate labels, general evidence rules, rejection requirement, and requested Reasoning/Confidence/Prediction format. It keeps the default 32-frame input and split. This simultaneously removes category descriptions and the Unknown examples, so its results concern the whole definition block rather than either component in isolation.

F Complete Reasoning-Evaluation Prompts

The placeholders `{annotation}` and `{reasoning}` below are populated with the stored reference and explanation. The doubled braces in JSON templates are retained from the Python format-

ting strings and become ordinary braces in the actual request. Line wrapping is adjusted only for readability.

F.1 EDC SYSTEM AND USER PROMPTS

System prompt.

```
You are a strict evaluator for video anomaly reasoning quality.
You judge whether the model-generated Reasoning correctly describes the key visual event indicated by the Video Annotation.
Return JSON only.
```

User prompt.

```
You are a strict evaluator for video anomaly reasoning recognition quality.
Your task is to judge whether the model-generated Reasoning recognizes and is grounded in the core event evidence described in the Video Annotation.
You are NOT judging whether the model's final prediction is correct. You are NOT judging the video directly. You are NOT requiring the Reasoning to reproduce every detail in the Annotation. You are ONLY judging whether the event and visual evidence described in the Reasoning correspond to the Video Annotation.
Input:
Video Annotation: {annotation} Reasoning: {reasoning}
Evaluation rules:
1. Use the Video Annotation as the reference evidence.
2. Identify the core event and key visual evidence described in the Annotation. If both a broad label and a detailed description are provided, prioritize the detailed description.
3. Evaluate semantic correspondence rather than exact wording. The Reasoning does not need to mention every annotated detail. Missing minor attributes, background details, locations, colors, or secondary cues should not reduce the score when the same core event is clearly recognized.
4. Focus on the main event described by the Reasoning. If the Reasoning captures the same core event and provides meaningful supporting visual evidence, treat it as full recognition even if some supporting details are missing.
5. If the Reasoning captures only part of the annotated event, a consequence, a secondary clue, or a broadly related event without sufficiently recognizing the core event, treat it as partial recognition.
6. If the Reasoning describes a substantially different event, contradicts the Annotation, describes an abnormal annotated event as normal, or invents an abnormal event for a Normal/routine Annotation, treat it as no reliable recognition.
7. For Normal/routine annotations, correctly recognizing ordinary activity or the absence of abnormal events counts as recognition.
8. Ignore category-space analysis, class names, and statements about whether the event belongs to predefined anomaly classes. Do not increase or decrease the score because the Reasoning assigns the event to a correct or incorrect class. Evaluate only the visual event content expressed in the Reasoning.
9. Do not infer visual evidence that is not explicitly stated in the Reasoning.
10. If the Reasoning is empty, says "No reasoning provided", is malformed, excessively repetitive, or impossible to evaluate, set all scores to 0 and use event_recognition_level = "unclear".
11. Keep the explanation concise and based only on the Annotation and Reasoning.
Before assigning scores, identify:
```

* annotation_core_event * reasoning_recognized_event * recognized_visual_evidence * missed_key_evidence

Also determine:

* annotation_is_normal_or_routine: true/false * reasoning_recognizes_normality: true/false * annotation_has_multiple_core_events: true/false * reasoning_recognizes_one_core_event: true/false * reasoning_describes_related_event_family: true/false * reasoning_has_concrete_visual_evidence: true/false

These intermediate judgments are diagnostic only. They should help explain the score but should not mechanically determine it.

Scoring rules:

Use integer scores from 0 to 2.

For event_recognition_score:

2 = full recognition: The Reasoning recognizes the same core event as the Annotation and provides meaningful visual evidence supporting that recognition. Minor missing details are acceptable.

1 = partial recognition: The Reasoning is meaningfully related to the annotated event but recognizes only part of it, a consequence, a secondary clue, or a broad related event without sufficiently capturing the core event.

0 = no reliable recognition: The Reasoning describes a substantially different or contradictory event, incorrectly normalizes an abnormal event, invents an anomaly for a Normal/routine video, or is not evaluable.

For key_evidence_coverage_score:

2 = strong coverage: The Reasoning captures the core event and sufficient key visual evidence.

1 = partial coverage: The Reasoning captures some relevant evidence but misses important evidence needed to characterize the core event.

0 = no meaningful coverage: The Reasoning provides little or no visual evidence corresponding to the annotated event.

For evidence_specificity_score:

2 = concrete: The Reasoning contains specific observable actions, objects, people, interactions, movements, or visible consequences.

1 = limited: The Reasoning contains relevant but broad, vague, or incomplete visual evidence.

0 = absent: The Reasoning contains no meaningful concrete visual evidence or is not evaluable.

Derive event_recognition_level from event_recognition_score:

* 2 -> "full" * 1 -> "partial" * 0 -> "none"

If the Reasoning is empty, malformed, excessively repetitive, or impossible to evaluate, use "unclear".

The recognition_failure_type should describe the main reason for incomplete or failed recognition:

* "none": the core event is recognized. * "minor_detail_missing": the core event is recognized but minor details are missing. * "consequence_only": only a consequence or secondary clue is recognized. * "broad_event_family_only": only a broadly related event is recognized. * "generic_only": only a vague abnormal situation is described. * "one_event_in_multi_event": only part of a multi-event annotation is recognized. * "wrong_event": a substantially different or contradictory event is described. * "normalization": an abnormal annotated event is incorrectly described as normal. * "no_reasoning": the Reasoning is empty or not evaluable. * "unclear": the issue cannot be determined.

The failure type is diagnostic. Do not determine the numeric score solely from the failure type.

Final consistency check:

```
Before returning the JSON, verify that the score, level, failure type,
recognized evidence, missed evidence, and explanation form a reasonable
and internally consistent judgment.
Return JSON only.  Do not output markdown or extra text.
{{
"event_recognition_level":  "full/partial/none/unclear",
"event_recognition_score":  0,
"key_evidence_coverage_score":  0,
"evidence_specificity_score":  0,
"annotation_is_normal_or_routine":  false,
"reasoning_recognizes_normality":  false,
"annotation_has_multiple_core_events":  false,
"reasoning_recognizes_one_core_event":  false,
"reasoning_describes_related_event_family":  false,
"reasoning_has_concrete_visual_evidence":  false,
"annotation_core_event":  "...",
"reasoning_recognized_event":  "...",
"recognized_visual_evidence":  ["..."],
"missed_key_evidence":  ["..."],
"recognition_failure_type":  "none/minor_detail_missing/consequence_only/
broad_event_family_only/generic_only/one_event_in_multi_event/wrong_
event/normalization/no_reasoning/unclear",
"explanation":  "short explanation within 30 words"
}}
```

F.2 LC SYSTEM AND USER PROMPTS

System prompt. `You are a strict evaluator for the logical quality of model-generated video anomaly reasoning.`

```
You judge whether the Reasoning is concrete, evidence-grounded within the
text, internally coherent, and logically connected.
Focus on the main reasoning chain rather than minor peripheral flaws.
Return JSON only.
```

User prompt. `You are a strict evaluator for the logical quality of model-generated video anomaly reasoning.`

```
Your task is to judge whether the model-generated Reasoning is logically
sound, evidence-grounded within the text, and internally coherent.
You are NOT judging whether the final prediction is correct.  You are NOT
judging whether the Reasoning matches the ground-truth annotation.  You
are NOT judging whether the described visual evidence actually appears in
the video.  You are NOT checking factual correctness against the video.
You are ONLY judging the quality of the given Reasoning text itself.
Input:  Reasoning:  {reasoning}
Evaluation rules:
1.  Evaluate only the Reasoning text itself.  Do not infer missing video
content.
2.  Judge whether the Reasoning provides concrete observable visual
evidence, such as actions, people, objects, interactions, movements,
scene changes, or visible consequences.
3.  Judge whether the main reasoning chain connects the stated evidence
to its interpretation or conclusion in a logically reasonable way.
4.  Judge whether the Reasoning is internally consistent and avoids major
contradictions.
5.  Judge whether the Reasoning relies on unsupported assumptions,
unsupported details, causal over-interpretation, or conclusions that
substantially exceed its own stated evidence.
6.  Distinguish the main reasoning chain from minor peripheral details.
A minor speculative phrase, uncertain attribute, or non-essential
```

```
unsupported detail should not by itself make an otherwise sound Reasoning
partially sound.
7. A short but concrete and logically valid Reasoning should not be
penalized merely because it is brief.
8. Generic claims such as "there is an anomaly", "the scene is
suspicious", "abnormal behavior occurs", or "this may be dangerous" are
not sufficient evidence unless supported by concrete observable details.
9. A Reasoning does not need to prove its conclusion with certainty. A
plausible interpretation that is reasonably supported by the stated
visual evidence can still be logically sound.
10. Do not treat an inference as unsupported merely because alternative
explanations are possible. Treat it as unsupported only when it requires
a substantial unstated assumption or clearly exceeds what the stated
evidence can reasonably justify.
11. Explicit absence of abnormal cues may be used as part of the
reasoning. For example, ordinary observable actions together with the
stated absence of aggression, force, restraint, unauthorized entry,
destruction, or other abnormal interactions may reasonably support a
normal or routine interpretation.
12. Do not penalize the Reasoning merely because the described evidence
might be wrong in the actual video. Since the video is not provided,
evaluate only whether the reasoning is coherent and adequately supported
by the evidence stated within the text itself.
13. If the Reasoning is empty, says "No reasoning provided", is
malformed, excessively repetitive, or impossible to evaluate, set
reasoning_quality_level to "unclear" and all scores to 0.
14. Keep the explanation concise and based only on the Reasoning text.
Before assigning scores, determine the following intermediate judgments:
* has_concrete_visual_evidence: true/false * has_clear_logical_
connection: true/false * has_unsupported_inference: true/false * has_
self_contradiction: true/false * has_irrelevant_content: true/false *
is_generic_or_vague: true/false * is_empty_or_malformed: true/false
Scoring principles:
1. Score each dimension independently according to its own definition.
2. reasoning_quality_score should reflect the overall quality of the
main reasoning chain.
3. Minor flaws, peripheral speculation, minor unsupported details, or
non-essential inaccuracies in wording should not automatically determine
the overall score.
4. If the Reasoning is empty, malformed, excessively repetitive, or
impossible to evaluate, set all scores to 0 and reasoning_quality_level
to "unclear".
Scoring rules:
Use integer scores from 0 to 2.
For reasoning_quality_score:
2 = sound reasoning: The main reasoning chain contains concrete and
relevant evidence, the interpretation or conclusion reasonably follows
from that evidence, and there is no major contradiction or unsupported
leap that undermines the main conclusion.
1 = partially sound reasoning: The Reasoning contains meaningful
evidence and/or a plausible logical connection, but the main reasoning
chain is incomplete, weak, ambiguous, noticeably over-interpreted, or
partly unsupported.
0 = unsound or not evaluable: The main reasoning chain is fundamentally
unsupported, contradictory, irrelevant, logically invalid, empty,
malformed, excessively repetitive, or impossible to evaluate.
For evidence_groundedness_score:
```

```
2 = strong evidence grounding: The Reasoning contains concrete and
specific observable evidence, such as specific actions, objects, people,
interactions, movements, scene changes, or visible consequences.
1 = partial evidence grounding: The Reasoning contains some relevant
observable evidence, but it is vague, incomplete, generic, or
insufficiently specific.
0 = no meaningful evidence grounding: The Reasoning provides little or
no concrete observable evidence, or is impossible to evaluate.
Important: Evaluate this dimension independently. Do NOT lower
evidence_groundedness_score merely because the conclusion is speculative,
logically weak, or contains an unsupported inference, as long as the
Reasoning itself contains concrete observable evidence.
For logical_coherence_score:
2 = strong logical coherence: The main reasoning chain is clear and
internally consistent, and the interpretation or conclusion reasonably
follows from the stated evidence.
1 = partial logical coherence: A logical connection is present, but it
is incomplete, weak, ambiguous, or contains a noticeable reasoning gap.
0 = poor or absent logical coherence: The main reasoning chain is
contradictory, irrelevant, logically invalid, or impossible to evaluate.
Important: Evaluate the main reasoning chain. Do not lower this score
solely because of a minor peripheral unsupported detail that does not
materially affect the main conclusion.
For speculation_score:
2 = substantial speculation: The Reasoning contains major unsupported
assumptions, invented causal claims, or over-interpretation that
materially affects the main reasoning chain.
1 = limited speculation: The Reasoning contains some unsupported
inference, uncertain interpretation, or over-interpretation, but the main
reasoning remains partly or mostly grounded.
0 = little or no meaningful speculation: The Reasoning stays close to
its stated observable evidence and avoids significant unsupported
assumptions.
If the Reasoning is empty, malformed, or impossible to evaluate, also use
speculation_score = 0 and mark is_empty_or_malformed = true.
Derive reasoning_quality_level from reasoning_quality_score:
* reasoning_quality_score = 2 -> "sound" * reasoning_quality_score = 1 ->
"partially_sound" * reasoning_quality_score = 0 and is_empty_or_malformed
= false -> "unsound" * reasoning_quality_score = 0 and is_empty_or_
malformed = true -> "unclear"
The main_issue should be the most important problem in the main reasoning
chain:
* "none": No major issue affects the main reasoning chain.
* "vague_evidence": The Reasoning lacks sufficiently concrete observable
evidence.
* "unsupported_inference": The main interpretation or conclusion
substantially exceeds what the stated evidence can reasonably support.
* "self_contradiction": A contradiction materially affects the main
reasoning chain.
* "unsupported_detail": The Reasoning contains unsupported details, but
they are more local or peripheral than the main conclusion.
* "irrelevant_reasoning": A substantial portion of the Reasoning focuses
on content that does not contribute to the main reasoning chain.
* "empty_or_malformed": The Reasoning is empty, excessively repetitive,
malformed, or impossible to evaluate.
* "unclear": The main issue cannot be determined.
Final consistency check:
```

```
Before returning the JSON, verify that the scores, reasoning_quality_
level, main_issue, intermediate judgments, and explanation are mutually
reasonable.
Do not force different dimensions to have the same score.
A Reasoning may have, for example:
* high evidence groundedness but lower logical coherence; * strong
logical coherence but some speculation; * concrete evidence with a minor
unsupported detail while still receiving reasoning_quality_score = 2 if
the main reasoning chain remains sound.
Use main_issue = "none" when no major issue materially affects the main
reasoning chain.
Use main_issue = "unsupported_detail" rather than "unsupported_inference"
when the unsupported content is peripheral and does not materially affect
the main conclusion.
Return JSON only.  Do not output markdown or extra text.
{{
"reasoning_quality_level":  "sound/partially_sound/unsound/unclear",
"reasoning_quality_score":  0,
"evidence_groundedness_score":  0,
"logical_coherence_score":  0,
"speculation_score":  0,
"has_concrete_visual_evidence":  true,
"has_clear_logical_connection":  true,
"has_unsupported_inference":  false,
"has_self_contradiction":  false,
"has_irrelevant_content":  false,
"is_generic_or_vague":  false,
"is_empty_or_malformed":  false,
"main_issue":  "none/vague_evidence/unsupported_inference/self_
contradiction/unsupported_detail/irrelevant_reasoning/empty_or_malformed/
unclear",
"explanation":  "short explanation within 30 words"
}}
```

G LIMITATIONS

Computational and API budgets limit evaluation to selected models and settings, while API updates may affect reproducibility. The three source datasets and limited category splits cover only part of real-world anomaly diversity; taxonomy, semantic relationships, and sample distributions also influence evaluation difficulty. Unknown is defined relative to the test-time candidate set, so our results do not establish generalization to categories absent from pretraining.

Reasoning evaluation depends on reference texts and an LLM judge. EDC measures agreement with annotations and is constrained by their accuracy and completeness. LC measures internal coherence, not factual correctness or agreement with the final prediction. Scores may also vary with the judge and prompt. Future work should broaden data and category coverage and assess evaluation reliability through human verification and alternative judges.