VALU: A Benchmark for Video Anomaly Temporal Localization and Understanding at Multiple Semantic Levels

Yixiao He*, **Menghao Zhang***, **Haifeng Sun**, **Jing Wang**, **Kangheng Lin**, **Jinghan Wang**, **Chenye Xu**, **Pengfei Ren**, **Qi Qi**†, **Jingyu Wang**
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
{heyixiao,zhangmenghao,hfsun,wangjing,linkangheng}@bupt.edu.cn
{wangjhbh,xuchenye,rpf,qiqi8266,wangjingyu}@bupt.edu.cn

Abstract

Video anomaly understanding (VAU) is critical for real-world scenarios. Recent advances in Video Large Language Models (Video-LLMs) enhance the ability of VAU models to describe and interpret anomalies. However, progress in anomaly localization is still limited by two key issues. First, most existing video anomaly datasets only annotate segments that are clearly inconsistent with the context, often omitting subsequent segments that are semantically part of the same abnormal event. Second, the field lacks systematic evaluation protocols. To bridge these gaps, we introduce **VALU**, a new benchmark that explicitly defines anomalies across five semantic levels and provides comprehensive temporal boundaries and detailed textual descriptions for each. Based on these annotations, we design three evaluation tasks that comprehensively assess models' capabilities across different dimensions, including temporal grounding, anomaly localization, and anomaly detail discrimination. Evaluation results reveal persistent challenges in current models' capabilities on VAU. We further analyze and discuss these findings, and hope that both VALU and insights will advance research in VAU and the development of Video-LLMs. Our benchmark will be publicly available here.

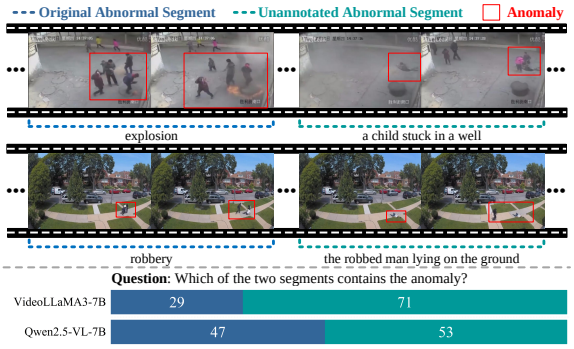


Figure 1: Examples of annotations in existing datasets. The top and middle videos are sourced from UCF-Crime (Sultani et al., 2018) and MSAD (Zhu et al., 2024), respectively. We selected 100 videos from these two datasets for evaluation. As shown blew, when asked "Which of the two segments contains the anomaly?", both models tend to select the unannotated segment.

1 Introduction

Video anomaly understanding (VAU) has extensive applications across various domains, such as security monitoring, industrial production inspection, violent content analysis, and disaster incident warnings (Cao et al., 2024; Zhu et al., 2024; Huang et al., 2025). Traditional research (Wu et al., 2023; Zhang et al., 2023; Wu et al., 2024b; Karim et al., 2024; Zhang et al., 2024b) in this field mainly focuses on video anomaly detection (VAD), which aims to identify abnormal frames in videos, but often lacks the capability to provide semantic descriptions and comprehensive analyses of the detected anomalies.

Recently, significant advances in Video Large Language Models (Video-LLMs) have propelled the development of VAU (Zhang et al., 2025c,a; Zanella et al., 2024; Yang et al., 2024; Ye et al., 2025; Ding et al., 2025; Shao et al., 2025). However, most existing efforts primarily focus on the semantic interpretation of abnormal events, while paying less attention to the anomaly temporal localization (Tang et al., 2024; Zhang et al., 2024a, 2025b). In fact, temporal grounding[1] and localization[2] are fundamental capabilities of Video-LLMs (Chen et al., 2024; Liu et al., 2024; Qian et al., 2024; Zhu et al., 2025; Cheng et al., 2025). Thoroughly investigating these abilities is of great significance for advancing the VAU field.

Systematic exploration of anomaly temporal un-

*Equal contribution.
†Corresponding author.

[1]Temporal grounding refers to aligning a given textual description to its corresponding time segment.
[2]Temporal localization focuses on determining the temporal boundary of the specific target event.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1252–1296
July 2-7, 2026

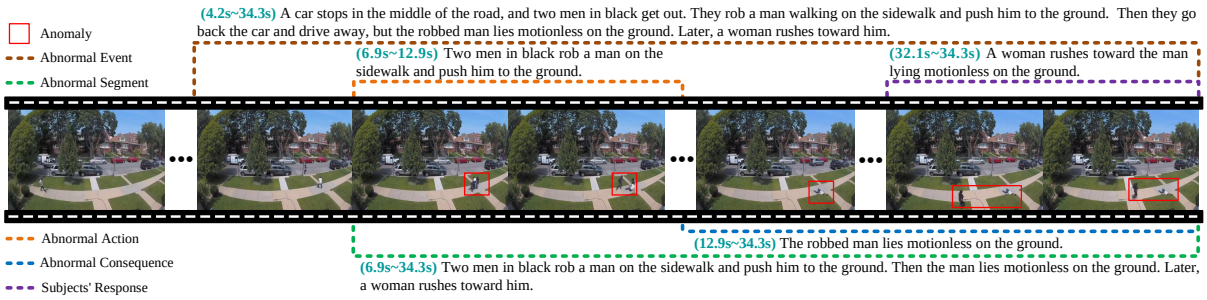


Figure 2: An example of the **multi-level** annotations of anomalies. The video is sourced from MSAD (Zhu et al., 2024), where the originally annotated anomaly boundaries range from 4.5s to 16s.

derstanding capability remains constrained by two major bottlenecks: ***semantic incompleteness of annotations*** and ***systematic lack of evaluation protocols***. **1)** Most video anomaly datasets are originally developed for VAD tasks, where only segments that are explicitly inconsistent with the video context are annotated as abnormal, while semantically related subsequent states are often ignored. As shown in Fig. 1, segments like explosion or robbery are labeled, but subsequent states such as a child stuck in a well, or a man lying on the ground remain unannotated. To examine the impact of incomplete annotations on Video-LLMs, we conducted a preliminary experiment: when prompted to select which segment contained an anomaly, both models tend to choose the unannotated segment (see bottom of Fig. 1). This suggests that current annotations cannot fully measure the ability to locate and understand anomalies, as models may identify abnormal segments beyond those annotated due to their broader semantic understanding. **2)** Existing evaluation protocols (Du et al., 2024b,a; Ma et al., 2025; Gani et al., 2025) assess models from a single dimension, lacking a comprehensive evaluation of anomaly temporal understanding.

To address the above issues, we propose the **VALU** (Video Anomaly Localize and Understand) benchmark, which defines anomalies across **five** distinct semantic levels and provides detailed manual annotations. Based on these, we design **three** tasks to systematically evaluate the VAU capabilities of Video-LLMs at different dimensions.

Specifically, we construct VALU based on UCF-Crime (Sultani et al., 2018), MSAD (Zhu et al., 2024), and ECVA (Du et al., 2024a). After processing stage, we obtain a total of **1,019** videos. Then we conduct multi-level annotations of anomalies: **1)** identifying semantically complete **abnormal events** within each video; **2)** annotating explicit **abnormal segments** within each event; **3)** further subdividing each segment into three semantic components: **abnormal actions**, **consequences**, and **subjects' responses**. Through this process, we provide manual temporal boundaries and detailed textual descriptions for all **five** semantic levels of anomalies (see Fig. 2 for an example). Based on these detailed annotations, we design **three** tasks to systematically evaluate models' VAU capabilities at different dimensions, including temporal grounding, anomaly localization, and anomaly detail discrimination.

We evaluate a range of models, from general Video-LLMs (Bai et al., 2025; Zhang et al., 2025a; Zhu et al., 2025; Yue et al., 2025; Wang et al., 2025) to VAU expert models (Tang et al., 2024; Zhang et al., 2025b), with sizes ranging from 1B to 78B. The results show that these models exhibit shortcomings across multiple dimensions. We further analyze and discuss these results, providing insights and directions for advancing the field.

The main contributions of this work can be summarized in three key aspects:

- We propose the VALU benchmark, which systematically categorizes anomalies into five semantic levels and provides detailed manual annotations for each level.
- We design three evaluation tasks to comprehensively assess VAU capabilities of Video-LLMs at different dimensions, including temporal grounding, anomaly localization, and anomaly detail discrimination.
- We benchmark a wide range of Video-LLMs, revealing that existing models face significant challenges in effectively handling comprehensive VAU tasks, and provide findings, analysis, and discussion to guide future work.

Dataset&Benchmark	#TV	#TAV	#TNV	Source	Anno.	Anomaly Localization	T.	D.	MC.
Subway (Adam et al., 2008)	2	2	0	Single Surveillance	Human	Frame-level	✗	✗	✗
UCSD Ped (Li et al., 2014)	48	48	0	Single Surveillance	Human	Bounding-box-level	✗	✗	✗
CUHK Avenue (Lu et al., 2013)	21	21	0	Single Surveillance	Human	Bounding-box-level	✗	✗	✗
SHTech Campus (Luo et al., 2017)	107	107	0	Single Surveillance	Human	Bounding-box-level	✗	✗	✗
UCF-Crime (Sultani et al., 2018)	290	140	150	Multiple Surveillance	Human	Frame-level	✗	✗	✗
XD-Violence (Wu et al., 2020)	800	500	300	Films/Online	Human	Frame-level	✗	✗	✗
UBnormal (Acsintoae et al., 2022)	211	158	53	Synthetic	Human	Pixel-level	✗	✗	✗
NWPU Campus (Cao et al., 2023)	242	124	118	Single Surveillance	Human	Frame-level	✗	✗	✗
UCA (Yuan et al., 2024)	310	206	104	Multiple Surveillance	Human	NA	✓	✓	✗
VAR (Wu et al., 2024a)	290	140	150	Multiple Surveillance	Human	NA	✗	✗	✗
MSAD (Zhu et al., 2024)	360	240	120	Multiple Surveillance	Human	Frame-level	✗	✗	✗
HAWK (Tang et al., 2024)	786	634	152	Multiple Surveillance	LLM	NA	✗	✓	✗
CUVA (Du et al., 2024b)	200	200	0	Surveillance/Films/Online	Human	Segment-level	✓	✓	✗
ECVA (Du et al., 2024a)	2,174	2,174	0	Surveillance/Films/Online	Human	Segment-level	✓	✓	✗
M-VAE (Ma et al., 2025)	200	200	0	Surveillance/Films/Online	Human	Segment-level	✓	✓	✗
HIVAU-70k (Zhang et al., 2025b)	150	98	52	Surveillance/Films/Online	LLM	NA	✗	✓	✗
VANE-Bench (Gani et al., 2025)	325	325	0	Surveillance/Synthetic	LLM	NA	✗	✗	✓
VALU	1,019	750	269	Multiple Surveillance	Human	Multiple Semantic-level	✓	✓	✓

Table 1: Comparison VALU with other video anomaly datasets and benchmarks. The abbreviations are defined as follows: **TV** (Test Videos), **TAV** (Test Abnormal Videos), **TNV** (Test Normal Videos), **Anno.** (Annotation Method), **T.** (Temporal Task), **D.** (Description Task), **MC.** (Multiple-Choice Task).

2 Related Work

Traditional Video Anomaly Datasets Most existing video anomaly datasets, such as SHTech Campus (Luo et al., 2017), UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), UBnormal (Acsintoae et al., 2022), NWPU Campus (Cao et al., 2023) and MSAD (Zhu et al., 2024), are designed for anomaly detection task. These datasets typically annotate only whether each frame is abnormal, and rarely provide semantic descriptions. In addition, anomalies are labeled only for segments that clearly inconsistent with the normal context, while semantically related subsequent states are ignored. As a result, these datasets are not suitable for assessing VAU capabilities.

Multimodal Benchmarks for VAU With the development of Video-LLMs (Lin et al., 2024; Maaz et al., 2024; Li et al., 2025; Yu et al., 2025), many benchmarks are constructed by enriching traditional datasets (e.g., UCF-Crime) with additional descriptions. These benchmarks typically evaluate models through video question answering (Tang et al., 2024; Zhang et al., 2024a, 2025b; Gani et al., 2025), video retrieval (Wu et al., 2024a; Yang et al., 2025), and video caption (Yuan et al., 2024). In addition, CUVA (Du et al., 2024b) and ECVA (Du et al., 2024a) aim to evaluate the causation understanding ability and collect videos from online video platforms. Ma et al. (2025) further construct the multi-scene abnormal event extraction task based on CUVA. Despite these advances, existing works still overlook the incompleteness of annotations, provide insufficient exploration of fundamental temporal grounding and localization.

Unlike these works, our proposed VALU defines anomalies across five semantic levels and provides comprehensive manual annotations for each, thereby addressing the incompleteness of annotations. In addition, VALU designs three tasks to systematically assess Video-LLMs across different dimensions, providing a comprehensive evaluation protocol. Table 1 summarizes the comparison between our benchmark and existing video anomaly datasets and benchmarks.

3 The Proposed VALU Benchmark

In this section, we introduce the VALU benchmark, including its multiple semantic levels definition of anomalies, construction process, and evaluation tasks. We leave more details in Appendix A and B.

3.1 Multiple Semantic Levels Definition

To more thoroughly evaluate the video anomaly temporal localization and understanding abilities, and to address the incompleteness of existing annotations, as shown in Fig. 2 and Appendix A.3, we define anomalies into five distinct semantic levels:

Abnormal Event The semantically complete event in the video involving anomalies. This level includes not only the abnormal occurrence itself, but also the actions and contexts leading up to the event, its aftermath, and the reactions of relevant

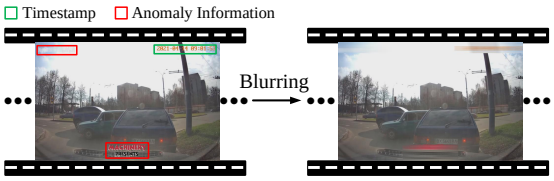


Figure 3: An example of watermark blurring. The video is sourced from ECVA (Du et al., 2024a).

subjects such as people, animals, or vehicles.

Abnormal Segment The visually explicit sub-event within an abnormal event where clear anomalies can be observed in the video frames. Each abnormal segment consists of three elements: abnormal actions, abnormal consequences, and subjects' responses. It excludes parts of the semantically complete abnormal event that do not visually display clear anomalies.

Abnormal Action The specific action or behavior that clearly deviates from the normal situation and indicates the precise occurrence of an anomaly. Abnormal action is the central component of an abnormal segment.

Abnormal Consequence The consequence or aftereffect directly resulting from the abnormal action. It provides further context and evidence for the identification of an anomaly. This is also a constituent part of an abnormal segment.

Subjects' Response The reaction or response of people, animals, vehicles, or other subjects to abnormal actions or their consequences. This reflects how subjects perceive and respond to the anomaly, and it is a component of an abnormal segment.

3.2 Benchmark Construction

Collection of Videos We select videos from two real-world surveillance datasets, UCF-Crime (Sultani et al., 2018) and MSAD (Zhu et al., 2024). Additionally, we filter real-world abnormal videos in the ECVA benchmark (Du et al., 2024a). We trim and segment repetitive sections within these videos and remove videos without obvious anomalies. Although VALU is surveillance-centered, it covers diverse real-world viewpoints. More details are provided in Appendix A.1. Overall, we collect 750 abnormal videos and 269 normal videos.

Blurring Watermarks Many videos contain watermarks showing timestamps or anomaly information, which may cause Video-LLMs to answer anomaly-related questions based on these cues rather than the video content. To address this, as shown in Fig. 3 and more examples in Appendix A.2, we have blurred these watermarks.

Manual Annotation We develop annotation guidelines and trained five annotators proficient in English to conduct the annotation work. Annotators are required to accurately mark the temporal boundaries of different anomaly levels with a precision of 0.1 seconds and provide detailed textual descriptions. For normal videos, annotators must annotate at least one normal event per video. Subsequently, annotators engage in cross-reviewing and correcting annotations to ensure the temporal boundaries and textual descriptions are accurate, comprehensive, and correspond accurately. Please refer to Appendix A.3 for more details.

3.3 Tasks Definition and Evaluation Metrics

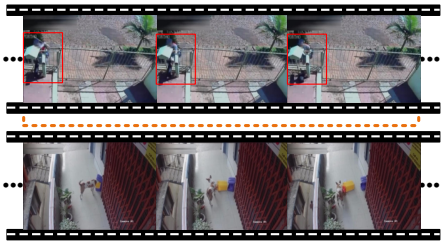
In this section, we introduce the evaluation tasks in VALU, along with the corresponding evaluation methods and metrics. Table 2 provides an overview of the three tasks and their corresponding metrics. Additional prompts and implementation details are included in Appendix B.

Temporal Description Grounding (TDG) This task is designed to evaluate models' ability to identify the start and end times of the video segment based on the given description. We design two subtasks: abnormal description grounding and normal description grounding. These aim to evaluate Video-LLMs' temporal grounding capabilities for anomalies of different semantic levels and for normal events, respectively. Consistent with previous work (Tang et al., 2022; Cheng et al., 2025), we report the mean temporal Intersection over Union (mIoU) and the average recall values (mR) at different IoU thresholds: R@0.3, R@0.5, and R@0.7.

Anomaly Localization and Description (ALD) This task assesses models' ability to localize and describe anomalies of different semantic levels based on given prompts. Models are required to predict the temporal boundaries and describe the content of detected anomalies. This directly measures their capacity for both temporal localization and semantic understanding of video anomalies. For evaluation, we use DeepSeek-V3 (DeepSeek-AI, 2024) to extract anomaly descriptions and corresponding temporal boundaries from model answers, while filtering out analysis components (more details are

Task	Model Output	Capability Assessed	Metrics
TDG	Start/end timestamps of the given description	Temporal grounding of abnormal/normal video segments	mIoU, R@0.3, R@0.5, R@0.7, mR
ALD	Temporal boundaries and anomaly descriptions	Autonomous anomaly localization and semantic understanding	mIoU, R@0.3, R@0.5, R@0.7, mR; Coverage, Consistency
ADC	Multiple-choice option selection	Anomaly detail comprehension and discrimination	Accuracy

Table 2: Summary of the three evaluation tasks in VALU and their corresponding metrics.



Matching

A. A postman throws a stack of letters forcefully at the black dog in the yard when it starts barking at him.
B. A postman takes out a stick and beats the black dog in the yard a few times in front of the mailbox.
C. A postman uses a spray bottle from his bag to spray the black dog in the yard as it barks at him near the mailbox.
D. A postman drops his mailbag on purpose onto the black dog in the yard while placing letters in the mailbox.
E. A postman aggressively kicks at the barking black dog in the yard several times.
F. There are abnormal actions or behaviors in the video, but none of the above options (A–E) accurately describe the specific abnormal actions or behaviors observed.
G. There is no abnormal occurrence in the video.

Exclusion

A. A postman throws a stack of letters forcefully at the black dog in the yard when it starts barking at him.
B. A postman uses a spray bottle from his bag to spray the black dog in the yard as it barks at him near the mailbox.
C. A postman drops his mailbag on purpose onto the black dog in the yard while placing letters in the mailbox.
D. A postman aggressively kicks at the barking black dog in the yard several times.
E. There are abnormal actions or behaviors in the video, but none of the above options (A–D) accurately describe the specific abnormal actions or behaviors observed.
F. There is no abnormal occurrence in the video.

Rejection

A. While the dog is rummaging through the trash bin, a masked individual suddenly appears, smashes the security camera, and runs away.
B. As the dog is near the trash bin, a fire suddenly breaks out inside the bin, sending thick smoke across the corridor.
C. While the dog is eating trash, it knocks over the bin, and a resident slips on spilled garbage, falls hard, and injures their leg.
D. As the dog approaches the trash bin, a fight erupts between two neighbors nearby, resulting in broken furniture and one person being pushed against the railing.
E. There is an abnormal event in the video, but none of the above options (A–D) accurately describe the abnormal event as observed in the footage.
F. There is no abnormal occurrence in the video.

Abnormal Action
Anomaly
Green Option: Correct Choice

Figure 4: Examples of ADC task. Both videos are from UCF-Crime (Sultani et al., 2018). The abnormal video above is used for Matching and Exclusion tasks, while the normal video below is used for Rejection task.

provided in Appendix B.2). This extraction process is fully automated, with no human intervention involved in the scoring pipeline. Anomaly localization performance is evaluated with the same metrics as the TDG task. To assess descriptive quality, we follow Wang et al. (2024) by decomposing both model-generated and ground-truth descriptions into atomic units and calculating ***1) Coverage***, the extent to which the model-generated description covers the ground-truth, and ***2) Consistency***, the extent to which the model-generated description correctly represents the ground-truth. Both scores have a maximum value of 2. For subsequent reporting, we normalize these scores to a percentage scale.

Anomaly Description Choice (ADC) This task is designed to evaluate Video-LLMs' ability for anomaly detail comprehension and discrimination at different semantic levels. For each video, we use DeepSeek-V3 to construct four distractor anomaly descriptions that differ from the ground truth in specific details. For normal videos, we create four descriptive options depicting non-existent abnormal events closely related to the scene context. All generated descriptions are reviewed and refined by annotators to ensure quality. We then assess Video-LLMs on three capabilities using a multiple-choice format (see examples in Fig. 4), namely ***1) Matching***, where models identify the ground-truth anomaly description from several distractors for abnormal videos, testing anomaly detail comprehension; ***2) Exclusion***, where models are presented only with incorrect descriptions for abnormal videos and must select the exclusion choice such as "anomaly present, but none of the descriptions are correct", measuring discrimination capability; and ***3) Rejection***, where models are presented with descriptions of non-existent abnormal events for normal videos and must correctly select the "no anomaly" option, testing their ability to distinguish between normal and abnormal content. We report the accuracy for each subtask.

4 Experiment Evaluation and Analysis

4.1 Experimental Setup

We evaluate a range of Video-LLMs, including general models such as VideoLLaMA3 (Zhang et al., 2025a), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025) (covering sizes from 1B to 78B parameters), as well as models equipped with reasoning capabilities, such as MiMo-VL (7B) (Yue et al., 2025) and TimeZero (7B) (Wang et al., 2025). We also benchmark VAU expert models, including HAWK (7B) (Tang et al., 2024) and HolmesVAU (2B) (Zhang et al., 2025b), both of which are trained on video anomaly datasets. For all models, we use the official weights and settings for evaluation, with 32 frames uniformly sampled per video except for HolmesVAU, where we follow the authors' setup using 16 frames. In addition, for MiMo-VL, we evaluate its performance with thinking mode turned on and off.

	Abnormal Event				Abnormal Segment				Abnormal Action				Abnormal Consequence				Subjects' Response				Normal Event		
	mR	mIoU	Mat.	Exc.	mR	mIoU	Mat.	Exc.	mR	mIoU	Mat.	Exc.	mR	mIoU	Mat.	Exc.	mR	mIoU	Mat.	Exc.	mR	mIoU	Rej.
1∼3B Video-LLMs																							
InternVL3-1B	15.0	17.7	34.0	6.1	11.2	14.3	40.7	16.7	8.8	11.9	40.1	15.4	2.3	5.3	46.9	7.4	4.8	7.2	39.8	16.8	5.0	7.3	0.0
InternVL3-2B	18.0	19.3	44.6	28.7	13.3	14.8	48.3	20.0	10.7	12.5	53.7	43.5	6.8	9.4	40.6	56.9	5.9	8.3	16.8	71.7	9.2	10.8	60.6
VideoLLaMA3-2B	7.2	6.7	39.3	2.3	8.0	7.5	37.1	0.7	5.5	5.9	51.3	1.6	2.8	2.9	46.9	0.9	7.3	6.7	23.0	1.8	2.3	2.4	52.8
HolmesVAU-2B†	0.2	3.0	43.2	0.0	0.1	1.7	49.7	2.9	0.6	2.0	57.7	10.0	0.1	0.1	46.6	18.9	0.0	0.2	31.9	5.3	0.4	2.4	0.0
Qwen2.5-VL-3B	46.6	47.0	58.8	7.5	34.7	37.1	53.2	14.7	30.7	32.9	66.8	43.6	11.3	15.2	44.3	37.7	11.3	15.6	31.9	53.1	34.3	35.4	60.6
7∼9B Video-LLMs																							
MiMo-VL (*w/o think*)	49.1	47.6	13.5	91.3	39.6	39.7	18.8	92.4	34.0	34.0	11.7	92.0	19.9	22.7	12.2	94.3	23.4	27.0	19.8	93.8	43.6	41.7	57.2
MiMo-VL (*w/ think*)	43.6	43.1	15.9	92.7	35.7	36.6	20.0	92.8	29.8	30.9	13.5	94.4	19.9	22.5	14.6	95.4	22.6	26.0	16.8	91.2	37.9	37.9	58.0
TimeZero	69.8	63.7	45.3	64.7	60.8	56.5	47.9	49.6	44.3	44.4	52.7	70.1	35.4	38.2	51.1	63.1	37.0	39.8	36.3	55.8	53.3	50.0	92.9
VideoLLaMA3-7B	46.4	48.1	56.9	15.1	45.2	46.0	56.3	11.6	40.5	41.8	68.7	23.1	36.6	37.6	62.9	23.4	31.1	34.3	28.3	36.3	39.1	41.2	91.8
Qwen2.5-VL-7B	61.7	57.7	35.6	75.5	51.9	49.5	40.0	60.9	37.7	38.9	47.2	75.1	23.7	27.4	44.6	68.6	23.4	26.0	33.6	66.4	52.0	49.7	97.8
HAWK-7B†	4.8	5.9	19.7	2.3	3.7	4.7	20.8	3.3	3.7	4.7	19.7	4.3	1.2	1.6	15.4	5.1	0.3	0.8	14.2	7.1	3.5	4.6	4.1
InternVL3-8B	25.9	28.7	64.0	5.9	21.1	23.7	57.7	2.9	18.3	20.0	66.2	15.1	8.8	12.0	64.3	8.3	9.3	12.9	58.4	13.3	31.2	32.2	95.9
InternVL3-9B	27.1	31.4	64.5	52.3	25.7	28.9	63.4	33.8	19.8	21.9	72.2	36.2	12.6	16.2	60.0	57.1	15.5	18.4	46.0	58.4	35.8	37.0	45.4
14∼38B Video-LLMs																							
InternVL3-14B	39.8	40.8	63.4	22.3	33.6	34.6	62.6	17.5	25.5	27.0	71.3	17.9	14.5	17.5	63.1	21.1	17.2	19.1	54.0	15.9	48.5	46.1	100.0
Qwen2.5-VL-32B	61.6	58.1	55.1	9.5	47.7	47.6	52.1	15.2	40.4	40.9	54.1	18.3	23.3	26.3	46.3	17.4	23.2	27.9	46.0	23.0	57.6	56.0	99.3
InternVL3-38B	42.2	42.7	70.4	16.0	34.0	35.6	70.1	13.4	25.8	28.1	77.7	18.7	14.0	16.5	70.9	21.1	15.5	18.0	60.2	42.5	44.7	44.3	100.0
72∼78B Video-LLMs																							
Qwen2.5-VL-72B	63.2	58.6	62.1	39.6	49.8	47.8	57.2	41.6	41.6	40.8	68.8	45.5	22.8	25.9	56.0	43.4	26.3	30.2	60.2	48.7	55.3	53.1	88.5
InternVL3-78B	29.1	32.7	72.4	19.9	25.4	28.1	66.8	20.0	20.2	22.6	75.0	27.2	9.4	13.0	68.6	28.3	11.9	14.9	69.9	30.1	24.0	27.2	96.3

Table 3: Evaluation results on TDG and ADC tasks in VALU. For TDG, we report mR (the average of R@0.3, 0.5, 0.7) and mIOU. For ADC, we report the accuracy of Matching (Mat.) and Exclusion (Exc.) across five levels of anomalies in abnormal videos, and the accuracy of Rejection (Rej.) in normal videos. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event				Abnormal Segment				Abnormal Action				Abnormal Consequence				Subjects' Response			
	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.
1∼3B Video-LLMs																				
InternVL3-1B	0.0	0.0	11.5	18.5	5.0	6.2	9.4	16.4	0.0	0.1	15.9	19.1	0.6	0.7	14.8	16.9	0.0	0.0	16.6	18.7
InternVL3-2B	0.0	0.0	12.3	22.7	0.6	0.9	10.0	25.0	0.0	0.0	4.4	8.1	0.0	0.0	11.0	15.3	0.0	0.0	4.3	6.2
VideoLLaMA3-2B	10.3	10.9	6.7	10.3	13.0	13.7	6.7	13.6	25.3	26.6	13.6	18.9	16.7	17.4	9.2	12.6	38.1	39.1	15.4	17.7
HolmesVAU-2B†	0.0	0.5	13.4	21.6	0.0	0.3	9.9	21.6	0.3	1.1	5.7	12.5	0.0	0.0	8.3	12.3	0.0	0.2	5.7	8.8
Qwen2.5-VL-3B	21.7	27.0	10.2	23.4	14.5	19.1	6.4	21.6	9.8	12.3	9.6	18.5	1.5	2.7	7.7	10.6	5.9	7.9	9.6	10.7
7∼9B Video-LLMs																				
MiMo-VL (*w/o think*)	3.7	4.0	16.0	24.7	14.7	15.9	17.1	31.0	0.6	0.8	17.3	18.5	13.2	14.9	16.8	18.5	2.1	3.0	10.8	10.8
MiMo-VL (*w/ think*)	1.8	2.0	18.8	25.0	14.7	15.6	16.9	30.7	2.9	3.1	18.5	20.5	11.0	12.9	16.1	19.2	6.2	7.5	14.8	10.5
TimeZero	26.2	25.6	17.5	24.9	27.6	28.7	16.1	27.9	12.5	13.3	17.4	21.2	14.5	17.8	14.0	13.5	11.5	12.4	12.7	12.5
VideoLLaMA3-7B	87.1	80.5	19.9	30.2	70.5	66.5	17.3	26.8	44.5	47.3	19.0	23.8	38.4	41.8	12.7	15.4	36.9	40.7	19.8	21.2
Qwen2.5-VL-7B	38.8	38.2	19.8	27.4	23.7	23.9	15.3	25.8	23.5	24.9	18.3	21.8	12.6	14.9	14.0	15.3	16.5	17.4	14.7	15.2
HAWK-7B†	4.7	6.0	10.3	13.3	2.8	4.2	9.2	11.9	2.8	4.2	9.4	10.4	1.3	2.5	7.2	8.1	2.4	2.6	10.4	9.1
InternVL3-8B	19.6	23.9	17.0	32.8	15.8	20.5	14.1	33.4	2.6	3.1	20.9	26.1	8.9	11.3	6.3	10.0	0.9	0.9	17.4	20.6
InternVL3-9B	3.8	4.2	20.2	30.0	19.4	23.2	14.8	32.0	7.9	10.2	17.3	23.0	7.7	9.8	15.9	20.9	1.2	1.4	13.1	16.5
14∼38B Video-LLMs																				
InternVL3-14B	23.0	28.2	11.0	29.9	21.7	25.9	12.1	30.3	16.9	19.5	21.0	30.4	14.2	16.8	10.6	21.3	1.5	2.0	18.7	16.7
Qwen2.5-VL-32B	53.6	52.1	20.3	25.4	39.8	41.4	15.2	27.4	39.4	40.8	17.9	19.3	23.5	26.2	14.1	13.9	31.0	34.1	15.5	16.4
InternVL3-38B	23.0	27.7	21.1	35.8	25.5	28.7	18.0	40.1	9.1	10.1	28.0	36.6	13.5	16.1	18.6	26.2	2.9	4.2	22.2	21.9
72∼78B Video-LLMs																				
Qwen2.5-VL-72B	24.8	24.6	21.4	28.2	39.2	39.5	18.1	30.8	42.6	42.9	18.6	24.0	18.8	21.4	18.2	20.1	26.0	28.8	17.6	16.1
InternVL3-78B	1.5	2.1	11.9	36.3	6.0	9.0	12.4	35.6	16.4	19.3	30.0	31.6	12.1	15.6	13.2	23.0	3.2	4.6	25.9	21.4

Table 4: Evaluation results on ALD task in VALU. We report mR (the average of R@0.3, 0.5, 0.7), mIOU, as well as the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

4.2 Evaluation Results and Findings

Table 3 and Table 4 present the evaluation results across different tasks. More detailed results are provided in Appendix D. Based on these results, we summarize the following findings.

Performance of VAU expert models. Although HAWK and HolmesVAU are expected to excel in VAU, their performance on VALU is unsatisfactory. Both models show almost no capability in temporal grounding and anomaly localization, with mR and mIoU scores close to zero in the TDG and ALD, respectively. Furthermore, their anomaly description performances do not surpass those of general Video-LLMs of similar size. Their performance on the ADC task is also weak, with scores of only 4.1 and 0 in the Rejection subtask, while most other models achieve higher results.

One main reason is that their training data is incompletely annotated, typically labeling only the

most obvious abnormal segments. This causes the models to focus on explicit anomaly actions but lack the ability to identify or represent entire coherent events and consequences. On VALU, which requires multi-level semantic localization and understanding of anomalies, their performance drops significantly. Furthermore, these models often overfit to narrow objectives, such as simple anomaly classification or fixed text generation, which limits their semantic reasoning and task adaptation. Overall, these results show that current datasets and training strategies are insufficient for developing robust VAU capabilities, as models trained in this way can detect salient anomaly patterns but often fail on comprehensive event-level reasoning, causal understanding, or nuanced semantic distinctions essential for real-world VAU.

Performance of reasoning models. MiMo-VL and TimeZero, which incorporate reinforcement learning algorithms such as GRPO (Shao et al., 2024) for post-training, demonstrate certain strengths in deep thinking and video understanding. Specifically, TimeZero achieves state-of-the-art performance in the TDG task on the VALU benchmark. However, neither model shows significant advantages over other models in other tasks; for example, MiMo-VL performs poorly in video anomaly localization. Moreover, enabling the reasoning module in MiMo-VL does not consistently improve performance across all VALU tasks. Overall, while current reasoning models achieve outstanding results in some tasks, they have yet to bring comprehensive improvements to VAU and still leave substantial room for advancement.

Performance on TDG task. As shown in Table 3, the Qwen2.5-VL series achieves competitive temporal grounding performance among similar-sized models, while all models show limitations for the levels of abnormal consequence and subjects' response. Notably, scaling up from small to medium models brings significant gains, but further increasing yield only marginal or even negative returns. These findings demonstrate that, although recent Video-LLMs have advanced in grounding textual descriptions with temporal segments, they still lack deep semantic understanding of complex anomaly scenarios, especially for less salient or context-dependent aspects. This suggests that increasing model size alone is insufficient, and targeted improvements in semantic and temporal reasoning are required to achieve robust VAU models.

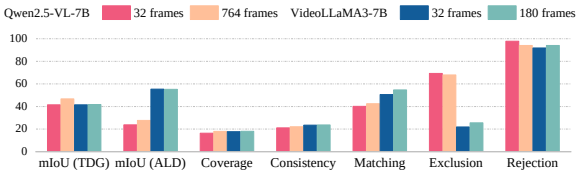


Figure 5: Impact of the number of input frames.

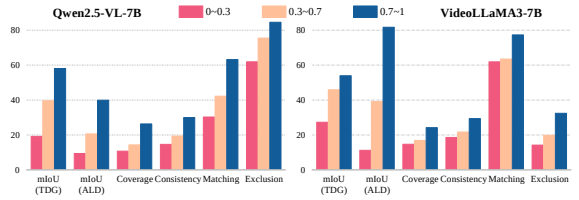


Figure 6: Impact of the proportion of abnormal action.

Performance on ADC task. According to Table 3, many models achieve higher accuracy on the Matching subtask than on Exclusion, showing they are relatively good at selecting the correct anomaly description but remain challenged in recognizing when all provided options are incorrect. Larger models tend to perform better on the Rejection subtask, more accurately distinguishing normal videos from irrelevant anomalies. These results reveal that, while Video-LLMs have some ability to match explicit descriptions, they still lack robust comprehension and discrimination of anomaly details across different semantic levels. This reflects persistent limitations in the models' fine-grained semantic understanding and flexible reasoning, which are crucial for reliable VAU in real-world scenarios.

Performance on ALD task. Table 4 shows that, except for VideoLLaMA3, most models perform worse on anomaly localization in ALD than on temporal description grounding in TDG. This gap reveals that, while grounding textual descriptions with time spans is already feasible, identifying and temporally localizing anomalies across different semantic levels remains much more complex for current models. The persistently low coverage and consistency scores for anomaly descriptions further indicate that these models struggle to interpret both the content and boundaries of anomalies in a comprehensive way. Overall, these findings underscore the necessity to advance beyond basic grounding, and pursue comprehensive temporal reasoning and semantic understanding, which are essential for capturing the complexity of anomalies in diverse real-world applications.

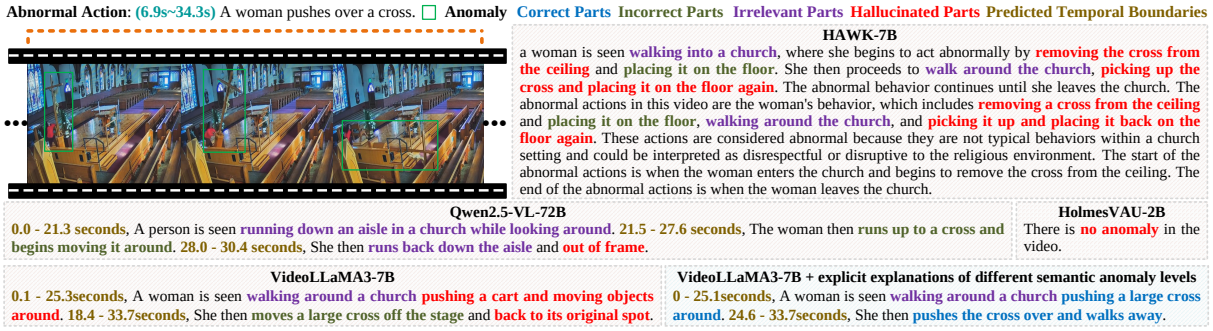


Figure 7: Qualitative results on ALD task. The video is sourced from MSAD (Zhu et al., 2024).

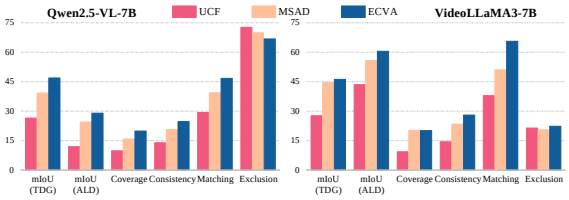


Figure 8: Performance differences across video sources.

	mR	mIoU	Coverage	Consistency
VideoLLaMA3-7B	**55.5**	**55.4**	17.7	23.5
w/ anomaly guidance	51.8	52.7	**18.3**	**25.3**
Qwen2.5-VL-7B	23.0	23.9	16.4	21.1
w/ anomaly guidance	**30.0**	**31.4**	**17.1**	**22.3**

Table 5: Impact of providing anomaly guidance in prompts on ALD task. "w/ anomaly guidance" means that, before the original prompt, we prepend explicit explanations and definitions for each semantic anomaly level to offer more semantic guidance to the model.

4.3 Further Analysis and Discussion

Based on the evaluation results and findings, we further analyze and discuss the following aspects.

How does the number of input frames affect Video-LLMs' performance? As shown in Fig. 5, increasing the number of input frames results in only slight improvements across multiple subtasks, and in some cases leads to marginal decreases. This indicates that the primary bottleneck for VAU performance lies in the models' own capability rather than the number of frames provided.

How do Video-LLMs perform on different proportions of anomaly? To further examine models' ability to localize and understand brief or sparse anomalies, we analyze model performance relative to the proportion of abnormal action duration in the video. All samples are grouped into (0, 0.3), (0.3, 0.7), and (0.7, 1) ranges by abnormal action proportion. As shown in Fig. 6, model performance in all tasks steadily increases with greater anomaly proportion. When anomalies occupy only a small part of the video, all metrics drop significantly. These results suggest current Video-LLMs are unsuitable for real-world surveillance, where anomalies typically appear as brief segments in long videos and require precise localization and understanding.

How do Video-LLMs perform on videos from different sources? We further investigate the performance of Video-LLMs on videos from different sources within VALU, as illustrated in Fig. 8 and detailed in Appendix D.1. Both models perform worst on UCF-Crime videos (except in the Exclusion subtask). This disparity may be attributed to the lower resolution, blurred visuals, and reduced frame rates of the surveillance videos in UCF-Crime. This finding highlights that current models are still limited for real-world VAU applications, where low-quality and challenging visuals are common in surveillance footage, thus restricting their effectiveness in practical scenarios.

Does prompting with detailed anomaly guidance enhance ALD performance? We further investigate whether providing Video-LLMs with explicit textual explanations of different semantic anomaly levels in the prompt improves ALD performance (see Table 5 and more details in Appendix D.2). Results show that such guidance leads to minor improvements in anomaly description scores for both models, but has inconsistent effects on localization, and sometimes even slightly hinder, localization performance. This suggests that simply supplying models with detailed semantic guidance is insufficient to fundamentally overcome their limitations in localizing and understanding anomalies. Deeper integration of semantic knowledge and more advanced anomaly temporal understand-

ing capabilities remain necessary for robust VAU in line with real-world needs.

Qualitative results. Fig. 7 presents qualitative results on ALD task, with more examples available in Appendix D.3. It can be observed that many models struggle to localize and describe the core abnormal actions in the videos. HolmesVAU incorrectly determines that there are no abnormal actions present, while HAWK fails to provide any temporal localization for the abnormal temporal boundary. Furthermore, they often describe behaviors or events unrelated to the actual anomalies, or produce descriptions of hallucinated abnormal behaviors that do not exist in the surveillance video.

5 Conclusion

In this paper, we propose the VALU benchmark, which systematically defines anomalies at five semantic levels and provides three evaluation tasks to comprehensively assess the VAU capabilities of Video-LLMs. Extensive experiments reveal that current models are still far from meeting the requirements of real-world VAU applications, struggling with localization and semantic comprehension across different tasks and anomaly semantic levels. We hope that the multi-level annotations of anomalies and multi-dimensional evaluation tasks in VALU can support future research and the development of more robust VAU models.

Limitations

This work still has several limitations. First, due to constraints on cost and computational resources, we are only able to evaluate open-source Video-LLMs ranging from 1B to 78B parameters, and have yet to assess state-of-the-art closed-source models (such as GPT-5.4 (OpenAI, 2026)) or larger-scale open-source models (like Qwen3-VL-235B-A22B (Qwen Team, 2025)). This limits the comprehensiveness of our evaluation and discussion.

Second, VALU is still primarily centered on surveillance-style anomaly videos. We make this choice because surveillance remains the dominant domain in real-world video anomaly understanding applications, such as public safety, traffic monitoring, and security inspection. However, this domain focus may limit the benchmark's generalization to other video scenarios. Moreover, due to the scarcity of publicly available raw surveillance footage, the number of videos that can be included in VALU remains limited. Although we have made our best effort to curate high-quality videos from existing datasets, broader source diversity is still needed.

Third, the ALD task adopts a free-description paradigm to better reflect realistic VAU settings, where models need not only localize anomalies but also describe them in open form. While this design enables richer evaluation of semantic understanding, it also introduces additional challenges for standardized assessment, since open-form descriptions may vary in wording and structure. To mitigate this issue, we combine the open-ended ALD task with more standardized TDG and ADC tasks, and use an automated evaluation pipeline for description scoring. Nevertheless, this trade-off still remains a limitation of the current design.

In future work, we plan to evaluate stronger Video-LLMs, expand VALU with more diverse real-world source videos and viewpoints, and further investigate how to improve Video-LLMs for VAU tasks in practical applications.

Ethical Considerations

We conducted rigorous quality control to eliminate any potentially discriminatory, biased, or otherwise inappropriate language and content. All descriptions were required to remain objective, neutral, and fact-based. Additionally, we ensured that all annotators received fair compensation for their work and that no exploitative or unfair labor practices occurred throughout the annotation process. All videos are collected from existing public datasets (Sultani et al., 2018; Zhu et al., 2024; Du et al., 2024a). When distributing and releasing VALU, we will strictly adhere to the original licenses and data use agreements associated with these datasets.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2024YFE0200800, the National Natural Science Foundation of China under Grants (62321001, 62471055, U23B2001, 62101064, 62201072), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM107), the High-Quality Development Project of the MIIT (2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the 2025 Education and Teaching Reform Project Funding at Beijing University of Posts and Telecommunications (2025YZ005).

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20111–20121. IEEE.

Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. 2023. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 20392–20401. IEEE.

Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. 2024. A survey on visual anomaly detection: Challenge, approach, and prospect. *CoRR*, arXiv:2401.16402.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. 2025. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *CoRR*, arXiv:2503.11495.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, arXiv:2412.19437.

Zongcan Ding, Haodong Zhang, Peng Wu, Guansong Pang, Zhiwei Yang, Peng Wang, and Yanning Zhang. 2025. Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model. *CoRR*, arXiv:2504.10320.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, and Jun Liu. 2024a. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *CoRR*, arXiv:2412.07183.

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. 2024b. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18793–18803. IEEE.

Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. 2025. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3123–3140. Association for Computational Linguistics.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. 2025. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models. In *Forty-second International Conference on Machine Learning*.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. 2024. Real-time weakly supervised video anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 6834–6842. IEEE.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2014. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):18–32.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. 2024. E.T. bench: Towards open-ended event-level video-language understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 FPS in MATLAB. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2720–2727. IEEE Computer Society.

Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 341–349. IEEE Computer Society.

Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. 2025. Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive LLM. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 4004–4013. ACM.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics.

OpenAI. 2026. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Qwen Team. 2025. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

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. *CoRR*, arXiv:2504.13092.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6479–6488. Computer Vision Foundation / IEEE Computer Society.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. 2024. HAWK: learning to understand open-world video anomalies. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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 Trans. Circuits Syst. Video Technol.*, 32(12):8238–8249.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Jiawei Wang, Liping Yuan, and Yuchen Zhang. 2024. Tarsier: Recipes for training and evaluating large video description models. *CoRR*, arXiv:2407.00634.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. 2025. Timezero: Temporal video grounding with reasoning-guided LVLM. *CoRR*, abs/2503.13377.

Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. 2024a. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Trans. Image Process.*, 33:2213–2225.

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. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 322–339. Springer.

Peng Wu, Xiaotao Liu, and Jing Liu. 2023. Weakly supervised audio-visual violence detection. *IEEE Trans. Multim.*, 25:1674–1685.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024b. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 6074–6082. AAAI Press.

Shuyu Yang, Yilun Wang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2025. Towards scalable video anomaly retrieval: A synthetic video-text benchmark. *CoRR*, arXiv:2506.01466.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. 2024. Follow the rules: Reasoning for video anomaly detection with large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pages 304–322. Springer.

Muchao Ye, Weiyang Liu, and Pan He. 2025. VERA: explainable video anomaly detection via verbalized learning of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8679–8688. Computer Vision Foundation / IEEE.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and Wenbing Tao. 2025. Unhackable temporal reward for scalable video mllms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22052–22061. IEEE.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Xiao-Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, and 54 others. 2025. Mimo-vl technical report. *CoRR*, abs/2506.03569.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18527–18536. 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. *CoRR*, arXiv:2501.13106.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024a. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal LLM. *CoRR*, arXiv:2406.12235.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13843–13853. Computer Vision Foundation / IEEE.

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. 2024b. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 17385–17394. IEEE.

Menghao Zhang, Jingyu Wang, Jing Wang, Qi Qi, Zirui Zhuang, Haifeng Sun, and Ning Xiao. 2023. Robust video anomaly detection framework via prior knowledge and multi-path frame prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. 2025c. Llava-video: Video instruction tuning with synthetic data.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, arXiv:2504.10479.

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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

A Details of Benchmark Construction

A.1 Details of Video Collection

Our proposed VALU benchmark is constructed from three representative video anomaly datasets: UCF-Crime (Sultani et al., 2018), MSAD (Zhu et al., 2024), and ECVA (Du et al., 2024a). Below, we provide comprehensive descriptions of these datasets and explain the specific selection protocols used in our benchmark.

UCF-Crime (Sultani et al., 2018) is the first large-scale, real-world surveillance video anomaly dataset, comprising 1,900 videos collected by annotators from the web. Its test set contains 140 abnormal videos and 150 normal videos. Over ten annotators independently labeled the temporal boundaries of abnormal segments in the test set, with the final labels obtained by averaging their annotations. However, as illustrated in Fig. 1 and further in Fig. 9, many semantically abnormal segments in UCF-Crime remain unlabeled. Furthermore, some videos exhibit repeated or replayed segments, as exemplified in Fig. 10, we have trimmed and split these sections.

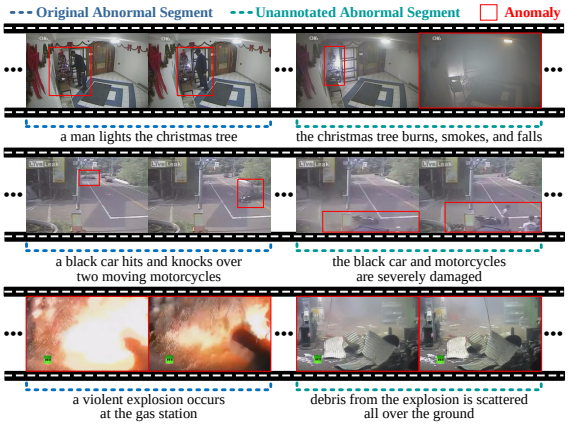


Figure 9: Examples of annotations in UCF-Crime (Sultani et al., 2018).

Figure 10: An example of the video with repeated sections in UCF-Crime (Sultani et al., 2018).

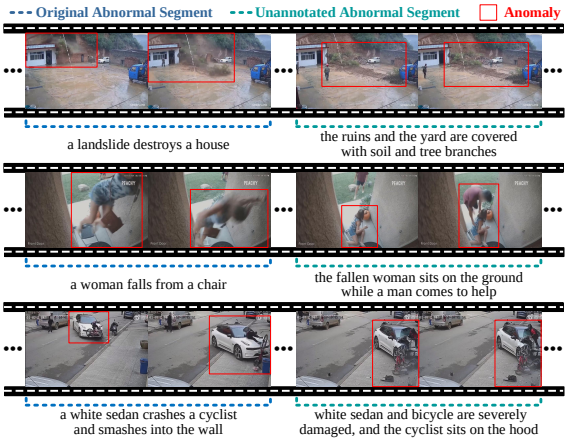


Figure 11: Examples of annotations in MSAD (Zhu et al., 2024).

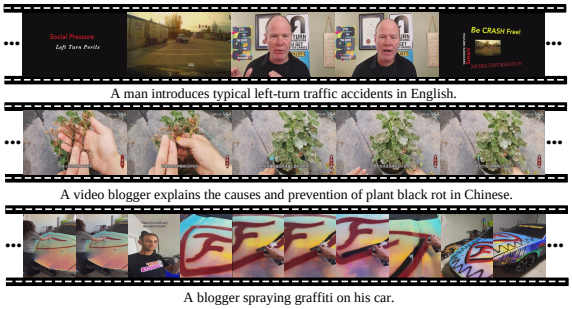


Figure 12: Examples of videos where anomalies cannot be identified from visual content in ECVA (Du et al., 2024a).

MSAD (Zhu et al., 2024) is a recently introduced multi-scenario real-world video anomaly dataset encompassing 14 distinct scenes. Under its semi-supervised protocol, the test set includes 240 abnormal and 120 normal videos; under the weakly supervised protocol, there are 120 abnormal and 120 normal videos. We adopt all 360 test videos from the semi-supervised setting for evaluation in VALU. Similar to UCF-Crime, MSAD is designed for the video anomaly detection task, resulting in some semantically abnormal segments being unannotated. Additional examples are shown in Fig. 11.

ECVA (Du et al., 2024a) extends the CUVA benchmark (Du et al., 2024b) and consists of 2,174 abnormal videos sourced from various domains, including news reports, movies, vlogs, social media, interviews, animations, and surveillance footage. ECVA focuses on evaluating Video-LLMs' causal reasoning abilities. As shown in Fig. 12, many videos contain anomalies that cannot be identi-

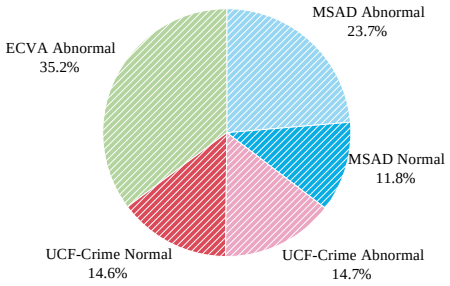


Figure 13: The distribution of video sources in VALU.

fied from visual information alone. To ensure that our benchmark focuses on visually accessible real-world scenarios, we select only those videos in which the anomaly is clearly observable from video content alone. This process results in a curated subset of 349 real-world abnormal videos from ECVA.

The overall source distribution of videos included in VALU is shown in Fig. 13.

A.2 Details of Watermark Blurring

We preprocess all videos to remove visible artifacts such as timestamps, news titles, and subtitles that could reveal anomaly-related information and potentially bias the outputs of Video-LLMs. Two annotators use standardized video editing software to conduct the blurring process, and three additional annotators subsequently review and refine the results to ensure consistency. More examples of this process are shown in Fig. 14.

Please note that videos presented in Fig. 1, Fig. 9, Fig. 10, Fig. 11, and Fig. 12 are sourced from the original datasets and have not been processed with watermark blurring.

A.3 Details of Manual Annotation

The annotation work is performed by five annotators with backgrounds in computer vision and multimodal research, all of whom are proficient in English. As illustrated in Fig. 15, all annotators utilize PotPlayer[3] to facilitate frame-level temporal localization. Prior to annotation, all annotators receive training on the annotation guidelines and are provided with definitions and examples covering all five semantic levels of anomalies.

We provide an example of the annotation format in Fig. 18. The distributions of average description length (in words) are illustrated in Fig. 19, Fig. 20, Fig. 21, and Fig. 22. The statistics for the average temporal length (in seconds) of the annotated boundaries are shown in Fig. 23, Fig. 24, Fig. 25, and Fig. 26. Additional annotation examples are presented in Fig. 27 and Fig. 28.

[3]https://potplayer.daum.net/

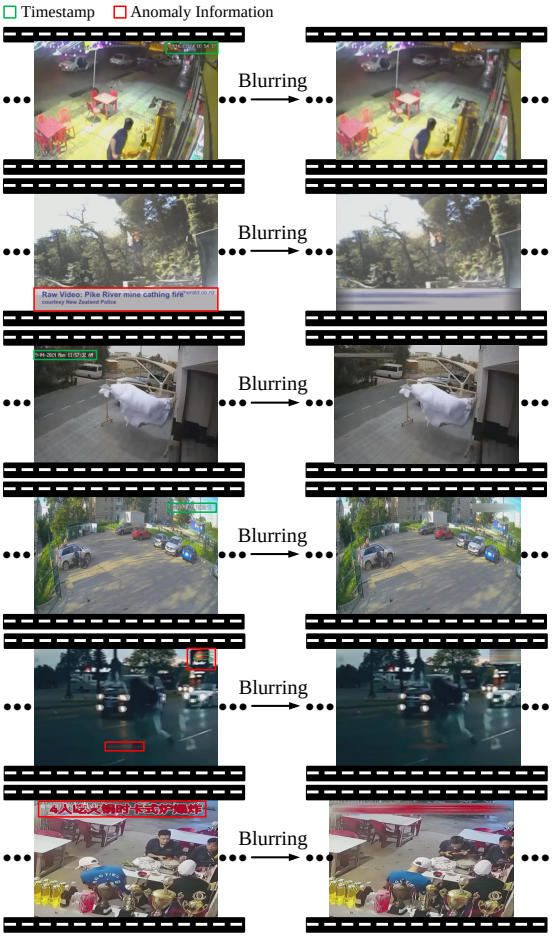


Figure 14: Examples of watermark blurring. The top two, middle two, and bottom two videos are sourced from UCF-Crime (Sultani et al., 2018), MSAD (Zhu et al., 2024), and ECVA (Du et al., 2024a), respectively.

B Details of Evaluation Tasks

In this section, we provide further details on the evaluation tasks included in VALU.

B.1 Temporal Description Grounding (TDG)

The prompt template used for the TDG task is shown in Fig. 29. For each instance, the "description" field is filled with the specific event description from our benchmark. Models are required to analyze the provided video and output the precise start and end times (in seconds) during which the described event occurs.

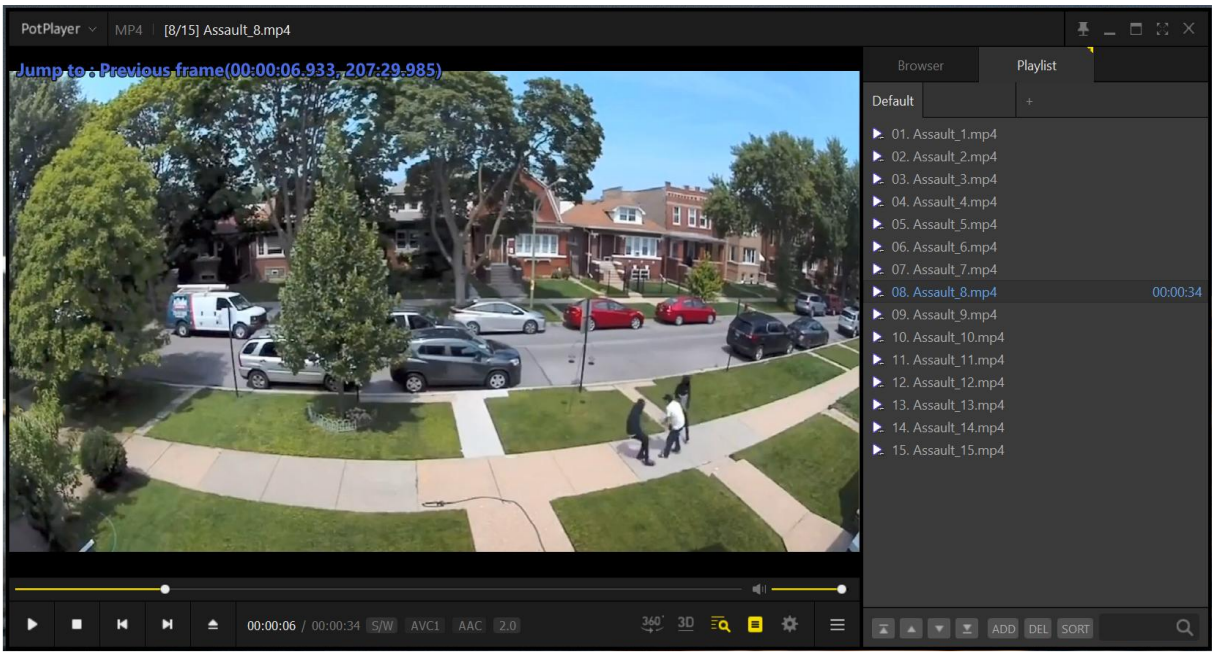


Figure 15: Annotation operation interface for frame-level temporal localization using PotPlayer.

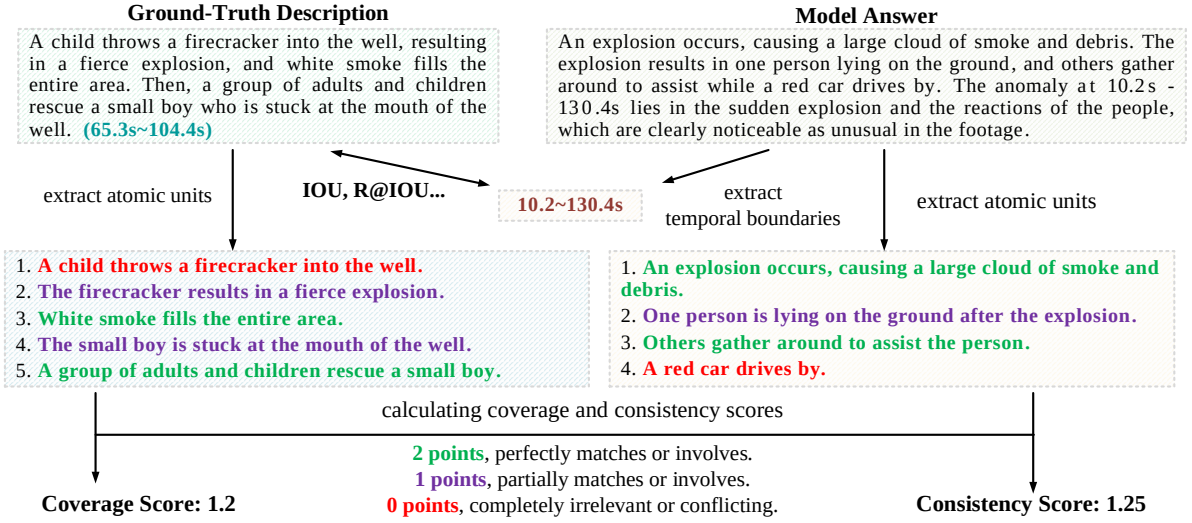


Figure 16: Example of the scoring pipeline for the ALD task. The relevant video is provided at the top of Fig. 1.

B.2 Anomaly Localization and Description (ALD)

We provide the example of the scoring pipeline in Fig. 16. The prompt templates for evaluating anomalies at different semantic levels are presented in Fig. 30, Fig. 31, Fig. 32, Fig. 33, and Fig. 34. Additionally, Fig. 35 and Fig. 36 show the prompt templates for extracting atomic units from ground-truth descriptions, as well as for extracting predicted temporal boundaries and atomic units from the responses of Video-LLMs. The prompt templates for the Coverage and Consistency score calculations are illustrated in Fig. 37 and Fig. 38, respectively.

Our pipeline for computing Coverage and Consistency scores is inspired by Wang et al. (2024). To further assess the reliability of these metrics, we conduct a human consistency experiment. Specifically, we select some model responses from Qwen2.5-VL-7B (Bai et al., 2025) and VideoLLaMA3-7B (Zhang et al., 2025a), and invite human experts to review each video and determine which model provides a better answer. For each case, we average the model's Coverage and Consistency scores to produce our overall evaluation score, then quantify align between these scores

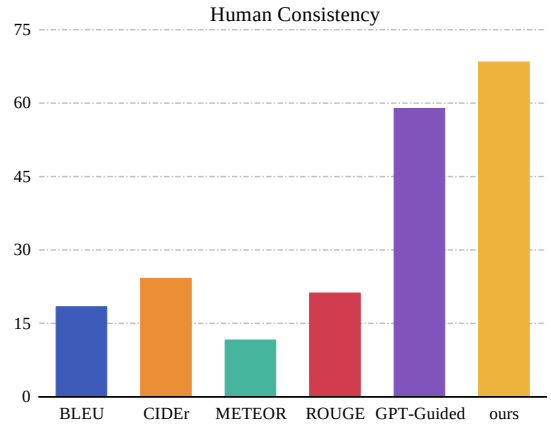


Figure 17: Human consistency result in ALD task.

and the human judgments. We compare our scoring approach to other metrics, including BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and the GPT-Guided score used in HAWK (Tang et al., 2024) (proposed by Liu et al. (2023)). As shown in Fig. 17, our scores demonstrate a higher level of align with human judgment.

B.3 Anomaly Description Choice (ADC)

We present the prompt templates for the Matching subtask in Fig. 39, Fig. 40, Fig. 41, Fig. 42, and Fig. 43; for the Exclusion subtask in Fig. 44, Fig. 45, Fig. 46, Fig. 47, and Fig. 48; and for the Rejection subtask in Fig. 49.

C Open-Source Policy

VALU is constructed based on UCF-Crime (Sultani et al., 2018), MSAD (Zhu et al., 2024), and ECVA (Du et al., 2024a). We strictly adhere to the relevant licenses and data use agreements of these source datasets in all aspects of construction and distribution. Upon acceptance of this paper, we will open-source the VALU annotation and evaluation data to facilitate further research. Before releasing any watermarked-blurred evaluation videos, we will consult with the authors of UCF-Crime, MSAD, and ECVA, and make these videos available only in a manner that is fully compliant with their licensing and stated terms of use.

D Details of Experiment Evaluation

D.1 Comprehensive Evaluation Results

The comprehensive evaluation results for the TDG task are presented in Table 6 and Table 7. In addition, detailed results for TDG on videos sourced from UCF-Crime, MSAD, and ECVA are shown in Table 8, Table 9, Table 10, Table 11, Table 12, and Table 13.

The comprehensive evaluation results for the ALD task are summarized in Table 14. Detailed results for ALD on videos from UCF-Crime, MSAD, and ECVA are reported in Table 15, Table 16, and Table 17.

The comprehensive evaluation results for the ADC task are shown in Table 18 and Table 19. Detailed results for ADC on videos from UCF-Crime, MSAD, and ECVA are provided in Table 20, Table 21, Table 22, Table 23, Table 24, and Table 25.

D.2 Implementation Details of Detailed Anomaly Guidance

We present the system prompt used to implement Detailed Anomaly Guidance in Fig. 50. Detailed results are also provided in Table 26.

D.3 More Qualitative Results

We provide more qualitative results for the TDG task in Fig. 51 and Fig. 52, for the ADC task in Fig. 53 and Fig. 54, and for the ALD task in Fig. 55 and Fig. 56.

```
{
    "video": "example.mp4",
    "abnormal_events": [
        {
            "description": "A car stops in the middle of the road, and two men in black get out. They rob
a man walking on the sidewalk and push him to the ground. Then they go back the car and drive away, but
the robbed man lies motionless on the ground. Later, a woman rushes toward him.",
            "timestamps": [
                4.2,
                34.3
            ]
        }
    ],
    "abnormal_segments": [
        {
            "description": "Two men in black rob a man on the sidewalk and push him to the ground. Then
the man lies motionless on the ground. Later, a woman rushes toward him.",
            "timestamps": [
                6.9,
                34.3
            ]
        }
    ],
    "abnormal_actions": [
        {
            "description": "Two men in black rob a man on the sidewalk and push him to the ground.",
            "timestamps": [
                6.9,
                12.9
            ]
        }
        ],
    "abnormal_consequences": [
        {
            "description": "The robbed man lies motionless on the ground.",
            "timestamps": [
                12.9,
                34.3
            ]
        }
    ],
    "abnormal_responses": [
        {
            "description": "A woman rushes toward the man lying motionless on the ground.",
            "timestamps": [
                32.1,
                34.3
            ]
        }
    ]
}
```


Figure 18: An example of the annotation format. The relevant video is provided in Fig.2.

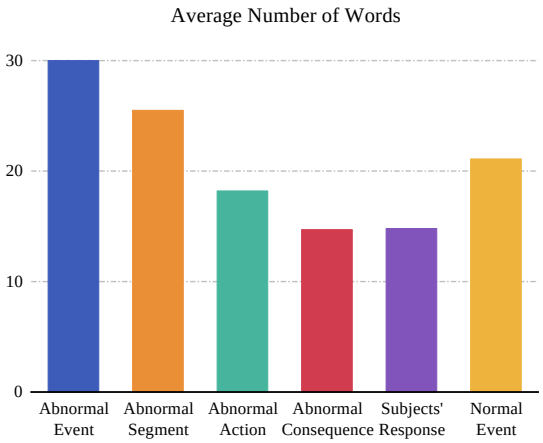


Figure 19: Average number of words in VALU.

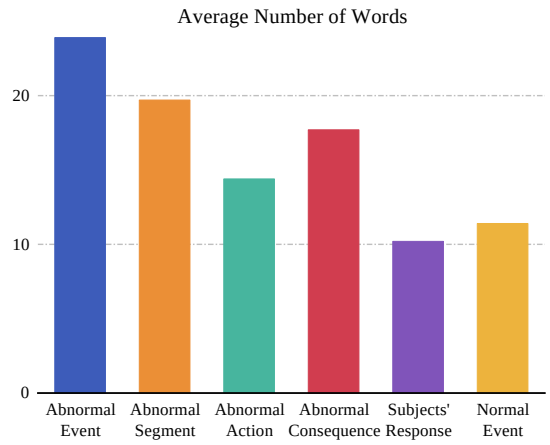


Figure 21: Average number of words in VALU for videos sourced from MSAD (Zhu et al., 2024).

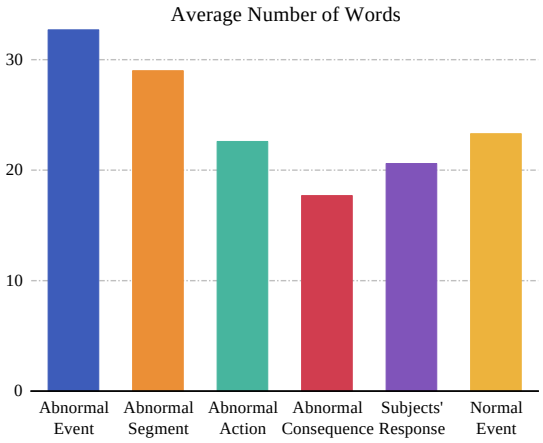


Figure 20: Average number of words in VALU for videos sourced from UCF-Crime (Sultani et al., 2018).

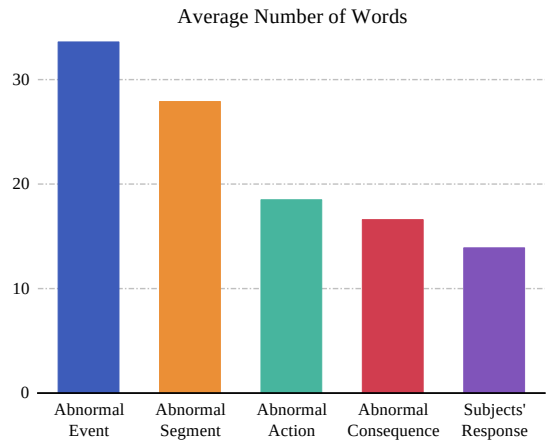


Figure 22: Average number of words in VALU for videos sourced from ECVA (Du et al., 2024a).

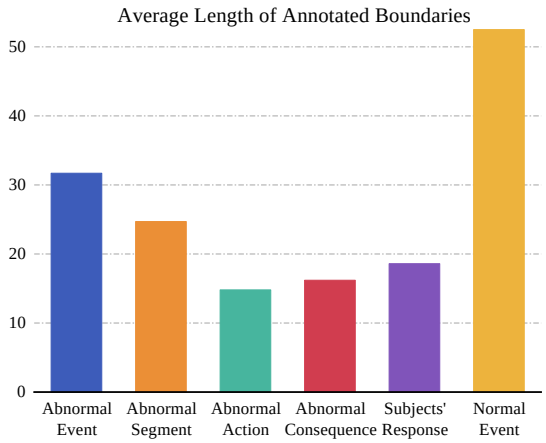


Figure 23: Average average temporal length of annotated boundaries in VALU.

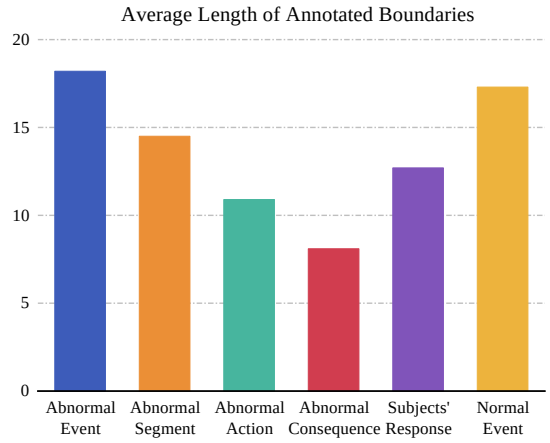


Figure 25: Average average temporal length of annotated boundaries in VALU for videos sourced from MSAD (Zhu et al., 2024).

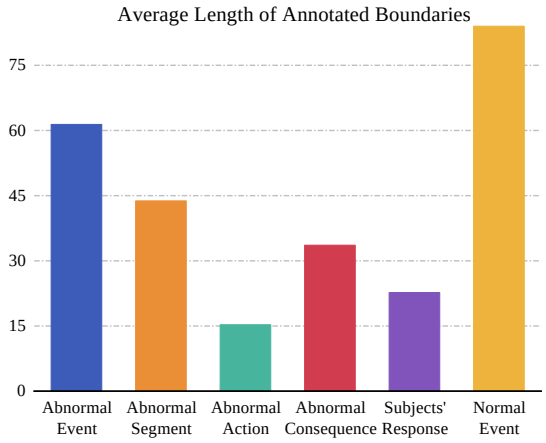


Figure 24: Average average temporal length of annotated boundaries in VALU for videos sourced from UCF-Crime (Sultani et al., 2018).

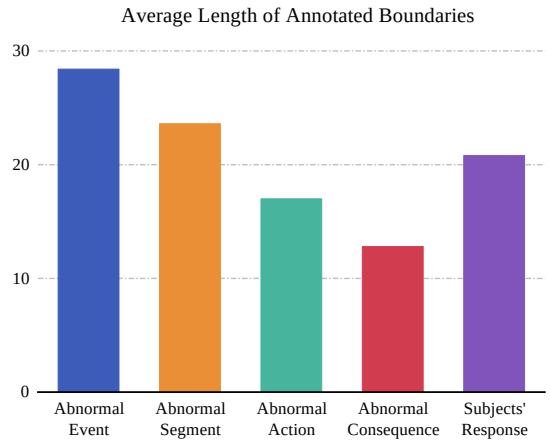


Figure 26: Average average temporal length of annotated boundaries in VALU for videos sourced from ECVA (Du et al., 2024a).

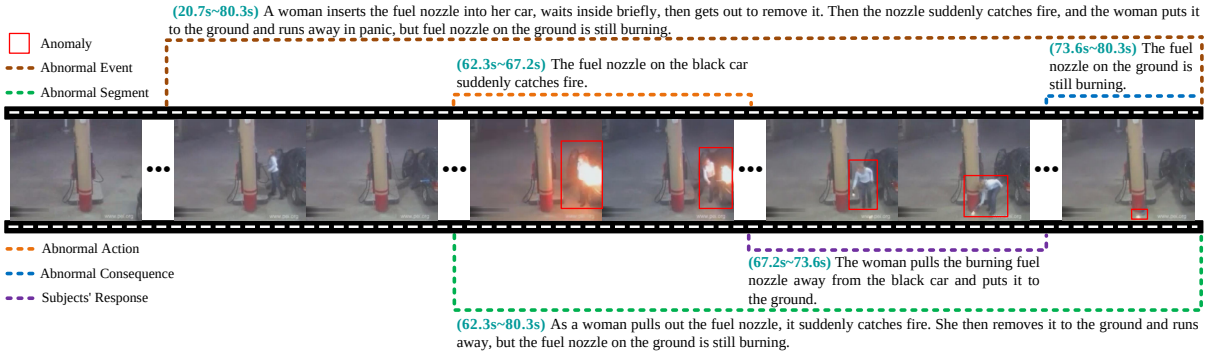


Figure 27: An example of the multi-level annotations of anomalies. The video is sourced from UCF-Crime (Sultani et al., 2018), where the originally annotated anomaly boundaries range from 61.0s to 67.3s.

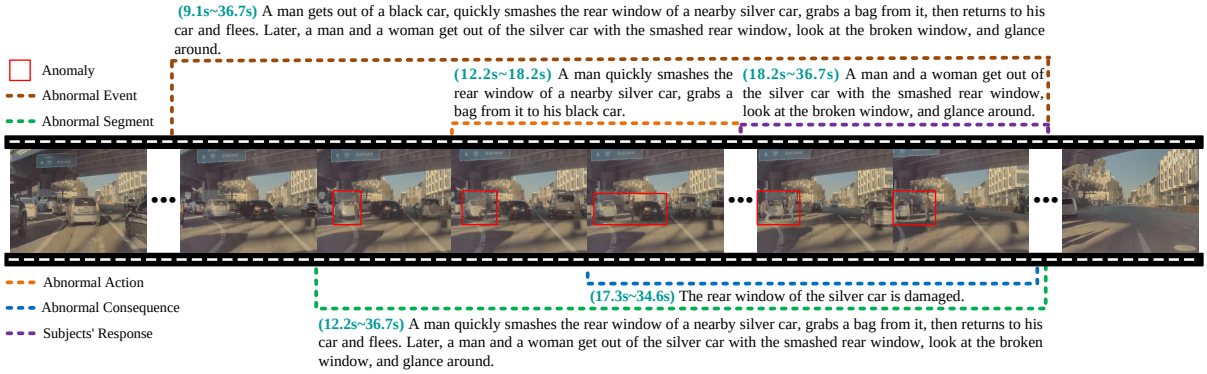


Figure 28: An example of the multi-level annotations of anomalies. The video is sourced from ECVA (Du et al., 2024a), where the originally annotated anomaly boundaries range from 10s to 22s.

```
Given a video and a description of an event, please analyze the video and accurately identify
the start and end times (in seconds) during which the described event occurs. The description
is:
{description}
```

Figure 29: Prompt template used for the TDG task in VALU.

```
Please identify and describe in detail the event involving an abnormal occurrence in the video.
Specify the start and end times (in seconds) for the event, and include the context, abnormal
actions or behaviors, and any subsequent developments or consequences observed.
```

Figure 30: Prompt template used for the abnormal event level in ALD task.

```
Please identify the abnormal segment that occurs in the video. Specify the start and end times
(in seconds) of the event and focus on describing the anomalous or unexpected aspect as seen in
the footage.
```

Figure 31: Prompt template used for the abnormal segment level in ALD task.

```
Please identify and provide detailed descriptions of the abnormal actions or behaviors observed
in the video. Specify the start and end times (in seconds) of abnormal actions or behaviors.
```

Figure 32: Prompt template used for the abnormal action level in ALD task.

```
Please identify and describe in detail the consequences or aftereffects resulting from the
abnormal behaviors or actions in the video.  Specify the relevant start and end times (in
seconds) for the consequences.
```

Figure 33: Prompt template used for the abnormal consequence level in ALD task.

```
Please identify and describe in detail how people, animals, vehicles, or other subjects respond
or react to the abnormal actions or consequences in the video. Specify the start and end times
(in seconds) of these responses.
```

Figure 34: Prompt template used for the subjects' response level in ALD task.

```
You are given a descriptive paragraph about a video, which may contain several events.
Your task is to extract and separate each atomic event described, following these instructions:
- Each event should be an atomic and brief sentence, describing a single action, with subject
and predicate (optionally object).
- Do NOT merge multiple actions into one event; split as needed.
- Replace pronouns with the nouns they refer to.
- If no describable event exists, return an empty list.
- Output a JSON with a key “events” and a list of sentences as values, and nothing else.

Examples:
{Examples 1}

{Examples 2}

{Examples 3}

Now, process the following paragraph and output ONLY the JSON list of events:
{description}
```

Figure 35: Prompt template for extracting atomic units from ground-truth descriptions in ALD task.

```
You are given a model-generated paragraph that attempts to localize and describe abnormal
events in a video. However, the paragraph may contain extraneous text such as explanation,
commentary, discussion, or repeated descriptions, rather than strictly reporting time-stamped
abnormal events.

Your task:
1. If the model’s output contains time-stamped (start_time, end_time) event descriptions of
anomalies, extract each time stamp and its corresponding event description.
2. If the model’s output contains abnormal event descriptions without time stamps, extract each
event description as a separate event, assign both start_time and end_time as -1.
3. Ignore all analytical discussion, speculative statements, repeated descriptions, and any
non-event commentary. Only extract explicit abnormal events.
4. Output a JSON list, where each item is a dict with keys “start_time”, “end_time”, “event”.
5. If multiple events are described in one sentence, split them into separate items. Do NOT
include duplicate events.
6. Output only the JSON list. Do not include any explanation, prefix, or extra text.

Examples:
{Examples 1}

{Examples 2}

{Examples 3}

Now, here is the model-generated paragraph for you to process:
{paragraph}

Please process it and output ONLY the JSON list according to the rules above.
```

Figure 36: Prompt template for extracting atomic units from model outputs in ALD task.

```
Given a list of ground-truth events (gt_events) and a list of predicted events (predicted_events)
generated by a model, analyze the relationship between each ground-truth event and the predicted
events. For each gt event, determine one of the following scores:

- 2: The predicted events contain a complete and precise coverage of this gt event.
- 1: The predicted events partially cover this gt event, i.e., contain a similar meaning but
are not fully complete.
- 0: The predicted events do not mention or cover this gt event at all.

For each gt event, output your score and provide a concise reason for your decision.

Output a JSON in the following format:
{
“results”: [
{“gt_event”: “copy gt event here”, “score”: put_score_here, “reason”: “your reason here”},
...
]
}

Input:

gt_events:
{gt_events}

predicted_events:
{predicted_events}

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON.

Output:
```

Figure 37: Prompt template for calculating Coverage score in ALD task.

```
Given a list of ground-truth events (gt_events) and a list of predicted events (predicted_events)
generated by a model, analyze the relationship between each predicted_event and the ground-truth
events. For each predicted_event, determine one of the following scores:

- 2: The predicted_event is fully and precisely covered or entailed by one of the gt_events.
- 1: The predicted_event is partially covered by one or more gt_events, i.e., its meaning is
similar but not fully complete.
- 0: The predicted_event is not covered at all by any gt_event.

For each predicted_event, output your score and provide a concise reason for your decision.

Output a JSON in the following format:
{
"results": [
{"predicted_event": "copy predicted_event here", "score": put_score_here, "reason": "your
reason here"},
...
]
}

Input:

gt_events:
{gt_events}

predicted_events:
{predicted_events}

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON.

Output:
```

Figure 38: Prompt template for calculating Consistency score in ALD task.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to G), and select **the single best option (A, B, C, D, E, F, or G)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Please identify and describe in detail the event involving an abnormal occurrence in the video. Which option best covers the context, abnormal actions, and subsequent developments or consequences? Please select the most accurate event description.

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: {E}
F: There is an event involving an abnormal occurrence in the video, but none of the above options (A-E) accurately describe the entire event and its context, abnormal actions, and consequences.
G: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 39: Prompt template for the Matching subtask in ADC at the abnormal event level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to G), and select **the single best option (A, B, C, D, E, F, or G)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option most accurately identifies and describes the abnormal segment as it occurred in the video? Focus on the anomalous or unexpected aspect as seen in the footage.

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: {E}
F: There is an abnormal segment in the video, but none of the above options (A-E) accurately describe the abnormal segment as observed in the footage.
G: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 40: Prompt template for the Matching subtask in ADC at the abnormal segment level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to G), and select **the single best option (A, B, C, D, E, F, or G)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option best describes the abnormal actions or behaviors observed in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: {E}
F: There are abnormal actions or behaviors in the video, but none of the above options (A-E) accurately describe the specific abnormal actions or behaviors observed.
G: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 41: Prompt template for the Matching subtask in ADC at the abnormal action level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to G), and select **the single best option (A, B, C, D, E, F, or G)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option provides the most accurate description of the consequence or aftereffect resulting from the abnormal behaviors or actions in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: {E}
F: There are consequences or aftereffects resulting from abnormal behaviors in the video, but none of the above options (A-E) accurately describe these consequences or aftereffects.
G: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 42: Prompt template for the Matching subtask in ADC at the abnormal consequence level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to G), and select **the single best option (A, B, C, D, E, F, or G)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option gives the best description of how people, animals, vehicles, or other subjects respond or react to the abnormal actions or consequences in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: {E}
F: There are responses or reactions by people, animals, vehicles, or other subjects to abnormal actions or consequences in the video, but none of the above options (A-E) accurately describe these responses.
G: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 43: Prompt template for the Matching subtask in ADC at the subjects' response level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Please identify and describe in detail the event involving an abnormal occurrence in the video. Which option best covers the context, abnormal actions, and subsequent developments or consequences? Please select the most accurate event description.

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There is an event involving an abnormal occurrence in the video, but none of the above options (A-D) accurately describe the entire event and its context, abnormal actions, and consequences.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 44: Prompt template for the Exclusion subtask in ADC at the abnormal event level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option most accurately identifies and describes the abnormal segment as it occurred in the video? Focus on the anomalous or unexpected aspect as seen in the footage.

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There is an abnormal segment in the video, but none of the above options (A-D) accurately describe the abnormal segment as observed in the footage.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 45: Prompt template for the Exclusion subtask in ADC at the abnormal segment level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option best describes the abnormal actions or behaviors observed in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There are abnormal actions or behaviors in the video, but none of the above options (A-D) accurately describe the specific abnormal actions or behaviors observed.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 46: Prompt template for the Exclusion subtask in ADC at the abnormal action level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option provides the most accurate description of the consequence or aftereffect resulting from the abnormal behaviors or actions in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There are consequences or aftereffects resulting from abnormal behaviors in the video, but none of the above options (A-D) accurately describe these consequences or aftereffects.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 47: Prompt template for the Exclusion subtask in ADC at the abnormal consequence level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option gives the best description of how people, animals, vehicles, or other subjects respond or react to the abnormal actions or consequences in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There are responses or reactions by people, animals, vehicles, or other subjects to abnormal actions or consequences in the video, but none of the above options (A-D) accurately describe these responses.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 48: Prompt template for the Exclusion subtask in ADC at the subjects' response level.

You are an expert in video analysis.
Please read the following question, review all the answer choices below (A to F), and select **the single best option (A, B, C, D, E, or F)** according to the video.
After your choice, **briefly explain your reasoning based only on the video information**.

Which option most accurately identifies and describes the abnormal event as it occurred in the video?

Options:
A: {A}
B: {B}
C: {C}
D: {D}
E: There is an abnormal event in the video, but none of the above options (A-D) accurately describe the abnormal event as observed in the footage.
F: There is no abnormal occurrence in the video.

Format your answer as:
Choice: <Your Option Letter>
Reason: <Short reasoning>

Figure 49: Prompt template for the Rejection subtask in ADC.

	Abnormal Event					Abnormal Segment					Abnormal Action				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	23.6	13.2	8.0	15.0	17.7	19.9	9.2	4.4	11.2	14.3	16.0	6.8	3.5	8.8	11.9
InternVL3-2B	28.0	17.1	8.9	18.0	19.3	22.6	12.0	5.2	13.3	14.8	19.0	9.7	3.4	10.7	12.5
VideoLLaMA3-2B	9.2	7.3	5.2	7.2	6.7	10.5	8.4	4.9	8.0	7.5	8.7	5.2	2.8	5.5	5.9
HolmesVAU-2B†	0.4	0.3	0.0	0.2	3.0	0.3	0.1	0.0	0.1	1.7	1.1	0.5	0.1	0.6	2.0
Qwen2.5-VL-3B	69.3	49.1	21.3	46.6	47.0	55.2	33.3	15.6	34.7	37.1	49.2	30.3	12.7	30.7	32.9
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	68.7	49.7	28.8	49.1	47.6	59.6	38.9	20.1	39.6	39.7	52.9	33.2	15.9	34.0	34.0
MiMo-VL (*w/ think*)	64.3	44.1	22.4	43.6	43.1	54.5	35.2	17.3	35.7	36.6	46.7	28.9	13.8	29.8	30.9
TimeZero	87.9	71.5	50.1	69.8	63.7	82.9	61.9	37.7	60.8	56.5	64.5	42.6	25.7	44.3	44.4
VideoLLaMA3-7B	71.2	44.4	23.6	46.4	48.1	70.3	41.9	23.6	45.2	46.0	66.4	37.9	17.2	40.5	41.8
Qwen2.5-VL-7B	76.0	64.8	44.3	61.7	57.7	70.8	52.3	32.7	51.9	49.5	55.1	36.9	21.1	37.7	38.9
HAWK-7B†	8.5	4.5	1.2	4.8	5.9	7.5	2.7	1.1	3.7	4.7	6.9	3.3	0.9	3.7	4.7
InternVL3-8B	42.5	23.9	11.5	25.9	28.7	36.3	18.3	8.5	21.1	23.7	30.5	17.0	7.4	18.3	20.0
InternVL3-9B	46.6	24.0	10.7	27.1	31.4	42.7	23.5	10.9	25.7	28.9	32.7	18.8	7.9	19.8	21.9
14∼38B Video-LLMs															
InternVL3-14B	61.1	39.4	18.8	39.8	40.8	53.0	33.4	14.4	33.6	34.6	40.2	25.6	10.8	25.5	27.0
Qwen2.5-VL-32B	77.7	63.1	43.9	61.6	58.1	67.6	47.3	28.3	47.7	47.6	57.5	40.9	22.9	40.4	40.9
InternVL3-38B	64.2	42.1	20.3	42.2	42.7	56.1	31.8	14.3	34.0	35.6	42.9	24.3	10.2	25.8	28.1
72∼78B Video-LLMs															
Qwen2.5-VL-72B	79.7	66.5	43.3	63.2	58.6	67.5	50.7	31.3	49.8	47.8	58.6	41.9	24.2	41.6	40.8
InternVL3-78B	49.8	27.9	9.7	29.1	32.7	44.7	22.7	8.7	25.4	28.1	36.1	18.2	6.3	20.2	22.6

Table 6: Detailed evaluation results on TDG task in VALU (Part I). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence					Subjects' Response					Normal Event				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	5.6	1.1	0.3	2.3	5.3	8.5	4.2	1.7	4.8	7.2	10.6	2.7	1.7	5.0	7.3
InternVL3-2B	12.7	6.2	1.4	6.8	9.4	11.9	5.1	0.8	5.9	8.3	17.3	8.0	2.3	9.2	10.8
VideoLLaMA3-2B	5.1	2.3	1.1	2.8	2.9	11.0	8.5	2.5	7.3	6.7	2.7	2.7	1.7	2.3	2.4
HolmesVAU-2B†	0.3	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.2	0.7	0.3	0.3	0.4	2.4
Qwen2.5-VL-3B	21.1	9.3	3.4	11.3	15.2	21.2	10.2	2.5	11.3	15.6	54.2	35.2	13.6	34.3	35.4
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	34.6	19.7	5.4	19.9	22.7	43.2	19.5	7.6	23.4	27.0	57.1	43.9	29.9	43.6	41.7
MiMo-VL (*w/ think*)	34.4	18.9	6.5	19.9	22.5	39.0	20.3	8.5	22.6	26.0	53.5	36.9	23.3	37.9	37.9
TimeZero	60.3	33.8	12.1	35.4	38.2	62.7	34.7	13.6	37.0	39.8	70.4	52.5	36.9	53.3	50.0
VideoLLaMA3-7B	57.7	36.9	15.2	36.6	37.6	55.9	26.3	11.0	31.1	34.3	66.8	32.6	17.9	39.1	41.2
Qwen2.5-VL-7B	41.1	22.3	7.6	23.7	27.4	42.4	22.0	5.9	23.4	26.0	65.1	52.8	38.2	52.0	49.7
HAWK-7B†	2.3	1.1	0.3	1.2	1.6	0.8	0.0	0.0	0.3	0.8	5.6	3.7	1.3	3.5	4.6
InternVL3-8B	17.2	6.8	2.5	8.8	12.0	18.6	7.6	1.7	9.3	12.9	48.8	30.6	14.3	31.2	32.2
InternVL3-9B	22.8	10.4	4.5	12.6	16.2	26.3	14.4	5.9	15.5	18.4	58.5	32.2	16.6	35.8	37.0
14∼38B Video-LLMs															
InternVL3-14B	23.7	14.4	5.4	14.5	17.5	27.1	17.8	6.8	17.2	19.1	63.8	49.8	31.9	48.5	46.1
Qwen2.5-VL-32B	38.9	22.3	8.7	23.3	26.3	39.0	23.7	6.8	23.2	27.9	68.8	57.8	46.2	57.6	56.0
InternVL3-38B	24.8	12.4	4.8	14.0	16.5	28.8	14.4	3.4	15.5	18.0	64.1	45.5	24.6	44.7	44.3
72∼78B Video-LLMs															
Qwen2.5-VL-72B	38.0	21.4	9.0	22.8	25.9	41.5	27.1	10.2	26.3	30.2	72.8	55.1	37.9	55.3	53.1
InternVL3-78B	18.6	7.3	2.3	9.4	13.0	23.7	9.3	2.5	11.9	14.9	40.5	21.3	10.3	24.0	27.2

Table 7: Detailed evaluation results on TDG task in VALU (Part II). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event					Abnormal Segment					Abnormal Action				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	8.0	5.3	2.7	5.3	7.8	5.3	2.0	0.7	2.7	5.5	6.0	2.2	1.1	3.1	4.4
InternVL3-2B	10.0	4.7	1.3	5.3	7.8	8.0	2.7	1.3	4.0	5.6	7.1	2.2	0.5	3.3	4.6
VideoLLaMA3-2B	2.7	2.7	2.0	2.4	2.2	3.3	3.3	2.7	3.1	2.9	2.2	1.1	1.1	1.4	2.0
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-3B	36.0	20.0	6.7	20.9	26.7	24.7	12.7	4.0	13.8	18.8	17.9	8.7	2.2	9.6	13.9
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	48.7	34.0	20.7	34.4	35.6	41.3	24.7	14.7	26.9	28.1	28.8	13.0	5.4	15.8	17.9
MiMo-VL (*w/ think*)	43.3	30.7	16.7	30.2	32.2	36.7	23.3	10.0	23.3	26.2	22.3	10.9	2.7	12.0	16.0
TimeZero	74.0	57.3	34.7	55.3	51.5	67.3	45.3	22.7	45.1	43.5	41.3	26.6	9.8	25.9	27.9
VideoLLaMA3-7B	42.0	21.3	11.3	24.9	30.5	45.3	17.3	8.7	23.8	29.5	44.0	23.9	10.3	26.1	28.0
Qwen2.5-VL-7B	52.7	43.3	26.0	40.7	40.8	48.7	32.7	16.7	32.7	33.3	31.0	18.5	7.1	18.8	22.4
HAWK-7B†	2.0	0.7	0.0	0.9	2.0	0.0	0.0	0.0	0.0	0.5	1.1	0.5	0.0	0.5	0.9
InternVL3-8B	21.3	10.7	6.0	12.7	16.1	14.0	6.0	4.0	8.0	11.3	15.2	7.1	3.3	8.5	9.8
InternVL3-9B	30.0	10.7	6.0	15.6	21.3	24.0	10.0	6.0	13.3	16.9	15.8	8.2	2.7	8.9	10.6
14∼38B Video-LLMs															
InternVL3-14B	34.0	21.3	10.0	21.8	25.6	26.0	17.3	4.7	16.0	18.5	16.8	10.9	3.3	10.3	11.5
Qwen2.5-VL-32B	62.0	47.3	30.0	46.4	46.5	52.0	30.0	12.7	31.6	34.5	34.2	19.6	6.5	20.1	22.9
InternVL3-38B	31.3	19.3	9.3	20.0	24.4	20.7	13.3	4.7	12.9	15.6	16.3	9.8	2.2	9.4	11.1
72∼78B Video-LLMs															
Qwen2.5-VL-72B	66.7	54.0	42.0	54.2	52.7	52.7	38.0	24.7	38.4	39.2	38.6	25.5	8.7	24.3	25.7
InternVL3-78B	29.3	16.7	6.0	17.3	21.5	23.3	14.0	6.7	14.7	17.0	16.8	9.2	2.7	9.6	11.5

Table 8: Detailed evaluation results on TDG task in VALU for videos sourced from UCF-Crime (Sultani et al., 2018) (Part I). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence					Subjects' Response					Normal Event				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	1.2	0.0	0.0	0.4	3.7	6.5	3.2	3.2	4.3	5.6	5.7	1.9	1.9	3.1	5.8
InternVL3-2B	4.8	3.6	0.0	2.8	5.1	6.5	0.0	0.0	2.2	3.9	16.4	8.8	2.5	9.2	10.7
VideoLLaMA3-2B	1.2	0.0	0.0	0.4	0.4	0.0	0.0	0.0	0.0	0.0	1.3	1.3	0.6	1.0	1.2
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.6	0.6	0.6	2.0
Qwen2.5-VL-3B	11.9	7.1	1.2	6.7	10.8	19.4	6.5	0.0	8.6	12.0	42.1	25.8	12.6	26.8	28.8
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	22.6	16.7	7.1	15.5	17.8	29.0	16.1	6.5	17.2	21.0	51.6	41.5	27.7	40.3	39.1
MiMo-VL (*w/ think*)	25.0	16.7	8.3	16.7	18.9	29.0	19.4	3.2	17.2	21.4	50.3	39.0	25.8	38.4	37.3
TimeZero	45.2	31.0	10.7	29.0	31.1	45.2	25.8	12.9	28.0	30.7	58.5	45.9	34.6	46.3	44.6
VideoLLaMA3-7B	38.1	22.6	10.7	23.8	26.8	29.0	16.1	9.7	18.3	23.6	48.4	27.7	15.1	30.4	32.5
Qwen2.5-VL-7B	22.6	15.5	2.4	13.5	19.0	22.6	9.7	3.2	11.8	17.1	59.1	48.4	36.5	48.0	45.8
HAWK-7B†	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.1	3.8	2.5	0.6	2.3	3.7
InternVL3-8B	13.1	7.1	2.4	7.5	9.8	9.7	3.2	3.2	5.4	8.7	39.0	25.8	10.1	24.9	27.0
InternVL3-9B	13.1	9.5	1.2	7.9	12.0	22.6	9.7	6.5	12.9	15.5	42.8	24.5	11.9	26.4	29.0
14∼38B Video-LLMs															
InternVL3-14B	14.3	11.9	4.8	10.3	12.2	19.4	9.7	6.5	11.8	14.2	49.1	34.0	20.1	34.4	34.8
Qwen2.5-VL-32B	33.3	19.0	1.2	17.9	22.1	45.2	19.4	9.7	24.7	28.6	61.6	54.1	39.6	51.8	50.3
InternVL3-38B	13.1	9.5	1.2	7.9	11.4	19.4	3.2	0.0	7.5	11.1	49.1	32.1	14.5	31.9	33.1
72∼78B Video-LLMs															
Qwen2.5-VL-72B	39.3	22.6	10.7	24.2	27.0	45.2	32.3	19.4	32.3	35.2	66.7	59.7	44.7	57.0	55.1
InternVL3-78B	13.1	7.1	1.2	7.1	10.3	16.1	3.2	0.0	6.5	10.7	30.2	17.6	5.7	17.8	21.3

Table 9: Detailed evaluation results on TDG task in VALU for videos sourced from UCF-Crime (Sultani et al., 2018) (Part II). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event					Abnormal Segment					Abnormal Action				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	25.3	17.0	12.0	18.1	19.9	18.3	10.0	5.0	11.1	14.3	14.6	6.7	3.3	8.2	11.8
InternVL3-2B	38.2	22.8	14.5	25.2	25.7	28.2	15.8	8.3	17.4	18.7	24.3	12.6	5.4	14.1	16.6
VideoLLaMA3-2B	12.4	10.4	7.1	10.0	9.3	18.3	13.3	6.2	12.6	12.2	18.4	10.0	3.8	10.7	11.4
HolmesVAU-2B†	0.4	0.0	0.0	0.1	4.7	0.4	0.0	0.0	0.1	2.4	2.1	0.4	0.0	0.8	2.9
Qwen2.5-VL-3B	78.8	61.0	22.4	54.1	52.8	61.4	36.9	16.2	38.2	39.9	60.3	36.4	13.4	36.7	38.6
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	74.7	52.7	28.6	52.0	50.6	66.4	40.7	17.8	41.6	42.7	62.4	38.4	16.5	39.1	39.0
MiMo-VL (*w/ think*)	73.9	49.8	22.4	48.7	47.8	65.1	40.2	16.6	40.7	41.2	57.4	33.9	15.7	35.7	36.4
TimeZero	73.9	49.8	22.4	48.7	47.8	65.1	40.2	16.6	40.7	41.2	57.4	33.9	15.7	35.7	36.4
VideoLLaMA3-7B	83.0	50.6	24.9	52.8	52.8	82.2	47.3	24.5	51.3	50.1	80.3	41.4	16.3	46.0	46.4
Qwen2.5-VL-7B	77.6	66.0	41.1	61.5	56.6	70.5	49.0	27.8	49.1	47.3	60.7	36.8	20.1	39.2	40.5
HAWK-7B†	7.1	5.8	1.7	4.8	4.8	5.8	2.9	0.8	3.2	3.7	4.1	2.1	0.8	2.3	3.1
InternVL3-8B	45.6	26.6	11.2	27.8	30.1	41.1	19.9	8.3	23.1	24.9	32.2	17.2	7.1	18.8	20.4
InternVL3-9B	44.0	23.7	11.6	26.4	30.7	40.2	24.1	11.6	25.3	28.5	28.9	18.4	7.1	18.1	21.1
14∼38B Video-LLMs															
InternVL3-14B	69.3	46.5	22.0	45.9	44.9	56.0	34.9	16.6	35.8	36.3	43.0	29.3	12.4	28.2	29.6
Qwen2.5-VL-32B	82.6	65.6	43.2	63.8	60.2	68.0	44.4	23.7	45.4	46.5	62.8	42.6	23.1	42.8	43.4
InternVL3-38B	69.3	44.4	19.9	44.5	44.8	59.8	33.6	14.5	36.0	37.7	46.7	24.8	8.7	26.7	30.1
72∼78B Video-LLMs															
Qwen2.5-VL-72B	82.2	64.3	29.0	58.5	55.2	67.6	41.1	18.3	42.3	42.6	60.3	35.1	16.9	37.5	38.0
InternVL3-78B	51.5	27.0	8.3	28.9	32.7	42.7	21.2	10.0	24.6	27.2	34.3	18.6	8.3	20.4	22.4

Table 10: Detailed evaluation results on TDG task in VALU for videos sourced from MSAD (Zhu et al., 2024) (Part I). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence					Subjects' Response					Normal Event				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	7.0	0.9	0.9	2.9	5.9	10.3	2.6	0.0	4.3	7.6	16.2	3.5	1.4	7.0	8.9
InternVL3-2B	14.9	4.4	1.8	7.0	9.0	15.4	7.7	0.0	7.7	10.3	18.3	7.0	2.1	9.2	10.9
VideoLLaMA3-2B	9.6	3.5	1.8	5.0	5.7	12.8	7.7	0.0	6.8	6.7	4.2	4.2	2.8	3.8	3.8
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.7	0.0	0.0	0.2	2.8
Qwen2.5-VL-3B	21.1	10.5	5.3	12.3	15.6	20.5	10.3	2.6	11.1	14.5	67.6	45.8	14.8	42.7	42.8
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	36.0	18.4	3.5	19.3	23.2	38.5	12.8	2.6	17.9	24.7	63.4	46.5	32.4	47.4	44.6
MiMo-VL (*w/ think*)	34.2	20.2	5.3	19.9	23.0	38.5	12.8	5.1	18.8	24.5	57.0	34.5	20.4	37.3	38.7
TimeZero	34.2	20.2	5.3	19.9	23.0	38.5	12.8	5.1	18.8	24.5	83.8	59.9	39.4	61.0	55.9
VideoLLaMA3-7B	60.5	35.1	7.0	34.2	36.8	69.2	20.5	7.7	32.5	36.6	87.3	38.0	21.1	48.8	51.0
Qwen2.5-VL-7B	41.2	19.3	9.6	23.4	27.8	41.0	20.5	0.0	20.5	23.9	71.8	57.7	40.1	56.6	54.1
HAWK-7B†	4.4	2.6	0.0	2.3	3.4	2.6	0.0	0.0	0.9	1.0	7.7	4.9	2.1	4.9	5.7
InternVL3-8B	16.7	7.9	3.5	9.4	11.5	20.5	12.8	2.6	12.0	15.0	59.9	35.9	19.0	38.3	38.2
InternVL3-9B	21.9	9.6	6.1	12.6	15.9	20.5	12.8	5.1	12.8	16.7	76.1	40.8	21.8	46.2	45.9
14∼38B Video-LLMs															
InternVL3-14B	21.1	14.9	4.4	13.5	17.0	30.8	17.9	0.0	16.2	17.2	80.3	67.6	45.1	64.3	58.8
Qwen2.5-VL-32B	35.1	16.7	7.9	19.9	23.3	30.8	23.1	2.6	18.8	23.8	76.8	62.0	53.5	64.1	62.4
InternVL3-38B	22.8	7.0	3.5	11.1	14.2	30.8	20.5	2.6	17.9	19.8	81.0	60.6	35.9	59.2	56.8
72∼78B Video-LLMs															
Qwen2.5-VL-72B	27.2	12.3	5.3	14.9	19.2	33.3	15.4	0.0	16.2	20.8	79.6	50.0	30.3	53.3	50.8
InternVL3-78B	19.3	7.0	1.8	9.4	12.6	20.5	15.4	2.6	12.8	15.6	52.1	25.4	15.5	31.0	33.7

Table 11: Detailed evaluation results on TDG task in VALU for videos sourced from MSAD (Zhu et al., 2024) (Part II). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event					Abnormal Segment					Abnormal Action				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs															
InternVL3-1B	29.1	14.0	7.5	16.9	20.4	27.1	11.7	5.6	14.8	17.9	21.8	9.2	4.9	11.9	15.8
InternVL3-2B	28.8	18.4	8.4	18.5	19.7	24.9	13.4	4.7	14.3	16.1	21.6	11.6	3.5	12.2	13.7
VideoLLaMA3-2B	9.7	7.2	5.3	7.4	6.8	8.4	7.2	5.0	6.9	6.2	5.6	4.0	3.0	4.2	4.3
HolmesVAU-2B†	0.6	0.6	0.0	0.4	2.8	0.3	0.3	0.0	0.2	1.8	1.1	0.8	0.3	0.7	2.3
Qwen2.5-VL-3B	76.9	53.2	26.7	52.3	51.7	63.8	39.6	20.1	41.1	42.9	57.5	37.1	17.5	37.4	38.6
7∼9B Video-LLMs															
MiMo-VL (*w/o think*)	73.0	54.3	32.3	53.2	50.7	62.7	43.7	24.0	43.5	42.5	58.6	39.8	20.7	39.7	38.8
MiMo-VL (*w/ think*)	66.6	46.0	24.8	45.8	44.6	54.9	36.8	20.9	37.5	37.9	51.9	34.7	18.0	34.9	34.7
TimeZero	90.5	76.0	56.5	74.4	67.6	87.5	69.4	45.1	67.3	61.4	71.0	50.0	34.4	51.8	50.5
VideoLLaMA3-7B	75.5	49.9	27.9	51.1	52.2	72.7	48.5	29.2	50.1	50.1	68.5	42.5	21.2	44.1	45.7
Qwen2.5-VL-7B	84.7	73.0	54.0	70.6	65.5	80.2	62.7	42.6	61.8	57.7	63.4	46.0	28.8	46.1	46.1
HAWK-7B†	12.3	5.3	1.4	6.3	8.3	11.7	3.6	1.7	5.7	7.2	11.6	5.4	1.3	6.1	7.6
InternVL3-8B	49.2	27.7	14.0	30.3	33.1	42.5	22.3	10.6	25.1	28.1	36.9	21.8	9.7	22.8	24.7
InternVL3-9B	55.3	29.9	12.0	32.4	36.0	52.2	28.8	12.6	31.2	34.2	43.7	24.3	11.1	26.3	27.9
14∼38B Video-LLMs															
InternVL3-14B	67.0	42.2	20.4	43.2	44.4	62.3	39.1	17.0	39.5	40.1	49.9	30.5	13.5	31.3	32.9
Qwen2.5-VL-32B	81.1	68.0	50.1	66.4	61.5	73.8	56.5	37.9	56.1	53.7	65.6	50.3	30.9	48.9	48.3
InternVL3-38B	74.6	50.0	25.1	49.9	49.0	68.4	38.3	18.2	41.6	42.6	53.6	31.3	15.1	33.3	35.2
72∼78B Video-LLMs															
Qwen2.5-VL-72B	83.6	73.3	53.5	70.1	63.4	73.5	62.4	42.9	59.6	54.9	67.5	54.3	36.6	52.8	50.0
InternVL3-78B	57.3	33.2	12.3	34.3	37.5	55.0	27.4	8.7	30.4	33.3	46.9	22.4	6.7	25.3	28.3

Table 12: Detailed evaluation results on TDG task in VALU for videos sourced from ECVA (Du et al., 2024a) (Part I). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence					Subjects' Response				
	R@0.3	R@0.5	R@0.7	mR	mIoU	R@0.3	R@0.5	R@0.7	mR	mIoU
1∼3B Video-LLMs										
InternVL3-1B	7.0	1.9	0.0	3.0	5.6	8.3	6.2	2.1	5.6	7.9
InternVL3-2B	15.3	8.9	1.9	8.7	11.9	12.5	6.2	2.1	6.9	9.4
VideoLLaMA3-2B	3.8	2.5	1.3	2.5	2.2	16.7	14.6	6.2	12.5	11.0
HolmesVAU-2B†	0.6	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.2
Qwen2.5-VL-3B	26.1	9.6	3.2	13.0	17.2	22.9	12.5	4.2	13.2	18.8
7∼9B Video-LLMs										
MiMo-VL (*w/o think*)	40.1	22.3	5.7	22.7	24.9	56.2	27.1	12.5	31.9	32.7
MiMo-VL (*w/ think*)	39.5	19.1	6.4	21.7	24.0	45.8	27.1	14.6	29.2	30.2
TimeZero	68.2	41.4	16.6	42.0	42.8	75.0	41.7	18.8	45.1	47.9
VideoLLaMA3-7B	66.2	45.9	23.6	45.2	43.9	62.5	37.5	14.6	38.2	39.3
Qwen2.5-VL-7B	51.0	28.0	8.9	29.3	31.6	56.2	31.2	12.5	33.3	33.6
HAWK-7B†	1.9	0.6	0.6	1.1	0.9	0.0	0.0	0.0	0.0	1.0
InternVL3-8B	19.7	5.7	1.9	9.1	13.5	22.9	6.2	0.0	9.7	13.8
InternVL3-9B	28.7	11.5	5.1	15.1	18.6	33.3	18.8	6.2	19.4	21.6
14∼38B Video-LLMs										
InternVL3-14B	30.6	15.3	6.4	17.4	20.7	29.2	22.9	12.5	21.5	23.8
Qwen2.5-VL-32B	44.6	28.0	13.4	28.7	30.8	41.7	27.1	8.3	25.7	30.8
InternVL3-38B	32.5	17.8	7.6	19.3	20.9	33.3	16.7	6.2	18.8	21.0
72∼78B Video-LLMs										
Qwen2.5-VL-72B	45.2	27.4	10.8	27.8	30.1	45.8	33.3	12.5	30.6	34.7
InternVL3-78B	21.0	7.6	3.2	10.6	14.7	31.2	8.3	4.2	14.6	17.0

Table 13: Detailed evaluation results on TDG task in VALU for videos sourced from ECVA (Du et al., 2024a) (Part II). We report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event		Abnormal Segment		Abnormal Action		Abnormal Consequence		Subjects' Response		Normal Event
	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Rej.
1∼3B Video-LLMs											
InternVL3-1B	34.0	6.1	40.7	16.7	40.1	15.4	46.9	7.4	39.8	16.8	0.0
InternVL3-2B	44.6	28.7	48.3	20.0	53.7	43.5	40.6	56.9	16.8	71.7	60.6
VideoLLaMA3-2B	39.3	2.3	37.1	0.7	51.3	1.6	46.9	0.9	23.0	1.8	52.8
HolmesVAU-2B†	43.2	0.0	49.7	2.9	57.7	10.0	46.6	18.9	31.9	5.3	0.0
Qwen2.5-VL-3B	58.8	7.5	53.2	14.7	66.8	43.6	44.3	37.7	31.9	53.1	60.6
7∼9B Video-LLMs											
MiMo-VL (*w/o think*)	13.5	91.3	18.8	92.4	11.7	92.0	12.2	94.3	19.8	93.8	57.2
MiMo-VL (*w/ think*)	15.9	92.7	20.0	92.8	13.5	94.4	14.6	95.4	16.8	91.2	58.0
TimeZero	45.3	64.7	47.9	49.6	52.7	70.1	51.1	63.1	36.3	55.8	92.9
VideoLLaMA3-7B	56.9	15.1	56.3	11.6	68.7	23.1	62.9	23.4	28.3	36.3	91.8
Qwen2.5-VL-7B	35.6	75.5	40.0	60.9	47.2	75.1	44.6	68.6	33.6	66.4	97.8
HAWK-7B†	19.7	2.3	20.8	3.3	19.7	4.3	15.4	5.1	14.2	7.1	4.1
InternVL3-8B	64.0	5.9	57.7	2.9	66.2	15.1	64.3	8.3	58.4	13.3	95.9
InternVL3-9B	64.5	52.3	63.4	33.8	72.2	36.2	60.0	57.1	46.0	58.4	45.4
14∼38B Video-LLMs											
InternVL3-14B	63.4	22.3	62.6	17.5	71.3	17.9	63.1	21.1	54.0	15.9	100.0
Qwen2.5-VL-32B	55.1	9.5	52.1	15.2	54.1	18.3	46.3	17.4	46.0	23.0	99.3
InternVL3-38B	70.4	16.0	70.1	13.4	77.7	18.7	70.9	21.1	60.2	42.5	100.0
72∼78B Video-LLMs											
Qwen2.5-VL-72B	62.1	39.6	57.2	41.6	68.8	45.5	56.0	43.4	60.2	48.7	88.5
InternVL3-78B	72.4	19.9	66.8	20.0	75.0	27.2	68.6	28.3	69.9	30.1	96.3

Table 14: Detailed evaluation results on ADC task in VALU. We report the accuracy of Matching (Mat.) and Exclusion (Exc.) across five levels of anomalies in abnormal videos, and the accuracy of Rejection (Rej.) in normal videos. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event		Abnormal Segment		Abnormal Action		Abnormal Consequence		Subjects' Response		Normal Event
	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Rej.
1∼3B Video-LLMs											
InternVL3-1B	24.7	6.7	29.3	33.3	36.7	30.0	34.9	13.3	38.7	19.4	0.0
InternVL3-2B	30.7	26.7	34.7	13.3	42.7	29.3	37.3	56.6	9.7	87.1	53.7
VideoLLaMA3-2B	17.3	2.0	16.0	0.0	31.3	0.0	49.4	0.0	29.0	0.0	36.9
HolmesVAU-2B†	36.7	0.0	42.0	1.3	50.7	8.0	51.8	22.9	22.6	6.5	0.0
Qwen2.5-VL-3B	50.0	8.7	45.3	17.3	60.7	42.7	41.0	54.2	22.6	58.1	51.7
7∼9B Video-LLMs											
MiMo-VL (*w/o think*)	13.3	96.0	7.3	97.3	9.3	97.3	9.6	98.8	9.7	90.3	41.6
MiMo-VL (*w/ think*)	14.0	97.3	8.7	96.0	12.0	98.0	8.4	100.0	9.7	83.9	41.6
TimeZero	32.0	68.7	39.3	48.7	42.7	72.0	39.8	81.9	29.0	71.0	90.6
VideoLLaMA3-7B	40.7	11.3	38.0	2.7	56.0	14.0	51.8	37.3	3.2	41.9	85.2
Qwen2.5-VL-7B	24.7	80.0	29.3	58.0	37.3	76.0	32.5	81.9	22.6	67.7	98.7
HAWK-7B†	20.7	1.3	22.7	2.0	22.7	1.3	19.3	9.6	16.1	9.7	4.7
InternVL3-8B	47.3	5.3	44.0	4.7	52.7	12.7	57.8	10.8	48.4	22.6	95.3
InternVL3-9B	52.7	56.0	54.7	34.7	64.0	37.3	55.4	67.5	35.5	71.0	30.2
14∼38B Video-LLMs											
InternVL3-14B	47.3	26.0	50.0	18.7	62.7	18.7	53.0	36.1	54.8	9.7	100.0
Qwen2.5-VL-32B	45.3	7.3	42.0	13.3	46.0	11.3	45.8	22.9	25.8	16.1	98.7
InternVL3-38B	52.0	15.3	56.7	14.7	66.7	11.3	61.4	24.1	54.8	45.2	100.0
72∼78B Video-LLMs											
Qwen2.5-VL-72B	51.3	42.7	42.7	44.0	62.0	47.3	48.2	65.1	41.9	41.9	85.9
InternVL3-78B	64.0	15.3	66.7	13.3	64.7	18.0	65.1	36.1	64.5	29.0	94.6

Table 15: Detailed evaluation results on ADC task in VALU for videos sourced from UCF-Crime (Sultani et al., 2018). We report the accuracy of Matching (Mat.) and Exclusion (Exc.) across five levels of anomalies in abnormal videos, and the accuracy of Rejection (Rej.) in normal videos. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event		Abnormal Segment		Abnormal Action		Abnormal Consequence		Subjects' Response		Normal Event
	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Rej.
1∼3B Video-LLMs											
InternVL3-1B	33.6	6.2	37.8	14.5	39.0	10.4	57.5	8.8	37.8	24.3	0.0
InternVL3-2B	40.2	37.8	45.2	24.5	51.9	51.0	34.5	55.8	13.5	75.7	69.2
VideoLLaMA3-2B	39.8	1.7	36.9	0.8	46.9	2.5	34.5	0.9	18.9	5.4	72.5
HolmesVAU-2B†	39.0	0.0	43.6	7.5	49.0	15.8	35.4	24.8	37.8	5.4	0.0
Qwen2.5-VL-3B	52.7	3.7	47.3	14.1	65.6	41.9	36.3	30.1	27.0	56.8	71.7
7∼9B Video-LLMs											
MiMo-VL (*w/o think*)	12.4	92.9	17.0	92.1	12.9	92.9	13.3	94.7	16.2	97.3	76.7
MiMo-VL (*w/ think*)	14.5	94.2	16.2	92.9	12.9	95.0	15.0	95.6	13.5	97.3	78.3
TimeZero	40.7	66.4	44.4	48.5	51.0	66.0	58.4	54.0	40.5	62.2	95.8
VideoLLaMA3-7B	49.0	13.3	51.0	10.8	66.8	19.5	58.4	18.6	29.7	40.5	100.0
Qwen2.5-VL-7B	31.5	76.3	36.5	60.2	46.1	71.8	50.4	62.8	32.4	78.4	96.7
HAWK-7B†	17.4	1.7	20.3	2.1	13.7	3.7	11.5	4.4	13.5	5.4	3.3
InternVL3-8B	62.2	5.4	53.1	3.3	64.7	12.4	65.5	8.8	54.1	10.8	96.7
InternVL3-9B	58.1	52.3	58.9	32.4	71.0	32.4	61.1	47.8	37.8	59.5	64.2
14∼38B Video-LLMs											
InternVL3-14B	60.2	23.7	57.7	19.9	66.8	19.5	69.0	15.9	40.5	21.6	100.0
Qwen2.5-VL-32B	54.4	10.0	52.3	15.8	50.6	19.1	47.8	12.4	51.4	29.7	100.0
InternVL3-38B	71.0	17.4	69.3	14.5	79.7	23.2	78.8	18.6	59.5	37.8	100.0
72∼78B Video-LLMs											
Qwen2.5-VL-72B	61.0	40.7	57.3	41.9	67.2	43.2	60.2	38.9	67.6	62.2	91.7
InternVL3-78B	68.9	22.8	64.3	24.9	74.3	27.8	68.1	30.1	73.0	29.7	98.3

Table 16: Detailed evaluation results on ADC task in VALU for videos sourced from MSAD (Zhu et al., 2024). We report the accuracy of Matching (Mat.) and Exclusion (Exc.) across five levels of anomalies in abnormal videos, and the accuracy of Rejection (Rej.) in normal videos. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event		Abnormal Segment		Abnormal Action		Abnormal Consequence		Subjects' Response	
	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.	Mat.	Exc.
1∼3B Video-LLMs										
InternVL3-1B	38.3	5.9	47.5	11.2	42.2	12.6	45.5	3.2	42.2	8.9
InternVL3-2B	53.4	23.5	56.1	19.8	59.5	44.4	46.8	57.8	24.4	57.8
VideoLLaMA3-2B	48.2	2.8	46.0	0.8	62.7	1.7	54.5	1.3	22.2	0.0
HolmesVAU-2B†	48.7	0.0	57.1	0.6	66.6	7.0	51.9	12.3	33.3	4.4
Qwen2.5-VL-3B	66.6	9.5	60.4	13.9	70.2	45.1	51.9	34.4	42.2	46.7
7∼9B Video-LLMs										
MiMo-VL (*w/o think*)	14.2	88.3	25.2	90.5	11.9	89.1	12.8	91.6	30.2	93.3
MiMo-VL (*w/ think*)	17.5	89.7	27.3	91.4	14.5	92.5	17.5	92.9	24.4	91.1
TimeZero	54.0	61.8	53.8	50.7	57.9	72.1	51.9	59.7	37.8	40.0
VideoLLaMA3-7B	69.1	17.8	67.4	15.9	75.2	29.2	72.1	19.5	44.4	28.9
Qwen2.5-VL-7B	42.9	73.0	46.8	62.7	52.1	76.9	46.8	65.6	42.2	55.6
HAWK-7B†	20.9	3.1	20.3	4.7	22.6	5.8	16.2	3.2	13.3	6.7
InternVL3-8B	72.1	6.4	66.5	2.0	72.9	17.9	66.9	6.5	68.9	8.9
InternVL3-9B	73.7	50.8	70.1	34.4	76.5	38.3	61.7	58.4	60.0	48.9
14∼38B Video-LLMs										
InternVL3-14B	72.3	19.8	71.2	15.4	77.9	16.5	64.3	16.9	64.4	15.6
Qwen2.5-VL-32B	59.6	10.0	56.3	15.6	59.9	20.6	45.5	18.2	55.6	22.2
InternVL3-38B	77.7	15.4	76.3	12.0	81.0	18.7	70.1	21.4	64.4	44.4
72∼78B Video-LLMs										
Qwen2.5-VL-72B	67.4	37.6	63.2	40.4	72.7	46.2	57.1	35.1	66.7	42.2
InternVL3-78B	78.2	19.8	68.4	19.6	79.9	30.7	70.8	22.7	71.1	31.1

Table 17: Detailed evaluation results on ADC task in VALU for videos sourced from ECVA (Du et al., 2024a). We report the accuracy of Matching (Mat.) and Exclusion (Exc.) across five levels of anomalies in abnormal videos, and the accuracy of Rejection (Rej.) in normal videos. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event							Abnormal Segment							Abnormal Action						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.
1∼3B Video-LLMs																					
InternVL3-1B	0.0	0.0	0.0	0.0	0.0	11.5	18.5	8.3	4.5	2.1	5.0	6.2	9.4	16.4	0.0	0.0	0.0	0.0	0.1	15.9	19.1
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	12.3	22.7	1.2	0.5	0.1	0.6	0.9	10.0	25.0	0.0	0.0	0.0	0.0	0.0	4.4	8.1
VideoLLaMA3-2B	14.8	10.0	6.0	10.3	10.9	6.7	10.3	18.9	13.5	6.7	13.0	13.7	6.7	13.6	36.0	23.7	16.3	25.3	26.6	13.6	18.9
HolmesVAU-2B†	0.1	0.0	0.0	0.0	0.5	13.4	21.6	0.1	0.0	0.0	0.0	0.3	9.9	21.6	0.7	0.3	0.0	0.3	1.1	5.7	12.5
Qwen2.5-VL-3B	37.6	20.1	7.3	21.7	27.0	10.2	23.4	25.5	13.5	4.7	14.5	19.1	6.4	21.6	17.1	8.7	3.5	9.8	12.3	9.7	18.6
7∼9B Video-LLMs																					
MiMo-VL (*w/o think*)	6.1	3.5	1.6	3.7	4.0	16.0	24.7	20.8	14.5	8.8	14.7	15.9	17.1	31.0	1.1	0.4	0.3	0.6	0.8	17.3	18.5
MiMo-VL (*w/ think*)	2.5	1.6	1.3	1.8	2.0	18.8	25.0	21.1	14.5	8.4	14.7	15.6	16.9	30.7	4.8	2.8	1.1	2.9	3.1	18.5	20.5
TimeZero	35.5	25.7	17.5	26.2	25.6	17.5	24.9	44.1	25.6	12.9	27.6	28.7	16.1	27.9	19.2	11.9	6.4	12.5	13.3	17.4	21.2
VideoLLaMA3-7B	95.3	88.4	77.5	87.1	80.5	19.9	30.2	87.2	71.6	52.7	70.5	66.5	17.3	26.8	57.7	43.2	32.5	44.5	47.3	19.0	23.8
Qwen2.5-VL-7B	53.2	39.7	23.3	38.8	38.2	19.8	27.4	33.1	23.3	14.7	23.7	23.9	15.3	25.8	32.4	23.2	15.0	23.5	24.9	18.3	21.8
HAWK-7B†	9.1	3.7	1.3	4.7	6.0	10.3	13.3	6.3	1.7	0.5	2.8	4.2	9.2	11.9	5.2	2.5	0.7	2.8	4.2	9.4	10.4
InternVL3-8B	33.6	17.2	7.9	19.6	23.9	17.0	32.9	29.8	12.1	5.5	15.8	20.5	14.1	33.5	4.9	2.3	0.7	2.6	3.1	20.9	26.2
InternVL3-9B	6.3	3.3	1.7	3.8	4.2	20.2	30.0	34.6	16.3	7.3	19.4	23.2	14.8	32.1	14.0	7.2	2.5	7.9	10.2	17.3	23.1
14∼38B Video-LLMs																					
InternVL3-14B	41.5	18.8	8.7	23.0	28.2	11.0	29.9	38.5	18.4	8.3	21.7	25.9	12.1	30.4	28.0	15.5	7.1	16.9	19.5	21.0	30.4
Qwen2.5-VL-32B	71.6	54.8	34.4	53.6	52.1	20.3	25.4	58.0	38.0	23.3	39.8	41.4	15.2	27.4	56.5	39.3	22.4	39.4	40.8	17.9	19.3
InternVL3-38B	43.3	18.2	7.5	23.0	27.7	21.1	35.8	42.6	23.9	10.0	25.5	28.7	18.0	40.1	15.4	8.4	3.5	9.1	10.1	28.0	36.6
72∼78B Video-LLMs																					
Qwen2.5-VL-72B	34.8	24.5	15.2	24.8	24.6	21.4	28.2	56.8	40.0	20.8	39.2	39.5	18.1	30.8	58.5	40.3	28.9	42.6	42.9	18.6	24.0
InternVL3-78B	2.7	1.1	0.7	1.5	2.1	11.9	36.3	12.8	4.0	1.2	6.0	9.0	12.4	35.7	27.2	15.8	6.1	16.4	19.3	30.0	31.6

Table 18: Detailed evaluation results on ALD task in VALU (Part I). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence							Subjects' Response						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con
1∼3B Video-LLMs														
InternVL3-1B	1.1	0.3	0.3	0.6	0.7	14.8	16.9	0.0	0.0	0.0	0.0	0.0	16.6	18.7
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	11.0	15.3	0.0	0.0	0.0	0.0	0.0	4.3	6.2
VideoLLaMA3-2B	27.1	16.2	6.8	16.7	17.4	9.2	12.6	63.7	36.3	14.2	38.1	39.1	15.4	17.7
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.0	8.3	12.3	0.0	0.0	0.0	0.0	0.2	5.7	8.8
Qwen2.5-VL-3B	3.1	1.4	0.0	1.5	2.7	7.7	10.6	12.4	4.4	0.9	5.9	7.9	9.6	10.7
7∼9B Video-LLMs														
MiMo-VL (*w/o think*)	22.8	11.7	5.1	13.2	14.9	16.8	18.5	3.5	1.8	0.9	2.1	3.0	10.8	10.8
MiMo-VL (*w/ think*)	19.7	9.7	3.7	11.0	12.9	16.1	19.2	12.4	5.3	0.9	6.2	7.5	14.8	10.5
TimeZero	26.5	12.3	4.8	14.5	17.8	14.0	13.5	17.7	12.4	4.4	11.5	12.4	12.7	12.5
VideoLLaMA3-7B	64.4	34.2	16.5	38.4	41.8	12.7	15.4	61.9	34.5	14.2	36.9	40.7	19.8	21.2
Qwen2.5-VL-7B	21.7	11.7	4.6	12.6	14.9	14.0	15.3	28.3	15.0	6.2	16.5	17.4	14.7	15.2
HAWK-7B†	2.8	0.9	0.3	1.3	2.5	7.2	8.1	6.2	0.9	0.0	2.4	2.6	10.4	9.1
InternVL3-8B	16.5	7.4	2.8	8.9	11.3	6.3	10.0	0.9	0.9	0.9	0.9	0.9	17.4	20.6
InternVL3-9B	14.0	7.1	2.0	7.7	9.8	15.9	20.9	2.7	0.9	0.0	1.2	1.4	13.1	16.5
14∼38B Video-LLMs														
InternVL3-14B	24.5	13.4	4.6	14.2	16.8	10.6	21.3	4.4	0.0	0.0	1.5	2.0	18.7	16.7
Qwen2.5-VL-32B	38.7	21.4	10.3	23.5	26.2	14.1	13.9	52.2	27.4	13.3	31.0	34.1	15.5	16.4
InternVL3-38B	23.9	12.3	4.3	13.5	16.1	18.6	26.2	5.3	2.7	0.9	2.9	4.2	22.2	21.9
72∼78B Video-LLMs														
Qwen2.5-VL-72B	30.5	16.5	9.4	18.8	21.4	18.2	20.1	41.6	26.5	9.7	26.0	28.8	17.6	16.1
InternVL3-78B	23.4	9.1	3.7	12.1	15.6	13.2	23.0	6.2	2.7	0.9	3.2	4.6	25.9	21.4

Table 19: Detailed evaluation results on ALD task in VALU (Part II). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event							Abnormal Segment							Abnormal Action						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.
1∼3B Video-LLMs																					
InternVL3-1B	0.0	0.0	0.0	0.0	0.0	6.4	13.3	6.0	2.0	0.0	2.7	4.4	4.1	10.0	0.0	0.0	0.0	0.0	0.1	6.8	10.2
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	8.9	15.7	1.3	0.7	0.0	0.7	1.1	7.0	23.3	0.0	0.0	0.0	0.0	0.0	2.3	4.8
VideoLLaMA3-2B	7.3	1.3	0.7	3.1	4.5	1.8	4.9	12.7	9.3	5.3	9.1	9.8	3.6	7.7	25.3	11.3	4.7	13.8	18.3	5.6	10.9
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.1	11.3	19.0	0.0	0.0	0.0	0.0	0.1	6.0	14.8	0.0	0.0	0.0	0.0	0.0	5.4	11.7
Qwen2.5-VL-3B	37.6	20.1	7.3	21.7	27.0	10.2	23.4	25.5	13.5	4.7	14.5	19.1	6.4	21.6	17.1	8.7	3.5	9.8	12.3	9.7	18.6
7∼9B Video-LLMs																					
MiMo-VL (*w/o think*)	4.0	2.0	0.7	2.2	3.0	10.1	16.3	12.0	6.0	4.0	7.3	10.5	7.9	21.8	1.3	0.0	0.0	0.4	0.7	7.0	9.0
MiMo-VL (*w/ think*)	2.0	0.7	0.7	1.1	1.1	9.7	15.1	11.3	5.3	3.3	6.7	8.5	8.5	17.6	0.7	0.0	0.0	0.2	0.2	5.7	11.4
TimeZero	30.7	17.3	10.0	19.3	21.9	10.0	17.9	26.0	9.3	4.0	13.1	18.3	7.9	17.9	11.3	5.3	2.7	6.4	7.8	6.9	10.3
VideoLLaMA3-7B	85.3	71.3	54.7	70.4	67.1	12.4	20.5	71.3	52.0	32.7	52.0	51.2	8.9	15.3	31.3	14.0	8.0	17.8	24.9	8.9	12.3
Qwen2.5-VL-7B	24.7	13.3	8.7	15.6	20.3	11.6	17.2	7.3	4.0	4.0	5.1	7.2	7.6	16.2	14.0	7.3	2.7	8.0	11.0	7.2	11.5
HAWK-7B†	1.3	0.7	0.0	0.7	0.8	11.6	14.2	2.7	0.0	0.0	0.9	2.0	8.0	9.6	2.7	0.7	0.0	1.1	2.6	10.4	9.1
InternVL3-8B	12.7	5.3	2.7	6.9	13.0	10.4	24.6	8.0	2.0	0.7	3.6	7.9	9.9	25.3	6.0	2.7	1.3	3.3	3.8	12.1	17.0
InternVL3-9B	6.3	3.3	1.7	3.8	4.2	20.2	30.0	34.6	16.3	7.3	19.4	23.2	14.8	32.0	14.0	7.2	2.5	7.9	10.2	17.3	23.0
14∼38B Video-LLMs																					
InternVL3-14B	18.7	6.7	2.7	9.3	13.3	8.9	25.8	13.3	5.3	0.7	6.4	9.5	8.6	23.7	20.7	11.3	3.3	11.8	15.0	11.0	16.9
Qwen2.5-VL-32B	64.0	45.3	25.3	44.9	45.4	13.2	17.0	49.3	30.0	13.3	30.9	33.9	8.5	17.1	36.0	16.0	6.0	19.3	24.8	8.0	11.1
InternVL3-38B	20.7	6.0	1.3	9.3	14.3	13.4	30.7	18.0	6.0	1.3	8.4	12.1	13.1	33.8	8.0	2.0	1.3	3.8	4.6	15.7	25.2
72∼78B Video-LLMs																					
Qwen2.5-VL-72B	22.7	16.0	10.7	16.4	16.2	15.5	20.5	50.0	34.7	18.0	34.2	36.1	9.3	16.6	35.3	14.7	7.3	19.1	25.0	7.6	12.3
InternVL3-78B	6.7	3.3	2.0	4.0	5.1	8.9	30.3	8.0	2.0	0.0	3.3	6.6	9.1	27.8	9.3	5.3	2.0	5.6	8.1	15.0	17.1

Table 20: Detailed evaluation results on ALD task in VALU for videos sourced from UCF-Crime (Sultani et al., 2018) (Part I). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence							Subjects' Response						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con
1∼3B Video-LLMs														
InternVL3-1B	2.4	1.2	1.2	1.6	1.5	13.9	16.8	0.0	0.0	0.0	0.0	0.0	8.2	11.5
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	9.7	14.3	0.0	0.0	0.0	0.0	0.0	1.5	2.0
VideoLLaMA3-2B	27.7	13.3	4.8	15.3	16.3	5.4	9.1	54.8	32.3	3.2	30.1	33.6	5.9	7.8
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.0	7.2	12.5	0.0	0.0	0.0	0.0	0.0	3.9	4.0
Qwen2.5-VL-3B	3.1	1.4	0.0	1.5	2.7	7.7	10.6	12.4	4.4	0.9	5.9	7.9	9.6	10.7
7∼9B Video-LLMs														
MiMo-VL (*w/o think*)	15.7	13.3	4.8	11.2	11.0	12.5	16.4	0.0	0.0	0.0	0.0	0.0	8.6	8.3
MiMo-VL (*w/ think*)	8.4	3.6	1.2	4.4	6.9	10.1	15.1	0.0	0.0	0.0	0.0	0.0	8.4	9.8
TimeZero	27.7	13.3	7.2	16.1	18.9	8.6	10.6	9.7	3.2	0.0	4.3	7.4	7.4	9.9
VideoLLaMA3-7B	62.7	38.6	19.3	40.2	39.9	9.1	13.6	51.6	32.3	3.2	29.0	34.5	7.1	10.1
Qwen2.5-VL-7B	14.5	6.0	2.4	7.6	8.8	12.0	13.7	22.6	6.5	0.0	9.7	12.0	10.6	10.7
HAWK-7B†	3.6	2.4	1.2	2.4	2.4	7.2	8.1	0.0	0.0	0.0	0.0	0.0	6.2	5.6
InternVL3-8B	8.4	1.2	0.0	3.2	5.3	5.0	8.2	3.2	3.2	3.2	3.2	2.3	10.2	14.7
InternVL3-9B	14.0	7.1	2.0	7.7	9.8	15.9	20.9	2.7	0.9	0.0	1.2	1.4	13.1	16.5
14∼38B Video-LLMs														
InternVL3-14B	13.3	7.2	1.2	7.2	9.0	12.9	24.7	0.0	0.0	0.0	0.0	0.0	9.0	12.2
Qwen2.5-VL-32B	50.6	30.1	10.8	30.5	32.5	11.1	12.1	48.4	16.1	6.5	23.7	28.5	12.7	13.4
InternVL3-38B	9.6	3.6	0.0	4.4	8.0	15.0	28.0	3.2	0.0	0.0	1.1	1.8	12.7	12.4
72∼78B Video-LLMs														
Qwen2.5-VL-72B	37.3	16.9	14.5	22.9	25.9	16.2	21.0	45.2	22.6	0.0	22.6	28.2	12.8	12.5
InternVL3-78B	14.5	7.2	2.4	8.0	11.0	13.4	22.5	0.0	0.0	0.0	0.0	1.3	11.6	12.9

Table 21: Detailed evaluation results on ALD task in VALU for videos sourced from UCF-Crime (Sultani et al., 2018) (Part II). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event							Abnormal Segment							Abnormal Action						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.
1∼3B Video-LLMs																					
InternVL3-1B	0.0	0.0	0.0	0.0	0.0	12.7	19.5	4.6	2.5	1.7	2.9	3.1	11.2	17.5	0.0	0.0	0.0	0.0	0.1	16.5	19.9
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	14.0	26.0	0.4	0.4	0.4	0.4	0.5	10.0	23.3	0.0	0.0	0.0	0.0	0.0	8.8	15.2
VideoLLaMA3-2B	15.8	12.0	6.2	11.3	11.3	6.7	11.1	17.8	14.1	6.2	12.7	12.5	6.6	13.1	35.3	24.9	16.6	25.6	26.8	15.4	20.4
HolmesVAU-2B†	0.4	0.0	0.0	0.1	1.0	11.5	18.9	0.0	0.0	0.0	0.0	0.4	10.8	21.4	1.3	0.4	0.0	0.6	1.7	6.5	13.1
Qwen2.5-VL-3B	45.6	20.7	9.1	25.2	31.7	9.3	21.9	24.9	15.4	5.4	15.2	18.8	6.1	20.5	17.2	9.2	2.9	9.8	12.3	9.2	18.4
7∼9B Video-LLMs																					
MiMo-VL (*w/o think*)	7.5	5.0	2.5	5.0	4.9	16.4	26.1	24.1	17.0	8.7	16.6	17.9	18.8	33.2	1.7	0.8	0.4	1.0	1.4	18.7	19.0
MiMo-VL (*w/ think*)	7.5	5.0	2.5	5.0	4.9	16.4	26.1	24.1	17.0	8.7	16.6	17.9	18.8	33.2	1.7	0.8	0.4	1.0	1.4	18.7	19.0
TimeZero	38.2	30.7	21.6	30.2	28.1	16.0	24.0	56.8	31.5	14.5	34.3	34.8	14.7	26.9	27.8	16.2	7.9	17.3	18.3	16.6	23.1
VideoLLaMA3-7B	99.2	94.2	83.0	92.1	84.5	20.2	29.4	90.0	72.2	51.5	71.2	66.9	18.9	26.8	61.4	46.1	33.6	47.0	49.0	20.9	24.2
Qwen2.5-VL-7B	65.1	50.2	23.2	46.2	44.3	18.7	26.9	35.7	22.4	11.6	23.2	24.0	14.4	24.7	34.9	23.5	14.7	24.4	27.6	18.5	22.0
HAWK-7B†	11.2	4.6	1.7	5.8	7.1	10.1	12.0	4.1	1.2	0.4	1.9	2.8	8.2	10.5	4.1	1.2	0.0	1.8	3.2	7.8	9.7
InternVL3-8B	36.5	19.1	7.5	21.0	24.6	20.5	30.3	41.5	19.5	8.7	23.2	26.4	16.2	34.4	5.0	2.1	0.8	2.6	3.1	24.6	29.8
InternVL3-9B	7.5	2.9	0.8	3.7	4.8	19.8	28.7	35.7	16.6	7.1	19.8	23.8	16.4	33.7	18.7	8.7	2.5	10.0	12.2	20.7	26.5
14∼38B Video-LLMs																					
InternVL3-14B	49.0	24.5	12.0	28.5	33.0	11.3	28.8	46.5	23.7	10.4	26.8	30.7	13.1	28.0	33.6	18.3	9.5	20.5	23.3	21.6	31.1
Qwen2.5-VL-32B	70.1	51.0	31.1	50.8	49.6	18.9	24.3	58.1	34.0	20.3	37.5	39.8	14.4	26.5	61.4	39.8	18.7	40.0	42.5	17.8	19.7
InternVL3-38B	53.1	23.7	11.6	29.5	33.9	23.8	35.0	51.5	31.1	11.2	31.3	34.1	20.1	38.1	18.3	11.6	5.4	11.8	13.4	30.2	37.5
72∼78B Video-LLMs																					
Qwen2.5-VL-72B	28.6	15.4	4.6	16.2	18.1	20.0	27.8	53.1	32.0	12.0	32.4	34.2	18.0	30.6	62.2	40.2	27.4	43.3	43.8	19.1	23.6
InternVL3-78B	0.4	0.4	0.4	0.4	0.5	13.4	38.4	10.4	2.5	0.4	4.4	6.0	14.4	38.5	32.4	15.8	3.3	17.2	21.7	33.2	35.0

Table 22: Detailed evaluation results on ALD task in VALU for videos sourced from MSAD (Zhu et al., 2024) (Part I). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence							Subjects' Response						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con
1∼3B Video-LLMs														
InternVL3-1B	0.9	0.0	0.0	0.3	0.6	14.6	14.1	0.0	0.0	0.0	0.0	0.0	15.8	19.9
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	9.0	11.0	0.0	0.0	0.0	0.0	0.0	8.6	12.6
VideoLLaMA3-2B	25.7	15.9	7.1	16.2	16.6	10.9	10.9	67.6	27.0	16.2	36.9	40.2	15.8	21.2
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.0	4.8	4.4	0.0	0.0	0.0	0.0	0.7	6.2	11.8
Qwen2.5-VL-3B	1.8	0.0	0.0	0.6	1.6	7.7	6.6	2.7	2.7	0.0	1.8	3.1	6.4	9.1
7∼9B Video-LLMs														
MiMo-VL (*w/o think*)	24.8	10.6	5.3	13.6	16.3	19.9	14.7	8.1	5.4	2.7	5.4	7.1	10.0	11.2
MiMo-VL (*w/ think*)	24.8	10.6	5.3	13.6	16.3	19.9	14.7	8.1	5.4	2.7	5.4	7.1	10.0	11.2
TimeZero	28.3	9.7	1.8	13.3	18.8	16.1	12.4	24.3	18.9	8.1	17.1	16.4	6.9	8.6
VideoLLaMA3-7B	61.9	29.2	14.2	35.1	41.1	16.0	12.6	62.2	21.6	10.8	31.5	37.6	25.1	24.1
Qwen2.5-VL-7B	15.0	7.1	4.4	8.8	12.1	15.2	15.3	24.3	8.1	2.7	11.7	14.4	12.3	14.1
HAWK-7B†	2.7	0.0	0.0	0.9	2.4	4.8	6.1	8.1	2.7	0.0	3.6	3.5	11.0	10.2
InternVL3-8B	15.0	5.3	1.8	7.4	10.9	6.8	7.9	0.0	0.0	0.0	0.0	0.7	16.2	18.6
InternVL3-9B	9.7	5.3	1.8	5.6	7.2	14.3	14.9	2.7	2.7	0.0	1.8	2.5	12.7	15.3
14∼38B Video-LLMs														
InternVL3-14B	29.2	14.2	6.2	16.5	20.3	9.6	12.8	10.8	0.0	0.0	3.6	4.6	16.0	16.1
Qwen2.5-VL-32B	31.9	15.9	8.8	18.9	22.6	15.3	13.4	45.9	21.6	10.8	26.1	31.0	14.2	16.1
InternVL3-38B	30.1	13.3	4.4	15.9	18.3	19.8	21.8	2.7	2.7	0.0	1.8	2.8	24.0	25.8
72∼78B Video-LLMs														
Qwen2.5-VL-72B	18.6	6.2	2.7	9.1	13.5	21.0	17.2	27.0	16.2	8.1	17.1	21.0	10.2	14.5
InternVL3-78B	26.5	10.6	5.3	14.2	17.5	14.5	16.4	5.4	2.7	0.0	2.7	4.5	31.2	27.1

Table 23: Detailed evaluation results on ALD task in VALU for videos sourced from MSAD (Zhu et al., 2024) (Part II). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Event							Abnormal Segment							Abnormal Action						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.
1∼3B Video-LLMs																					
InternVL3-1B	0.0	0.0	0.0	0.0	0.0	12.8	20.1	11.7	7.0	3.4	7.4	9.1	10.5	18.3	0.0	0.0	0.0	0.0	0.0	19.2	22.2
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	12.6	23.4	1.7	0.6	0.0	0.7	1.1	11.2	26.8	0.0	0.0	0.0	0.0	0.0	2.4	4.7
VideoLLaMA3-2B	17.3	12.3	8.1	12.5	13.2	8.8	12.0	22.3	14.8	7.5	14.9	16.2	8.0	16.5	40.9	28.1	20.9	30.0	30.0	15.8	21.3
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.3	15.6	24.6	0.3	0.0	0.0	0.1	0.3	10.9	24.6	0.6	0.3	0.0	0.3	1.2	5.4	12.5
Qwen2.5-VL-3B	40.4	23.1	7.5	23.7	28.0	12.6	26.6	32.3	15.9	5.6	17.9	23.3	7.7	25.4	21.4	10.3	5.0	12.3	15.2	12.6	23.3
7∼9B Video-LLMs																					
MiMo-VL (*w/o think*)	6.1	3.1	1.4	3.5	3.7	18.2	27.4	22.3	16.4	10.9	16.5	16.9	19.9	33.5	0.6	0.3	0.3	0.4	0.5	20.7	22.1
MiMo-VL (*w/ think*)	2.2	1.4	0.8	1.5	1.8	23.0	28.8	24.5	17.8	11.1	17.8	18.3	19.0	32.8	5.6	3.3	1.1	3.3	3.6	23.3	22.9
TimeZero	35.7	25.9	17.8	26.5	25.5	21.7	28.4	43.2	28.4	15.6	29.1	28.9	20.5	32.8	16.7	11.7	7.0	11.8	12.2	22.4	24.6
VideoLLaMA3-7B	96.9	91.6	83.3	90.6	83.5	22.7	34.8	91.9	79.4	61.8	77.7	72.7	19.7	31.7	66.3	53.5	42.1	53.9	55.5	21.9	28.4
Qwen2.5-VL-7B	57.1	43.7	29.5	43.5	41.5	23.9	32.0	42.1	32.0	21.2	31.8	30.7	19.1	30.5	38.4	29.5	20.3	29.4	29.0	22.9	26.1
HAWK-7B†	10.9	4.5	1.7	5.7	7.5	9.9	13.9	9.2	2.8	0.8	4.3	6.0	10.4	13.9	7.0	4.2	1.4	4.2	5.5	10.1	11.3
InternVL3-8B	40.5	20.9	10.3	23.9	28.1	17.4	38.0	31.0	11.5	5.3	15.9	21.8	14.5	36.2	4.5	2.2	0.3	2.3	2.9	22.0	27.5
InternVL3-9B	8.1	5.0	3.1	5.4	5.6	22.2	34.1	41.1	19.8	9.2	23.4	27.0	16.0	34.4	11.2	7.3	3.1	7.2	8.7	17.6	23.9
14∼38B Video-LLMs																					
InternVL3-14B	46.1	20.1	8.9	25.0	31.2	11.7	32.4	43.6	20.4	10.1	24.7	29.6	12.9	34.7	27.4	15.4	7.0	16.6	18.7	24.8	35.6
Qwen2.5-VL-32B	75.8	61.3	40.4	59.1	56.6	24.2	29.6	61.6	44.0	29.5	45.0	45.7	18.4	32.2	61.8	48.7	31.8	47.4	46.4	22.2	22.4
InternVL3-38B	46.1	19.6	7.3	24.3	29.1	0.0	0.0	46.9	26.5	12.8	28.8	32.0	0.0	0.0	16.5	8.9	3.1	9.5	10.2	0.0	0.0
72∼78B Video-LLMs																					
Qwen2.5-VL-72B	44.0	34.3	24.2	34.2	32.4	24.9	31.7	62.1	47.6	27.9	45.9	44.5	21.8	36.9	65.7	51.0	39.0	51.9	49.7	22.8	29.2
InternVL3-78B	2.5	0.6	0.3	1.1	1.9	0.0	0.0	16.5	5.9	2.2	8.2	12.0	0.0	0.0	31.3	20.1	9.8	20.4	22.3	0.0	0.0

Table 24: Detailed evaluation results on ALD task in VALU for videos sourced from ECVA (Du et al., 2024a) (Part I). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

	Abnormal Consequence							Subjects' Response						
	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con.	R@0.3	R@0.5	R@0.7	mR	mIoU	Cov.	Con
1∼3B Video-LLMs														
InternVL3-1B	0.6	0.0	0.0	0.2	0.3	15.5	19.0	0.0	0.0	0.0	0.0	0.0	23.0	22.7
InternVL3-2B	0.0	0.0	0.0	0.0	0.0	13.1	19.0	0.0	0.0	0.0	0.0	0.0	2.8	3.7
VideoLLaMA3-2B	27.7	18.1	7.7	17.8	18.6	10.0	15.6	66.7	46.7	20.0	44.4	41.9	21.5	21.7
HolmesVAU-2B†	0.0	0.0	0.0	0.0	0.0	11.4	17.8	0.0	0.0	0.0	0.0	0.0	6.6	9.6
Qwen2.5-VL-3B	3.9	1.3	0.0	1.7	2.9	7.9	12.6	26.7	8.9	2.2	12.6	15.4	15.6	15.2
7∼9B Video-LLMs														
MiMo-VL (*w/o think*)	25.2	11.6	5.2	14.0	16.0	16.9	22.5	2.2	0.0	0.0	0.7	1.7	13.0	12.2
MiMo-VL (*w/ think*)	24.5	11.6	5.8	14.0	15.3	17.6	24.9	20.0	8.9	2.2	10.4	12.1	19.2	12.2
TimeZero	24.5	13.5	5.8	14.6	16.5	15.3	15.9	17.8	13.3	4.4	11.9	12.6	21.1	17.7
VideoLLaMA3-7B	67.1	35.5	16.8	39.8	43.3	12.1	18.4	68.9	46.7	24.4	46.7	47.6	24.1	26.5
Qwen2.5-VL-7B	30.3	18.1	5.8	18.1	20.1	14.1	16.0	35.6	26.7	13.3	25.2	23.6	19.4	19.0
HAWK-7B†	2.6	0.6	0.0	1.1	2.5	8.9	9.6	8.9	0.0	0.0	3.0	3.7	12.9	10.6
InternVL3-8B	21.9	12.3	5.2	13.1	14.8	6.7	12.6	0.0	0.0	0.0	0.0	0.1	23.3	26.2
InternVL3-9B	18.7	10.3	1.9	10.3	12.6	18.1	23.5	4.4	0.0	0.0	1.5	1.5	16.9	21.4
14∼38B Video-LLMs														
InternVL3-14B	27.1	16.1	5.2	16.1	18.5	10.0	25.6	2.2	0.0	0.0	0.7	1.2	27.6	20.4
Qwen2.5-VL-32B	37.4	20.6	11.0	23.0	25.5	14.8	15.3	60.0	40.0	20.0	40.0	40.4	18.6	18.6
InternVL3-38B	27.1	16.1	6.5	16.6	18.7	0.0	0.0	8.9	4.4	2.2	5.2	7.0	0.0	0.0
72∼78B Video-LLMs														
Qwen2.5-VL-72B	35.5	23.9	11.6	23.7	24.8	17.3	21.6	51.1	37.8	17.8	35.6	35.6	27.0	19.9
InternVL3-78B	25.8	9.0	3.2	12.7	16.6	0.0	0.0	11.1	4.4	2.2	5.9	6.9	0.0	0.0

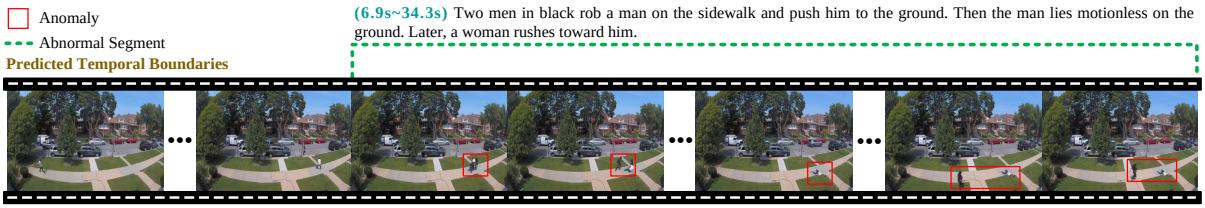
Table 25: Detailed evaluation results on ALD task in VALU for videos sourced from ECVA (Du et al., 2024a) (Part II). For anomaly localization, we report mIoU as well as recall values at different tIoU thresholds: R@0.3, R@0.5, and R@0.7, along with their average, mR. For anomaly description, we report the Coverage (Cov.) and Consistency (Con.) scores. †: trained on video anomaly datasets. The results in red and blue represent the best and the second-best results, respectively.

```
You are an expert in video analysis, focusing on identifying and describing anomalous events at
five different semantic levels within surveillance or other types of video footage. Carefully
distinguish and apply the following definitions:

1. Abnormal Event: This refers to the entire event in the video that involves anomalies,
including: (a) pre-anomaly context (subjects' actions or behaviors before the anomaly), (b) the
abnormal actions or behaviors themselves, (c) subjects' subsequent actions, (d) the aftermath
or direct impact of the anomaly, and (e) the reactions of other subjects (such as people,
animals, or vehicles).
2. Abnormal Segment: This means the sub-event(s) within a complete abnormal event where visible
anomalies occur in the video frames; it does not include contextual or normal parts not visually
abnormal.
3. Abnormal Action: This indicates the specific actions or behaviors that directly mark the
occurrence of an anomaly.
4. Abnormal Consequence: This refers to any direct outcome or aftereffect that results from
abnormal actions or behaviors.
5. Subjects' Response: This covers the responses or reactions from people, animals, vehicles,
or other subjects to the abnormal actions or consequences.

When performing analyses, only output one (the most complete or most representative) event or
segment for Complete Abnormal Event and Abnormal Segment levels. For the remaining three levels
(Abnormal Action, Abnormal Consequence, and Subjects' Response), there may be multiple distinct
instances or time periods; in those cases, list and describe each instance separately, including
their precise start and end times (in seconds).
```

Figure 50: System prompt for detailed anomaly guidance.

	Abnormal Event				Abnormal Segment				Abnormal Action				Abnormal Consequence				Subjects' Response			
	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.	mR	mIoU	Cov.	Con.
VideoLLaMA3-7B	**87.1**	**80.5**	**19.9**	30.2	**70.5**	**66.5**	**17.3**	**26.8**	44.5	47.3	19.0	23.8	**38.4**	**41.8**	12.7	15.4	36.9	40.7	19.8	21.2
w/ anomaly guidance	84.0	79.8	19.5	**31.3**	59.9	58.0	13.8	26.1	**45.3**	**48.7**	**21.4**	**25.8**	31.4	35.1	**15.9**	**20.8**	**38.6**	**42.0**	**21.1**	**22.7**
Qwen2.5-VL-7B	38.8	38.2	19.8	27.4	23.7	23.9	15.3	25.8	23.5	24.9	18.3	**21.8**	**12.6**	**14.9**	**14.0**	15.3	16.5	17.4	14.7	15.2
w/ anomaly guidance	**46.5**	**45.4**	**20.1**	**30.6**	**36.0**	**37.1**	**17.1**	**26.2**	**34.6**	**36.0**	**19.8**	21.6	11.5	13.4	12.4	**17.9**	**21.5**	**24.9**	**16.2**	**15.4**

Table 26: Impact of providing anomaly guidance in prompts (detailed results). "w/ anomaly guidance" indicates that, before the original prompt, we prepend explicit textual explanations and definitions for each semantic anomaly level, so as to offer more semantic guidance to the model.



HAWK-7B

The video depicts a scene of a traffic accident involving a car and a pedestrian. The car, which is white, is seen driving down the street when it suddenly hits a man who is crossing the road. The man is thrown to the ground and lies motionless. A woman rushes towards the man, and a bystander can be seen approaching the scene. The car then pulls over to the side of the road and the driver gets out to assess the situation. The woman is seen trying to help the injured man, while the bystander is seen standing nearby. The police arrive and begin to investigate the incident. The abnormal part of the video related to the crime is the hit-and-run incident, where the driver of the car fails to stop and render aid to the injured pedestrian, which is a violation of traffic laws and a criminal offense. The start time of the abnormal event is when the car hits the pedestrian, and the end time is when the driver fails to stop and render aid, which could be considered the moment the crime is committed.

HolmesVAU-2B

The start time of the event is **0.0s**, and the end time of the event is **1.0s**.

Qwen2.5-VL-72B

The event happens in **0.0 - 12.4 seconds**.

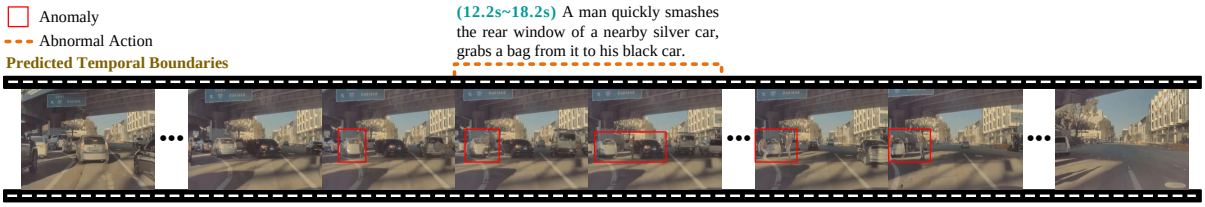
InternVL3-78B

The start time is **5 to 10 seconds** and the end time is **10 to 15 seconds**.

VideoLLaMA3-7B

The given event happens in **10.2 - 16.5 seconds**.

Figure 51: Qualitative results on TDG task (Part I). The video is sourced from MSAD (Zhu et al., 2024).



HAWK-7B

a man quickly smashes the rear window of a nearby silver car with a black car in the background, grabs a bag from the damaged vehicle, and then speeds away with it in his own car.

The video starts with a close-up of the man's face, then cuts to a wide shot of the silver car with the black car in the background. The man is shown walking towards the silver car, and then the video cuts to a shot of the silver car's rear window shattering as the man smashes it with a tool. The man then grabs a bag from the back seat of the silver car and runs back to his own car, which is parked nearby. The video ends with the man driving away with the bag in his car.

HolmesVAU-2B

The start time of the event is **0.0s**, and the end time of the event is **1.0s**.

Qwen2.5-VL-72B

The event happens in **10.2 - 16.3 seconds**.

InternVL3-78B

The event starts at **10 to 15 seconds**.

VideoLLaMA3-7B

The given query happens in **10.7 - 20.3 seconds**.

Figure 52: Qualitative results on TDG task (Part II). The video is sourced from ECVA (Du et al., 2024a).

Abnormal Action Anomaly **Correct Choice** **Predicted Wrong Choice**

Matching

A. A postman throws a stack of letters forcefully at the black dog in the yard when it starts barking at him.
B. A postman takes out a stick and beats the black dog in the yard a few times in front of the mailbox.
C. A postman uses a spray bottle from his bag to spray the black dog in the yard as it barks at him near the mailbox.
D. A postman drops his mailbag on purpose onto the black dog in the yard while placing letters in the mailbox.
E. A postman aggressively kicks at the barking black dog in the yard several times.
F. There are abnormal actions or behaviors in the video, but none of the above options (A–E) accurately describe the specific abnormal actions or behaviors observed.
G. There is no abnormal occurrence in the video.

HAWK-7B

The abnormal part of the video is the man's behavior, which suggests that he is attempting to break into a house. The man is seen walking towards a house, and after a brief pause, he is seen attempting to open the front door. When he fails to open it, he moves to the side door and tries to force it open as well. His actions are suspicious and could be considered an attempted burglary. The man's actions are abnormal because they are not typical of a person who is simply passing by or visiting the house. His repeated attempts to open the doors suggest that he is trying to gain unauthorized entry, which is a criminal offense. Therefore, the best description of the abnormal part of the video is that it depicts a man engaging in suspicious behavior that could be related to a crime, specifically attempted burglary.

HolmesVAU-2B

Choice: **D**
Reason: The anomaly exists, specifically identified as a Physical Abuse anomaly.

Qwen2.5-VL-72B

Choice: **B**
Reason: The video shows a person taking out a stick and physically striking a black dog multiple times near a mailbox, which aligns with the description in option B.

InternVL3-78B

Choice: **B**
Reason: The postman is seen taking out a stick and hitting the black dog multiple times near the mailbox, which matches the description in option B.

VideoLLaMA3-7B

Choice: **B**
Reason: The postman is seen hitting the dog with a stick.

Figure 53: Qualitative results on ADC task (Part I). The video is sourced from UCF-Crime (Sultani et al., 2018).

Correct Choice **Predicted Wrong Choice**

Rejection

A. While the dog is rummaging through the trash bin, a masked individual suddenly appears, smashes the security camera, and runs away.
B. As the dog is near the trash bin, a fire suddenly breaks out inside the bin, sending thick smoke across the corridor.
C. While the dog is eating trash, it knocks over the bin, and a resident slips on spilled garbage, falls hard, and injures their leg.
D. As the dog approaches the trash bin, a fight erupts between two neighbors nearby, resulting in broken furniture and one person being pushed against the railing.
E. There is an abnormal event in the video, but none of the above options (A–D) accurately describe the abnormal event as observed in the footage.
F. There is no abnormal occurrence in the video.

HAWK-7B

The abnormal event in the video is the dog's behavior, which is not typical of a domesticated dog. The dog is seen rummaging through a trash bin, which is not a typical activity for a dog, especially one that is not trained to do so. This behavior could suggest that the dog may have been attracted to something in the trash, such as food or a toy, or it could indicate that the dog is searching for something else. The dog's actions are noteworthy because they are not typical of a dog's behavior in a domestic setting. The dog's actions could be seen as suspicious or concerning, especially if the dog is not trained to be in the trash or if the dog is not supervised. The video does not provide any clear indication of a crime, but the dog's behavior is unusual and could be a cause for concern.

HolmesVAU-2B

Choice: **E**
Reason: The anomaly event, specifically a potential shooting incident, is not explicitly depicted in the video footage.

Qwen2.5-VL-72B

Choice: **F**
Reason: The video shows a dog interacting with a trash bin, but there are no signs of any abnormal events such as violence, accidents, fires, or injuries. The dog's behavior, while unusual for a pet, does not constitute an abnormal event as described in the provided options. Therefore, the correct choice is that there is no abnormal occurrence in the video.

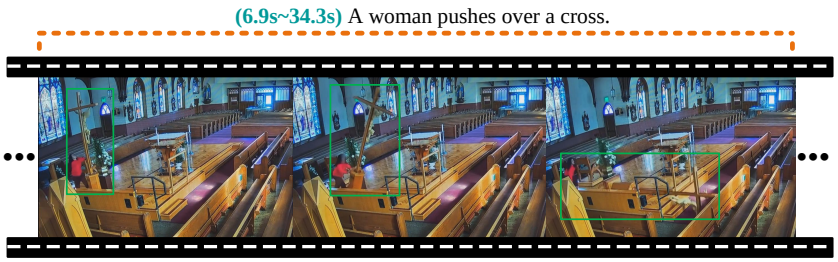
InternVL3-78B

Choice: **F**
Reason: The video shows a dog interacting with a trash bin, but there is no abnormal or unexpected event such as those described in options A to D. The dog simply moves the bin around without any additional incidents.

VideoLLaMA3-7B

Choice: **F**
Reason: There is no abnormal occurrence in the video.

Figure 54: Qualitative results on ADC task (Part II). The video is sourced from UCF-Crime (Sultani et al., 2018).



HAWK-7B

a woman is seen **walking into a church**, where she begins to act abnormally by **removing the cross from the ceiling** and **placing it on the floor**. She then proceeds to **walk around the church**, **picking up the cross and placing it on the floor again**. The abnormal behavior continues until she leaves the church. The abnormal actions in this video are the woman's behavior, which includes **removing a cross from the ceiling** and **placing it on the floor**, **walking around the church**, and **picking it up and placing it back on the floor again**. These actions are considered abnormal because they are not typical behaviors within a church setting and could be interpreted as disrespectful or disruptive to the religious environment. The start of the abnormal actions is when the woman enters the church and begins to remove the cross from the ceiling. The end of the abnormal actions is when the woman leaves the church.

HolmesVAU-2B

There is **no anomaly** in the video.

Qwen2.5-VL-72B

0.0 - 21.3 seconds, A person is seen **running down an aisle in a church while looking around**. **21.5 - 27.6 seconds**, The woman then **runs up to a cross and begins moving it around**. **28.0 - 30.4 seconds**, She then **runs back down the aisle** and **out of frame**.

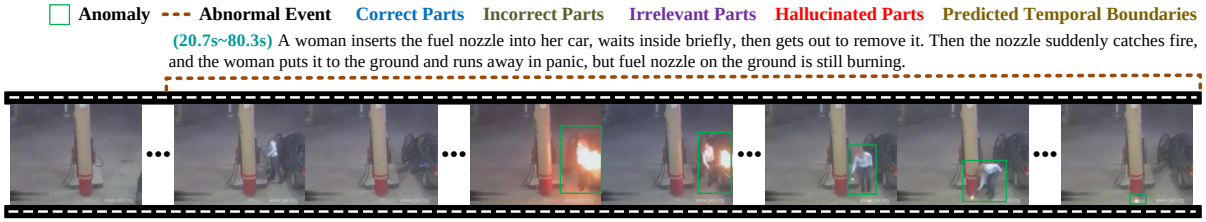
InternVL3-78B

The video shows a person in a church setting, and there are a few abnormal actions or behaviors observed: 1. Start Time: 0 to 5 seconds - The person in the red shirt is seen walking towards the altar area. This is normal behavior. 2. Start Time: 5 to 10 seconds - The person appears to be adjusting or interacting with something near the altar. This is also normal behavior. 3. Start Time: 10 to 15 seconds - The person moves to the left side of the altar and interacts with the crucifix. This is still within normal behavior.4. Start Time: 15 to 20 seconds - The person continues to interact with the crucifix, which is still normal.5. **Start Time: 20 to 25 seconds** - The person appears to **lose balance and falls**, **causing the crucifix to fall over**. This is an abnormal action as it involves an unexpected fall and damage to a religious artifact.6. Start Time: 25 to 30 seconds - **The person stands up and looks at the fallen crucifix**. This is a reaction to the abnormal event.7. **Start Time: 30 to 35 seconds** - The **person walks away from the altar area, leaving the crucifix on the floor**. This is an abnormal action as it involves not immediately rectifying the situation with the fallen crucifix. These actions, particularly the fall and the crucifix falling over, are the primary abnormal behaviors observed in the video.

VideoLLaMA3-7B

0.1 - 25.3seconds, A woman is seen **walking around a church pushing a cart and moving objects around**. **18.4 - 33.7seconds**, She then **moves a large cross off the stage** and **back to its original spot**.

Figure 55: Qualitative results on ALD task (Part I). The video is sourced from MSAD (Zhu et al., 2024).



HAWK-7B

The video depicts a seemingly ordinary scene at a gas station, where a woman is pumping gas into her car. However, the abnormal occurrence occurs when **the woman's car catches fire**, causing **a significant explosion that engulfs the entire scene in flames**. The **woman is seen running away from the burning vehicle**, and **the firefighters arrive to extinguish the flames**. The event is significant because it involves a dangerous and potentially deadly situation, which could have resulted in serious injury or loss of life. The woman's actions, such as running away, suggest that she may have been in shock or panic due to the sudden and unexpected nature of the fire. The subsequent arrival of firefighters and their efforts to extinguish the flames indicate that the situation was potentially life-threatening and required immediate attention. Overall, the video captures a moment of chaos and danger, highlighting the importance of fire safety and the need for quick response in emergency situations.

HolmesVAU-2B

The anomaly event begins with **the woman in white putting her hands on the fuel pump**, and then she suddenly **throws a fire extinguisher at the fuel pump**, causing **a fire to erupt** and **smoke to fill the air**. The subsequent events show the **woman running away from the fire**, and **the fuel pump being engulfed in flames**, with **the fire spreading rapidly and causing significant damage to the structure**.

Qwen2.5-VL-72B

The abnormal event in the video occurs between **64.0 and 72.0 seconds**. The unexpected aspect is when **a child suddenly appears from behind the car and runs across the gas station area**. This is unusual because children should not be playing near fuel pumps, which can be dangerous due to the risk of fire or other accidents. The sudden appearance of the child also disrupts the normal activity at the gas station, where adults are typically expected to be present and attentive.

InternVL3-78B

The event is "**a woman lights a cigarette and throws it on the ground**".

VideoLLaMA3-7B

0.0 - 12.3seconds, **A car is parked at a gas station**. **12.3 - 47.5seconds**, **A woman gets out of the car and begins to pump gas into it**. **47.5 - 81.6seconds**, **She leaves the pump hanging from the nozzle and walks away**.

Figure 56: Qualitative results on ALD task (Part II). The video is sourced from UCF-Crime (Sultani et al., 2018).