Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly

Hang Du[1]*, Sicheng Zhang[1]*, Binzhu Xie[1]*, Guoshun Nan[1]†, Jiayang Zhang[1], Junrui Xu[1], Hangyu Liu[1], Sicong Leng[2], Jiangming Liu[3], Hehe Fan[4], Dajiu Huang[5], Jing Feng[5], Linli Chen[5], Can Zhang[1], Xuhuan Li[1], Hao Zhang[1], Jianhang Chen[1], Qimei Cui[1], Xiaofeng Tao[1]

[1]Beijing University of Posts and Telecommunications [2]Nanyang Technological University

[3]Yunnan University [4]Zhejiang University [5]China Telecom Co., Ltd. Sichuan Branch

{7597892, zhangsicheng, xbz_nicous, nanguo2021, bbxst0371}@bupt.edu.cn,
{hangyuliugk50, Lengsicong, jmliunlp, crane.h.fan}@gmail.com, 19150109713@189.cn,
19113436101@189.cn, 15308007327@189.cn, {zhangcan_bupt, lixuhuan}@bupt.edu.cn,
zh242653915@gmail.com, {jh_chen, cuiqimei, taoxf}@bupt.edu.cn

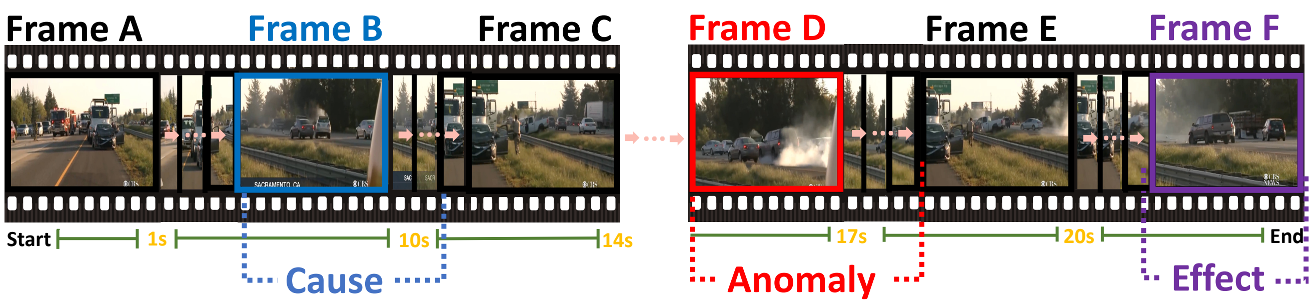

Figure 1. **Illustration of causations of video anomaly.** The clip started at Frame D refers to a traffic accident, which was caused by the event indicated with Frame B 7 seconds before. The clip in Frame F shows the effect of such an anomaly. A model needs to understand such a long-range relation in the video to yield correct text-based explanations.

Abstract

Video anomaly understanding (VAU) aims to automatically comprehend unusual occurrences in videos, thereby enabling various applications such as traffic surveillance and industrial manufacturing. While existing VAU benchmarks primarily concentrate on anomaly detection and localization, our focus is on more practicality, prompting us to raise the following crucial questions: "what anomaly occurred?", "why did it happen?", and "how severe is this abnormal event?". In pursuit of these answers, we present a comprehensive benchmark for Causation Understanding of Video Anomaly (CUVA). Specifically, each instance of the proposed benchmark involves three sets of human annotations to indicate the "what", "why" and "how" of an anomaly, including 1) anomaly type, start and end times, and event descriptions, 2) natural language explanations for the cause of an anomaly, and 3) free text reflecting the effect of the abnormality. In addition, we also introduce MMEval, a novel evaluation metric designed to better align with human preferences for CUVA, facilitating the measurement of existing LLMs in comprehending the underlying cause and corresponding effect of video anomalies. Finally, we propose a novel prompt-based method that can serve as a baseline approach for the challenging CUVA. We conduct extensive experiments to show the superiority of our evaluation metric and the prompt-based approach. Our code and dataset are available at https://github.com/fesvhtr/CUVA.

1. Introduction

Anomalies represent occurrences or scenarios that deviate from the norm, defying expectations and straying from routine conditions [2, 3, 8, 13]. These events are typically characterized by their unique, sudden, or infrequent nature, of-

*Equal Contribution
†Corresponding Author

ten demanding special attention or intervention [44].

Recently proliferated video anomaly understanding (VAU) [35, 58] aims at automatically comprehending such abnormal events in videos, thereby facilitating various applications such as traffic surveillance, environmental monitoring, and industrial manufacturing. In this direction, video anomaly detection and localization, which refer to identifying abnormal occurrences, and localizing temporally or spatially locate anomalous events in videos, has attracted enormous attention [11, 25, 29, 37, 54, 55, 57, 59, 64].

Existing VAU benchmarks [9, 20, 52] and approaches [10, 15, 16, 19, 38, 49, 63, 67, 71] primarily focus on the aforementioned anomaly detection and localization tasks, while the underlying cause and the corresponding effect of these occurrences, are still largely under-explored. These cues are crucial for perceiving the abnormality and making decisions based on human-interpretable explanations. Figure 1 demonstrates a scene of a traffic accident involving many vehicles. "The accident occurred because a white car parked by the roadside, and a dark gray car traveled at high speed to swerve and rear-end the black car next to it." Challenges of comprehending such a cause of the accident include: 1) *capturing key cues in the long video:* a model needs to recognize the white car at the moment indicated by Frame B, which is 7 seconds before the accident in the clip indicated by Frame D. It is challenging for a model to capture such a long-range relation. 2) *building a logic chain of the cause-effect:* a model needs to further learn rich interactions among clips in the video, indicated by Frame B, Frame C, and Frame D, to build a logic chain of causation of the anomaly, facilitating the generation of the explanations and results. The above two challenges require the development of causation understanding methods that specifically take these characteristics of video anomaly into consideration.

Previous works have demonstrated the great importance of leveraging large, high-quality, and challenging benchmarks to develop and evaluate the state-of-the-art deep learning methods for the VAU task [1, 18, 30, 39, 47, 53]. Along this line, existing benchmarks have shown their promise [8, 44, 59]. Towards VAU in more practical real-world scenarios, they have some limitations: 1) *Lack of cause and effect explanations.* Existing annotations involve the periods when anomalies occur, without providing an explanation of the underlying cause and the effect, as well as the descriptions of targeting anomaly. 2) *Lack of proper evaluation metrics.* Some remotely related metrics to measure the text-based explanation or description of the video anomaly, such as BLEU [42] and ROUGE [26], can not be directly applied to measure multimodal VAU tasks, as they are designed only for text modality. 3) *Limited length of videos.* In real-world scenarios, a piece of video may include more than 1.5 minutes [4]. However, samples in existing VAU usually have fewer than 30 seconds, which greatly simplifies the challenges of VAU in real-world cases.

The above limitations of existing datasets call for a benchmark of Causation Understanding of Video Anomaly. Towards that, we present CUVA, a comprehensive benchmark that contains high-quality annotations of $1,000$ videos from the real world, covering 10 major categories, and 42 subcategories of different anomaly types, each involving a 117-second long video and "65.7" tokens across "4.3" sentences on average. Specifically, we manually write free-text explanations to detail the underlying cause and the corresponding effects, the descriptions of these events, and the relationships among them. Moreover, we come up with a novel evaluation metric to measure the capability of a method on the challenging CUVA. We also propose a novel prompt-based approach based on video large language model (VLM) [24, 36, 65]. Experiments show the superiority of the metric and the proposed method. The main contributions of our work can be summarised as follows:

- We develop CUVA, a new benchmark for causation understanding of video anomaly. To the best of our knowledge, CUVA is the first large-scale benchmark focused on the causation of video anomalies. Compared with existing datasets, our dataset is more comprehensive and more challenging with much higher-quality annotations.
- We present a novel metric to measure the challenging CUVA in a human-interpretable manner, and introduce a prompt-based method to capture the key cues of anomalies and build a logic chain of the cause-effect.
- We conduct extensive experiments on the proposed CUVA. Results show that CUVA enables us to develop and evaluate various VLM methods for causation understanding of video anomalies closer to real-world cases.

2. Related Work

Anomaly Datasets: Existing VAU datasets primarily focus on anomaly detection and localization, and can be broadly categorized into weakly-supervised ones [51, 59], and semi-supervised ones [2, 34, 44, 46]. These datasets emphasize the time points or time periods of anomalous events based on frame-level or pixel-level annotations. Our CUVA significantly differs from the existing datasets in these aspects, More detailed comparisons are available in Table 1.

Evaluation Metrics: VAU evaluation metrics [62] include, reference-based ones such as ROUGE [26] and BLEURT [48], answer-based ones such as BLEU [42], Rankgen [22] and QAFactEval [14], and others such as Longformer [6], UniEval [70] and MoverScore [69]. Recently, various GPT-based metrics [5, 7, 61] have been developed. The key difference between our proposed MMEval and the above ones is: MMEval aims to evaluate the video and text anomaly understanding based on a large language model, while the existing one focuses on a single modality.

Methods: Video large language models (VLM) have been

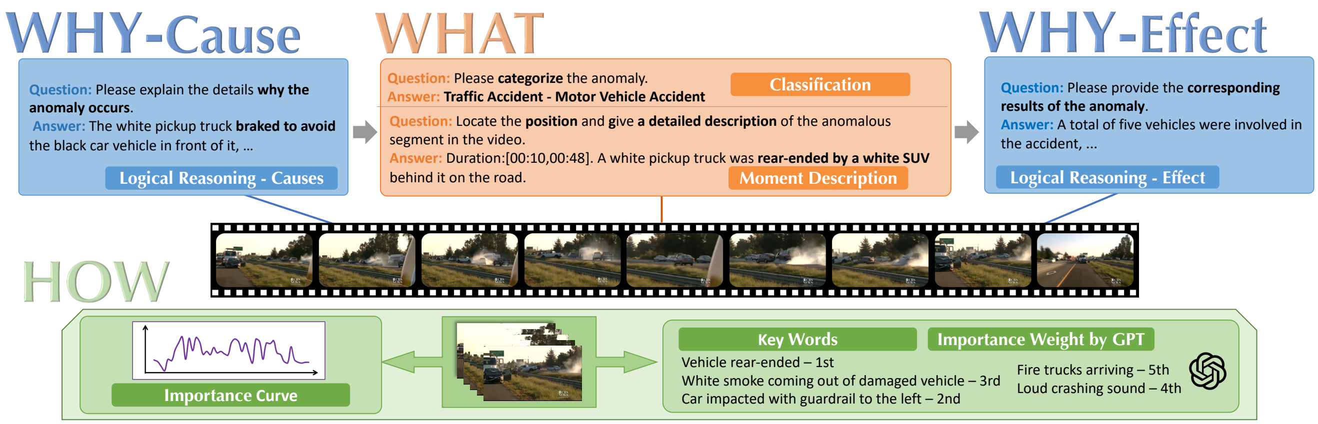

Figure 2. **Overview of the proposed CUVA benchmark.** Our CUVA benchmark consists of manual text-based annotation, including detailed explanations of cause (Why) and effect (Why), anomaly types (What), detailed event descriptions (What), as well as importance scores that can form a curve of events (How).

widely used for text generation based on videos [23, 36, 50, 68], exploring prompts to unlock the capability of VLMs. Prompt-based methods can be categorized into "hard prompt" and "soft prompt" [12, 31, 32, 45]. For the challenging CUVA task, we proposed a novel method that leverages both hard prompts and soft ones to tackle two challenges raised at the beginning, i.e., capturing the key cues and building a logic chain of anomaly causation.

3. The Proposed CUVA Benchmark

In this section, we first introduce some CUVA sub-tasks. Then we show how we collect and annotate data. We also provide a quantitative analysis of the benchmark. The overview of our CUVA is demonstrated in Figure 2.

3.1. Task Definition

What anomaly occurred: This task includes two objectives: anomaly classification and anomaly description. *Anomaly Classification* includes all the anomaly classes present in the video, which are taken from our database of predefined anomaly classes, as shown in Figure 4 (a). Each video has multiple anomaly classes at different levels, and this task will challenge the model's ability to detect anomaly classes at multiple levels of granularity. *Anomaly Moment Description* includes the timestamp in which the anomaly occurs and a detailed description of the anomalous event.
Why this anomaly happened: This task aims to describe the causal relationships within the video. Anomaly reasoning describes the reasons for the occurrence of anomalies in the video. This task requires the model to infer the cause of the anomaly based on the video content and describe it in natural language, which challenges the model's ability of video comprehension and reasoning. Anomaly results primarily describe the impacts caused by anomalous events in the video. It mainly tests the model's ability to handle details of anomalous events in the video.
How severe this anomaly: This task aims to reflect the changing trends in the severity of anomalies within the video. Thus, we propose a novel annotation approach called the importance curve. Details of our importance curve's pipeline can be found in Figure 3. This approach has three advantages: 1) It provides an intuitive representation of the temporal variation in anomaly severity within the video. 2) It offers a more intuitive depiction of the inherent causal relationships among anomalous events in the video. 3) Such an approach enables us to unify various Video Temporal Grounding labels and tasks (e.g. Moment Retrieval, Highlight Detection, Video Summarization) under the same framework.*

3.2. Dataset Collection

We crawled data from prominent video platforms such as Bilibili and YouTube†. And we discarded videos that encompass sensitive themes such as pornography and politics. Throughout the data collection process, we thoroughly analyze the quantity and quality of videos in each category, which in turn lead to the selection of the final 11 categories of anomalous videos. These videos are then categorized into 11 main categories, such as "robbery", "traffic accident" and "fire". Each major category is further divided into subcategories. For example, we divided the "fire" category into the "commercial building fire", "forest fire", "factory fire" and "residential fire" subcategories. In this way, we obtain 42 subcategories in total.

*More details are available in Section 2 of Appendix A.
†We have obtained permission from Bilibili www.bilibili.com and YouTube www.youtube.com to use their video data for non-commercial purposes.

Dataset	Domain	Video				# Anomaly Types	QA			
		# Total Frames	Total Length	A.C.L	Audio		Localization	Description	Reasoning	Outcome
UCF-Crimes [51]	Crime	13,741,393	128.0h	242.5s	No	13	Frame	NA	NA	NA
XD-Violence [59]	Volence	114,096	21.07h	164.3s	Yes	6	Frame	NA	NA	NA
ShanghaiTech [34]	Pedestrian	317,398	-	-	No	13	Bounding-box	NA	NA	NA
UCSD Ped1 [56]	Pedestrian	14,000	0.1h	6.6s	No	5	Bounding-box	NA	NA	NA
UCSD Ped2 [56]	Pedestrian	4,560	0.1h	6.6s	No	5	Bounding-box	NA	NA	NA
CUHK Avenue [33]	Pedestrian	30,652	0.5h	1.4s	No	5	Bounding-box	NA	NA	NA
TAD [64]	Traffic	721,280	1.2h	36.8s	Irrelevant	4	Bounding-box	NA	NA	NA
Street Scene [44]	Traffic	203,257	380.6s	3.7s	No	17	Bounding-box	NA	NA	NA
CamNuvem [11]	Robbery	6,151,788	57h	192.2s	No	1	Frame	NA	NA	NA
Subway Entrance [3]	Pedestrian	86,535	1.5h	-	No	5	Frame	NA	NA	NA
Subway Exit [3]	Pedestrian	38,940	1.5h	-	No	3	Frame	NA	NA	NA
UCF–Crime Extension [41]	Crime	734,400	7.5h	112.5s	No	1	Frame	NA	NA	NA
BOSS [54]	Multiple	48,624	0.5h	660.0 s	No	11	Frame	NA	NA	NA
UMN [37]	behaviors	3,855	0.1h	29.1s	No	1	Frame	NA	NA	NA
UBnormal [2]	Multiple	236,902	2.2h	14.6s	No	22	Pixel-level	NA	NA	NA
CUVA (Ours)	Multiple	3,345,097	32.5h	117.0s	Yes	**42**	**Time Duration**	**Free-text**	**Free-text**	**Free-text**

Table 1. **Comparisons between the proposed CUVA and existing VAU datasets.** Our CUVA is the first large-scale benchmark for causation understanding of video anomaly. It encompasses samples from 42 domains, such as vandalism, traffic accidents, fire incidents, and pedestrian incidents, etc. CUVA sub-tasks primarily focus on the evaluation of causation understanding of video anomaly, and these tasks answer the "What", "Why" and "How" of an anomaly. All textual descriptions or explanations are annotated in **free-text** format. Here **A.C.L.** typically stands for "Average Clip Length."

3.3. Annotation Pipeline

Our dataset construction pipeline involves three stages: pre-processing, manual annotation, and importance curve processing. The whole process takes about 150 hours with over 20 annotators.[‡]

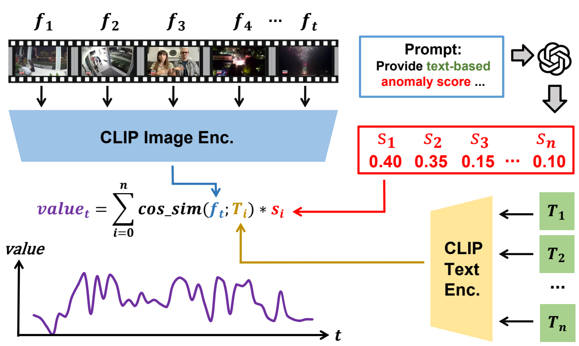

Figure 3. **Pipeline of generating an importance curve.** Annotators need to consider previous tasks (e.g., Logical Description, Moment Description) and video content to create 3 to 6 short sentences T_i describing all events in the video. We rank these sentences' anomaly severity by ChatGPT [40] and obtain anomaly scores s. Simultaneously, we sample frames f_t from the video and use CLIP [43] to measure the similarity between sentences and frames. The resulting similarity scores are multiplied by the anomaly scores for each sentence to get $value_t$ for each frame.

3.3.1 Pre-processing

First, we crawl videos from Bilibili and YouTube. Then, we manually cut the collected videos to ensure the quality of video content and exclude non-ethical content and sensitive information through manual screening.[§] Throughout the dataset collection and annotation process, we strictly follow the ethical requirement of the website.[¶] Finally, $1,000$ anomaly video clips are obtained.

3.3.2 Manual Annotation

We annotate the videos in English according to the designed annotation document, and the annotation is divided into two rounds. We employ a mechanism similar to kappa [60] to screen and train annotators, ensuring the consistency of their annotation content. In the first round, We ask annotators to annotate all videos according to the task definition. In the second round, we ask these annotators to review and supplement the annotation results of the first round.

3.3.3 Post-processing of Importance Curve

Due to the limited capabilities of the CLIP model and sampling intervals, the initial curve may fail to accurately reflect the time periods of anomalies, which significantly impacts the effectiveness of downstream tasks. Thus, we incorporate the following three tasks to optimize the importance curve, such as Video Captioning [24], Video Entailment [66], and Video Grounding [27] respectively. Based on these tasks, we employ a voting mechanism to precisely identify the time segments in the video corresponding to the given key sentences.[‖]

[‡]More details of our dataset are available in Section 3 of Appendix A.

[§]Detailed screening criteria can be found in Section 4 of Appendix A.

[¶]More details about ethical consideration are presented in Section 5 of Appendix A.

[‖]Details can be found in Section 6 of Appendix A.

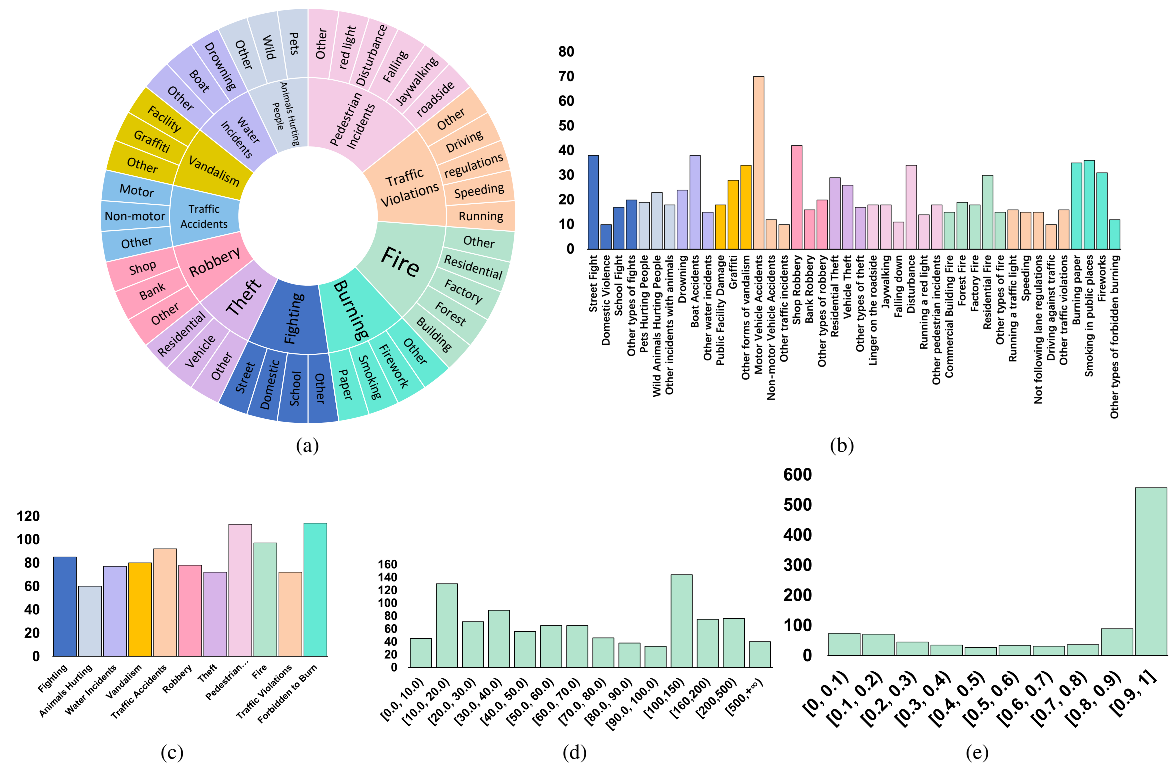

Figure 4. **Statistics of our CUVA dataset.** Figure (a) shows all anomaly types in CUVA. Figure (b) and (c) show the number of videos in each anomaly type. Figure (d) shows the distribution of video length. Figure (e) shows the temporal distribution of anomalous segments.

3.4. Dataset Statistics

Our CUVA dataset contains $1,000$ video clips and $6,000$ question-answer pairs, the total length of these videos is 32.46 hours, and the average frames of videos is $3,345$. The frames are extracted from the original videos at a rate of 60 FPS. The videos encompass a wide range of domains. Then, we categorize anomaly events into 11 scenarios, resulting in a total of 42 types of anomalies, as illustrated in Figure 4 (a). The distribution of video categories is illustrated in Figure 4 (b) and 4 (c). The distribution of video lengths can be found in Figure 4 (d), along with the percentage of video time proportions shown in Figure 4 (e).

4. The Proposed Method: Anomaly Guardian

In this section, we introduce a novel prompt-based method named Anomaly Guardian (A-Guardian), which is designed to address the two challenges presented by our dataset. By leveraging the exceptional logical reasoning capabilities of VLM, we select it as the foundation of our method to *build a logic chain of the cause-effect*. *To effectively capture crucial cues within lengthy videos*, we present a novel prompt mechanism aimed at guiding VLMs to concentrate on pivotal clues in the video pertinent to the provided questions.

4.1. Design of Hard Prompts

We use ChatGPT [40] to assist in confirming and supplementing user prompts first, enabling the VLM to better understand the user's intent. Specifically, we first utilize an instruction prompt containing an example to correct misleading guidance and standardize the output format. Due to the presence of numerous events in long videos, we employ a multi-turn dialog mechanism to assist VLM in identifying events relevant to anomaly occurrences in the video. After multiple rounds, VLM can focus on segments more relevant to the anomaly, providing more accurate answers.**

4.2. Design of Soft Prompts

We leverage a selector in MIST [17] to better capture spatio-temporal features relevant to the given questions processed by ChatGPT [40]. We first divide the video into N segments of uniform length, with each segment comprising T frames. To better capture interactions among different granularities of visual concepts, we divide each frame into M patches. Furthermore, we leverage $[CLS]$ token to represent each segment and frame. Specifically, We first use the CLIP [43] with frozen parameters to extract patch-level features denoted as $\mathbf{P} = \{p^1, p^2, ..., p^m\}$, where $p^m \in \mathbb{R}^{T\times M\times D}$

**Details of the hard prompts are available in Section 1 of Appendix B.

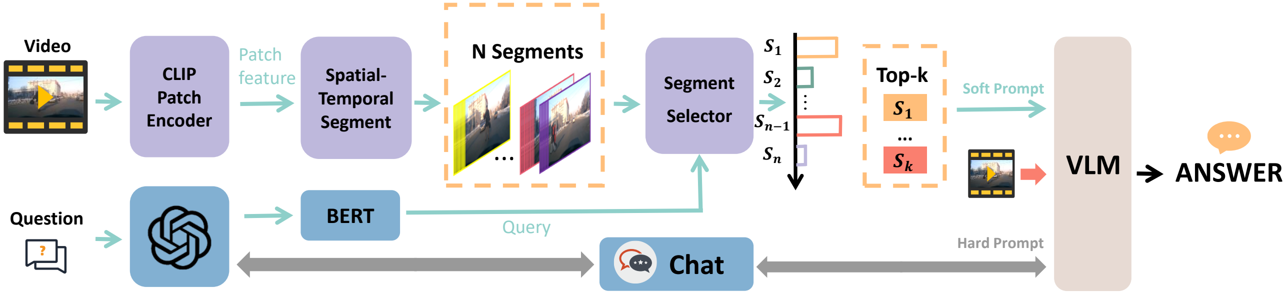

Figure 5. **Architecture of the proposed prompt-based method A-Guardian.**

and D is the dimension of each patch-level feature. Then, we perform pooling operations on patch features' spatial dimensions to obtain frame features.

$$f_{kt} = Pooling(p_{kt,1}, p_{kt,2}, \ldots, p_{kt,M}) \tag{1}$$

where $p_{kt,m}$ indicates the m-th patch at the t-th frame of the k-th segment. Then, the segment features are obtained by pooling frame features along the temporal dimension, where $f_{kt} \in \mathbb{R}^{T \times D}$:

$$s_k = Pooling(f_{k1}, f_{k2}, \ldots, f_{kT}) \tag{2}$$

Similarly, the question feature is obtained by pooling the word features, where $w_z \in \mathbb{R}^{Z \times D}$ and $q \in \mathbb{R}^D$

$$\boldsymbol{q} = Pooling(w_1, ..., w_z) \tag{3}$$

After that, we select the patch features of the top k segments using cross-modal temporal attention and top-k selection from MIST [17], as expressed by the following formulation. The term "selector" corresponds to a top-k selection function utilized to pick the video segment features from the Top_k segments considering the question.

$$\mathbf{X}_t = \underset{Top_k}{\textbf{selector}}\left(\text{softmax}\left(\frac{\boldsymbol{q} \cdot \mathbf{s}^T}{\sqrt{d_k}}\right), \mathbf{S}\right) \tag{4}$$

4.3. Answer Prediction

Finally, we follow a previous work [21] to concatenate the hard prompts and soft prompts and feed them into the VLM for inference. During the training phase, we employ GPT to generate candidate answers and data augmentation. We only finetune the selector by optimizing the softmax cross-entropy loss, aligning the predicted similarity scores with the ground truth.[††]

5. Experiment

5.1. The Proposed MMEval Metric

Given that our dataset extensively employs free-text descriptions to delineate both anomalous events with their

[††]Details can be found in Section 2 of Appendix B.

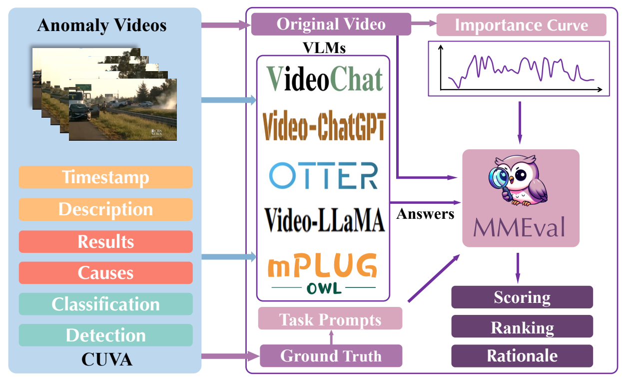

Figure 6. **Overview of our MMEval metric.**

causal relationships, and recognizing CUVA is a multimodal dataset (integrating video, text, and appended comments), which necessitates a shift from solely relying on Natural Language Generation (NLG) metrics to a broader consideration that encompasses the rich multimodal input information. Thus, we introduce a novel evaluation metric namely mmEval as depicted in Figure 6. In order to assess the model's performance from a multimodal perspective and infuse human-like reasoning abilities into the evaluation metric, we choose Video-ChatGPT [36] as our foundation model. We utilize natural language prompts to guide mmEval in specifying the task types to be evaluated and design three natural language prompts, each corresponding to one of the three free-text descriptions in the dataset. To enhance the robustness of the model, we utilize curve labels to help VLM focus more on segments of anomalies within the video. Specifically, by setting thresholds to extract periods of important events in the curve, we perform dense sampling on that segment of the video, helping the VLM focus more on the crucial parts of the video. Our MMEval metric can be used for scoring, ranking, and providing rationale explanations.

5.2. Implementation Details

We follow Video-ChatGPT [36] to adopt CLIP-L/14 visual encoder to extract both spatial and temporal video fea-

Method	Metric	Description	Causes	Effect
mPLUG-owl [65]	BLEU	0.55	0.65	0.47
	ROUGE	12.58	13.54	8.83
	BLEURT	40.66	43.28	37.95
	MoverScore	51.97	**52.71**	50.06
	UniEval	67.46	62.29	**59.07**
	MMEval (Ours)	73.42	17.15	44.31
Video-LLaMA [68]	BLEU	0.60	0.53	0.35
	ROUGE	13.15	12.36	8.02
	BLEURT	40.55	43.02	39.68
	MoverScore	51.32	51.25	49.48
	UniEval	52.28	47.29	43.03
	MMEval (Ours)	65.65	16.24	32.84
PandaGPT [50]	BLEU	0.66	0.51	0.30
	ROUGE	13.33	14.09	8.79
	BLEURT	38.23	43.95	39.95
	MoverScore	51.73	51.54	49.62
	UniEval	57.05	54.88	50.84
	MMEval (Ours)	74.19	22.47	**69.45**
Otter [23]	BLEU	**1.07**	**1.09**	**1.11**
	ROUGE	**15.19**	**15.87**	**11.40**
	BLEURT	29.92	32.52	28.94
	MoverScore	**53.54**	54.25	**51.91**
	UniEval	45.14	49.05	47.51
	MMEval (Ours)	76.30	3.53	39.21
Video-ChatGPT [36]	BLEU	0.30	0.29	0.41
	ROUGE	9.75	9.08	8.23
	BLEURT	46.83	**49.52**	37.24
	MoverScore	50.73	50.70	49.83
	UniEval	**70.82**	**70.77**	54.35
	MMEval (Ours)	78.55	44.57	46.08
Video-ChatGPT [36] + **A-Guardian (Ours)**	BLEU	0.55	0.51	0.38
	ROUGE	14.35	9.08	8.23
	BLEURT	**47.10**	48.13	**48.28**
	MoverScore	52.25	52.28	49.95
	UniEval	68.18	63.41	51.87
	MMEval (Ours)	**79.65**	**58.92**	50.64

Table 2. **Main results on the proposed CUVA benchmark.** We test the **Description**, **Cause**, and **Effect** tasks on our CUVA benchmark using multiple VLMs and Video-ChatGPT equipped with the proposed **A-Guardian**. We conduct evaluations using both traditional metrics and our **MMEval** metric. The scores of all metrics range from 0 to 100.

Methods	Detection	Classification	Timestamp
mPLUG-Owl [65]	89.4%	11.5%	9.0%
Video-LLaMA [68]	25.0%	13.1%	0.7%
PandaGPT [50]	100.0%	32.6%	N/A
Otter [23]	64.3%	41.3%	N/A
Video-ChatGPT [36]	60.0%	21.3%	3.2%

Table 3. **Secondary results on the proposed CUVA benchmark.** We use the accuracy metrics to evaluate the **Detection** and **Classification** tasks. We also use IOU to evaluate the **Moment** task, **N/A** to indicate the model lacks the ability to answer the question.

tures. In our approach, we utilize the Vicuna-v1.1 model, comprised of 7B parameters, and initialize it with weights from LLaVA [28]. All experiments were conducted on four NVIDIA A40 GPUs, and each task took around 8 hours.

5.3. Consistency evaluation of MMEval

Our MMEval metric can better align with human's preference on causation understanding of video anomaly. To validate the consistency of our evaluation metric with human judgment, we conducted a human consistency experiment. Using the ranking of answers from first-round annotations, second-round annotations, and GPT-generated answers as the ground truth (1. *Second round* 2. *First round* 3. *ChatGPT*). we employ various evaluation metrics and human beings who view the videos to rank these answers based on the corresponding questions, as shown in Table 4.

Metrics	Answer Pool Ranking		
	Description	Cause	Effect
Human Evaluation	87.3%	77.3%	87.3%
BLEU [42]	67.8%	60.4%	63.2%
ROUGE [26]	54.4%	55.5%	52.1%
BLEURT [48]	80.4%	73.2%	76.7%
MoverScore [69]	67.8%	60.4%	63.2%
UniEval [70]	78.2%	70.1%	74.3%
MMEval (Ours)	**82.3%**	**80.2%**	**89.1%**

Table 4. **Human consistency evaluation**

5.4. Quantitative evaluation of A-Guardian

Our A-Guardian model achieves state-of-the-art performance in both the description and cause tasks. We conducted experiments on all tasks involved in our dataset, and the results are summarized in Table 2. For free-text tasks (e.g. Cause, Effect, Description), we evaluated the performance of various VLMs and our model under different evaluation metrics. Our model also outperforms the majority of models in the effect task. For the other tasks (e.g. Detection, Classification, Timestamp), we set a uniform prompt and use string matching to extract answers relevant to the questions from the inference results of VLMs. Table 3 shows the results of these tasks.

Model	MMEval (%)		
	Description	Cause	Effect
Ours	79.65	58.92	50.64
- Soft Prompt	78.92	54.22	49.11
- Hard Prompt	78.55	44.57	46.08

Table 5. **Ablation Experiment**

5.5. Ablation Study

Both hard and soft prompts significantly improve the VLM's understanding of the video's causation. This section investigates the influence of soft prompts and hard prompts on our method. As shown in Table 5, the design of hard prompts achieves a greater improvement than that of soft prompts, indicating that the hard prompts are more intuitively effective in uncovering VLM's reasoning capabilities compared to the soft prompt.

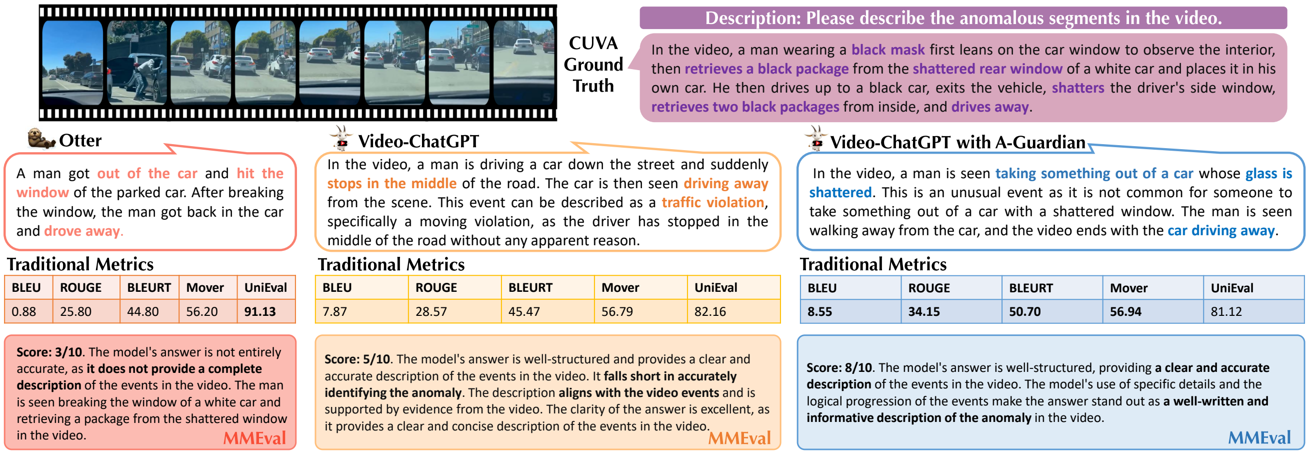

Figure 7. **Case study**. Comparisons with and without our proposed prompt-based method on the CUVA using MMEval.

5.6. Case Study

In Figure 7, we illustrate the performance of Otter, Video-ChatGPT, and Video-ChatGPT with A-Guardian, showcasing the different answers they provide for the anomaly causation task. In terms of the model's response, it can be observed that Video-ChatGPT provides descriptions that are generally correct, but it does not focus on describing the anomaly event. Instead, it pays attention to describing the actions of the vehicles. However, with the addition of our A-Guardian model, its descriptions become more accurate, specifically highlighting the theft as an anomaly event and providing detailed descriptions such as *"taking something out of a car"* and *"glass is shattered"*. Otter and Video-ChatGPT achieve similar scores based on traditional metrics, but their answers convey completely different meanings. Otter's description does not align with the video, while Video-ChatGPT incorrectly describes the anomaly subject. As MMEval possesses the ability to evaluate from the multimodal perspective, it is able to identify the parts that pertain to the description of the anomaly in the videos, which shows highly consistent conclusions with human beings.

5.7. Result Discussion

Through experiments, we have discovered and summarized the following conclusions: 1) For free-text tasks, most VLMs excel in the description of anomalies but perform poorly on the task of causation analysis. This is because the tasks of description only require the VLM to comprehend the content of the videos, but causation analysis requires the VLM to possess a certain level of reasoning capability to build a logic chain of the cause-effect. 2) Timestamp localization task is the most challenging. Due to the relatively simplistic temporal and spatial relationships between video frames, VLM performs poorly on fine-grained tasks such as timestamp localization but excels relatively in coarse-grained tasks such as anomaly detection and classification. 3) Traditional metrics are poor at evaluating reasoning tasks. As shown in Figure 7, they generate similar evaluations for these answers, making it difficult to distinguish between them. However, MMEval is able to distinguish these answers' inner differences and generate more accurate evaluation results.

6. Conclusion

This paper presents CUVA, a novel benchmark for causation understanding of video anomaly. To the best of our knowledge, our CUVA is the first benchmark in the field. Compared with the existing datasets, CUVA is more comprehensive and more challenging with much higher-quality annotations. We believe the proposed CUVA will encourage the exploration and development of various downstream tasks such as anomaly detection, anomaly prediction, anomaly reasoning, etc. We also present MMEval, a novel evaluation to measure the challenging CUVA in a human-interpretable manner. Furthermore, we put forward a prompt-based approach that can serve as a baseline approach for CUVA. Such an approach can capture the key cues of anomalies and build a logic chain of the cause-effect. Experimental results show that CUVA enables us to develop and evaluate various VLM methods. In the future, we plan to apply our CUVA to more practical scenarios for anomaly understanding and other VLM-based tasks.

Acknowledgement

This work was partially supported by the joint funds for Regional Innovation and Development of the National Natural Science Foundation of China (No.U21A20449), the National Natural Science Foundation of China (No.62171045), the Fundamental Research Funds for the Central Universities (No.2242022k60006) and the Joint Research Fund for Beijing Natural Science Foundation and Haidian Original Innovation (No.L232001).

References

[1] Armstrong Aboah. A vision-based system for traffic anomaly detection using deep learning and decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4212, 2021. 2

[2] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20143–20153, 2022. 1, 2, 4

[3] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 1, 4

[4] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863, 2021. 2

[5] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023. 2

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 2

[7] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 2

[8] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20392–20401, 2023. 1, 2

[9] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20392–20401, 2023. 2

[10] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition*, 122:108213, 2022. 2

[11] Davi D de Paula, Denis HP Salvadeo, and Darlan MN de Araujo. Camnuvem: A robbery dataset for video anomaly detection. *Sensors*, 22(24):10016, 2022. 2, 4

[12] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*, 2021. 3

[13] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4037–4042, 2020. 1

[14] Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*, 2021. 2

[15] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2448–2460, 2023. 2

[16] Fan Fei, Yean Cheng, Yongjie Zhu, Qian Zheng, Si Li, Gang Pan, and Boxin Shi. Split: Single portrait lighting estimation via a tetrad of face intrinsics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(02):1079–1092, 2024. 2

[17] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 5, 6

[18] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021. 2

[19] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[20] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35: 32142–32159, 2022. 2

[21] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland, 2022. Association for Computational Linguistics. 6

[22] Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. Rankgen: Improving text generation with large ranking models. *arXiv preprint arXiv:2205.09726*, 2022. 2

[23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3, 7

[24] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 4

[25] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware

transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10055–10069, 2023. 2
[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2, 7
[27] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 4
[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7
[29] Rui Liu, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation, 2023. 2
[30] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24500–24510, 2023. 2
[31] Jochem Loedeman, Maarten C. Stol, Tengda Han, and Yuki M. Asano. Prompt Generation Networks for Input-based Adaptation of Frozen Vision Transformers. *arXiv e-prints*, art. arXiv:2210.06466, 2022. 3
[32] IV Logan, Robert L., Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. *arXiv e-prints*, art. arXiv:2106.13353, 2021. 3
[33] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 4
[34] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017. 2, 4
[35] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2023. 2
[36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2, 3, 6, 7
[37] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition*, pages 935–942. IEEE, 2009. 2, 4
[38] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2764–2774, 2021. 2
[39] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019. 2
[40] OpenAI. Chatgpt. https://www.openai.com/gpt-3, 2022. Accessed: November 12, 2023. 4, 5
[41] Halil Ibrahim Ozturk and Ahmet Burak Can. Adnet: Temporal anomaly detection in surveillance videos. *arXiv preprint arXiv:2104.06653*, 2021. 4
[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2, 7
[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 5
[44] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020. 2, 4
[45] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. *arXiv e-prints*, art. arXiv:2112.01518, 2021. 3
[46] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020. 2
[47] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019. 2
[48] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020. 2, 7
[49] Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18717–18726, 2023. 2
[50] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 3, 7
[51] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 4
[52] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140:109567, 2023. 2

[53] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 2

[54] Sergio A Velastin and Diego A Gomez-Lira. People detection and pose classification inside a moving train using computer vision. In *Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017, Proceedings 5*, pages 319–330. Springer, 2017. 2, 4

[55] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10873–10883, 2023. 2

[56] Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1220–1223. IEEE, 2010. 4

[57] Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. Lana: A language-capable navigator for instruction following and generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19048–19058, 2023. 2

[58] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022. 2

[59] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 2, 4

[60] Yinglin Xia. Chapter eleven - correlation and association analyses in microbiome study integrating multiomics in health and disease. In *The Microbiome in Health and Disease*, pages 309–491. Academic Press, 2020. 4

[61] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension, 2023. 2

[62] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering, 2023. 2

[63] Li Xu, Mark Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta compositional referring expression segmentation. pages 19478–19487, 2023. 2

[64] Yajun Xu, Chuwen Huang, Yibing Nan, and Shiguo Lian. Tad: A large-scale benchmark for traffic accidents detection from video surveillance, 2022. 2, 4

[65] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 7

[66] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 4

[67] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754, 2022. 2

[68] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 7

[69] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, 2019. Association for Computational Linguistics. 2, 7

[70] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 2, 7

[71] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3082–3092, 2023. 2