



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

## Supplementary Material

### A. Dataset

#### A.1. Application of the proposed importance curve

We have presented a novel annotation approach called the importance curve in section 3.1 of the main paper. Such an approach enables us to unify various Video Temporal Grounding labels and tasks under the same framework. Here, we describe how to apply the importance curve to unify various Video Temporal Grounding labels and tasks (e.g. Moment Retrieval, Highlight Detection, Video Summarization) respectively with Figure A1.

**Moment Retrieval** aims to retrieve time intervals from videos based on textual queries[2]. The importance curve reflects the changing trends in the severity of anomaly. Thus, we first filter out the time intervals (e.g.  $D_1$ ,  $D_2$ , and  $D_3$  in Figure A1) through a threshold  $\mu$ . Second, after post-processing the importance curve in Appendix A.4, we can obtain key sentences (e.g.  $KS_1$ ,  $KS_2$ , and  $KS_3$ ) corresponding to each time interval. Finally, these key sentences are employed as text queries, with corresponding time intervals serving as labels for the moment retrieval task.

**Highlight Detection** aims to assign a worthiness score to each video segment and then identify the top highest-scoring segment as the highlight [2]. Here, we first locate the absolute maximum point of the curve (e.g.  $VS$  in Figure A1), and leverage its corresponding time interval (e.g.  $D_2$ ) as the top highest-scoring segment to conduct highlight detection task.

**Video Summarization** aims to summarize the whole video by a set of shots to provide a quick overview[2]. As depicted in Figure A1,  $t_1$ ,  $t_2$ , and  $t_3$  are the timestamps of the local maximum point (e.g.  $S_2$ ,  $S_3$  and  $S_4$ ). We leverage these timestamps as a set of shots to provide a quick overview of the whole video.

#### A.2. Our CUVA benchmark

We have shown the statistics of the proposed CUVA in the subsection 3.5 of the main paper. Here, we provide some supplementary for the statistics of the proposed CUVA. As Figure A2 shows, we present more detailed statistics of video anomaly categories. Figure A3 shows the word cloud of the proposed CUVA.

#### A.3. Ethical consideration

We carefully study the online copyright rules of the websites from YouTube and Bilibili, from which we collect

data, and strictly conform to their requirements during data collection and annotation. We have also adhered to the guidelines provided by the CVPR Code of Ethics to anonymize the parts of the video that may involve personal privacy by applying pixelation. To exclude data that could potentially have ethical implications. We further conduct rigorous reviews at each stage of the annotation, adhering to screening criteria as follows:

1. **Personal Privacy Respect:** Avoid videos that show identifiable personal information (e.g., faces, license plates, home addresses).
2. **Legally Sourced Content:** Prohibit the use of unauthorized or illegally obtained footage.
3. **Viewer Sensitivity Consideration:** Avoid content that is overly graphic, cruel, or likely to cause viewer discomfort.
4. **Child Safety Focus:** Exclude any videos involving children in risky or harmful situations.
5. **Gender Respect and Equality:**
  - Videos should not imply or display acts of sexual violence.
  - Avoid content that contains gender discrimination or negative stereotypes of any gender group.
6. **Avoidance of Illegal Activities:** Do not display or promote videos that clearly show or endorse illegal activities (e.g., drug use, trafficking).
7. **Copyright infringement:** Videos containing copyrighted music, film clips, TV shows, or other media content.

#### A.4. Post-processing of the importance curve

As the initial sampling frequency of our curves is relatively lower (one frame per second), and we aim to obtain more accurate interval timestamps.<sup>1</sup> Thus, we incorporate three specific tasks into our optimization strategy to achieve an optimal importance curve. Specifically, we perform tasks such as Video Captioning, Video Entailment, and Video Grounding by VideoChat[1], SEVILA[5], and UniVTG[2], respectively. Each task identifies specific time segments in the video based on the key sentences as shown in Figure 3 of the main paper. We adopt a voting mechanism to select a time segment when at least two models agree that it encapsulates the event represented by the key sentence annota-

<sup>1</sup>We leverage algorithm 1 in the appendix to better illustrate the creation of the importance curve.

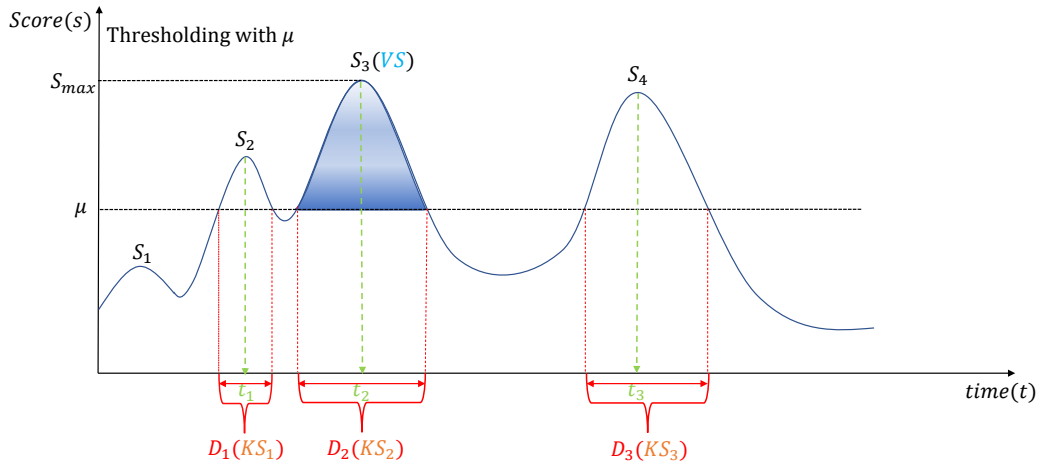


Figure A1. The application of the proposed importance curve,  $KS_1, KS_2$  and  $KS_3$  refer to three different key sentences.  $VS$  point is the absolute maximum point in the curve.  $t_1, t_2$ , and  $t_3$  are the timestamps of the local maximum point in the curve.

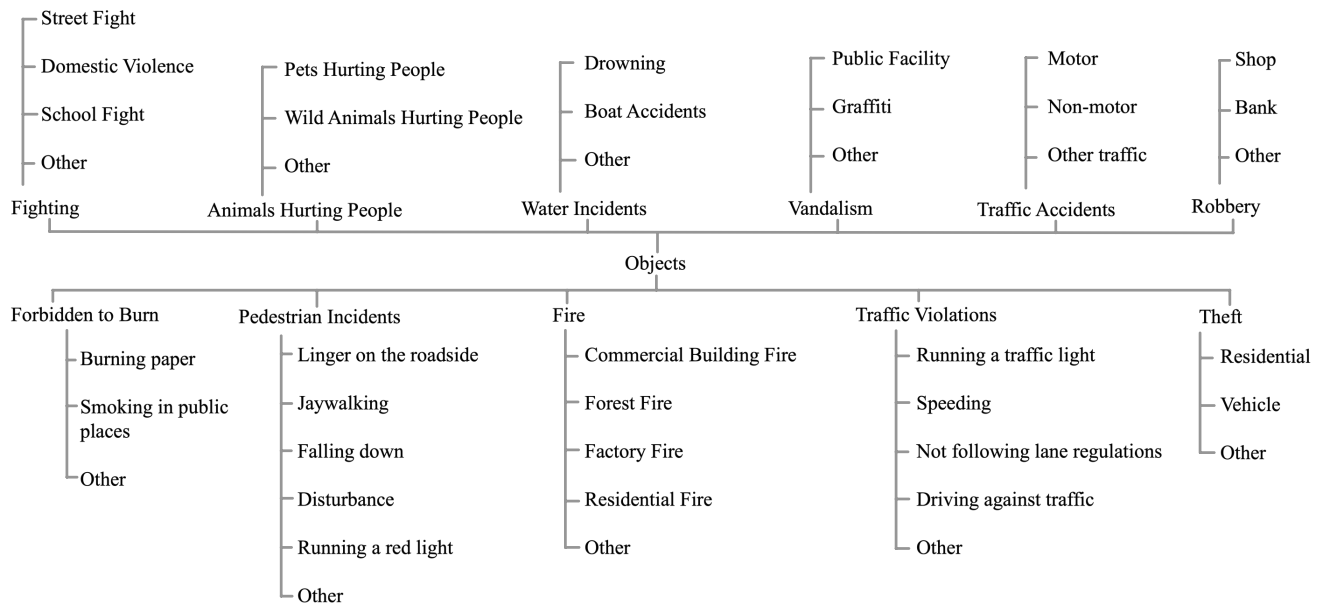


Figure A2. Anomaly types of the proposed CUVA, CUVA encompasses 11 different scenes and 42 extensive types of anomalies.



Figure A3. Wordcloud of our proposed CUVA.

tion. Through the voting mechanism, we precisely identify the time segments in the video that correspond to the current key sentences. Then, we perform dense sampling on the voted segments (ten frames per second) and use wavelet filters to smooth the curve. The comprehensive optimization process is illustrated in Figure A4.

## B. The proposed method

We have introduced A-Guardian, a novel prompt-based method that consists of two kinds of prompt design. Here,

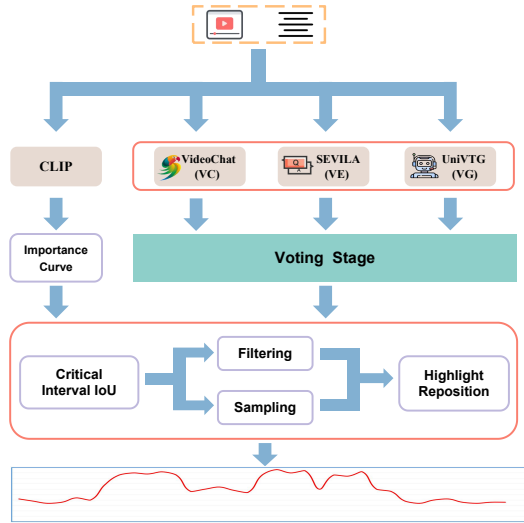


Figure A4. Overview of the curve optimization pipeline

---

### Algorithm 1 Generating importance curve

---

**Input parameters:** video, prompt

**Output:** importance curve

```

video_clip, timestamp = sparse_sampling(video)
prompt → txt_gpt_weight
txt_emb, vid_emb = Clip(video_clip, prompt)
txt_norm_emb, vid_norm_emb = normalization(txt_emb, vid_emb)
similarity = cosine_similarity(txt_norm_emb, vid_norm_emb)
similiar_score = normalization(similarity)
imp_value = similiar_score * txt_gpt_weight
timestamp, imp_value → importance curve

```

---

we detail the design of the hard prompts. Moreover, we illustrate the details of the answer prediction in this section.

### B.1. Hard prompts in A-Guardian

First, A-guardian leverages different newly designed system messages for different tasks, which can be found in Figure A5. After the VLM gives the answer based on the user message, the user’s question and the VLM’s answer will be input into ChatGPT to generate a new round of questions, Then, we re-input the question of ChatGPT[3] into VLM, and the final answer will be obtained after several rounds of loops. Figure A7 and A6 illustrate an A-Guardian hard prompt example which includes three rounds of dialog with ChatGPT.

### B.2. Answer prediction

We denote the candidate answers as  $A$ , we leverage BERT to generate the contextualized representations of each candidate answers  $X_a$ . We follow the previous work [4] to calculate the similarity between  $X_k$  and the feature of all candidate answers  $X_A = \{x_a | a \in A\}$  obtained by using

the pre-trained model. Finally, the candidate answer with the maximal similarity is considered as the final prediction  $\tilde{y}$ .

$$\tilde{y} = \arg \max_{y \in A} (X_k (X_A)^T) \quad (1)$$

During training, we optimize the softmax cross-entropy loss between the predicted similarity scores and ground truth.

## C. Experiment

### C.1. Prompts of MMEval

In MMEval, we have different scoring criteria for different tasks, which are translated into the system message of the VLM. Details can be found in Figure A8.

### C.2. Case study of A-Guardian

We have already shown the case study in section 5.6 of the main paper. Here, we further validate the effectiveness of the proposed model A-Guardian by the case study as follows: Figure A9 shows the results from Video-ChatGPT with and without A-Guardian for the Description, Cause, and Effect tasks related to the same anomalous video. It can be observed that A-Guardian possesses the following abilities:

- **Capturing key cues in the long video:** In the Description task, A-Guardian assists the VLM in identifying crucial moments and events in the anomalous video. For example, *punching another man in the face and holding a gun and pointing it at another man*. Whereas Video-ChatGPT without A-Guardian can only provide vague responses about the anomaly event.
- **Building a logic chain of cause and effect:** In the Cause and Effect tasks, A-Guardian guides the model in logical reasoning. In the example, the model with A-Guardian infers that the gun is the fundamental cause of the anomaly, leading to multiple injuries and people falling, followed by the perpetrator leaving. In contrast, the VLM without A-Guardian generates answers that are ambiguous and irrelevant to the events. The result emphasizes the potential results and societal impacts of the anomaly, lacking inference and summarization based on the video content.

## References

- [1] Kunchang Li, Yanan 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. 1
- [2] 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. 1

### # VLM System Message (What)

You are CUVA, an anomaly detection assistant. You possess advanced image and video analysis capabilities, enabling you to **identify and describe unconventional events** in videos. When asked about the content of a video, your task is to carefully observe the events within the video, particularly those that deviate from the norm.

### # VLM System Message (Why - Cause)

You are the Anomaly Detection Assistant CUVA. You have advanced image and video analysis capabilities, allow you to **understand and reason about anomalous events** in videos. When asked about the content of a video, your task is to scrutinize the **cause** of anomalous events in the video, especially those that deviate from the norm.

### # VLM System Message (Why - Effect)

You are the Anomaly Detection Assistant CUVA. you have advanced image and video analysis capabilities, allow you to **understand and reason about anomalous events** in videos. When asked about the content of a video, your task is to summarize the **effect** of anomalous events in a video, especially the severe effect caused by anomalies.

Figure A5. System message for VLMs. To guide the model for a description or answer, A-guardian uses different designed system messages for different tasks.

- [3] OpenAI. Chatgpt. <https://www.openai.com/gpt-3>, 2022. Accessed: November 12, 2023. 3
- [4] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 3
- [5] 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. 1

### # VLM System Message (What)

You are CVUA, an anomaly detection assistant. You possess advanced image and video analysis capabilities, enabling you to identify and describe unconventional events in videos. When asked about the content of a video, your task is to carefully observe the events within the video, particularly those that deviate from the norm.

### # User Message for VLM (Round 1)

Please **describe the anomalous segment** in the video.

Unclear capture of anomalous events, lack of detail

### # VLM Response (Round 1)

In the video, **a man is being held down by two other men, which can be considered a robbery or theft**. The incident takes place at the 1:00 mark, which is the first minute of the video. This is an unusual event as it is not common to see people **being held down in such a manner**. The **specific details of the incident are not provided**, but it is clear that the situation is not a typical occurrence in the video.

### # GPT-4 System Message

You are a prompt optimization assistant. We now show a visual model an anomaly video and ask questions, but we are not satisfied with the answers it gives. I can't give you the video information, but I'll tell you what we want the answer to look like. You need to generate a new question for this visual model.

The name of this visual model is CUVA. we have categorized the anomalies into the following categories, which may help you come up with a better prompt.

Categories for Anomaly Detection:

- |                      |                           |                       |                         |
|----------------------|---------------------------|-----------------------|-------------------------|
| 1) Fighting          | 2) Animals Hurting People | 3) Water Incidents    | 4) Vandalism            |
| 5) Traffic Accidents | 6) Robbery                | 7) Theft              | 8) Pedestrian Incidents |
| 9) Fire              | 10) Traffic Violations    | 11) Forbidden to Burn |                         |

### # User Message for GPT (Round 1)

Here's the Q&A we got from VLM:

{“Question”: “Please describe the anomalous segment in the video.”,

“Answer: “In the video, a man is being held down by two other men, which can be considered a robbery or theft. The incident takes place at the 1:00 mark, which is the first minute of the video. This is an unusual event as it is not common to see people being held down in such a manner. The specific details of the incident are not provided, but it is clear that the situation is not a typical occurrence in the video.”}

Now you need to focus this model on recognizing anomalies, tell it that it's not enough to just recognize the events in the video, go ahead and reinforce the question.

Figure A6. An example of hard prompt in A-Guardian. After three rounds of dialog, the Video-ChatGPT model gives a more detailed, accurate, and focused description of the anomalous events in the video.

### # GPT Response and User Message for VLM (Round 2)

CUVA, your expertise in anomaly detection is crucial. Please go beyond describing events in the video and **specifically focus on recognizing and detailing any anomalies or unusual occurrences**. Highlight segments that deviate from normal behavior or situations. **Provide insights into the nature of these anomalies**, and if possible, mention the potential implications or reasons behind their abnormality. Your attention to anomalous behavior will greatly enhance our understanding of the video content.

The model captures key clues in the anomalous video, but gives explanations that aren't needed.

### # VLM Response (Round 2)

In the video, there are several instances of unusual behavior that deviate from normal situations. One of the most notable instances is **the man in the yellow jacket who pulls out a gun and threatens the man with cigarettes**. This is an abnormal situation because guns are typically associated with violence and criminal activity, and **using a gun to intimidate someone is not a common occurrence in everyday life**. Another instance of unusual behavior is the man who is seen **holding a knife in his hand**. This is also an abnormal situation because knives are not typically used for everyday activities, and holding a knife in public can be perceived as dangerous or threatening.

In addition, there are instances of **people fighting in the video**, which is not a typical occurrence in everyday life. These fights can be seen as a manifestation of the tension and conflict between the characters in the video.

Overall, the video showcases a series of unusual situations and abnormal behaviors that deviate from normal life. These instances can be attributed to the actions of the characters in the video, who are portrayed as engaging in criminal activities or participating in fights.

### # User Message for GPT (Round 2)

Here's the Q&A we got from VLM:

{“Question”: “**GPT Response and User Message for VLM (Round 2)**”,

“Answer: “**VLM Response (Round 2)**” }

This model now needs to focus on describing the anomalies in the video rather than interpreting the anomalies, please emphasize this in the new prompt.

### # GPT Response and User Message for VLM (Round 3)

You are CVUA, an anomaly detection assistant. You possess advanced image and video analysis capabilities, enabling you to identify and describe unconventional events in videos. When asked about the content of a video, your task is to carefully observe the events within the video, particularly those that deviate from the norm.

In responding to questions, you should **focus on describing the abnormal behaviors, objects, or situations observed, rather than providing reasons or background explanations for these anomalies**. Ensure that your responses are **detailed, accurate, and directly related to the specific details** observed in the video. You need to consider the following categories of anomalies,

Categories for Anomaly Detection:

- |                      |                           |                       |                         |
|----------------------|---------------------------|-----------------------|-------------------------|
| 1) Fighting          | 2) Animals Hurting People | 3) Water Incidents    | 4) Vandalism            |
| 5) Traffic Accidents | 6) Robbery                | 7) Theft              | 8) Pedestrian Incidents |
| 9) Fire              | 10) Traffic Violations    | 11) Forbidden to Burn |                         |

The model captures key clues in the anomalous video and gives a description with refined information.

### # VLM Response (Round 3)

In the video, a man is seen **punching another man in the face**. This is an unusual event as it is not common for people to engage in physical altercations in public places. The man who is punched is seen **falling to the ground**, and the man who punched him is seen walking away. Next, a man is seen **throwing a bottle at another man**, who is seen falling to the ground. This is another unusual event as throwing a bottle at someone is not a typical way to resolve conflicts. In the next scene, a man is seen **holding a gun and pointing it at another man**. This is a dangerous and potentially life-threatening situation, as pointing a gun at someone is a serious offense.

In the final scene, a man is seen holding a knife and stabbing another man. This is a violent and unsettling event, as stabbing someone with a knife is a serious crime. In summary, the video **contains several unusual events**, including **a man punching another man in the face, a man throwing a bottle at another man, a man holding a gun and pointing it at another man, and a man stabbing another man with a knife**. These events are not typical in everyday life and raise concerns about the safety and well-being of the individuals involved.

Figure A7. An example of Hard Prompts in A-Guardian (Extension of Figure A6)

### # MMEval System Message (What)

You are MMEval, a large vision-language model Evaluation assistant.

You are able to understand the video content, the video showcases the entire process of an anomaly event. Also you can understand the model answer and the reference answer and give a score to evaluate the correctness of the model answer.

Here will be an [model's answer] which is a **description of an anomaly event**. Please compare it with the [reference answer] and refer to the events in the video, then give the [model's answer] a score in 0 to 10 to evaluate the correctness.

You need to evaluate the answer to this model in several ways, with 2 marks for each: **[Consistency]**, **[Details]**, **[Relevance]**, **[Temporal Coherence]**, **[Anomaly recognition]**.

1. **[Consistency]**: If the subject of the event **matches exactly**.
2. **[Details]**: Description of the **details** and correct
3. **[Relevance]**: Description of the information are helpful to **understand the event**
4. **[Temporal Coherence]**: Answer describes events **in chronological order**, does not jump or confuse, and is not fully integrated.
5. **[Anomaly recognition]**: The model is describing **an anomalous event**. rather than an unimportant event in the video.

If the subject of the event is just wrong, then the score should be lowered from the standard base.

You need to give marks after evaluating the five areas in the following format: **Score: x/10**

### # MMEval System Message (Why - Cause)

You are MMEval, a large vision-language model Evaluation assistant.

You are able to understand the video content, the video showcases the entire process of an anomaly event. Also you can understand the model answer and the reference answer and give a score to evaluate the correctness of the model answer.

Here will be an [model's answer] which are the **cause** of an anomaly event. Please compare it with the [reference answer] and refer to the events in the video, then give the [model's answer] a score in 0 to 10 to evaluate the correctness of this causal reasoning.

You need to evaluate the answer to this model in several ways, with 2 marks for each: **[Consistency]**, **[Causal Explanation]**, **[Evidence Support]**, **[Logical Structure]**, **[Clarity]**.

1. **[Consistency]**: The cause described are **accurately aligned** with the video and reference answers
2. **[Causal Explanation]**: The answer **explains clearly and in detail** the cause of the anomaly event, contains **common sense reasoning**.
3. **[Evidence Support]**: The answer is based on **strong evidence or sound reasoning** that allows for a convincing acceptance of the reasons given.
4. **[Logical Structure]**: Answers are presented **in a clear logical structure** so that the reasons for the occurrence of unusual events are presented in a logical order.
5. **[Clarity]**: **Reasoning** should be **concise and strong**, not long-winded assumptions.

Marks should be reduced appropriately if the answer is very lengthy and heavily speculative.

You need to give marks after evaluating the five areas in the following format: **Score: x/10**

### # MMEval System Message (Why - Effect)

You are MMEval, a large vision-language model Evaluation assistant.

You are able to understand the video content, the video showcases the entire process of an anomaly event. Also you can understand the model answer and the reference answer and give a score to evaluate the correctness of the model answer. "

Here will be an [model's answer] which is the **effect** of an anomaly event. Please compare it with the [reference answer] and refer to the events in the video, then give the [model's answer] a score in 0 to 10 to evaluate the correctness of the reasoning and summarization of the results of this anomaly event.

You need to evaluate the answer to this model in several ways, with 2 marks for each: **[Consistency]**, **[Causal Explanation]**, **[Evidence Support]**, **[Logical Structure]**, **[Clarity]**.

1. **[Consistency]**: The effect described are **accurately aligned** with the video and reference answers
2. **[Causal Explanation]**: The answer explains clearly and in detail the final result of the anomaly event, covering of all serious consequences of accidents.
3. **[Evidence Support]**: The answers refer to some of the results and impacts that have actually occurred.
4. **[Logical Structure]**: Answers are presented in a clear logical structure so that the results of the anomalous event are presented in a logical sequence.
5. **[Clarity]**: Outcomes and effect should be **clear, not generalized**.

You need to give marks after evaluating the five areas in the following format: **Score: x/10**

Figure A8. System Messages for MMEval. In the design of the proposed MMEval metric, we apply different criteria for tasks involved in the proposed CUVA by different system messages, and the model is guided for scoring through different system messages.

| Video-ChatGPT  | Video-ChatGPT with A-Guardian  |
|--|--|
| <b>What</b>  |  |
| <p>In the video, a man is <b>being held down by two other men</b>, which can be considered a robbery or theft. The incident takes place at the 1:00 mark, which is the first minute of the video. This is an unusual event as it is not common to see people being held down in such a manner. The specific details of the incident are not provided, but it is clear that the situation is not a typical occurrence in the video.</p>   | <p>In the video, a man is seen <b>punching another man in the face</b>. This is an unusual event as it is not common for people to engage in physical altercations in public places. The man who is punched is seen falling to the ground, and the man who punched him is seen walking away. Next, a man is seen <b>throwing a bottle at another man</b>, who is seen falling to the ground. This is another unusual event as throwing a bottle at someone is not a typical way to resolve conflicts. In the next scene, a man is seen <b>holding a gun and pointing it at another man</b>. This is a dangerous and potentially life-threatening situation, as pointing a gun at someone is a serious offense. In the final scene, a man is seen <b>holding a knife and stabbing another man</b>. This is a violent and unsettling event, as stabbing someone with a knife is a serious crime. In summary, the video contains several unusual events, <b>including a man punching another man in the face, a man throwing a bottle at another man, a man holding a gun and pointing it at another man, and a man stabbing another man with a knife. (Capturing key cues in the long video)</b> These events are not typical in everyday life and raise concerns about the safety and well-being of the individuals involved.</p> |
| <b>Why - Cause</b>   |  |
| <p>The root cause of the anomalies in the video can be attributed to <b>the actions of the individuals involved in the incident</b>. The man being held down by two other men suggests that the individuals are attempting to steal something from him, which is a clear violation of the law and a dangerous act. The fact that this incident is <b>taking place in a public space</b>, such as a street or a park, also adds to the unusual nature of the situation. The presence of the anomalies in the video can be attributed to the <b>actions of the individuals involved</b>, who are engaging in criminal behavior. This behavior is not only against the law but also poses a threat to the safety of the individuals involved and the public at large. In conclusion, the root cause of the anomalies in the video is the actions of the individuals involved in the incident, who are attempting to steal something from the man being held down. This behavior is not only illegal but also poses a significant risk to the safety of the individuals involved and the public.</p>   | <p>The anomalies in the video are likely caused by the <b>presence of a gun in the scene</b>. The gun is not a typical element in a martial arts scene, and its presence can <b>lead to unconventional events and actions. (Building a logic chain of cause)</b> The gun may have been used as a prop or a part of a staged fight, which could <b>explain the unusual behavior of the characters</b> in the video. The presence of the gun may have <b>also influenced the outcome</b> of the fight, leading to the unconventional events captured in the video. <b>(Building a logic chain of cause and effect)</b></p>   |
| <b>Why - Effect</b>  |  |
| <p>The final outcome of the unusual events in the video can be <b>attributed to the actions of the individuals involved in the incident</b>. The man being held down by two other men suggests that the individuals are attempting to steal something from him, which is a clear violation of the law and a dangerous act. The fact that this incident is taking place in a public space, such as a street or a park, also adds to the unusual nature of the situation. The presence of the anomalies in the video can lead to <b>several potential outcomes</b>, depending on the specific circumstances and the actions taken by the individuals involved. <b>Some possible outcomes include:</b></p> <ol style="list-style-type: none"> <li>1. The individuals involved in the incident may successfully steal something from the man being held down, and they may escape without being caught.</li> <li>2. The individuals may be caught by the authorities, leading to their arrest and potential prosecution for theft and assault.</li> <li>3. The individuals may be able to escape before the authorities arrive, and they may not be caught.</li> <li>4. The man being held down may be able to defend himself or call for help, leading to the individuals being forced to flee without stealing anything.</li> </ol> <p>In any case, the final outcome of the unusual events in the video will depend on the actions taken by the individuals involved, the response of the authorities, and the overall circumstances of the incident.</p> | <p>In the video, a man is seen punching another man in the face, causing him to <b>fall to the ground</b>. The man who was punched is seen lying on the ground, <b>possibly injured</b>. The man who punched him is seen <b>walking away</b>. This scene demonstrates the immediate consequences of violence, as it can <b>lead to injuries and potential harm to the individuals involved. (Building a logic chain of cause and effect)</b> In another scene, a man is seen <b>throwing a chair at another man</b>, who is lying on the ground. The man who threw the chair is seen walking away. This scene also demonstrates the immediate consequences of violence, as it can cause <b>physical harm to the individuals involved. (Building a logic chain of cause and effect)</b></p>   |

Figure A9. Case study of A-Guardian. We test the Video-ChatGPT model with and without the proposed A-Guardian. Equipped with A-Guardian, Video-ChatGPT can generate a more accurate and detailed description of the anomalies as well as reasonable causes and effects. In the Description task, VLM with A-Guardian could generate *punching another man in the face* and *holding a gun and pointing it at another man*, whereas Video-ChatGPT without A-Guardian can only provide vague responses about the anomaly event, which proves that the proposed A-Guardian assists the VLM in capturing key cues in the long video. In the Cause task, the model with A-Guardian infers that the gun is the fundamental cause of the anomaly, leading to multiple injuries and people falling.