Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity

Supplementary Material

A. Details of the Data Engine.

To construct a dataset with hierarchical annotations with both short-term and long-term anomalies, we developed a semi-automated annotation engine that combines manual efforts with the generative capabilities of LLM. In the main paper, we present the complete annotation workflow. Below, we provide additional details about the data engine.

A.1. Hierarchical Video Decoupling

Before annotation, we collected videos from the training sets of the UCF-Crime [5] and XD-Violence [7] datasets. From UCF-Crime, we selected 758 normal videos and 735 anomaly videos, while from XD-Violence, we selected 1,904 normal videos and 2,046 anomaly videos. The anomaly videos included their original video-level labels, *e.g.*, *Abuse*, *Explosion*. For the anomaly videos, we organized a team of five annotators to label each anomaly event within the videos. The annotation process took approximately 20 hours to complete. For the normal videos, we considered all segments to be normal and randomly cropped segments of varying lengths to serve as normal event-level video segments. These anomaly and normal event-level videos were further divided into shorter clip-level segments. For UCF-Crime, we adopted the clip-level divisions from UCA [9]. For XD-Violence, we performed uniform division.

A.2. Hierarchical Free-text Annotation

Clip Captioning. For videos in UCF-Crime, we fully utilized the manually annotated captions from UCA [9]. For videos in XD-Violence, we used LLaVA-Next-Video-7B [11] as our captioner to generate textual descriptions for clip-level videos. The specific prompt is as follows:

'Please provide a short and brief description of the video clip, focusing on the main subjects and their actions.'

Event Summary. We combined all captions and video-level category labels to generate anomaly-related summaries for each event using an LLM. We selected LLaMA3-70B [1] as our LLM due to its strong summarization capabilities. The specific prompt is as follows:

'The dense caption of the video is: {clip captions}. There are (is no) abnormal events ({video-level label}) in the video. Your response should include the following three parts: 1. Whether the anomaly exists and the specific name of the anomaly. 2. A summary of the anomaly events. 3. Brief explanation of the basis for judging the anomaly.'

Video Summary. Similar to generating event summaries, we generated video-level summaries by analyzing the event-level summaries. The specific prompt is as follows:

'Below is a summary of all the events in the video: {event summaries}. There are (is no) abnormal events ({video-level label}) in the video. Your response should include the following three parts: 1. Whether the anomaly exists and the specific name of the anomaly. 2. Detailed description of the video anomaly event from start to end. 3. Brief analysis of the basis for judging the anomaly.'

Annotation Format. In Fig.A, we present an example of the hierarchical free-text annotations for a video.

A.3. Hierarchical Instruction Data Construction

To construct the instruction dataset, we designed question prompts tailored to different tasks, including **Caption**, **Judgment**, **Description**, and **Analysis**. For each instruction item, we randomly selected one prompt from the pool and matched it with the corresponding content from the free-text annotations as the answer.

Caption.

1. "Describe the video briefly."
2. "Describe the main events that take place in this video."
3. "Give a short description of the video."
4. "What happened in this video?"
5. "Generate a brief caption for the video."
6. "Can you provide a brief description of the video?"
7. "Briefly describe the main subjects and their actions in the video."
8. "Provide a short overview of what happens in the video?"
9. "Describe the key moments that showcase the subjects' activities in the video."
10. "Describe the sequence of events involving the main subjects in the video."
11. "What activities happen throughout the video?"
12. "Describe the main subjects and their roles in the video."
13. "What key moments stand out in the video?"
14. "What are the primary activities showcased in the video?"
15. "What happens to the main subjects as the video progresses?"
16. "What is a brief overview of what happens in the video?"
17. "Describe the main subjects and their contributions to the video."
18. "Describe the key events in the video."
19. "Describe the video's main activities."
20. "Can you describe the main action in this video briefly?"
21. "Describe the video clip concisely."
22. "Provide a brief description of the given video clip."
23. "Summarize the visual content of the video clip."
24. "Give a short and clear explanation of the subsequent video clip."

Judgement.

1. "What types of anomalies are shown in the video clip?"
2. "Are there any anomaly events detected in the video?"
3. "Detect and classify the anomaly events in the video."
4. "Identify any abnormal behaviors depicted in the video."
5. "Determine whether there are anomaly events in the video and the specific name of the anomaly."
6. "What anomalies can be identified in the video?"
7. "What categories of anomalies can be found in the video?"
8. "Could you point out any abnormal actions in the video?"
9. "Point out the abnormal actions in the video."


```
{
    "video": "v=2rfyeR-YaJw__#1_label_G-0-0",
    "n_frames": 1940,
    "fps": 24.0,
    "label": ["Explosion"],
    "clips": [[[5.583, 11.903], [11.903, 18.222], [18.222, 24.542]],
              [[36.167, 43.48], [43.48, 50.792]]],
    "clip_captions":[
        [
            "A military tank moving across a barren landscape with low-rise buildings and sparse vegetation. the sky is overcast, and the overall color palette is muted with earthy tones.",
            "A series of images depicting a barren landscape with a few buildings in the background. the foreground consists of a rocky terrain with sparse vegetation. the sky is overcast, and there are no visible people or moving objects.",
            "A silhouette of a person operating a large, mounted weapon on a rocky terrain under a clear sky. the individual appears to be adjusting or aiming the weapon."
        ],
        [
            "A series of images depicting a hazy, dusty environment with buildings in the background. the focus is on a red light source that appears to be moving or flashing, possibly indicating some form of activity or event occurring in the area.",
            "A series of images depicting a military tank moving across a landscape with buildings in the background. the tank is seen from different angles, and there are visible explosions occurring in the distance."
        ]
    ],
    "events": [[5.583, 24.542], [36.167, 50.792]],
    "event_summary": [
    "The anomaly exists, specifically an Explosion anomaly. The basis for judging the anomaly is the presence of an unexpected and sudden release of energy, likely accompanied by a loud noise and visible effects such as smoke, fire, or debris.",
    "The anomaly exists and its specific name is Explosion. The anomaly event is a series of visible explosions occurring in the distance as a military tank moves across a landscape with buildings in the background. The basis for judging the anomaly is the sudden and intense release of energy and light in the form of explosions, which is an unusual and notable event in the otherwise desolate and quiet war-torn environment depicted in the video"],
    "video_summary": "The anomaly exists, specifically named as Explosion. The anomaly event depicts a series of visible explosions occurring in the distance as a military tank moves across a landscape with buildings in the background, characterized by sudden and intense releases of energy and light, accompanied by loud noise and visible effects such as smoke, fire, or debris. The basis for judging the anomaly lies in the unusual and notable nature of these explosions, which stand out against the otherwise desolate and quiet war-torn environment depicted in the video, making them an unexpected and sudden release of energy that grabs attention."
}
```


Figure A. **An example of hierarchical free-text annotations.** For each labeled video, the hierarchical free-text annotations include clip-level captions, event-level, and video-level anomaly analysis. Additionally, the temporal boundaries for each event and clip are annotated.

10. "Are there anomalies observed in the video clip?"

Description.

1. "Describe the anomaly events observed in the video."
2. "Could you describe the anomaly events observed in the video?"
3. "Could you specify the anomaly events present in the video?"
4. "Give a description of the detected anomaly events in this video."
5. "Could you give a description of the anomaly events in the video?"
6. "Provide a summary of the anomaly events in the video."
7. "Could you provide a summary of the anomaly events in this video?""
8. "What details can you provide about the anomaly in the video?"
9. "How would you detail the anomaly events found in the video?"
10. "How would you describe the particular anomaly events in the video?"

Analysis.

1. "Why do you judge this event to be anomalous?"
2. "Can you provide the reasons for considering it anomalous?"
3. "Can you give the basis for your judgment of this event as an anomaly?"
4. "What led you to classify this event as an anomaly?"
5. "Could you provide the reasons for considering this event as abnormal?"
6. "What evidence do you have to support your judgment of this event as an anomaly?"
7. "Can you analyze the factors contributing to this anomalous event?"
8. "Could you share your analysis of the anomalous event?"
9. "What patterns did you observe that contributed to your conclusion about this event being an anomaly?"
10. "How do the characteristics of this event support its classification as an anomaly?"

A.4. Data Samples.

To facilitate understanding, we provide the final constructed instruction data at various temporal granularities, including **clip-level**, **event-level**, and **video-level**, as shown in Fig. B, Fig. C, and Fig. D.

B. Details of the Anomaly Scorer

B.1. Model Architecture

We use UR-DMU [12] as the anomaly scorer in our Anomaly-focused Temporal Sampler. As shown in Fig. E, UR-DMU utilizes a Global and Local Multi-Head Self Attention (GL-MHSA) module to capture both long-range and short-range temporal relationships among video snippets. Furthermore, UR-DMU introduces two memory banks to store and differentiate abnormal and normal prototypes, thereby maximizing the margins between these two representations. In order to learn discriminative representations, UR-DMU uses triplet loss to increase the feature distance

```
{
    "id": "Abuse007_x264_E0C2",
    "type": "clip",
    "video":
    ↪  "ucf-crime/clips/train/Abuse007_x264_E0C2.mp4",
    "conversations": [
    {"from": "human",
    "value": "Describe the key events in the video."},
    {"from": "gpt",
    "value": "The man in shorts was slapped to the
    ↪  ground by the strong man opposite."}
    ]
}
```

Figure B. **An example of clip-level instruction data.**

```
{
    "id": "Abuse018_x264_E0",
    "type": "event",
    "video":
    ↪  "ucf-crime/events/train/Abuse018_x264_E0.mp4",
    "conversations": [
    {"from": "human",
    "value": "Identify any abnormal behaviors depicted
    ↪  in the video."},
    {"from": "gpt",
    "value": "An anomaly exists, specifically an
    ↪  instance of Police Brutality."},
    {"from": "human",
    "value": "Could you provide a summary of the anomaly
    ↪  events in this video?"},
    {"from": "gpt",
    "value": "The anomaly event involves the police
    ↪  officer kicking the man who rushed towards him,
    ↪  causing the man to fall heavily to the ground
    ↪  with his head on his back."},
    {"from": "human",
    "value": "Could you share your analysis of the
    ↪  anomalous event?"},
    {"from": "gpt",
    "value": "The excessive use of force by the police
    ↪  officer, which deviates from standard protocol
    ↪  and may be considered an abuse of authority."}
    ]
}
```

Figure C. **An example of event-level instruction data.**

after interacting with different memories. Simultaneously, it utilizes KL loss to constrain the normal memory to follow a Gaussian distribution, accounting for the variance introduced by noise. Furthermore, We leveraged the event-level anomaly boundaries obtained during the annotation phase to generate frame-level labels and computed a binary cross-entropy loss, *i.e.*, $\mathcal{L}_{AS}$, which is simple yet effective. Thus, the loss function for the anomaly scorer is defined as follows:

$$\mathcal{L} = \mathcal{L}_{AS} + \mathcal{L}_{triplet} + \mathcal{L}_{kl} \tag{1}$$

```
{
    "id": "v=xLFbqPzSI2Y__#1_label_B4-0-0",
    "type": "video",
    "video": "xd-violence/videos/train/
    v=xLFbqPzSI2Y__#1_label_B4-0-0.mp4",
    "conversations": [
    {"from": "human",
    "value": "Are there any anomaly events detected in
    ↪  the video?"},
    {"from": "gpt",
    "value": "The anomaly exists and its specific name
    ↪  is the Riot anomaly."},
    {"from": "human",
    "value": "Could you give a description of the
    ↪  anomaly events in the video?"},
    {"from": "gpt",
    "value": "The anomaly event depicts a chaotic and
    ↪  intense urban unrest scene, starting with a
    ↪  chaotic street scene where individuals are
    ↪  running away from an area of fire and smoke,
    ↪  with people appearing to be in a state of panic
    ↪  or urgency, and some being sprayed with water by
    ↪  individuals in uniform. The scene escalates to
    ↪  show a growing fire in the background, police
    ↪  officers in riot gear standing on the street,
    ↪  and a formation of police officers spread out
    ↪  across the road, suggesting an ongoing riot or
    ↪  civil unrest. "},
    {"from": "human",
    "value": "Can you give the basis for your
    ↪  judgment?"},
    {"from": "gpt",
    "value": "The presence of multiple indicators of a
    ↪  riot or emergency situation, including fire,
    ↪  smoke, and people fleeing in panic, as well as
    ↪  the sudden appearance of flames, smoke, and
    ↪  police in riot gear, and the presence of violent
    ↪  and destructive events, such as fires and
    ↪  explosions, and the response of authorities,
    ↪  including the use of tear gas or pepper spray,
    ↪  which deviate significantly from normal,
    ↪  peaceful urban activity."}
    ]
}
```

Figure D. **An example of video-level instruction data.**

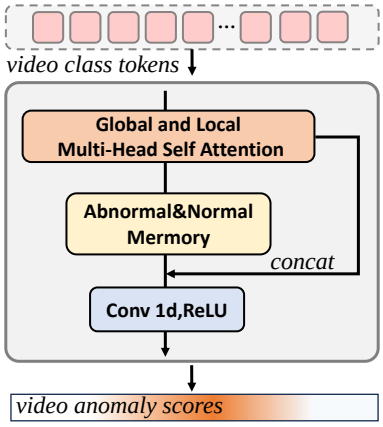


Figure E. **Architecture of the Anomaly Scorer (UR-DMU [12]).**

B.2. Visualization Results

In Fig. F, we present visualized results of anomaly scores and sampled frames on the UCF-Crime and XD-Violence test sets. These results demonstrate the accuracy of our

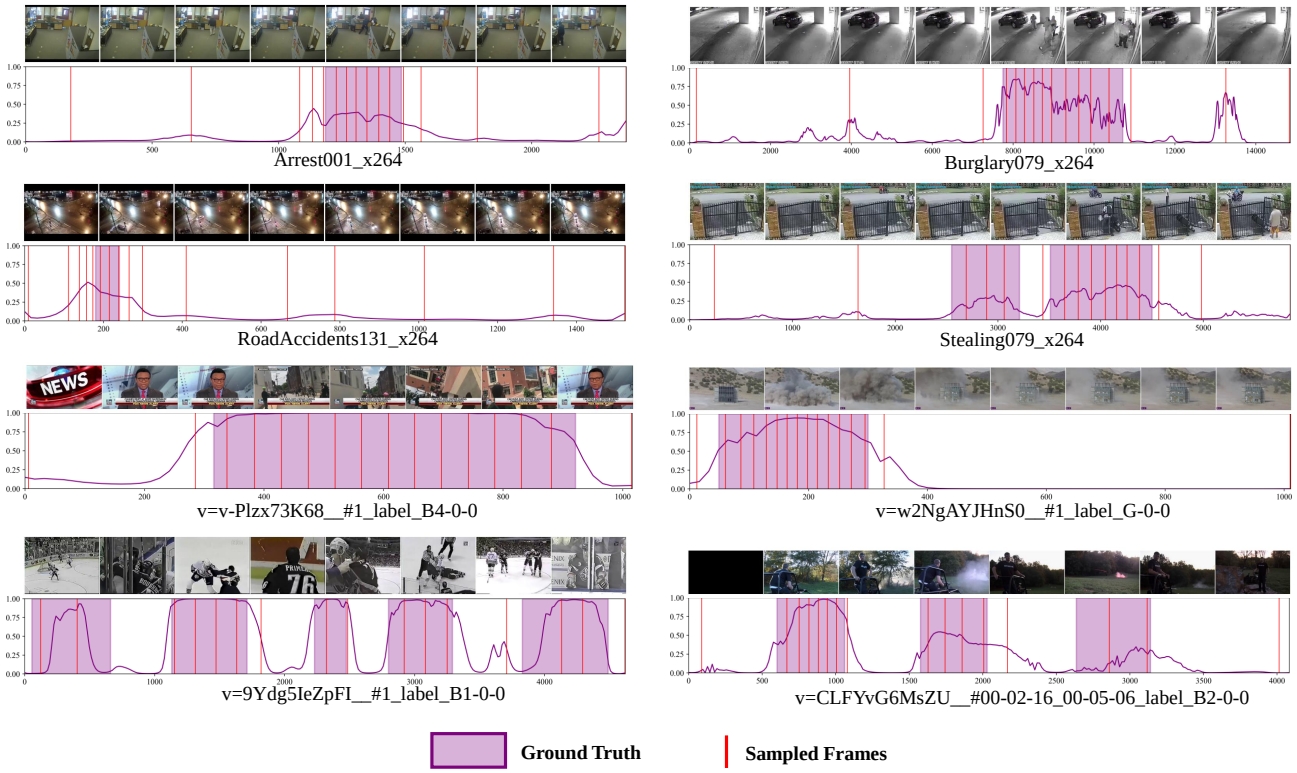


Figure F. **Visualization results of anomaly scores and sampled frames output by the Anomaly-focused Temporal Sampler.**.

Table A. **Comparision with related multimodal/explanable VAU methods and benchmarks.** HIVAU-70k provides accurate temporal annotations and hierarchical anomaly-related free-text annotations.

Methods	#Catogories	#Samples	Text			Temp. Anno.	MLLM tuning
			clip-level	event-level	video-level		
UCA [8]	13	23,542	✓	✗	✗	✓	✗
LAVAD [10]	N/A	N/A	✓	✗	✓	✗	✗
VAD-VideoLLama [4]	13/7	2,400	✗	✗	✓	✗	projection
CUVA [3]	11	6,000	✗	✗	✓	✗	✗
Hawk [6]	-	16,000	✗	✗	✓	✗	projection
HIVAU-70k (Ours)	**19**	**70,000**	✓	✓	✓	✓	LoRA

method in anomaly detection within complex real-world scenarios, with the sampled frames being concentrated in anomalous regions.

C. Discussion with related works.

In Table A, we provide a comprehensive comparison with related works in terms of benchmarks and methods.

Summary of related works: Recently, there has been substantial research on multi-modal Video Anomaly Understanding, making significant contributions to advancing open-world anomaly understanding. LAVAD [10] utilized several pre-trained foundational models to offer a training-free explainable VAD process. VAD-VideoLLaMA [4], designed a three-phase training method to finetune Video-LLaMA in the VAD domain. CUVA [3] introduced a dataset and metric for evaluating causation understanding of video anomalies. Hawk [6] constructed an instruction dataset and finetuned a video-language framework that incorporates both motion and video information.

Difference and Advantages of our proposed benchmark and method:

- We develop a semi-automated annotation engine that scales hierarchical anomaly annotation efficiently, combining manual refinement with LLM-based annotation to maintain high-quality data across multiple granularities, resulting in over **70,000** annotations at clip, event, and video levels, which significantly surpasses previous

datasets in scale.

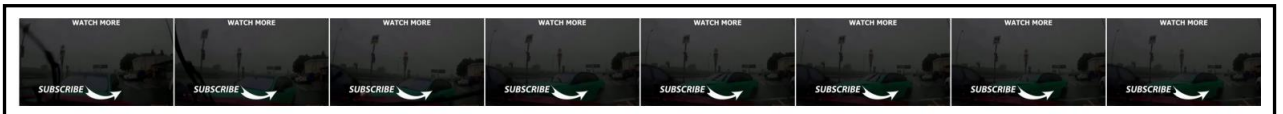
- UCA [9] only provides clip-level captions, overlooking the understanding of anomalies across longer time spans. CUVA [3] and Hawk [6], on the other hand, only offer video-level instruction data, neglecting finer-grained visual perception and anomaly analysis. In contrast, our proposed HIVAU-70k takes a multi-temporal granularity perspective, offering more comprehensive and diverse anomaly annotations for open-world anomaly detection. It enables progressive and comprehensive learning, from short-term visual perception to long-term anomaly reasoning.
- We propose the **Anomaly-focused Temporal Sampler (ATS)**, integrated with a multi-modal visual-language model. Benefiting from the precise temporal annotations we provide, the ATS is able to focus on anomaly-dense video segments. This integration significantly improves efficiency and accuracy in long-video anomaly detection.

D. More Qualitative Results.

As shown from Fig. G to Fig. J, we present the output of explainable text generated by Holmes-VAU compared with the base model, *i.e.*, InternVL-2B [2]. The results demonstrate significant improvements in the model's visual perception and anomaly analysis capabilities after fine-tuning on HIVAU-70k.

E. Limitations and Future Work.

While our work demonstrates significant strides in multi-granular video anomaly understanding, several limitations present opportunities for future enhancement. First, optimizing for real-time streaming remains a challenge. Our sparse sampling approach improves efficiency, but further refinement is necessary for seamless deployment in streaming contexts. Additionally, our work has so far focused on surveillance data, extending our framework to other domains, such as industrial monitoring and medical diagnostics, will help validate its generalization capabilities. Lastly, integrating additional sensory data, like audio, and scalable hierarchical annotation could enhance anomaly detection and broaden applicability.



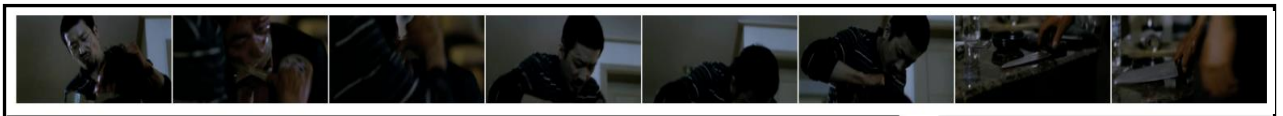
Prompt: What activities happen throughout the video?

clip-level <Caption>

Ground Truth:
A green car driving on a road in rainy weather. the windshield wipers are in motion, and the road appears wet. there are traffic signs and other vehicles visible in the background.

InternVL2
Throughout the video, we see cars driving on the road, a green car driving on the road, a green car driving on the road, a green car driving on the road, a green car driving on the road

Holmes-VAU (Ours)
A view from inside a vehicle, looking out through the windshield at a street scene. the sky is overcast, and the visibility is low due to the weather conditions. there are several cars on the road, including a green car in the foreground. the road is lined with traffic signs and streetlights, and there are buildings and trees visible in the background.



Prompt: Describe the main events that take place in this video.

clip-level <Caption>

Ground Truth:
A man in a striped shirt holding another person by the hair, appearing to be in distress. the scene transitions to a close-up of the man's hands as he picks up a knife from a cluttered kitchen counter.

InternVL2
The video clip shows a man in a striped shirt cutting a piece of meat with a knife on a kitchen counter. The camera focuses on the man's hands and the knife as he cuts the meat. The background is blurred, emphasizing the action in the foreground.

Holmes-VAU (Ours)
A close-up of two individuals engaged in a physical altercation. one person is wearing a striped shirt and appears to be grappling with another individual who is dressed in dark clothing. the setting is dimly lit, suggesting an indoor environment, possibly a room or a kitchen

Figure G. **Qualitative comparison of anomaly understanding explanation with our baseline model, *i.e.*, InternVL-2B.** Correct and wrong explanations are highlighted in green and red, respectively.

Prompt: Summarize the visual content of the video clip.

clip-level <Caption>

Ground Truth:
An aerial view of a rural area with fields and a road. a building is shown in the center of the frame, surrounded by a fence. suddenly, an explosion occurs at the building, causing a massive fireball and a plume of smoke to rise into the sky.

InternVL2
The video shows an aerial view of a building that is on fire. The flames and smoke are visible in the video.

Holmes-VAU (Ours)
A large explosion occurring in an industrial area, with thick plumes of smoke and fire rising into the sky. the explosion is massive and appears to be the central focus of the scene.

Prompt: Describe the video clip.

clip-level <Caption>

Ground Truth:
A series of intense action scenes involving a group of individuals in a chaotic environment. the main subject, dressed in dark clothing, is seen engaging in combat with multiple opponents wielding sticks and other weapons. the setting appears to be an industrial or construction site, with visible scaffolding and construction materials

InternVL2
A group of people are fighting in a factory.

Holmes-VAU (Ours)
A group of individuals engaged in a physical altercation in an industrial setting. the participants are wearing various types of clothing, including helmets and protective gear, and are wielding sticks and other objects. the environment is cluttered with equipment and machinery, suggesting a workshop or factory floor.

Figure H. **Qualitative comparison of anomaly understanding explanation with our baseline model, *i.e.*, InternVL-2B.** Correct and wrong explanations are highlighted in green and red, respectively.

Prompt: Are there any anomaly events detected in the video?

event-leve <Judgment>

Ground Truth:
An anomaly exists, specifically an Assault anomaly.

InternVL2
No, there are no anomaly events detected in the video.

Holmes-VAU (Ours)
The anomaly exists and the specific name of the anomaly is Assault.

Prompt: How would you detail the anomaly events found in the video?

event-level <Description>

Ground Truth:
The anomaly event involves a white car colliding with a black car, causing the black car to be hit onto the sidewalk, and people waiting for the bus to scatter in all directions

InternVL2
There are no anomaly events found in the video.

Holmes-VAU (Ours)
The anomaly event involves a sudden and unexpected collision between a car and a bus, resulting in a cloud of dust and debris being kicked up into the air, followed by people running away from the scene, indicating a sudden and unexpected event that disrupts the normal flow of traffic and pedestrian activity.

Prompt: Could you provide the reasons for considering this event as abnormal?

event-level <Analysis>

Ground Truth:
The sudden and unexpected impact between the two vehicles, deviating from the normal flow of traffic, indicating a hazardous and unusual event.

InternVL2
The presence of multiple vehicles involved in the incident suggests that it was not an isolated incident but rather part of a larger event or series of events.

Holmes-VAU (Ours)
The sudden and unexpected collision between the two vehicles, deviating from the normal flow of traffic, which is a clear indication of an anomaly event.

Figure I. **Qualitative comparison of anomaly understanding explanation with our baseline model, *i.e.*, InternVL-2B.** Correct and wrong explanations are highlighted in green and red, respectively.

Prompt: Could you point out any abnormal actions in the video?

video-level <Judgment>

Ground Truth:
The anomaly exists and is specifically named Physical Abuse.

InternVL2
There are several categories of anomalies that can be observed. Firstly, there is a scene where a man is seen standing in front of a toilet, and another man is standing behind him. Additionally, there is a scene where a man is standing in front of a toilet, and another man is standing behind him.

Holmes-VAU (Ours)
The anomaly exists, and the specific name of the anomaly is Abuse.

Prompt: Provide a summary of the anomaly events in the video.

video-level <Description>

Ground Truth:
The anomaly event depicts a chaotic and tense atmosphere in Paris, with scenes of protests, vandalism, and arson, including the use of tear gas, smoke bombs, and water cannons, as well as vehicles and buildings being set on fire and vandalized with graffiti.

InternVL2
Throughout the video, there are several anomalies that can be observed. The first anomaly is the presence of yellow vests worn by some of the protesters, which is not a typical attire for such gatherings. The second anomaly is the use of water cannons to disperse the crowd, which is not a typical method of crowd control. The third anomaly is the fact that some of the protesters are wearing helmets, which is not a typical attire for such gatherings.

Holmes-VAU (Ours)
The anomaly events involve chaotic street scenes with people running, smoke in the air, and police officers in riot gear, indicating a state of unrest and potential violence.

Prompt: Why do you judge this event to be anomalous?

video-level <Analysis>

Ground Truth:
The presence of violent or destructive behavior, law enforcement intervention, and disruption of public order in the urban environment depicted in the video.

InternVL2
The presence of riot police, which are typically used for crowd control and maintaining order during protests or demonstrations, suggests that this is not a regular occurrence in the area

Holmes-VAU (Ours)
The presence of violent and destructive behavior, which deviates from normal and peaceful urban activity.

Figure J. **Qualitative comparison of anomaly understanding explanation with our baseline model, *i.e.*, InternVL-2B.** Correct and wrong explanations are highlighted in green and red, respectively.

References

[1] AI@Meta. Llama 3 model card. 2024. 1

[2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5

[3] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18793–18803, 2024. 4, 5

[4] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024. 4

[5] 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. 1

[6] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. *arXiv preprint arXiv:2405.16886*, 2024. 4, 5

[7] 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. 1

[8] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023. 4

[9] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22052–22061, 2024. 1, 5

[10] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 4

[11] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1

[12] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3769–3777, 2023. 2, 3