VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs

Anonymous ACL submission

Abstract

The recent advancements in Large Language Models (LLMs) have greatly influenced the development of Large Multi-modal Video Models (Video-LMMs), significantly enhancing our ability to interpret and analyze video data. Despite their impressive capabilities, current Video-LMMs have not been evaluated for anomaly detection tasks, which is critical to their deployment in practical scenarios e.g., towards identifying deepfakes, manipulated video content, traffic accidents and crimes. In this paper, we introduce VANE-Bench, a benchmark designed to assess the proficiency of Video-LMMs in detecting and localizing anomalies and inconsistencies in videos. Our dataset comprises an array of videos synthetically generated using existing state-of-the-art text-to-video generation models, encompassing a variety of subtle anomalies and inconsistencies grouped into five categories: unnatural transformations, unnatural appearance, pass-through, disappearance and sudden appearance. Additionally, our benchmark features real-world samples from existing anomaly detection datasets, focusing on crime-related irregularities, atypical pedestrian behavior, and unusual events. The task is structured as a visual question-answering challenge to gauge the models' ability to accurately detect and localize the anomalies within the videos. We evaluate nine existing Video-LMMs, both open and closed sources, on this benchmarking task and find that most of the models encounter difficulties in effectively identifying the subtle anomalies. In conclusion, our research offers significant insights into the current capabilities of Video-LMMs in the realm of anomaly detection, highlighting the importance of our work in evaluating and improving these models for real-world applications. Our code and data will be publicly available.

Large Language Models (LLMs) like ChatGPT have ushered in a new era of real-world AI applications in varied and diverse sectors like manufacturing, legal services, space exploration, transportation, retail, healthcare, education, and technology (Abdullah et al., 2022; Marr, 2023). Further, the current trend in development of these LLMs have been to introduce multi-modal capabilities like vision, and audio to these models along with text (et al., 2024; OpenAI, 2024). This motivates us to ask the question whether the current Large Multi-modal Models (LMMs) are capable and accurate in tackling the problem statement of Video Anomaly Detection (VAD) which has immense practical applications in factories, autonomous driving, crime warning, and traffic management (Liu et al., 2024a).

Further, we have recently observed superior visual quality of various AI-generated videos due to the improvements in the underlying algorithms, which are based on diffusion models, and transformers (Brooks et al., 2024; HPCAI Tech, 2024; Peebles and Xie, 2023). The current state-of-the-art (SOTA) AI text-to-video model is SORA from OpenAI (Brooks et al., 2024). The videos produced from SORA are of extremely high-fidelity that makes it nearly indistinguishable from real-life footage. Thus, SORA brings in new challenges in tackling misinformation, identifying deepfakes, and distinguishing real from fake videos especially during crucial events like democratic elections. Therefore, developing automated solutions to identify AI-generated videos has become the need of the hour.

Motivated by the above mentioned points, we propose a novel and challenging benchmark, **VANE-Bench: Video ANomaly Evaluation Benchmark**, to evaluate various closed-source, and open-source Video-LMMs on their ability to detect anomalies in the videos. Our VANE-Bench consists of both real-world video anomalies from diverse surveillance footage capturing unusual pedestrian behaviour, criminal activities, and unusual

Figure 1: Samples showing the AI-Generated video category of VANE-Bench. We collect these synthetic videos from SORA (Brooks et al., 2024), Open-Sora (HPCAI Tech, 2024), Runway Gen2 (Runway Research, 2024), ModelScopeT2V (Wang et al., 2023a), and VideoLCM (Wang et al., 2023b). The correct option in each question is highlighted in bold. Note that many of these anomalies are extremely subtle and difficult for humans to detect since the changes happen in rapid succession, with the entire video played in under a second. Anomalies identified with red bounding boxes for clarity. Note that our actual dataset does not contain bounding box overlays.

events, as well as subtle and challenging anomalies and inconsistencies present in various AI-generated videos (See Fig 1). These AI-generated videos, especially from SOTA models like SORA, have subtle and hard to detect anomalies, which makes this a challenging task even for many humans. However, automatically detecting and identifying the anomalies in these synthetic video clips serves as an important step towards identifying AI-generated videos in the wild. We reformulate the problem statement of VAD into a visual question-answering (VQA) task to facilitate easier evaluation of LMMs. However, despite evaluating over nine recent Video-LMMs on VANE-Bench, we find that most current LMMs still struggle on this benchmark (see Fig.2), making VANE-Bench a challenging and a useful benchmark for tracking the progress of Video-LMMs for the foreseeable future.

Our contributions can be summarized as follows:

1. We present VANE-Bench: Video ANomaly Evaluation Benchmark, consisting of 325 video clips, and 559 challenging question-answer pairs from both real-world video surveillance, and AI-generated videos.

2. We perform detailed evaluation of over nine state-of-the-art closed-source and open-source Video-LMMs on VANE-Bench, and show that most models exhibit poor performance, highlighting the challenging nature of our proposed benchmark.

3. We conduct detailed result analysis, and also perform human evaluation on VANE-Bench to set a reasonable benchmark target.

4. We will open-source our code, and describe

the data construction process of VANE-Bench along with making our data publicly available.

We hope that VANE-Bench serves as a strong benchmark to improve the performance and capabilities of Video-LMMs on anomaly detection.

1 Related Work

Video-LMMs: LMMs integrate linguistic and visual data to process videos, leveraging LLMs like Llama (Meta, 2024) and connecting them with modality-specific encoders via interfaces like Q-former (Zhang et al., 2024; Dai et al., 2023; Yin et al., 2024). Notable open-source Video-LMMs include VideoChat (Li et al., 2023b), which uses a chat-centric system, and VideoChatGPT (Maaz et al., 2023), which combines a visual encoder with an LLM for detailed video conversations. Video-LLaMA (Zhang et al., 2023a) integrates audio and visual signals using Q-formers, while LLaMA-VID (Li et al., 2023c) represents each frame with context and content tokens for efficient processing. Despite these advancements, our work shows current LMMs perform poorly on VANE-Bench, highlighting the need for stronger models in anomaly detection.

Video-LMMs Benchmarking: Benchmarks like SEED-Bench (Li et al., 2023a) and MV-Bench (Li et al., 2024b) assess general comprehension through multiple-choice questions but lack focus on anomaly detection in AI-generated videos. CVRR-ES (khattak et al., 2024) evaluates real-world scenarios with open-ended questions but doesn't address AI-generated inconsistencies. VANE-Bench specifically evaluates VAD in both real-world and AI-generated videos, providing a targeted benchmark for this task. While Perception Test (Pătrăucean et al., 2023) focuses on lower-level perception in real-world videos, VANE-Bench targets subtle anomalies in AI-generated content, making it essential for assessing Video-LMM robustness.

Video Anomaly Detection: Traditional VAD methods typically rely on hand-crafted features and statistical models to identify deviations from normality. CUVA (Du et al., 2024) is a comprehensive benchmark which focuses on the causation of video anomalies. A survey on generalized VAD (Liu et al., 2024b) categorizes various methodologies and highlights benchmark limitations. These methods often fail with complex AI-generated videos. VANE-Bench addresses this by focusing on VAD in such videos, complementing existing benchmarks and targeting subtle inconsistencies in high-fidelity AI-generated content.

2 Dataset & Benchmark

Recent advancements in multi-modal Large Language Models (LLMs) have enabled these models to process text, image, and video data, presenting new opportunities and challenges in Video Anomaly Detection (VAD) (Liu et al., 2024a). Motivated by this progress, we aim to benchmark the capabilities of these multi-modal models (LMMs) on VAD.

To address VAD, we propose **VANE-Bench: Video ANomaly Evaluation Benchmark for Conversational LMMs**, comprising 325 video clips and 559 challenging ground-truth question-answer (QA) pairs. We have adapted the VAD problem into a Multiple-Choice Video Question Answering (MC-Video QA) (Tapaswi et al., 2016; Lei et al., 2019; Yu et al., 2019) task to facilitate the evaluation of LMMs, allowing for a more granular assessment of their video content understanding.

We evaluate the latest closed-source and open-source LMMs on VANE-Bench. Sec. 2.1 provides an overview of VANE-Bench, Sec. 2.2 describes the dataset categories, and Sec. 2.3 outlines our data collection methodology.

2.1 Overview

VANE-Bench consists of 325 video clips spanning real-world and synthetic video anomalies. We adapted standard VAD surveillance datasets such as CUHK Avenue (Lu et al., 2013), UCF-Crime (Sultani et al., 2018), and UCSD Pedestrian (Li et al., 2014) to our MC-Video QA problem. Additionally, we included 197 video clips from various open-source and state-of-the-art closed-source text-to-video diffusion models (Brooks et al., 2024; HPCAI Tech, 2024; Runway Research, 2024; Wang et al., 2023a,b).

The diverse data backgrounds and varied difficulty levels in VANE-Bench make it ideal for evaluating the reasoning and understanding capabilities of video LMMs. Benchmarking these models on a range of real-world and synthetic anomalies helps us understand their strengths and limitations, guiding future multi-modal AI research.

Overall, VANE-Bench aims to push the boundaries of what LMMs can achieve in video anomaly detection, providing a rigorous standard for evalu-

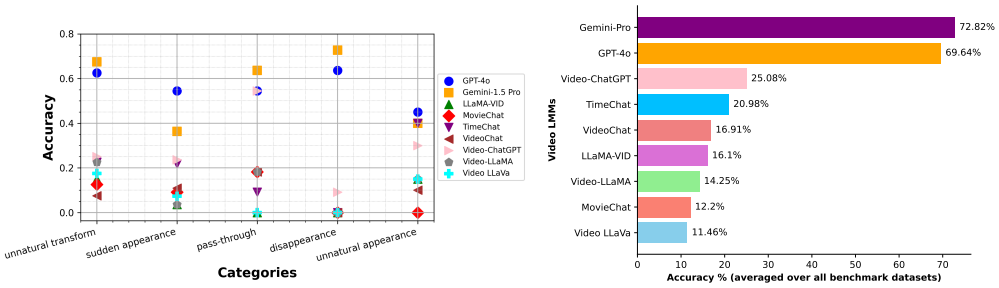

Figure 2: **Left:** Performance of Video-LMMs on five anomaly categories of SORA dataset. **Right:** Overall performance of Video-LMMs averaged across all the benchmark datasets including AI-generated and real-world anomaly datasets.

ating their performance on this challenging task.

2.2 Categories

The VANE-Bench dataset encompasses a variety of categories derived from both real-world surveillance footage and AI-generated video clips. Each category represents a distinct source and type of video anomaly. Below, we detail the different categories included in the dataset:

Real-World anomalies: The videos with these anomalies are sourced from several established real-world anomaly datasets, encompassing diverse anomaly types. The distribution of these anomalies is depicted in Fig. 3 (left). Fig. 3 (right) provides the total number of anomaly clips along with corresponding QA pairs for each dataset in this category. Detailed descriptions of each dataset within this category follow below.

1. **CUHK Avenue (Lu et al., 2013):** This category consists of 11 video clips with 33 associated question-answer (QA) pairs. The clips capture anomalous events in a campus environment, which shows individuals commuting in a university campus, and walking in and out of buildings. **Anomaly types.** The anomalies include unusual pedestrian behavior like randomly throwing bags and papers, or performing weird actions or dance moves.

2. **UCF-Crime (Sultani et al., 2018):** Comprising 95 video clips with 95 QA pairs, this category includes real-world surveillance footage. **Anomaly types.** The videos depict various criminal activities, such as arrest, assault, burglary, robbery, stealing, and vandalism.

3. **UCSD-Ped1 (Li et al., 2014):** This category contains 10 video clips with 30 QA pairs. The videos focus on pedestrian walkways. The Ped1 dataset is captured by a camera facing perpendicular to the road. **Anomaly types.** The anomalous events are due to the presence of non pedestrian entities (i.e. bikers, skaters, small carts, and wheelchairs) in the walkways.

4. **UCSD-Ped2 (Li et al., 2014):** Similar to UCSD-Ped1, this category includes 12 video clips with 36 QA pairs. In contrast with Ped1, the Ped2 dataset uses camera which is parallel to the road. **Anomaly types.** Abnormal events are due to non pedestrian entities in the walkways including bikers, skaters, small carts, and people walking across a walkway.

AI-Generated anomalies: The videos with these anomalies are obtained from various closed-source, and open-source text-to-video diffusion models. The anomalies in these clips are usually subtle, and hard to detect, which makes our VANE-Bench benchmark challenging. **General anomaly types:** The anomalies include sudden appearance of objects, unnatural transformation of solid physical objects, disappearance of objects, objects passing through other solids, and unnatural appearance of objects (i.e. distorted and deformed facial features, or other unnatural appearance like presence of extra fingers). The distribution of these anomalies in the dataset is shown in Fig. 3 (middle) and statistics about the number of clips and corresponding QA pairs is presented in Fig. 3 (right). Below we describe type of video samples in this category.

1. **SORA (Brooks et al., 2024):** This category consists of 46 video clips with 138 QA pairs. The video clips are generated using SORA, a state-of-the-art AI text-to-video model. Due to the high quality, and almost realistic looking videos generated by SORA, it becomes quite difficult to accurately identify the inconsistencies or anomalies present in the videos.

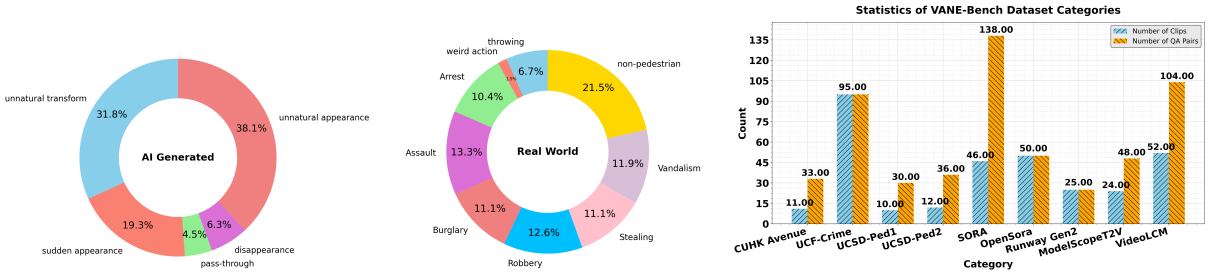

Figure 3: VANE-Bench dataset statistics: **Left and Middle:** Composition and type of anomalies present in AI-generated and real-world videos. **Right:** Number of samples and QA pairs present in each type of video dataset.

2. **OpenSora (HPCAI Tech, 2024):** With 50 video clips and 50 QA pairs, this category features AI-generated videos from the open-source version of SORA.

3. **Runway Gen2 (Runway Research, 2024):** This category includes 25 video clips with 25 QA pairs, created using a commercial text-to-video AI model.

4. **ModelScopeT2V (Wang et al., 2023a):** This category comprises of 24 video clips with 48 QA pairs, leveraging the video diffusion model trained by (Wang et al., 2023a) to produce videos from text captions. The videos were generated with 50 diffusion steps with 16 fps.

5. **VideoLCM (Wang et al., 2023b):** This category features 52 video clips with 104 QA pairs, generated using latent consistency models (Wang et al., 2023b) designed to create videos with high variability and with less latency. We used 20 diffuson steps to generate the videos with 16 fps. The videos were further post-processed by an LCM model trained on higher resolution videos to obtain better quality videos.

By including a wide range of video sources and anomaly types, the VANE-Bench dataset provides a comprehensive benchmark for evaluating the capabilities of large multi-modal models in video anomaly detection.

2.3 Constructing VANE-Bench

Fig. 4 describes the construction process of the VANE-Bench dataset. Since the synthetic AI generated videos from state-of-the-art models like SORA (Brooks et al., 2024) have subtle, and hard to detect inconsistencies, we require high quality captions describing all of the specific inconsistencies present in the given video. Our pipeline first annotates the anomalies using the frame annotation module (FAM). The caption generating module (CGM) then utilizes these annotations to produce captions, followed by the question-answer generation module (QAGM) creating QA pairs based on the annotated frames and captions. Annotating the clips before caption generation is crucial for focusing the model on the specific anomaly regions in the video (Shtedritski et al., 2023; Zhang et al., 2023b; Yang et al., 2023). Without annotations, the CGM often fails to reference the anomalies in the captions, as demonstrated in Sec. C of supplementary material. We briefly describe all the three stages involved in the semi-automatic dataset construction pipeline below.

2.3.1 Frame Annotation Module (FAM)

As described in Sec. 2.2, we first collect raw videos from existing VAD datasets like CUHK Avenue (Lu et al., 2013), UCF-Crime (Sultani et al., 2018), UCSD-Ped (Li et al., 2014), and also add additional challenging AI generated videos to the mix. For the VAD datasets, the bounding box annotations were already provided for a subset of the videos from these datasets. Thus, we only annotate the anomalies present in the AI-generated videos. In this stage, we first break down the raw videos into their constituent image frames. Second, we select and filter 10 consecutive frames from the video, which contains the inconsistency. We annotate these selected frames with a bounding box mentioning the type of the inconsistency. We consider the following inconsistency types: 'Sudden Appearance', 'Unnatural Transform', 'Disappearance', 'Pass-through, and 'Unnatural Appearance'. Fig. 4 shows the annotated 'Unnatural Transform' inconsistency affecting the kangaroo's legs and tails.

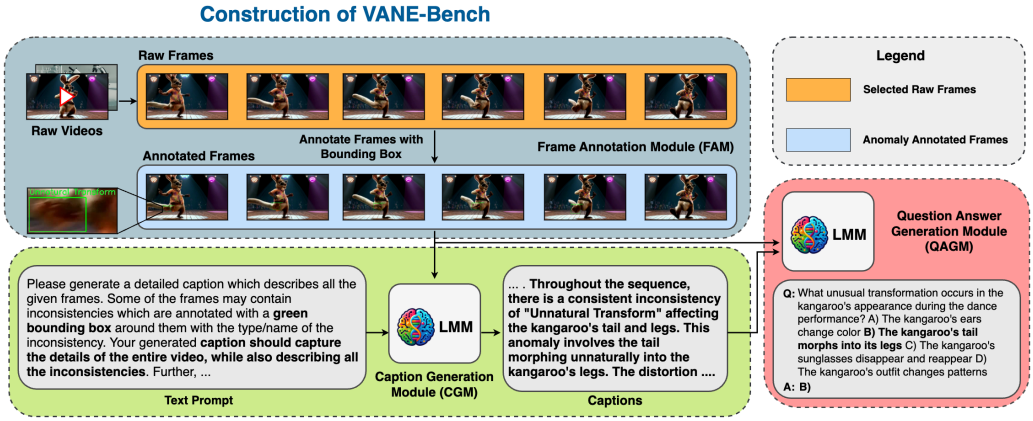

Figure 4: Flow diagram showing the semi-automatic construction process of our VANE-Bench dataset. The entire process can be divided into 3 interconnected stages/modules, i.e., i. Frame Annotation Module (FAM), ii. Caption Generation Module (CGM), iii. Question Answer Generation Module (QAGM).

2.3.2 Caption Generation Module (CGM)

The second stage of our data collection process involves the Caption Generation Module (CGM), which uses the annotated video frames from FAM, to generate a high quality and detailed caption which describes the inconsistency, along with the general events in the video. To generate the caption, we design a specialised custom prompt, and use the recently released GPT-4o (OpenAI, 2024) LMM, which has shown both impressive performance gains and cost savings. Thus, GPT-4o model takes in our custom prompt, along with the annotated frames to generate the descriptive video caption as shown in Fig. 4.

2.3.3 Question Answer Generation Module (QAGM)

The final stage of our VANE-Bench construction process involves using the generated caption from CGM, and the annotated frames from FAM to output the final high-quality, and challenging Question and Answer (QA) pairs. We create another custom prompt which we pass to the GPT-4o model, along with caption, and the annotated frames as input to generate the QA pairs. The selected raw frames containing the inconsistency, and their corresponding generated QA pairs form our VANE-Bench dataset.

3 Experiments and Results

Video-LMMs. We evaluate the anomaly detection and comprehension capabilities of both open-source and closed-source models. Among the open-source models, we evaluate 7 recent Video-LMMs, including Video-LLaVA (Lin et al., 2023), TimeChat (Ren et al., 2023), MovieChat (Song et al., 2023), LLaMA-ViD (Li et al., 2023c), VideoChat (Li et al., 2023b), Video-ChatGPT (Maaz et al., 2023), and Video-LLaMA-2 (Zhang et al., 2023a). For evaluating closed-source models, we use Gemini-1.5 Pro (Google, 2023) and GPT-4o (OpenAI, 2024).

Evaluation Protocol. For the evaluation of Gemini and GPT-4o, we utilize their respective official APIs, with each model receiving 10 video frames as input. The 10 frames are selected in a manner that encompasses all or the majority of the inconsistencies present in the video. In cases where an anomaly spans a longer duration, we sample multiple sets of 10 frames to ensure comprehensive coverage. As GPT-4o does not inherently support videos, we input the video clips as 10 frames to the GPT API, accompanied by the corresponding Visual Question-Answering (VQA) query. For each model under assessment, we generate responses to the questions independently and without retaining the conversation history. Few models such as Moviechat, output hallucinated responses when instructed to answer the query. In such cases, we consider the hallucinated responses as incorrect answer, due to the inability of the model to comprehend the situation in the video.

Evaluation metric. For the evaluation results of the Video-LMMs on our proposed VANE-Bench benchmark, we employ the standard VQA accuracy measure, which assigns a score of 1 to each correct answer and a score of 0 to each incorrect answer.

Benchmark Category	Video-LLaMA	VideoChat	Video-ChatGPT	Video-LLaVA	MovieChat	LLaMA-VID	TimeChat	Gemini-1.5 Pro	GPT4o
SORA	11.59	10.74	26.47	10.86	8.69	7.97	21.73	51.45	55.80
OpenSORA	18.00	28.00	22.00	18.00	10.00	14.00	26.00	84.00	68.00
Runway Gen2	16.00	4.00	12.00	16.00	1600	20.00	28.00	28.00	40.00
VideoLCM	10.57	17.64	18.26	19.23	14.42	19.23	22.11	49.04	50.96
Modelscope-T2V	10.41	20.83	16.66	16.66	6.25	14.58	20.83	75.00	64.58
Avenue	30.00	32.25	39.39	3.03	18.18	27.27	24.20	100.00	84.85
UCFCrime	9.47	11.57	31.57	10.52	18.51	15.78	7.30	76.84	83.16
UCSD-Ped1	16.66	13.33	40.00	2.77	6.66	6.66	27.58	96.67	93.33
UCSD-Ped2	5.55	13.88	19.44	6.06	11.11	19.44	11.11	94.44	86.11

Table 1: Evaluation results of Video-LMMs across different types of video samples on the VANE benchmark. We present results for both open-source and closed-source models. The first five rows show results on AI-generated videos and last four contain results on real world anomaly datasets.

3.1 Main Evaluation Results

3.1.1 Evaluation on Video-LLMs

AI-Generated anomalies. The AI-generated videos in our dataset are derived from five distinct models: SORA, OpenSORA, Runaway Gen-2, VideoLCM, and Modelscope-T2V. In the majority of these videos, the anomalies are subtle and not readily apparent, even to the human eye. As previously stated in section 2.2, the synthetic anomalies can manifest in five different forms. As shown in Table 1, the performance of open-source models in detecting anomalies in these videos is subpar. Although closed-source models outperform their open-source counterparts, their overall comprehension and detection of anomalies in the videos remain inadequate. This indicates that even robust closed-source models encounter difficulties in identifying subtle anomalies within the videos.

Real-world anomalies. Our real-world anomaly datasets benchmark, as discussed in section 2.2, comprises four real-world datasets and focuses on detecting crime-related irregularities, atypical pedestrian behavior, and unusual events. These anomalies are prevalent in real-world scenarios. In our analysis, we find that open-source models encounter difficulties in locating and identifying these anomalies. As shown in Table 1, these models perform poorly on these datasets. Conversely, we observe that closed-source models excel at detecting such real-world anomalies, indicating that they can effectively differentiate between the unusual events in the real-world scenarios. This can be attributed to the fact that these models are trained on a vast amount of existing real-world, internet-scale data.

We provide results on additional latest Video-LMMs in Section A.1 of Supplementary.

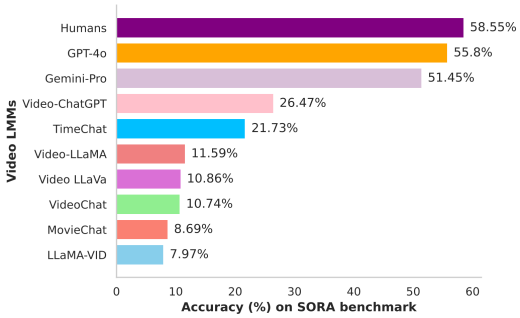

Figure 5: **Human vs Video-LMMs' performance on SORA.** Performance comparison of humans vs Video-LMMs on VQA task of detecting anomalies in SORA dataset. We find that closed-source Video-LMMs perform comparably to humans while open-source Video-LMMs struggle to detect the subtle anomalies.

3.1.2 Human Evaluation

We conducted a human evaluation on SORA-generated videos, which contain subtle and challenging anomalies that are difficult for humans to detect (see Fig. 1 top row) in a single viewing. Moreover, most of the video clips contain a multitude of foreground and background characters and elements, which makes it difficult for humans to focus on the inconsistencies within the short time frame. Some of the questions also specifically inquire about inconsistencies present in the background characters of the clips, rather than the foreground ones. To ensure fairness, our human evaluation was conducted under a set of rules, which include showing all 10 frames of the video to the human evaluator only once, followed by the question. Our human evaluation comparisons are presented in Fig 5. While humans outperform open-source models in detecting these subtle anomalies, their

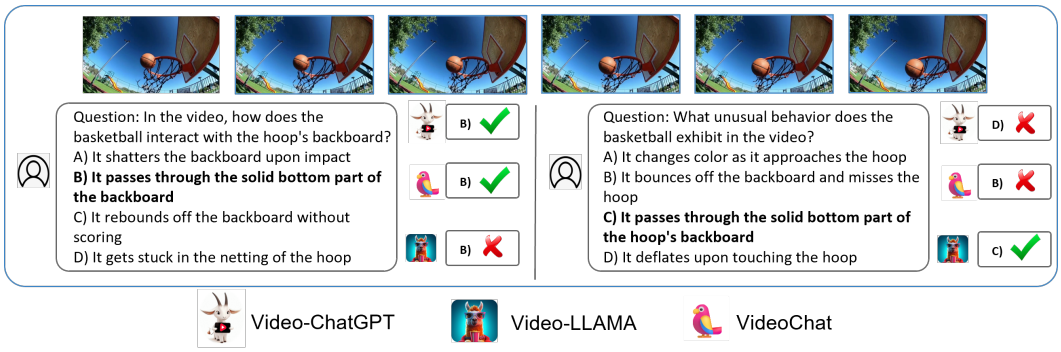

Figure 6: Inconsistency in Predictions: **Left:** Video-ChatGPT and VideoChat predict accurately, while Video-LLAMA selects incorrectly. **Right:** With a rephrased query, predictions shift. Video-ChatGPT and VideoChat err, whereas Video-LLAMA predicts correctly. This indicates the sensitivity of Video-LMMs towards query rephrasing.

performance remains sub-optimal. This indicates that, with the advancements in video generation techniques, there is a pressing need for more sophisticated and effective Video-LMMs capable of assisting in the detection of such challenging cases capable of evading human eyes as well.

3.2 Additional Analysis

Inconsistencies in Predictions. We find that, in the majority of cases, open-source Video-LMMs generate different results when prompted to answer the same query multiple times. Fig. 6 illustrates a sample example where the same questions were posed twice to the corresponding Video-LMMs, yielding different responses. In some instances, the answers generated by the Video-LMMs in both rounds were dissimilar and incorrect. However, we also found cases where Video-LMMs initially produced the correct answer, followed by an incorrect answer to the same query, albeit phrased slightly differently. This suggests that the majority of these open-source Video-LMMs struggle to comprehend the same query when presented in a different manner, leading to inconsistent and paradoxical predictions. In contrast, closed-source Video-LMMs are less prone to such inconsistent predictions and consistently produce the same output for the same queries, regardless of how they are phrased, indicating a superior comprehension of language. Refer to supplementary Section B for additional results.

Performance Analysis on SORA anomalies. The overall performance of open-source Video-LMMs on anomaly categories in synthetically generated SORA videos is subpar. To gain further insights, as depicted in Figure 2 (left), all open-source Video-LMMs exhibit less than 10% accuracy in detecting the “disappearance" anomaly, indicating that this particular type is the most difficult to identify for the majority of Video-LMMs. Among the open-source models, Videochat demonstrates above par performance compared to its open-source counterparts on most anomaly types, with the exception of the "unnatural appearance" category, where Timechat outperforms it. The remaining models display a fluctuating trend, with accuracy levels ranging from extremely low to moderately low across all anomaly types. The closed source-models, on the other hand, demonstrate superior performance compared to open-source models across all anomaly types.

We provide more insights and discussions in Section A.4 of Supplementary material.

4 Conclusion

We introduced VANE-Bench, a comprehensive benchmark for evaluating Video LMMs in VAD tasks, featuring real-world and AI-generated video clips. The AI-generated content, especially from advanced models like SORA, includes subtle inconsistencies, making VANE-Bench particularly challenging. Our evaluation of nine recent Video-LMMs on VANE-Bench shows significant gaps in detecting video anomalies, with even robust closed-source models struggling with nuanced discrepancies. Human assessments on SORA-generated videos confirm these subtle anomalies are challenging to identify, highlighting the need for advanced Video-LMMs. VANE-Bench is vital for advancing Video-LMMs in anomaly detection. As high-fidelity AI-generated content rises, our benchmark is crucial for developing models to identify subtle inconsistencies, aiding in the fight against misinformation and deepfakes. We hope VANE-Bench will guide future research to enhance the robustness and capability of Video-LMMs in this critical area.

5 Limitations

Our VANE-Bench is the first benchmark for evaluating Video-LMMs on anomalous videos from both AI-generated and real-world sources. While we have done our best to ensure a high quality evaluation of these Video-LMMs, certain limitations still manifest.

Our Question-answer pairs are designed to have 4 options. We design the instruct prompt to ensure that each Video-LMM outputs one out of 4 options. However, in some instances, model outputs hallucinated response and do not follow the instructions. As a result, we employ a post-response human-based filtration process, which involves an exhaustive verification and rectification of these errors. In our current setup, we mark these cases as wrong. We believe that future Video-LLMs will be more aligned with human intent and will follow the human instructions appropriately.

Additionally, the video samples from the SORA are limited in VANE-Bench. This is due to the fact that SORA model is not open-source yet, hence we rely on publicly available samples of SORA for evaluation.

References

Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. Ieee.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024a. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. 2024b. Goldfish: Vision-language understanding of arbitrarily long videos. *Preprint*, arXiv:2407.12679.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 2024. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. *Preprint*, arXiv:2405.00181.

Gemini Team et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Google. 2023. Gemini.

HPCAI Tech. 2024. Open-sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora.

Muhammad Uzair khattak, Muhammad Ferjad Naeem, Jameel Hassan, Naseer Muzzamal, Federcio Tombari, Fahad Shahbaz Khan, and Salman Khan. 2024. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms. *arXiv:2405.03690*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering. *Preprint*, arXiv:1809.01696.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2014. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023c. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. 2024a. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Comput. Surv.*, 56(7).

Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. 2024b. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *Preprint*, arXiv:2302.05087.

Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Bernard Marr. 2023. 15 amazing real-world applications of ai everyone should know about. https://bit.ly/4f2nrTd. Accessed: 26 Mar, 2024.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Preprint*, arXiv:2305.13786.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*.

Runway Research. 2024. Gen-2: The next step forward for generative ai. https://research.runwayml.com/gen2.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. *Preprint*, arXiv:1512.02902.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023a. Modelscope text-to-video technical report. *Preprint*, arXiv:2308.06571.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. 2023b. Videolcm: Video latent consistency model. *Preprint*, arXiv:2312.09109.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *Preprint*, arXiv:2306.13549.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Preprint*, arXiv:2401.13601.

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

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023b. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Appendix

In the following sections, we provide additional information for the paper: **VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs**. The contents are organized in the following order.

- Additional Findings and Results (Appendix A)
- Additional Results on Prediction Inconsistency (Appendix B)
- Importance of Frame Annotation Module (FAM) (Appendix C)
- Implementation Details (Appendix D)
- Distribution of VANE-Bench dataset (Appendix E)

A Additional Findings and Qualitative Results

A.1 Additional Quantitative Results

Video-LMMs are fewer in number compared to image-based multi-modal models, which limits the range of options available for evaluation. Given this scarcity, we selected 7 open-source and 2 closed-source LMMs that are currently among the most widely used. To ensure that our benchmark remains representative, we have also included additional results from other latest open-source LMMs: (Ataallah et al., 2024a,b; Li et al., 2024a), as shown in Table 2. Our findings reveal that open-source models still lag behind their closed-source counterparts in performance, indicating that simply adding more models wouldn't necessarily improve the overall representativeness of our benchmark. Our selected set, which includes both open-source and closed-source models, is already comprehensive, featuring state-of-the-art models like GPT-4o and Gemini-1.5 Pro. Given the limited number of Video-LMMs available, our reliance on this specific set of models is justified, as it accurately represents the current landscape of Video-LMM capabilities.

A.2 Qualitative Results

In Fig. 8, we showcase the response of both open-source and closed-source Video-LMMs on anomalous video samples from our VANE-benchmark. The query to the Video-LMMs contains the video and a question with multiple options associated with the specific anomaly present in the video. The anomalies in Fig. 8 constitute pass through (first row), unnatural appearance (second row), sudden appearance (third row), disappearance (fourth row) and unnatural transformation (fifth row).

Benchmark Category	LLaVA-NeXT	MiniGPT4-Video	Goldfish
SORA	11.59	10.74	26.47
OpenSORA	18.00	28.00	22.00
Runway Gen2	16.00	4.00	12.00
VideoLCM	10.57	17.64	18.26
Modelscope-T2V	10.41	20.83	16.66
UCFCrime	9.47	11.57	31.57
UCSD-Ped1	16.66	13.33	40.00
UCSD-Ped2	5.55	13.88	19.44

Table 2: Evaluation results of additional latest Video-LMMs across different types of video samples on the VANE benchmark. We present results for both open-source and closed-source models. The first five rows show results on AI-generated videos and last three contain results on real world anomaly datasets.

A.3 VANE-Bench frequent instances

Figure 7: **Frequent keywords:** Illustration of the most frequent keywords in the correct option set of VANE benchmark. These keywords signify the objects or human attributes in the videos that are most likely to exhibit anomalous behavior

Figure 7 presents a word cloud visualization, highlighting the most frequently occurring keywords within the correct option set of the VANE-Benchmark dataset. These prominent words are indicative of objects or human attributes in the videos that are most likely to exhibit anomalous behavior. From the figure, the most frequently occurring keyword is "Face" which indicated that the synthetically generated videos most likely struggle to generate a perfect human face.

A.4 Additional Discussions on Experimental Results

Per anomaly performance: To give further insights, Figure 2 (left) of the main paper illustrates the performance of LMMs on each type of anomaly present in the AI-generated videos. We can observe that closed-source models like GPT-4o and Gemini-1.5 Pro consistently exhibit strong performance across all five anomaly categories compared to their open-source counterparts. This likely stems from their access to significantly larger training datasets and model parameters, allowing for a more robust understanding of visual anomalies. Conversely, open-source models exhibit fluctuating performance depending on the anomaly type. We also note that open-source models struggle especially with the "disappearance" anomaly. We believe that it might be because of the fact that these models are trained on datasets focusing on the presence of objects and actions, and hence being more biased towards presence. Further, we believe that open-source models suffers from limited temporal reasoning capability and often use short-term mechanisms that limit their ability to track objects over time. The lack of datasets focusing on anomalies like "disappearance" also limits the model's capability to detect such patterns.

Higher performance of some LMMs: As seen in Table 1 and Figure 5 of main paper, we notice that some open-source LMMs perform better than their counterparts. For instance, we notice Video-ChatGPT achieves higher performance compared to other open-source models. We believe that it might be because of the following two reasons: 1. Training Data: While most open-source models rely solely on web-scraped video captioning data, Video-ChatGPT incorporates a manually-annotated video instruction dataset specifically designed for video understanding. This provides the model with a more direct and targeted learning experience, potentially enhancing its sensitivity to anomalies. 2. Two-Stage Training: Video-ChatGPT employs a two-stage training process, involving both video-language pre-training and instruction tuning. This enables the model to first develop a strong understanding of general video semantics and then refine its ability to follow user instructions and reason about specific events within videos.

B Additional results on Prediction inconsistency

As discussed in section 3.2 almost all Video-LMMs generate different results when prompted to answer the same query rephrased multiple times. While it is most common in open-source Video-LMMs, we found that closed-source Video-LMMs occasionally suffer from this problem as well. Fig. 9 illustrates additional sample examples where the same questions (phrased slightly differently) were posed twice to the corresponding Video-LMMs, yielding different responses.

We find that, in the majority of cases, open-source Video-LMMs generate different results when prompted to answer the same query multiple times. Fig. 6 illustrates a sample example where the same questions were posed twice to the corresponding Video-LMMs, yielding different responses. In some instances, the answers generated by the Video-LMMs in both rounds were dissimilar and incorrect. However, we also found cases where Video-LMMs initially produced the correct answer, followed by an incorrect answer to the same query, albeit phrased slightly differently. This suggests that the majority of these open-source Video-LMMs struggle to comprehend the same query when presented in a different manner, leading to inconsistent and paradoxical predictions. In contrast, closed-source Video-LMMs are less prone to such inconsistent predictions and consistently produce the same output for the same queries, regardless of how they are phrased, indicating a superior comprehension of language.

C Importance of Frame annotation module

Since the video inconsistencies present in state of the art AI models like SORA are quite subtle, and hard to detect, our Frame Annotation Module (FAM) ensures that we are able to generate high-quality and accurate captions for these videos. As shown in Fig. 10, without FAM, the generated caption is not able to describe the sudden appearance of the kangaroo's right foot near its tail. Further, the caption generated without our FAM is also not able to describe the extra set of paws which appear suddenly from the legs of the cat. Thus, FAM plays an important role in curating high-quality and accurate video captions.

D Implementation Details

We use the official code of each open-source Video-LMM for evaluation. Each of these codes are implemented in pytorch framework. We evaluate each one of them on a single NVIDIA A100 40GB GPU. For closed source Video-LMMs, we use their respective API for evaluation. We use GPT-4o (OpenAI, 2024) as our LMM to generate the captions and the final QA pairs in VANE-Bench. Next, we describe the prompts used in our Caption Generation Module (CGM), Question Answer Generation Module (QAGM), and in evaluating various Video-LMMs on VANE-Bench in the subsequent subsections.

D.1 Caption Generation Module (CGM)

System Prompt: You are a helpful and intelligent AI assistant which can generate informative captions for a given input of 10 consecutive images/frames from a video. The video is generated from an AI text-to-video diffusion model and has some obvious inconsistencies or anomalies in the form of various deformations, unrealistic physical transformations, unnatural appearance of objects, human faces, body parts, or sudden appearance, disappearance, or merging of objects. Your task is to generate a descriptive caption for the given input video, highlighting the inconsistencies or anomalies present in the video.

Text Prompt: Please generate a detailed caption which describes all the given frames. Some of the frames may contain inconsistencies which are annotated with a green bounding box around them with the type/name of the inconsistency. Your generated caption should capture the details of the entire video, while also describing all the inconsistencies. Thus, properly look at all the given frames and the region marked by the green bounding boxes when describing the inconsistencies. Further, make sure to mention specific details about each of the inconsistencies, and mention the exact names of the inconsistencies from the marked green bounding box. Also, while describing the inconsistency please be as specific and detailed as possible, don't be vague or general about the inconsistency. The reader of the caption should perfectly understand what inconsistencies/anomalies are in the video and what the video is about. Do not mention the green bounding box in your response; it is only for you to identify the inconsistencies. Make sure to describe all the inconsistencies in your caption. Do not analyze the impact of the inconsistencies; you should only describe them. There is no need to mention when the inconsistencies start or end, just describe them.

D.2 Question Answer Generation Module (QAGM)

System Prompt: You are a helpful and intelligent AI assistant which can curate high-quality and challenging question and their corresponding answers, which are used to test the video understanding capabilities of an multi-modal LLM model capable of taking videos as their inputs.

Text Prompt: You are given a video input, which is generated by a state of the art AI algorithm. Thus, these videos look very natural and almost realistic, but they are actually synthetic and generated by an AI algorithm. The videos may have some inconsistencies or anomalies present in them, which are generally localized to only a specific location in the video as identified by the green bounding boxes in the video. The rest of the video appears completely natural or realistic. These specific inconsistency may last for only a few frames of the video, or may last for the entire video itself. The inconsistency or anomalies in the video are generally events and phenomenon which is not observed in real world and physical scenarios. You will also be given a caption as input which describes the video, along with the specific inconsistency present in the video. Based on the given video and caption input, your task is to formulate 3 diverse and misleading questions to test whether the multi-modal LLM model can correctly identify the options based on the inconsistencies present in the video or not. So, your generated questions should give the model few options to choose from to make its answer, and these options should be of high quality and also have misleading choices so that you can test deeper level of understanding of these multi-modal LLM models. Thus, the goal of these questions is to accurately assess the multi-modal LLM's ability to accurately identify the inconsistencies present in the video. Generate questions that comprise both interrogative and declarative sentences, utilizing different language styles, and provide an explanation for each. Your response should be presented as a list of dictionary strings with keys 'Q' for questions and 'A' for the answer. Follow these rules while generating question and answers:

1. Do not provide answers in the question itself. For example, the ground-truth attribute or compo-

nent which makes the video scene unusual should never be mentioned in the question itself.

2. Ensure the questions are concrete and specific, and not vague or ambiguous.

3. The questions should be formed based on your deep understanding of the video and the caption. Thus, properly read the caption and look at the given video to generate the questions.

4. The questions should only pertain to the inconsistencies present in the video, and not about the video in general.

5. You may also ask the model some misleading questions talking about non-existent inconsistencies in the video, to test the model's ability to differentiate between real and fake inconsistencies.

6. Do not ask vague questions, and the answer should only contain one of the correct option mentioned in the question.

7. In your question itself you must provide multiple choice options for the answer, and the answer should be one of the options provided in the question. Please ensure you provide option choices and their corresponding letters in the question itself.

8. In your answer, only mention the correct option letter from the question. Make sure that the correct option letter is not always the same, and randomly shuffle the correct option letter for each question.

9. You must only follow the below output format, and strictly must not output any other extra information or text. Your output format should be strictly as follows, without any additional information or text:

["Q": 'first question A) <option1> B) <option2> C) <option3> D) <option4>', "A": 'Pick the correct option letter from A) B) C) D)', "Q": 'second question A) <option1> B) <option2> C) <option3> D) <option4>', "A": 'Pick the correct option letter from A) B) C) D)', ... }]

Given below is the caption input which describes the given video along with the specific inconsistency present in the video. The caption is: {caption}

D.3 Evaluating Video-LMMs

System Prompt: You are a helpful and intelligent multi-modal AI assistant, capable of performing visual question answering (VQA) task. You will be given as input 10 consecutive frames from a video, and a corresponding question related to the video, you have to answer the given question after analyzing and understanding the given input video. The question itself will present you with 4 lettered options like A) B) C) D), your task is to only output single letter corresponding to the correct answer (i.e. string literal 'A', 'B', 'C', or 'D'), and you should not output anything else.

Text Prompt: {question}

E Distribution of VANE-Bench dataset

How to view the dataset? The dataset alongside metadata will be hosted on the Hugging Face platform for download post acceptance of the paper. Users can directly load the dataset using Hugging Face Datasets library or download the zip file in the same Hugging Face repository. All instructions and code files to reproduce the experiments of the paper will be provided in a github repository.

How will the dataset be distributed? The dataset will be distributed to the public using the Hugging Face Dataset Hub. We have publicly released the codebase alongside instructions to reproduce and evaluate models on GitHub.

Dataset License. This work and dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The videos in VANE-Bench dataset are collected from publicly available sources and existing real-world datasets and are for academic research use only. The video generative models used to synthesize data samples in our VANE-Bench benchmark are open to use publicly and do not pose any privacy concerns as the persons or objects present in the generated videos are synthetic and do not exist in the real world. The real-world surveillance datasets - UCFCrime (Sultani et al., 2018), UCSD Pedestrian (Li et al., 2014), Aveneue (Lu et al., 2013); on the other hand, used in our work are all existing well-known and publicly available datasets that are released under open-source licenses. Thus, the original creators of these datasets have collected the data after taking informed consent from the stakeholders. By using VANE-Bench, you agree not to use the dataset for any harm or unfair discrimination. Please note that the data in this dataset may be subject to other agreements. Video copyrights belong to the original dataset providers, video creators, or platforms.

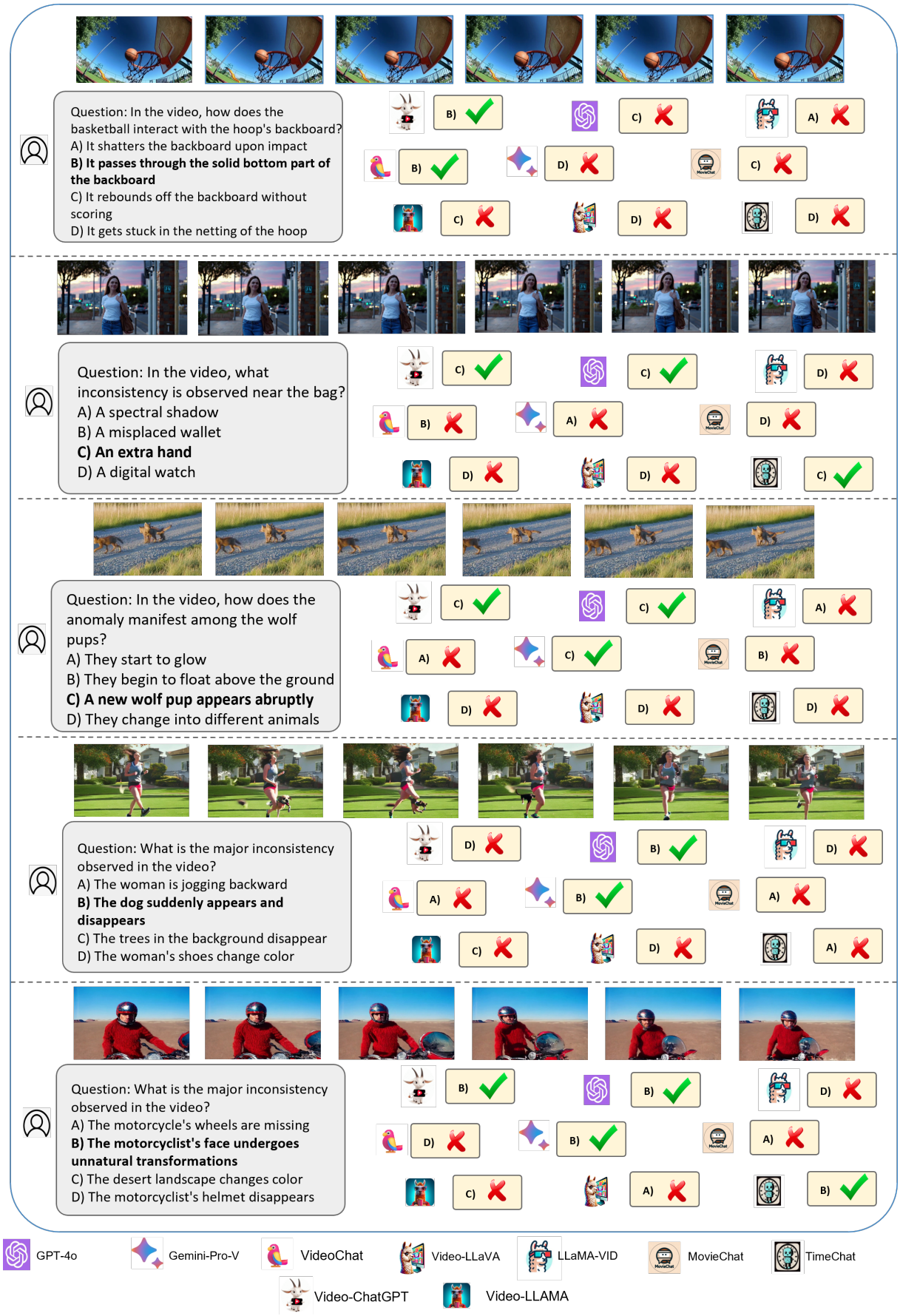

Figure 8: **Qualitative examples:** Figure shows the response of Video-LMMs to the VQA task of detecting anomalies in the video. The correct answer is written in bold in the user query. We find that majority of Video-LMMs struggle to answer the questions correctly.

Figure 9: **Prediction Inconsistency:** Figure shows the response of Video-LMMs to the VQA task of detecting anomalies in the video. The correct answer is written in bold in the user query. We find that majority of Video-LMMs struggle to answer the questions correctly.

Figure 10: Example showcasing the importance of our Frame Annotation Module (FAM). We note that without FAM, the LMM responsible for generating the captions is not able to identify or describe the accurate anomaly present in the video. However, by providing the bounding box annotation for the inconsistency, we are able to ensure that the generated caption accurately describes the anomaly in the video.