How Good is my Video LMM? Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs

Muhammad Uzair Khattak[1] **Muhammad Ferjad Naeem**[2] **Jameel Hassan**[1]
Muzammal Naseer[1] **Federico Tombari**[3,4] **Fahad Shahbaz Khan**[1,5] **Salman Khan**[1,6]
[1]Mohamed Bin Zayed University of AI [2]ETH Zurich [3]Google
[4]TU Munich [5]Linköping University [6]Australian National University

Abstract

Recent advancements in Large Language Models (LLMs) have led to the development of Video Large Multi-modal Models (Video-LMMs) that can handle a wide range of video understanding tasks. These models have the potential to be deployed in real-world applications such as robotics, AI assistants, medical surgery, and autonomous vehicles. The widespread adoption of Video-LMMs in our daily lives underscores the importance of ensuring and evaluating their robust performance in mirroring human-like reasoning and interaction capabilities in complex, real-world contexts. However, existing benchmarks for Video-LMMs primarily focus on general video comprehension abilities and neglect assessing their reasoning capabilities over complex videos in the real-world context, and robustness of these models through the lens of user prompts as text queries. In this paper, we present the Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES), a novel benchmark that comprehensively assesses the performance of Video-LMMs across 11 diverse real-world video dimensions. We evaluate 11 recent models, including both open-source and closed-source variants, and find that most of the Video-LMMs, especially open-source ones, struggle with robustness and reasoning when dealing with complex videos. Based on our analysis, we develop a training-free Dual-Step Contextual Prompting (DSCP) technique to effectively enhance the performance of existing Video-LMMs on CVRR-ES benchmark. Our findings provide valuable insights for building the next generation of human-centric AI systems with advanced robustness and reasoning capabilities. Our dataset and code are publicly available at: mbzuai-oryx.github.io/CVRR-Evaluation-Suite/.

1 Introduction

Recently, Large Language Models (LLMs) [30, 38, 12] have demonstrated impressive reasoning and planning capabilities while simultaneously handling a wide range of NLP tasks [33, 2]. Consequently, their integration with the vision modality, specifically for video understanding tasks, has given rise to Video Large Multi-modal Models (Video-LMMs) [15]. These models act as visual chatbots that accept both text and video as input and handle a diverse set of tasks, including video comprehension [21], detailed video understanding [18], and action grounding [37]. As these models directly capture video data, they hold substantial potential for deployment in real-world applications such as robotics, surveillance, medical surgery, and autonomous vehicles.

However, as these models assume an expanding role in our everyday lives, assessing their performance in comprehending complex videos and demonstrating reliable reasoning and robustness capabilities across diverse real-world contexts becomes essential. Video-LMMs with such capabilities will be

Submitted to the NeurIPS 2024 Track on Datasets and Benchmarks. Do not distribute.

Benchmark	Textual Robustness	Complex Reasoning	In the wild (OOD)	Contextual Dependency	Multiple Actions	Temporal Order & Fine-grained
MSVD-QA [35]	✗	✗	✗	✗	✗	✗
MSRVTT-QA [35]	✗	✗	✗	✗	✗	✗
TGIF-QA [11]	✗	✓	✗	✗	✓	✓
Activity Net-QA [36]	✗	✗	✗	✗	✗	✓
VideoChat-GPT [21]	✗	✗	✗	✓	✗	✓
MVBench [16]	✗	✓	✗	✗	✓	✓
SEED-Bench [14]	✗	✗	✗	✗	✓	✓
CVRR-ES (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of CVRR-ES with existing benchmarks for video question answering. The CVRR-ES benchmark represents an initial effort to assess Video-LMMs in the context of their applicability and suitability in real-world contexts.

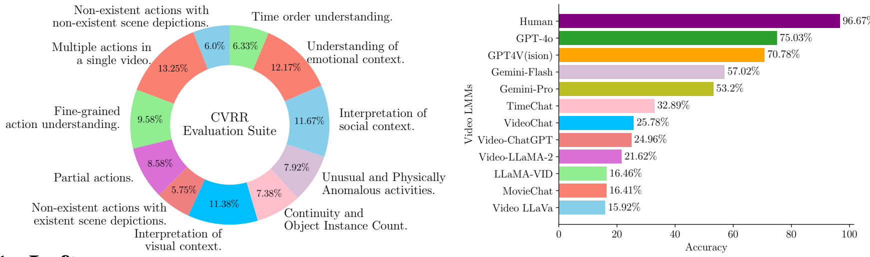

Figure 1: **Left:** CVRR-ES comprises of 11 diverse complex video evaluation dimensions encompassing a variety of complex, real-world contexts. **Right:** Overall performance of Video-LMMs on the CVRR-ES benchmark. Results for each Video-LMM are averaged across 11 video dimensions.

more effective when integrated into our daily lives for solving perception tasks and will be a promising step towards building trustworthy human-centric AI-assistive systems.

Several attempts in literature have been made to benchmark Video-LMMs. SEED-Bench [14] curated a MCQ-based dataset including 3 evaluation dimensions for videos. Similarly, MV-Bench [16] constructed the Video-LMM benchmark and assembled 20 video tasks for evaluating the spatial and temporal understanding of these models. While these methods aim at benchmarking Video-LMMs, they predominantly evaluate video and/or temporal comprehension abilities and overlook the complex reasoning aspects of Video-LMMs for real-world context, and their robustness towards user input text queries; both of which are crucial to ensure their responsible engagement with humans in various real-world situations in the wild. While some studies have explored similar areas such as hallucinations in image-based LLMs [19, 24], no such comprehensive study exists for the case of Video-LMMs.

Motivated by the wide-scale applications of Video-LMMs and the lack of world-centric complex video benchmarking efforts, we present a new benchmark, Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES), to comprehensively assess the performance of Video-LMMs. As shown in Tab. 1, CVRR-ES evaluates Video-LMMs on key aspects of robustness and reasoning in videos, encompassing video domains that more accurately test models in real-world scenarios such as videos having contextual dependency and in-the-wild aspects. CVRR-ES is an open-ended video QA benchmark comprising 11 real-world video category dimensions (Fig. 1, left) that encompass diverse evaluation aspects. These dimensions span from context-dependent (e.g., social, emotional, etc.) categories to ones that often take place in the wild such as videos containing physically anomalous activities. We comprehensively evaluate a representative set of 11 recent Video-LMMs (Fig. 1, right) including both open-source and closed-source models on the CVRR-ES benchmark using a LLM-assisted automatic evaluation framework [21, 4].

The performance of Video-LMMs on the CVRR-ES benchmark reveals that these models struggle to correctly comprehend complex videos indicating their weak reasoning and lack of robustness to the textual user queries (Fig. 2). For instance, state-of-the-art Video-LLaVA [18] achieves only 15.92% performance averaged across 11 video dimensions of CVRR-ES. In contrast, closed-source models including GPT4V(vision) [23] and Gemini-Vision-Pro [9] exhibit relatively stronger performance but still lag behind the performance of humans. Using CVRR-ES benchmark, we extensively perform quantitative and qualitative analysis and formulate important insights about these Video-LMMs based on their failure cases and individual performances across the diverse video dimensions.

Based on our analysis, we note that standard prompting struggles in steering Video-LMMs' focus for complex video understanding. Additionally, their limitations in reasoning and robust video understanding of real-world scenarios are dominantly driven by the quality of textual inputs (i.e., user questions). Based on these insights, we develop a training-free Dual-Step Contextual Prompting (DSCP) technique, which effectively steers the model's behavior during inference to elicit video-specific reasoning and improved robustness within Video-LMMs. With DSCP, Video-LMMs substantially improve on our benchmark, suggesting the potential of prompting methods for Video-LMMs.

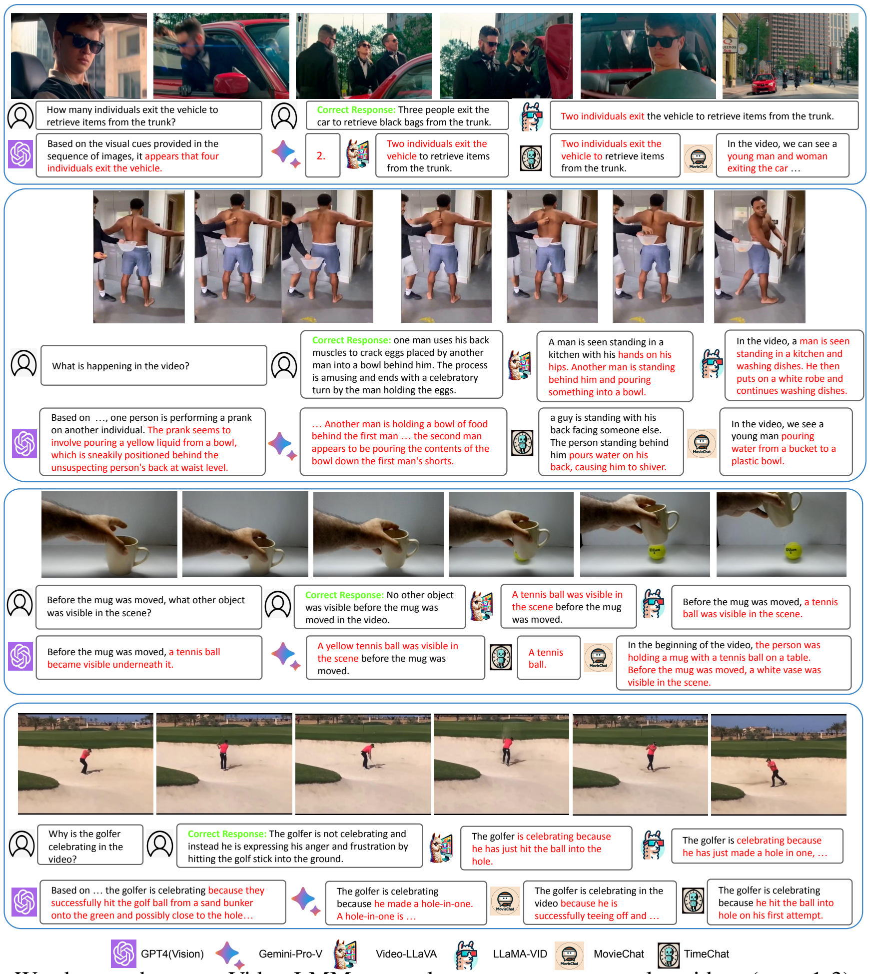

Figure 2: We observe that most Video-LMMs struggle to reason over complex videos (rows 1-3) and exhibit weak robustness and rectification abilities when answering user questions that can sometimes be confusing (row 4). The QA pairs in Comprehensive Video Reasoning and Robustness Evaluation Suite (CVRR-ES) benchmark assess the performance of Video-LMMs beyond general video comprehension. (best viewed zoomed in)

Our main contributions are as follows: (1) We present Complex Video Robustness and Reasoning Evaluation suite (CVRR-ES), a Video Question Answering benchmark designed to assess the reasoning and robustness capabilities of Video-LMMs on 11 diverse world-centric complex video dimensions (§3). (2) We extensively evaluate both open-source and closed-source Video-LMMs on the CVRR-ES benchmark and find that most models exhibit weak performance, highlighting their limited reasoning in complex videos and lack of robustness towards user text queries (§5.1). (3) We conduct comprehensive analysis and formulate important conclusions about Video-LMMs based on their failure cases and performance on the CVRR-ES benchmark. Our findings provide key insights for building the next generation of human-centric AI systems with improved robustness and reasoning capabilities (§5.4). (4) To improve Video-LMMs' reasoning and robustness abilities, we design a model-agnostic and training-free prompting method that effectively enhances their performance (§4).

2 Related Works

Video Large Multi-modal models (Video-LMMs). Video-LMMs [18, 17, 37] are visual chatbots capable of performing a wide range of video tasks, including video comprehension and captioning, video question-answering, and action grounding. These models accept both video and textual inputs and generate textual responses. From an architectural perspective, Video-LMMs combine pre-trained vision backbones [25, 6, 32] with large language models [30, 38] using connector modules such as MLP adapters, Q-former [5], and gated attention [1]. VideoChat [15] and VideoChat-GPT [17] presented initial open-source efforts in this direction and were trained with two stages of alignment and video-instruction following objectives. Recently, more advanced Video-LMMs have emerged in the field, with some models focusing on improving model architectures [17], expanding to new tasks

[22], and enabling support for long videos [28, 26]. In this work, we aim to develop a comprehensive benchmarking framework to assess the reasoning and robustness capabilities of these Video-LMMs and develop a training-free prompting technique to improve their performance on these fronts.

Benchmarking Video-LMMs. With the growing number of Video-LMMs emerging in the research community, several works have presented evaluation frameworks to assess and quantify these models for benchmarking and analysis purposes. SEED-Bench [14] evaluates the visual capabilities in both image and Video-LMMs across 12 unique dimensions. MV-Bench [16] curates 20 video tasks to evaluate the spatial and temporal understanding of Video-LMMs. Video-ChatGPT [21] develops a quantitative evaluation framework to assess model understanding on five aspects of general video comprehension, such as the correctness and consistency of model captions. While these evaluation frameworks provide effective insights, their assessments do not extend beyond general video-comprehension metrics to more advanced aspects of reasoning and robustness, particularly for real-world context cases. In contrast, our work focuses on providing a complex video reasoning and robustness benchmark and offers a thorough assessment of Video-LMMs in practical applications.

Training-free Prompting Techniques. Steering model behavior at inference time using prompting has become a common paradigm in the NLP domain. Prompting [34, 31] refers to the set of instructions given as a prefix to the language model to better align model responses with human intent without the need for task-specific fine-tuning. Prompting techniques can be as simple as a single sentence (e.g., "Let's think step by step") such as zero-shot chain of thought [34] prompting, to more detailed techniques such as combining chain-of-thought prompting with few-shot learning [2] and self-consistency chain of thought prompting [31]. Surprisingly, training-free prompting techniques for Video Large Multi-modal Models (Video-LMMs) have been minimally explored. In this work, we develop a dual-step prompting technique based on principled prompt instructions specifically designed to steer the model's behavior for improved reasoning and robustness over complex videos.

3 Complex Video Reasoning and Robustness Evaluation Suite

As Video-LMMs are touching new real-world applications, it is essential to ensure that they robustly handle the user inputs, comprehend the visual world, and exhibit human-like reasoning capabilities. In this work, our goal is to establish a comprehensive benchmark, Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES) to assess the *robustness* and *reasoning* capabilities of Video-LMMs over complex and contextual videos. We first provide an overview of CVRR-ES and then detail the video evaluation dimensions in Sec. 3.1. Subsequently, we discuss benchmark creation process in Sec. 3.2. We provide details on the human performance on CVRR-ES in Appendix C.

Overview. CVRR-ES encompasses evaluation dimensions that cover diverse video categories related to real-world scenarios, ranging from context-dependent (e.g., social, emotonal) categories to video types that often take place in the wild (e.g., anomalous activities). Specifically, we have compiled 11 video evaluation dimensions and curated 2,400 high-quality open-ended question-answer (QA) pairs, spanning 214 high-quality videos. The average video duration is 22.3 seconds, with maximum and minimum durations of 183 and 2 seconds, respectively. Fig. 2 shows some qualitative examples of collected videos for the CVRR-ES benchmark. Refer to Appendix C for additional statistical details.

3.1 CVRR-ES Video Category definitions.

For curating the CVRR-ES benchmark, we carefully select 11 diverse benchmark evaluation categories. As shown in Fig. 1 (left), these categories encompass a wide range of real-world complex and contextual video types. Below, we define each video evaluation dimension in detail.

1) Multiple actions in a single video. This category involves videos with 2-4 different human activities. We curate questions in this category to assess the model's ability to understand and reason about multiple actions and their interrelations in a single video.

2) Fine-grained action understanding. We collect videos that encompass fine-grained activities performed by humans, such as pushing, opening, closing, spreading, sitting, etc. This category tests the model's ability to interpret subtle and fine-grained actions through carefully crafted questions.

3) Partial actions. We observe that Video-LMMs produce content that is relevant to a video's context and likely to occur next. We collect videos with actions likely to be followed by other actions but not shown in the video e.g., cracking an egg in a kitchen suggests the next action of cooking the egg.

4) Time order understanding. Accurately recognizing the temporal sequence of activities in videos

is crucial for distinguishing between atomic actions, such as pushing and pulling. We collect videos of fine-grained actions occurring in a particular temporal direction and curate challenging questions.
5) Non-existent actions with existent scene depictions. This category examines the model's robustness and reasoning behavior in scenarios where we introduce non-existent activities into the video without altering the physical and spatial scenes or environmental details in it.
6) Non-existent actions with non-existent scene depictions. In this category, we increase the difficulty of the QA task by including questions containing both non-existent activities and scenes. We alter the details of objects, attributes, and background for non-existent scene comprehension. This tests the model's ability to correct misleading questions and avoid generating imaginary content.
7) Continuity and object instance count. This category contains videos (real-world and simulations) designed to test the models' ability to accurately recognize the number of instances of objects, people, etc., and distinguish between existing objects and new ones introduced later in the same video scene.
8) Unusual and physically anomalous activities. We collect videos depicting unusual actions that seemingly defy the laws of physics, such as a person floating in the air or driving a motorbike on a running river. Assessing Video-LMMs in such scenarios is crucial, as it allows us to determine whether they can generalize to understand actions in out-of-distribution videos in practical situations.
9) Interpretation of social context. We test Video-LMMs' ability to understand actions influenced by social contexts, such as helping an elderly person cross the road. Video-LMMs are assessed to determine their ability to accurately infer the rationale behind actions using the social context.
10) Understanding of emotional context. Similar to social context, humans can accurately understand and interpret each other's actions by considering the emotional context. We test Video-LMMs' ability to understand actions based on emotional context, e.g., a person crying due to joy.
11) Interpretation of visual context. This category tests the model's ability to understand actions by leveraging the overall visual contextual cues in the video. For example, to identify the number of people present based on the presence of shadows, one must utilize the visual context of shadows.

3.2 Building CVRR-ES Benchmark

Stage 1: Data collection and Annotation. We first collect high-quality videos and annotate each video via human assistance. To ensure that each evaluation dimension captures relevant attributes and information, we meticulously select videos that are representative of specific characteristics associated with that dimension. Overall, 214 unique videos are selected covering 11 dimensions with around 20 videos per evaluation dimension. Around 60% of these videos are collected from public academic datasets. To introduce diversity in the benchmark distribution, we select videos from multiple datasets including Something-Something-v2 [10], CATER [8], Charades [27], ActivityNet [3], HMDB51 [13], YFCC100M [29]. The remaining 40% of videos are collected from the internet. Following the video collection process, two experienced human annotators are assigned to generate captions for each video. For videos where initial captions or metadata are available from academic datasets, the captions are generated by the annotators based on them. For videos collected from the internet, captions are entirely generated by human annotators. To ensure consistency and high quality, we provide annotation instructions to annotators, who generate captions accordingly. Personalized annotation guidelines are used for each video category. Refer to additional details in Appendix C.
Stage 2: Question-Answer Generation. The first challenge is to select an evaluation setting to assess Video-LMMs. Humans typically engage in free-form conversation to interact with each other in day-to-day life. Inspired by this, we aim to simulate a similar style of interaction with Video-LMMs by curating open-ended QA pairs to evaluate these models for robustness and reasoning. We feed detailed ground-truth video captions to GPT-3.5 LLM, which is utilized to generate open-ended questions. The QA pairs covers both the reasoning and robustness aspects as detailed below.
Reasoning QA pairs: With Video-LMMs beginning to interact more directly with humans in our lives, it's crucial to validate the reasoning abilities of Video-LMMs for more reliable Human-AI interaction. When evaluating the reasoning capabilities of Video-LMMs, we aim to determine whether these models can understand the input video not only by analyzing spatial content but also by grasping the underlying rationale behind the occurring activities and their relationships with the surrounding context. This involves creating questions that go beyond simple video comprehension and scene description and require the model to engage in complex logical inference, contextual understanding, and reasoning about counterfactual and hypothetical scenarios.

Robustness QA pairs: In addition to evaluating the reasoning capabilities of LLMs, it is important to assess Video-LMMs to ensure their robust and responsible performance in real-world scenarios. In the context of Video-LMMs, robustness can be evaluated from both visual (video input) and textual interfaces. Our focus in this work lies on textual interface robustness by particularly testing the model's comprehension abilities when posed with misleading or confusing questions. This scenario mirrors realistic situations where users, based on their expertise levels, may pose irrelevant, misleading, or confusing questions. It is crucial for models to demonstrate reliability and robustness in handling such queries and avoid generating unreal or hallucinated content for input videos.

We curate specific prompts for each evaluation dimension to instruct LLM in generating QA pairs. Example prompts used as an instruction to LLMs for curating QA pairs for robustness and reasoning aspects are provided in Fig. 14 in the Appendix E.

Stage 3: QA Pairs Filtration. After generating the QA pairs, we employ a manual filtration step, with human assistance to verify each generated QA pair. Approximately 30% of the QA pairs generated by GPT-3.5 are found to be noisy, containing questions that are unrelated to the video evaluation dimensions or unanswerable based on the provided ground-truth captions. Additionally, many questions contain answers within the question itself. Therefore, an exhaustive filtering process is conducted which involves QA rectification and removing those samples which are not relevant to the video or evaluation type. This process results in a final set of 2400 high-quality QA pairs for the CVRR-ES benchmark. Examples of the final QA pairs are shown in Tab. 4 in the Appendix.

Stage 4: Evaluation Procedure. Previous methods in the literature [21, 4, 19, 24] have explored using LLM models as judges for quantifying results in open-ended QA benchmarks. We adopt a similar approach and instruct LLMs to act as teachers to assess the correctness of predicted responses from Video-LMMs compared to ground-truths. We generate open-ended predictions from Video-LMMs by providing video-question pairs as inputs and then present the model predictions and their ground-truth responses to the LLM Judge using the evaluation prompt. The Judge determines whether the prediction is correct or incorrect with a binary judgment, assigns a score from 1 to 5 representing the quality of the prediction, and provides a reasoning to explain its decision. Our ablative analysis in the Appendix. E demonstrates that reasoning-constrained LLM-based evaluation aligns the most with human-based judgment. Our evaluation prompt for LLM Judge is shown in Fig. 13 in Appendix E.

Quality of QA pairs. We show examples of QA pairs from CVRR-ES benchmark in Table 4 in Appendix C. Our QA pairs are of high quality and aim to test the understanding of Video-LMMs against reasoning and robustness criteria on multiple evaluation dimensions. To quantitatively assess the quality of the benchmark, we establish a quality assessment procedure [7]. We randomly sample 1120 QA pairs, which encompass all videos of the CVRR-ES benchmark, and request human experts to evaluate the quality of each QA pair by answering the following questions: **(1)** "*Does the QA pair correctly represent the evaluation dimension category under which it falls?*" (possible answers: "Yes", "No") **(2)** *Can the question be correctly answered given only the video content?* (possible answers: "Agree", "Disagree") and **(3)** *Is the corresponding paired ground-truth answer correct? (which will be used during evaluation as ground truth)* (possible answers: "Yes", "No"). On average, the answer of experts for the first question was "**Yes**" for 98.84% of the times. For the second and third questions, the averaged answer was "**Agree**" and "**Yes**" for 100% and 99.91% of the times, respectively.

4 Dual-Step Contextual Prompting for Video-LMMs.

Given their wide-scale potential in practical applications, new Video-LMMs are frequently introduced by the research community. Despite the availability of numerous Video-LMMs, the majority of them are trained using only positive examples and video-conversational templates that are primarily limited to tasks such as video-captioning and video question answering [15, 21, 26, 28]. This leads to highly over-affirmative behavior and a lack of self-rectification abilities in these models (Sec. 5.4).

Additionally, the templates have minimal focus on enhancing reasoning and robustness capabilities through reasoning instruction-tuning pairs, resulting in their weak performance against robustness and reasoning based evaluations in CVRR-ES. Consequently, enabling direct interaction of Video-LMMs with users in real-world scenarios can result in undesired responses when the user question is confusing and deceiving. Moreover, curating reasoning-based instruction fine-tuning datasets requires meticulous data curation steps, and retraining these models are computationally expensive [17, 26].

Alternatively, training-free prompting techniques in NLP literature have shown effectiveness in eliciting reasoning abilities in LLMs such as chain of thought and self-consistency prompting [34, 31]. Inspired by these, we present a Dual Step Contextual Prompting (DSCP) technique, which steers Video-LMM focus for enhanced reasoning while simultaneously encouraging the models to provide robust and grounded answers. DSCP is a two-step prompting method that **1)** ensures that the model comprehends the video while reasoning over crucial aspects of complex video understanding such as contextual information and decoding the complex relationships between objects and motions, etc., and **2)** encourages robustness by generating the response against the question while conditioning both on video and the unbiased context retrieved in the first step. Below we discuss each step of DSCP in detail.

Step 1: Video reasoning. We prompt Video-LMMs to interpret video from a reasoning perspective using ten principled instructions (Fig. 3, in **blue**) to direct the models to understand general video content, reason over the rationale behind actions and their relationships with the context, and consider factors like contextual priors, the temporal order of actions, instance count, and attributes. The prompting technique also includes instructions to ensure conciseness and factuality to mitigate hallucinations. Given a Video-LMM $\mathcal{F}$ and input video $\mathcal{V}$, we retrieve contextual reasoning information $I_{\texttt{context}}$ by providing principled reasoning prompt $P_{\texttt{reason}}$ along with the video to the LMM, $I_{\texttt{context}} = \mathcal{F}(P_{\texttt{reason}}|\mathcal{V})$. This contextual information is then used in the second step of DSCP to generate a grounded response to user question.

Dual Step Contextual Prompting for Video-LMMs

Retrieving Contextual reasoning information (Step 1)

As an intelligent video comprehension model, focus on these guidelines:
1. Differentiate recurring objects, count accurately, and identify movements and poses.
2. Understand directional movements and temporal order.
3. Pay attention to fine-grained actions with precision.
4. Assess incomplete actions without assuming completion.
5. Detect emotional, social, and visual cues.
6. Capture and analyze all relevant actions.
7. Identify unusual actions accurately.
8. Disagree with incorrect information given in question.
9. If you do not find the evidence in the frames, you can give a definite answer by assuming that the asked action/attribute is not present.
10. Provide to the point and concise response.
Now, proceed with answering the following question faithfully while keeping above guidelines in mind:
Question: What is happening in the video?

Context conditioned question-answering (Step 2)

Context for the given video is: {step 1 response}. Now answer a question truthfully based on the video and the provided context.
Question: {User question}

Figure 3: Principled prompt instructions in DSCP for Video-LMMs.

Step 2: Context conditioned question answering. To address the challenges of over-affirmative behavior and hallucinations in Video-LMMs when prompted with confusing or misleading questions, we propose an additional inference step. We note that Video-LMMs often possess factual knowledge about the video content but become distracted and hallucinate when prompted with confusing or misleading questions (Appendix D). Our DSCP technique conditions the model to first comprehend the video without attending to the user question and, therefore eliminates its influence. This complex video comprehension information, $I_{\texttt{context}}$ (formulated in step 1) is then used to condition the model on both the video and $I_{\texttt{context}}$. Finally, we pose the user question using prompt $P_{\texttt{user}}$ which combines the user question and the contextual reasoning information (Fig. 3, in **green**). The final response is $\mathcal{F}(P_{\texttt{user}}|\mathcal{V})$, where $P_{\texttt{user}} = [\text{question}; I_{\texttt{context}}]$. Here $[\ ;\ ]$ denotes the text prompt concatenation. The factual content generated in step 1 guides the model towards a robust response in step 2, producing factual and correct responses even with noisy or misleading user questions. We show the qualitative results of DSCP technique in Fig. 11 in Appendix D. This approach leads to responses that are better grounded in the actual video content and are robust against lower-quality user queries. The DSCP technique effectively enhances the performance of Video-LMMs on CVRR-ES (Sec. 5.2).

5 Evaluation Experiments on CVRR-ES.

Video-LMMs. Among the open-source models, we evaluate 7 recent Video-LMMs, including Video-LLaVA [18], TimeChat [26], MovieChat [28], LLaMA-ViD [17], VideoChat [15] Video-ChatGPT [21], and Video-LLaMA-2 [37]. For evaluating closed-source models, we use Gemini-Pro, Gemini-Flash, [9], GPT-4V and recent GPT-4o [23]. Refer to Appendix B for additional details.

5.1 Main Experiments on CVRR-ES.

Tab. 2 shows the evaluation results of Video-LMMs on CVRR-ES. Below, we discuss main results.

Open Source Video-LMMs struggles on CVRR-ES benchmark. All open-source LMMs show inferior performance across the different evaluation dimensions of CVRR-ES. Interestingly, some of the earlier developed open-source Video-LMMs, like Video-LLaMA, VideoChat, and Video-ChatGPT, exhibit higher performance compared to more recent models such as Video-LLaVA, MovieChat, and LLaMA-VID. Overall, TimeChat achieves the highest performance of 32.89% averaged across the 11 evaluation dimensions among open-source LMMs, followed by VideoChat with a score of 25.78%.

Humans rank highest in CVRR-ES benchmark. Human evaluation achieves the highest perfor-

Table 2: Evaluation results of Video LLMs across various video-evaluation categories on the CVRR-ES benchmark. We present results for both open-source and closed-source models and human evaluation.

Benchmark Category	Video-LLaMA-2	VideoChat	Video-ChatGPT	Video-LLaVA	MovieChat	LLaMA-VID	TimeChat	Gemini-V Pro	Gemini-V Flash	GPT4V	GPT4o	Human
Multiple Actions in single video.	16.98	23.90	27.67	15.72	12.58	17.92	28.30	43.08	44.65	57.55	62.89	**93.40**
Fine-grained action understanding.	29.57	33.48	26.96	25.22	23.48	26.09	39.13	51.61	64.78	77.39	80.43	**95.65**
Partial actions.	24.76	33.01	22.82	13.59	21.36	14.56	49.51	67.48	62.14	73.79	77.67	**98.54**
Time order understanding.	16.45	31.58	27.63	21.05	16.45	19.74	34.21	45.39	55.26	57.89	71.05	**97.37**
Non-existent actions with existent scene.	10.14	15.22	23.19	5.07	5.07	2.90	23.19	57.25	60.14	71.01	83.33	**97.10**
Non-existent actions with non-existent scene.	13.19	14.58	17.36	3.47	11.81	6.94	13.89	49.64	56.30	75.00	70.14	**100.00**
Continuity and Object instance Count.	28.25	24.29	28.41	21.47	19.77	24.86	34.46	36.16	43.50	62.71	62.71	**96.49**
Unusual and Physically Anomalous activities.	18.95	18.42	18.95	15.79	17.89	16.32	27.37	60.00	60.53	74.74	78.42	**96.84**
Interpretation of social context.	25.00	31.07	32.50	18.93	17.14	13.93	39.29	64.29	69.64	79.64	83.57	**97.51**
Understanding of emotional context.	21.92	23.63	21.23	15.07	13.70	14.73	27.40	47.26	52.74	66.44	70.89	**95.55**
Interpretation of visual context.	32.60	34.43	27.84	19.78	21.25	23.08	45.05	63.00	57.51	82.42	84.25	**94.87**
Average	21.62	25.78	24.96	15.92	16.41	16.46	32.89	53.20	57.02	70.78	75.03	**96.67**

Prompting Method	VideoChat	Video-LLaVA	MovieChat	LLaMA-VID	TimeChat
Standard prompting	25.78	15.92	16.41	16.46	32.89
Chain of Thought (CoT) prompting	22.44	25.87	15.89	29.68	**39.57**
DSCP (Stage 1)	38.07	32.12	28.05	25.13	33.04
DSCP (Both stages)	**47.92**	**37.93**	**35.87**	**46.85**	39.45

Table 3: **Prompting methods.** DSCP stage 1 uses only principled instructions of step 1 and DSCP (Both stages) uses complete dual-step technique.

mance on the CVRR-ES benchmark, with over 95% accuracy across all evaluation dimensions. These results suggest that the CVRR-ES QA pairs are reasonable and suitable for benchmarking.

Closed source models perform competitively on CVRR-ES. As shown in Tab. 2, both Gemini and GPT variants improve over open-source models and achieve high gains across all evaluation dimensions. The competitive results of GPT4o and Gemini-Flash on complex video evaluation dimensions such as partial actions, non-existent action/scene depiction, and context-dependent categories show that these models have a more sophisticated understanding of the complex visual contents of videos and have strong capabilities to rectify misleading and confusing user questions. Overall, GPT4o improves over Gemini-Flash by 18.01% and provides the highest average accuracy of 75.03%.

5.2 Effectiveness of DSCP method for improving Video-LMMs performance

We next integrate DSCP technique with Video-LMMs and present results for CVRR-ES in Fig. 4. DSCP improves the model's performance compared with models that use standard prompting (i.e., using only the question itself). These results also suggest that prompting techniques in Video-LMMs can better guide models for improved reasoning and robustness. With DSCP, initially low-performing Video-LMMs like Video-LLaVa, MovieChat, and LLaMA-Vid show much better relative gains and become competitive with other models. The highest relative gain of 184% is achieved by LLaMA-ViD, which moves from 7th place in the leaderboard to 2nd among the open-source models after using the DSCP technique. We observe similar overall positive trends of using DSCP with closed-source model Gemini, which improves on the benchmark by an absolute overall gain of 5.02%. We provide more detailed results comparisons in Appendix D.

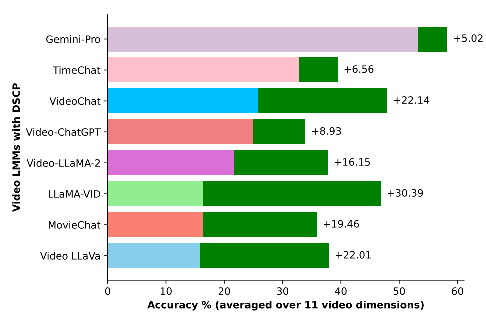

Figure 4: Video-LMMs with DSCP technique effectively improves their performance (gains are shown in **green**) on CVRR-ES benchmark.

5.3 Different prompting techniques.

We now study the contribution of each step of DSCP and compare it with chain-of-thought (CoT) prompting [34]. Results for the top 5 performing open Video-LMMs are shown in Tab. 3. CoT prompting improves over standard prompting in 3 out of 5 Video-LMMs, suggesting that prompting techniques from NLP literature can also guide multi-modal Video-LMMs to enhance reasoning and robustness. Next, we ablate on the first step of DSCP prompting, which uses principled instructions of DSCP step 1 as a prefix alongside the actual user question. DSCP step 1 notably improves model performance on all Video-LMMs, suggesting the effectiveness of the principled prompt instructions designed specifically for Video models. DSCP with both steps, which additionally uses the initial context in the second step, shows additional gains and achieves highest results on 4 out of 5 models.

5.4 Main findings and Qualitative Results

We now present key insights that can guide the development of the next generation of robust and reliable Video-LMMs. We show qualitative results and additional analysis in the Appendix A.

Models excelling at standard VQA benchmarks struggle on CVRR-ES. Our analysis in Sec. 5.1 reveals that the latest open-source Video-LMMs, like Video-LLaVA, MovieChat, and LLaMA-VID, perform less effectively on CVRR-ES compared to Video-LMMs that were introduced earlier in the community, such as VideoChat and Video-ChatGPT. Interestingly, the same recent models demonstrate superior performance on general video comprehension benchmarks. This suggests that current VQA benchmarks, like ActivityNet-QA [36] and MSRVTT [35], do not adequately correlate with the complex video reasoning and robustness scenarios highlighted in our benchmark. Consequently, this also indicates that most newer Video-LMMs are heavily trained to excel on the general video benchmarks while reducing their generalizability, reasoning, and robustness capabilities.

Over-affirmative behavior of open-source Video-LMMs. We observe that open-source models exhibit positive and over-affirmative responses. Open-source Video-LMMs consistently respond with "Yes" even when faced with confusing questions that describe non-existent actions and objects (Fig. 5 in Appendix. A). This highlights the vulnerability of these models when interacting with users in real-world scenarios. In our CVRR-ES benchmark, open-source models are notably vulnerable to evaluation dimensions of "*Non-existent actions with the existent scene*" and "*Non-existent actions with the non-existent scene*" compared to closed models. These models lack negation and self-rectification capabilities, especially when users provide misleading or confusing questions. We conjecture that such behavior arises due to the absence of negative instruction tuning pairs during training.

Tendency towards activity completion. Most open-source Video-LMMs have shown lower results on the evaluation dimension of partial actions, which focuses on incomplete or atomic actions. We note that most open-source models tend to complete actions, even when only part of the action is provided in the video (Fig. 6 in Appendix A). Upon examining the fine-tuning strategies [21, 20], we find that almost all models are trained on end-to-end actions-based instruction-tuning data, causing them to generate complete action descriptions at inference. This tendency highlights the vulnerability of Video-LMMs after deployment, as real-world scenarios often involve atomic, sub-atomic, and general actions alike. To improve the performance of Video-LMMs, it is crucial to incorporate diverse action types during the training phase, including partial and incomplete actions.

Video-LMMs struggles in understanding the emotional and social context. For more reliable interaction with humans in practical scenarios, Video-LMMs models should comprehend the video scenes with social and contextual reasoning capabilities similar to humans. The lower performance of Video-LMMs on social and emotional contextual dimensions in CVRR-ES highlights their limitations and lack of understanding of scenes based on contextual cues (Fig. 9 in Appendix A).

6 Conclusion

Given the expanding role of Video-LMMs in practical world-centric applications, it is crucial to ensure that these models perform robustly and exhibit human-like reasoning and interaction capabilities across various complex and real-world contexts. In this work, we present the CVRR-ES benchmark for Video-LMMs, aiming to evaluate Video-LMMs on these very fronts. Through extensive evaluations, we find that Video-LMMs, especially open-source ones, exhibit limited robustness and reasoning capabilities over complex videos involving real-world contexts. Based on our analysis, we formulate a training-free prompting technique that effectively improves the performance of Video-LMMs across various evaluation dimensions of the CVRR-ES benchmark. Furthermore, we analyze and investigate the failure cases of Video-LMMs on the CVRR-ES benchmark and deduce several important findings. We hope that the CVRR-ES benchmark, accompanied by our extensive analysis, will contribute towards building the next generation of advanced world-centric video understanding models.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. 2022. 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 4

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 5, 24

[4] Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*, 2023. 2, 6

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023. 3

[6] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv:2303.11331*, 2023. 3

[7] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[8] Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020. 5, 24

[9] Google. Gemini, 2023. 2, 7

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 5, 24

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2

[12] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[13] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5, 24

[14] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2, 4

[15] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1, 3, 6, 7

[16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023. 2, 4

[17] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 3, 6, 7

[18] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 2, 3, 7

[19] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2, 6

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 9, 22

[21] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 2, 4, 6, 7, 9, 22

[22] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 4

[23] OpenAI. GPT-4V(ision) System Card, 2023. 2, 7

[24] Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2024. 2, 6

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 3

[26] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023. 4, 6, 7

[27] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 5, 24

[28] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 4, 6, 7, 22

[29] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5, 24

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1, 3

[31] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 4, 7

[32] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 3

[33] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4, 7, 8

[35] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2, 9

[36] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2, 9, 14

[37] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 3, 7

[38] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv:2306.05685*, 2023. 1, 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   Justification: Yes, we have ensured that the main claims in the abstract and introduction accurately reflect the paper's contributions and scope.
   (b) Did you describe the limitations of your work? [Yes]
   Justification: We have discussed the limitations of our work in the Appendix. F.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   Justification: This is a dataset paper aimed at studying and benchmarking the reasoning of Video-LMMs in real-world context and robustness from the lens of user text queries. Therefore, to the best of our knowledge, there are no potential negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
   Justification: Yes we have read the ethics review guidelines and ensured that our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results [N/A]
   Justification: There is no theoretical result in this paper that requires a full set of assumptions and correct proof.
   (b) Did you include complete proofs of all theoretical results? [N/A]
   Justification: There is no theoretical result in this paper that requires a full set of assumptions and correct proof.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   Justification: We have attached the code, link to data, and all instructions to reproduce the main experimental results in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   Justification: We have provided implementation details in the Appendix. B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] .
   Justification: We did not have enough compute resources to completely re-run all the experiments for different seeds and report error bars for different runs. We are currently re-running the error bar experiments, and we plan to include all the experiments with different seeds in the final version.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
   Justification: We have provided details on the compute resources in the Appendix. B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   Justification: We have cited the creators of datasets used in our benchmark in the main paper in Sec. 3.2.
   (b) Did you mention the license of the assets? [Yes]
   Justification: Our dataset is released for educational and research purposes under the CC-BY-4.0 license. We have mentioned the license of assets in the files in our supplemental material as well as on our GitHub dataset hosting platform.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
Justification: Yes we have included the assets in the supplemental material and also on the public URL. Our assets can be publically accessed at mbzuai-oryx.github.io/CVRR-Evaluation-Suite/.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
Justification: We collected most of the videos from academic datasets while respecting their license information. The videos obtained from the web from YouTube are subject to the copyright of the original owners and are used only for research and academic purposes, consistant with previous works and benchmarks such as ActivityNet [36] etc.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
Justification: In our initial analysis using the subset (50%) of our CVRR-ES dataset, we noted that no video contained specific personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
Justification: The instructions to humans for the benchmark quality assessment are provided in Appendix C.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
Justification: Not applicable.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
Justification: The annotation process was carried out by the authors of this manuscript. As a result, the aspect of compensation for human subjects does not apply in this case.