

578 Appendix

579 In the following sections, we provide additional information for the paper: **Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs**. The contents are organized in the
580 following order.
581

- 582 • Additional findings and qualitative results (Appendix A)
- 583 • Implementation details (Appendix B)
- 584 • Additional details on CVRR-ES Benchmark (Appendix C)
- 585 • Analysis and additional results for DSCP technique (Appendix D)
- 586 • Additional Ablation Experiments (Appendix E)
- 587 • Limitations (Appendix F)

588 A Additional findings and qualitative results

589 Below we discuss additional observations about closed-source and open-source Video-LMMs based
590 on the evaluation and qualitative results on the CVRR-ES benchmark.

591 **Weak Generalization to extreme OOD videos.** The evaluation dimension of unusual and physically
592 anomalous activities in CVRR-ES resembles extreme out-of-distribution video examples. With
593 the exception of GPT4V and Gemini, Video-LMMs struggle with this dimension, indicating weak
594 generalizability towards OOD videos containing the coexistence of unusual objects and activities
595 that are extremely rare in typical videos. For instance, Video-LLaVA in Fig. 7 describes a person
596 falling on the street, while the video actually shows the person performing an optical illusion. To
597 be responsibly deployed in real-world applications, where OOD actions occur more frequently,
598 Video-LMMs needs to be trained to perform more robustly on OOD samples. This may involve
599 incorporating diverse and atypical examples in the training data to improve the model’s ability to
600 handle unusual situations.

601 **Limited understanding of temporal order in complex videos.** The CVRR-ES benchmark results
602 show that Video-LMMs perform relatively better on the fine-grained action dimension compared to
603 the time-order understanding dimension. While these models can accurately identify fine-grained
604 actions, they struggle with comprehending the correct temporal order of these actions within a video.
605 This limitation can lead to misinterpretations of the underlying information depending on temporal
606 order. We present failure cases of this dimension in Fig. 8. For building more advanced world-centric
607 Video-LMMs, it is crucial to enhance their ability to process and interpret event sequences accurately.

608 B Implementation details

609 For open-source models, we follow their default best inference settings and hyperparameters. To
610 evaluate Gemini and GPT-4V, we utilize their official APIs. Full videos are directly passed to Gemini
611 Vision-Pro, as its API (using Google Cloud vertexai framework) inherently supports video inputs.
612 However, as GPT-4V does not inherently support videos, we uniformly sample 8 frames for each
613 video which are passed into GPT API along with user questions. For each model under evaluation,
614 we generate responses to the questions independently and without retaining the chat history. For the
615 evaluation results of Video-LMMs on the CVRR-ES QA pairs, we utilize GPT-3.5 as a judge in all of
616 our experiments. For benchmarking the Video-LMMs, we used NVIDIA A100 40 GB GPU provided
617 by our university lab at MBZUAI.



Figure 5: **Over affirmative behaviour.** Most open-source Video-LMMs exhibit overly affirmative behavior by consistently agreeing with user questions, even when the questions are confusing or inaccurate.



Figure 6: **Action completion tendency.** Most open-source Video-LMMs tend to generate captions corresponding to complete actions and struggle with determining incomplete or partial actions.

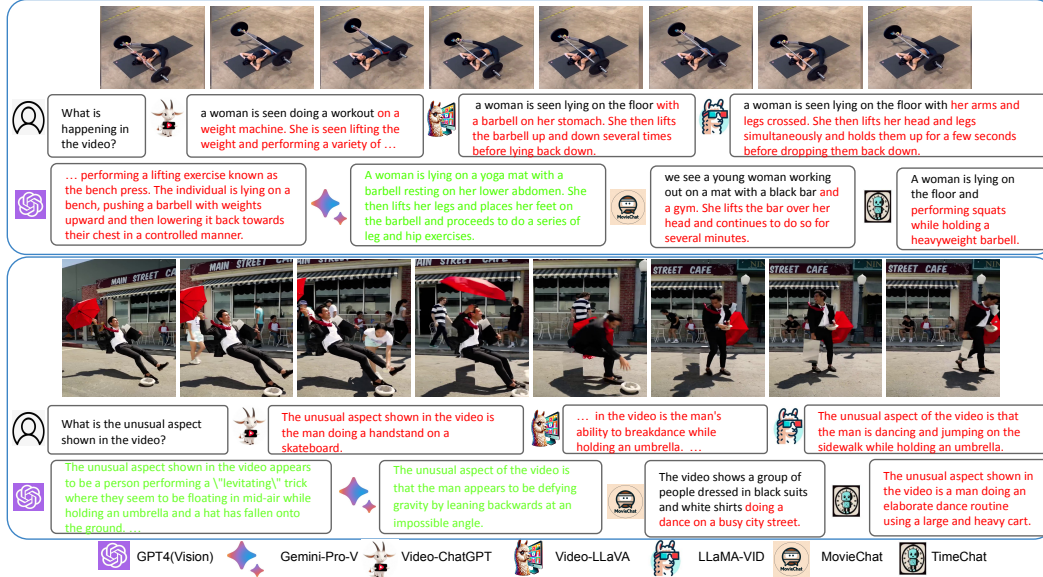


Figure 7: **Weak generalization on OOD videos.** Open-source Video-LMMs struggle to correctly reason over videos containing rare and unusual actions.

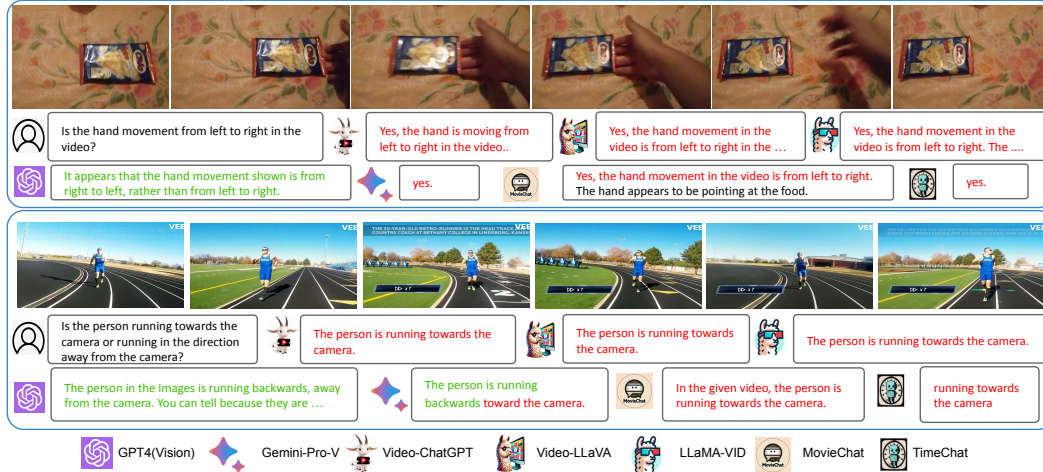


Figure 8: **Limited temporal understanding.** Most Video-LMMs struggle to accurately determine the temporal order of actions in videos. The bottom video shows a man running backward along a track.

Table 4: Examples of the question-answer pairs in the CVRR-ES benchmark for various complex video evaluation dimensions.

Evaluation Dimensions	Sample Question-Answer pairs
1. Multiple actions in a single video	Q. Does the person stand up to welcome the cat or remain seated throughout their interaction? A. The person remains seated throughout their interaction with the cat. Q. What is the next action performed by the person after using the laptop? A. The action directly after using the laptop is placing a bag in the refrigerator.
2. Fine-grained action understanding	Q. At any point in the video, does the man use the thread to sew fabric? A. No, the man uses the thread to create loops and demonstrate tying a knot; there is no depiction of sewing fabric. Q. What action is performed by the person’s hands in the video? A. The person’s hands are shown plugging a black USB charging cable into the charging port.
3. Partial actions	Q. What is happening in the video? A. The video shows the door of a red car and a person’s hand reaching to the handle of the car ... Q. Does the video include a moment where the snack is replaced to its original position on the right? A. No, the video concentrates on the initial action of moving the snack from the right to the left, without ...
4. Time order understanding	Q. Is the video showing the activity of taking out liquid from the soda can? A. No, the video does not show the activity of taking out the liquid from the soda can. The video shows ... Q. Is the person running in clockwise direction or anticlockwise direction on the race track? A. The person is running in anticlockwise direction in the video.
5. Non-existent actions with existent scene depictions	Q. After going through the bag, does the person meticulously clean the area around the sink? A. No, the person does not clean the area around the sink after going through the bag. The video focuses ... Q. What is the reaction of the audience when the keynote speaker delivers his speech? A. The scene does not include a moment where a keynote speaker is delivering a speech ...
6. Non-existent actions with non-existent scene depictions	Q. How do the children interact with the flowers in the video? A. There are no children interacting with the flowers depicted in the video. The footage is committed to displaying ... Q. What is the reaction of the child playing in the corner when the dog runs past? A. There is no child playing in the corner or any reaction to the dog running past ...
7. Continuity and Object Instance Count	Q. How many unique sunglasses appear throughout the video? A. As there are 4 persons in the car wearing the sunglasses, the number of unique sunglasses is 4. Q. Did the attire of both men remain the same upon re-entering the frame the second time? A. No, the attire of both men did not remain the same upon re-entering ...
8. Unusual and Physically Anomalous activities	Q. Is the person showcasing walking or running movements to reach an elevated position in the video? A. No, the person did not walk or run; they ascended and floated in the air through what ... Q. How the person is able to fly over the water? A. The person is using a flyboard system attached to his shoes using which he is flying over the water.
9. Interpretation of social context	Q. What was the response of the crowd when the girl landed the water bottle vertically? A. the crowd applauded to showcase appreciation for her perseverance and success. Q. What is the primary reason the boy touches the ashes before placing his hand on the goat? A. The boy uses the ashes to warm the goat, indicating his primary motive is care and providing warmth.
10. Understanding of emotional context	Q. Identify if the emotional context of the video is negative, based on the described actions and reactions? A. The emotional context of the video is not negative; it is overwhelmingly positive. The indicators of happiness, ... Q. Identify the nature of the interaction between the two individuals. Is it professional, hostile, or friendly? A. The interaction is friendly. This is evidenced by the warm hug and the handshake, ...
11. Interpretation of visual context	Q. Does the person in the video undergo a real physical transformation? A. No, ... They simply remove a rubber mask that made them look like a man, revealing that they are actually a woman. Q. Identify the unusual behavior depicted between a predator and its usual prey in the video. A. A cat plays and sleeps with chicks instead of hunting them. This showcases an unusual peace ...

633 based on the frequency of keywords in the answer set of CVRR-ES in Fig. 10. The frequent
634 words correspond to objects and attributes with which Video-LMMs could most likely interact when
635 deployed in practical scenarios.

636 **Human Evaluation.** To verify that the QA pairs in the CVRR-ES benchmark are reasonably
637 answerable and to establish a benchmark for human performance, we conduct a human evaluation.
638 Two human experts (authors) are instructed to watch the video corresponding to each question and
639 provide a free-form answer. The predictions of the human experts for all QA pairs are assessed using
640 an LLM-assisted evaluation. Individual final scores are averaged to mitigate potential bias from a
641 single human evaluator. The results of the human evaluation are presented in the main paper in the
642 experiments section (Sec. 5.1).

643 D Further Analysis and Additional Results for DSCP Method.

644 D.1 Discussion on DSCP method.

645 We note that Video-LMMs are often able to correctly comprehend the video content and produce
646 factual responses. However, they are extremely sensitive to user textual prompt inputs and exhibit
647 highly over-affirmative behavior. These attributes of Video-LMMs can lead to hallucinations and



Figure 11: **Qualitative results of DSCP prompting method.** Using our DSCP approach, Video-LMMs demonstrate enhanced robustness and reasoning capabilities over complex videos.

648 wrong output responses, especially when the user asks reasoning-based, low-quality, confusing, or
 649 misleading questions.

650 Our Dual-Step Contextual Prompting technique aims to address these limitations of Video-LMMs by
 651 explicitly delineating the contextual reasoning information retrieval from the user question answering
 652 using a two-step prompting technique. This strategy effectively eliminates the influence of the
 653 question, leading to more grounded overall responses.

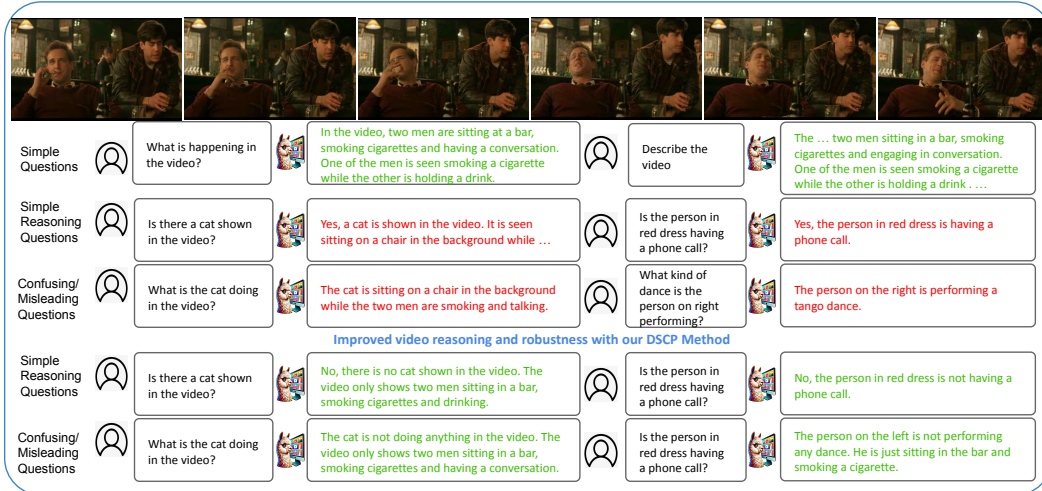


Figure 12: Effect of different prompts on Video-LLaVA. **Row 1:** Video-LLaVA often provides factual and correct information about the input video when prompted with simple and clear questions. **Row 2 & 3:** However, the model struggles to remain factual when the question becomes reasoning-based, confusing, or misleading, mainly due to its over-affirmative behavior. **Row 4 & 5:** Our DSCP method utilizes contextual reasoning information in the first step prompting, independent of the user question, and uses it as conditioning information in the second step, leading to more grounded and factual responses to user questions.

In Fig. 12, we show the sensitivity of Video-LMMs to textual prompts and the impact of each step in the DSCP prompting technique. It can be observed that prompting the model with simple questions, such as 'Describe the video content' or 'What is happening in the video?' leads to correct responses. However, as the user asks a reasoning-based question or a tricky question, the model struggles to reason properly and hallucinates due to an over-affirmative response. Finally, we generate the response using the DSCP method. The first step independently retrieves contextual reasoning information using principled prompt instructions, followed by asking the user a question conditioned on both the factual information retrieved earlier and the input video. We observe that integrating both steps of DSCP prompting injects improved reasoning and self-rectification capabilities into Video-LMMs.

D.2 Detailed comparison results.

In the main paper, we presented overall results comparisons between Video-LMMs utilizing the Dual-Step Contextual Prompting (DSCP) technique. Here, we show the per evaluation dimension performance of Video-LMMs when utilizing DSCP technique in Tab. 5. The results indicate that Video-LMMs with DSCP technique provide substantial performance improvements across various evaluation dimensions in the CVRR-ES benchmark.

While DSCP prompting reduces the performance for the evaluation dimension of time-order understanding for a few Video-LMMs such as VideoChat, Video-ChatGPT, and Gemini, the overall relative performance improvements are notable for the majority of the models. DSCP technique improves the performance of Video-LMMs across most evaluation dimensions. In particular, DSCP shows the highest gains for the evaluation dimensions of physically anomalous, contextual videos, fine-grained actions, and partial actions, demonstrating the model’s improved reasoning capabilities without any additional training. For evaluation dimensions involving explicit misleading user questions, such as non-existent actions with non-existent scene depiction, DSCP substantially improves the model’s performance. For instance, VideoChat improves from 14.38% to 58.33% on the same evaluation dimension, corresponding to relative gains of over 300%. This suggests that DSCP prompting acts as an additional filter layer that guides the model towards robust and grounded behavior.

The overall performance improvements of Video-LMMs with DSCP suggest that prompting techniques can effectively steer the behavior of Video-LMMs for enhanced reasoning and robustness over videos. Although DSCP shows promising results, the net performance of Video-LMMs is still far from satisfactory, which demands more advanced techniques to further enhance their capabilities, especially for open-source models.

E Ablation Studies.

Our CVRR-ES evaluation benchmark utilizes key design choices. In this section, we present several ablation studies to validate the effectiveness of these design choices.

Alignment of LLM as the Judge with Human evaluators.

We utilize LLMs such as GPT-3.5 as a judge for evaluating Video-LMMs on the CVRR-ES benchmark. In this study, we compare how closely LLM accuracy scores align with human evaluations. We assign two expert human evaluators to independently evaluate human performance by manually evaluating and scoring each candidate’s answer. We observe that the human evaluation results by LLM have an alignment percentage of 95.36%. This means that for 4.64% of QA pairs, there was a mismatch between LLM judgment and human judgment. The 95%+ alignment rate with GPT-3.5 is encouraging, and we conjecture that future LLMs will exhibit further alignment with human evaluations.

LLM Judgement improves by generating explanations. Our default evaluation prompt as shown in Fig. 13 requires the Judge LLM to generate a correct/incorrect flag, an answer quality score (ranging from 0 to 5), and the rationale behind the quality score and the correct/incorrect flag. The alignment score with human evaluators for this instruction prompt is 95.36%. Previously, we utilized

Table 5: Video LMMs evaluation results using our Dual-Step Contextual Prompting (DSCP) Technique. Video LMMs with DSCP technique effectively improves their reasoning and robustness capabilities on complex video-evaluation dimensions in CVRR-ES. Absolute gains over the standard prompting are shown in **green**.

Benchmark Category	Video-LLaMA2	VideoChat	Video-ChatGPT	Video-LLaVA	MovieChat	LLaMA-VID	TimeChat	Gemini-V Pro
Multiple Actions in single video.	32.39 (+15.41)	38.99 (+15.09)	32.70 (+5.03)	37.74 (+22.01)	27.36 (+14.78)	39.62 (+21.70)	32.08 (+3.77)	49.37 (+6.29)
Fine-grained action understanding.	35.65 (+6.09)	39.57 (+6.09)	28.26 (+1.30)	33.48 (+8.26)	41.74 (+18.26)	41.74 (+15.65)	40.87 (+1.74)	51.15 (-0.46)
Partial actions.	39.32 (+14.56)	50.49 (+17.48)	34.95 (+12.14)	47.57 (+33.98)	33.98 (+12.62)	52.91 (+38.35)	55.34 (+5.83)	61.17 (-6.31)
Time order understanding.	28.29 (+11.84)	28.95 (-2.63)	23.68 (-3.95)	30.26 (+9.21)	23.68 (+7.24)	31.58 (+11.84)	32.24 (-1.97)	43.42 (-1.97)
Non-existent actions with existent scene.	39.86 (+29.71)	65.94 (+50.72)	31.16 (+7.97)	47.10 (+42.03)	39.13 (+34.06)	51.45 (+48.55)	30.43 (+7.25)	68.12 (+10.87)
Non-existent actions with non-existent scene.	40.97 (+27.78)	58.33 (+43.75)	30.56 (+13.19)	42.36 (+38.89)	35.42 (+23.61)	56.94 (+50.00)	29.17 (+15.28)	71.94 (+22.30)
Continuity and Object instance Count.	31.07 (+2.82)	38.42 (+14.12)	31.64 (+3.23)	32.77 (+11.30)	35.59 (+15.82)	37.85 (+12.99)	38.98 (+4.52)	46.33 (+10.17)
Unusual and Physically Anomalous activities.	38.95 (+20.00)	50.00 (+31.58)	33.16 (+14.21)	31.58 (+15.79)	40.53 (+22.63)	40.53 (+24.21)	37.89 (+10.53)	65.26 (+5.26)
Interpretation of social context.	47.50 (+22.50)	58.21 (+27.14)	48.93 (+16.43)	43.93 (+25.00)	44.29 (+27.14)	64.29 (+50.36)	52.86 (+13.57)	72.14 (+7.86)
Understanding of emotional context.	35.27 (+13.36)	41.10 (+17.47)	30.14 (+8.90)	24.66 (+9.59)	32.88 (+19.18)	37.67 (+22.95)	33.56 (+6.16)	50.68 (+3.42)
Interpretation of visual context.	47.50 (+13.55)	58.21 (+22.71)	48.93 (+19.78)	43.93 (+26.01)	44.29 (+18.68)	64.29 (+37.73)	52.86 (+5.49)	72.14 (-2.20)
Average	37.77 (+16.15)	47.92 (+22.14)	33.89 (+8.93)	37.93 (+22.01)	35.87 (+19.46)	46.85 (+30.39)	39.45 (+6.56)	58.22 (+5.02)

Evaluation Prompt to LLM as a Judge

You are an intelligent chatbot designed for evaluating the correctness of AI assistant predictions for question-answer pairs.
Your task is to compare the predicted answer with the ground-truth answer and determine if the predicted answer is correct or not. Here's how you can accomplish the task:

###INSTRUCTIONS:

- Focus on the correctness and accuracy of the predicted answer with the ground-truth.
- Consider predictions with less specific details as correct evaluation, unless such details are explicitly asked in the question.

Please evaluate the following video-based question-answer pair:

Question: {CVRR-ES Question}

Ground truth correct Answer: {CVRR-ES GT answer}

Predicted Answer: {Video LMM prediction}

Provide your evaluation as a correct/incorrect prediction along with the score where the score is an integer value between 0 (fully wrong) and 5 (fully correct). The middle score provides the percentage of correctness.

Please generate the response in the form of a Python dictionary string with keys 'pred', 'score' and 'reason', where value of 'pred' is a string of 'correct' or 'incorrect', value of 'score' is in INTEGER, not STRING and value of 'reason' should provide the reason behind the decision.

Only provide the Python dictionary string.

For example, your response should look like this: {'pred': 'correct', 'score': 4.8, 'reason': reason}.

Figure 13: Prompt used to instruct LLM as a judge for evaluating Video-LMM responses on CVRR-ES benchmark. We employ GPT-3.5 turbo as the choice of LLM. The system prompt is shown in **blue** while the main prompt is shown in **green**.

the LLM Judge instruction prompt based on prior works [21, 20, 28], which do not request the model to provide the decision rationale. With their prompt, we observe that the Judge’s alignment with human evaluators is 89.63%. This suggests that requiring LLM Judge decisions to be accompanied by corresponding reasons yields more reliable evaluation results.

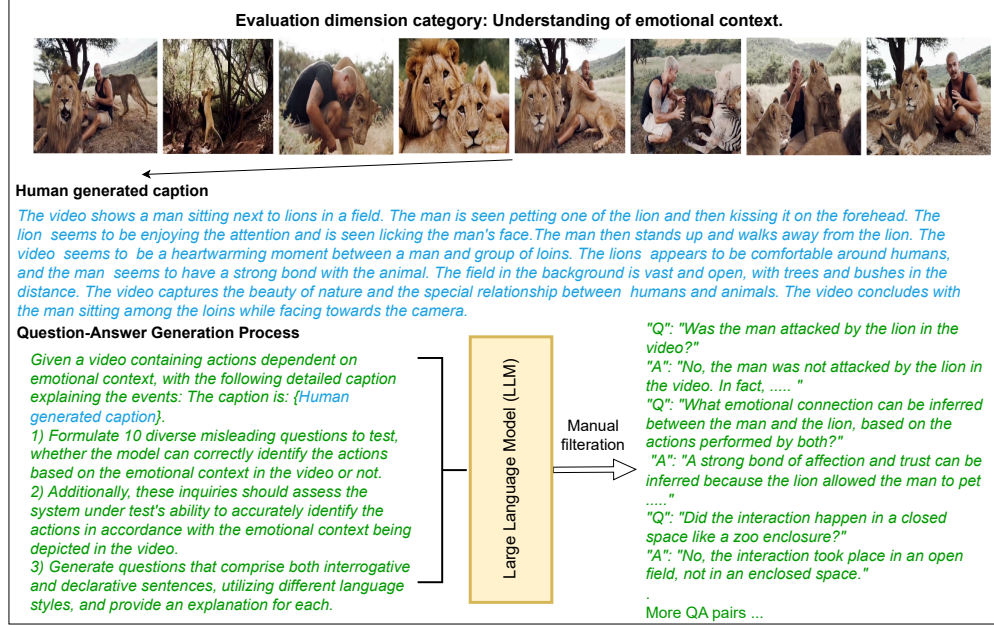


Figure 14: An illustration of the QA pair generation process using LLMs for our CVRR-ES benchmark. Human-generated video captions are input to LLMs which are instructed to generate diverse QA pairs encompassing both textual robustness and reasoning dimensions.

F Limitations

While we aimed to uncover several important insights about the practicality of Video-Language Models (Video-LMMs) in real-world contexts using our CVRR-ES benchmark, there are several limitations which are discussed below.

In the curation of the CVRR-ES benchmark, we used Language-Model-generated (LLM) QA pairs based on video captions to evaluate the Video-LMMs. We noted that using LLMs to generate the question-answer pairs can be challenging, as the LLM may produce straightforward questions sometimes and does not always adhere to the input prompt used to create QA pairs. To address this issue, we employed a human-based filtration process, which involved an exhaustive verification and rectification of the CVRR-ES benchmark's questions, as described in Stage 3 of the benchmark curation (Sec. 3.2). We believe that future LLMs will be more aligned with human intent for generating benchmark question-answer pairs to further minimize the need for manual filtration.

Secondly, our dual-step contextual prompting (DSCP) technique, while effective in improving the reasoning and robustness capabilities of Video-LMMs in our benchmark, it introduces additional inference latency due to the model's two-time forward pass. We plan to investigate remedies towards improving the compute efficiency of the DSCP technique in our future work.

722 Additional supplementary material for CVRR-ES.

723 A Dataset documentation and intended uses.

724 **Motivation and Purpose of the Dataset.** The widespread adoption of Video-LMMs in our daily
725 lives underscores the importance of ensuring and evaluating their robust performance in mirroring
726 human-like reasoning and interaction capabilities in complex, real-world contexts. In this work,
727 we present the Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES), a novel
728 benchmark dataset that comprehensively assesses the performance of Video-LMMs across 11 diverse
729 real-world video dimensions. The evaluation results of our CVRR-ES dataset provide valuable
730 insights for building the next generation of human-centric AI systems with advanced robustness and
731 reasoning capabilities.

732 **Who created the dataset?** The authors of this work created/curated the dataset and formulated the
733 overall benchmarking protocols.

734 **Overview of CVRR-ES dataset.** CVRR-ES benchmark consists of 2400 open-ended question-
735 answer (QA) pairs spanning over 214 unique videos (224 videos in total as some videos are used
736 for multiple evaluation dimensions) for evaluating Video-LMMs. The benchmark aims to assess
737 their robustness to user textual queries (e.g., confusing, misleading questions etc.) and reasoning
738 capabilities in a variety of complex and contextual videos covering 11 diverse evaluation dimensions.
739 For more details, refer to the main paper (Sec. 3.2).

740 **Collection Process.** The authors of this work have collected the videos manually for the CVRR-ES
741 benchmark. We first collect high-quality videos and annotate each video via human assistance. To
742 ensure that each evaluation dimension captures relevant attributes and information, we meticulously
743 select videos that are representative of specific characteristics associated with that dimension. Over-
744 all, 214 unique videos are selected covering 11 dimensions with around 20 videos per evaluation
745 dimension. Around 60% of these videos are collected from public academic datasets. To introduce
746 diversity in the benchmark distribution, we select videos from multiple datasets including Something-
747 Something-v2 [10], CATER [8], Charades [27], ActivityNet [3], HMDB51 [13], YFCC100M [29].
748 The remaining 40% of videos are collected from the internet.

749 **Preprocessing/cleaning/labeling.** The main filtration step was formulated for the cleaning and re-
750 labeling the LLM-generated question-answer pairs. Specifically, a manual filtration step is employed,
751 with human assistance to verify each generated QA pair. Approximately 30% of the QA pairs
752 generated by GPT-3.5 are found to be noisy, containing questions that are unrelated to the video
753 evaluation dimensions or unanswerable based on the provided ground-truth captions. Additionally,
754 many questions contain answers within the question itself. Therefore, an exhaustive filtering process
755 is conducted which involves QA rectification and removing those samples which are not relevant to
756 the video or evaluation type. This process results in a final set of 2400 high-quality QA pairs for the
757 CVRR-ES benchmark. Examples of QA pairs are shown in Tab. 4 in the Appendix.

758 **Primary use of dataset.** The dataset is primarily used to evaluate Video-LMMs on open-ended
759 video question-answer pairs covering a diverse set of evaluation dimensions over complex real-world
760 contextual videos. The benchmark evaluation results reflects the reasoning and robustness capabilities
761 of Video-LMMs.

762 B Distribution of CVRR-ES Dataset.

763 **How to view the dataset?** The final dataset files alongside code repository and instruction manual
764 are publically hosted at <https://mbzuai-oryx.github.io/CVRR-Evaluation-Suite/>. Additionally, all
765 instructions and code files to reproduce the experiments of the paper are present on the github
766 repository at this link: <https://github.com/mbzuai-oryx/CVRR-Evaluation-Suite/>.

767 **How will the dataset be distributed?** The dataset is distributed to the public using GitHub and
768 Onedrive platforms. We have publically release the code-base alongside instructions to reproduce
769 and evaluate models on GitHub.

770 **Dataset License.** This work and dataset is licensed under a Creative Commons Attribution-
771 NonCommercial-ShareAlike 4.0 International License. The videos in CVRR-ES dataset are collected
772 from public academic benchmarks (refer to main paper for more details) and crawled from YouTube
773 and are for academic research use only. By using CVRR-ES, you agree not to use the dataset for
774 any harm or unfair discrimination. Please note that the data in this dataset may be subject to other
775 agreements. Video copyrights belong to the original dataset providers, video creators, or platforms.

776 **URL to Croissant metadata record documenting the dataset/benchmark available for viewing**
777 **and downloading by the reviewers.** Dataset files can be downloaded and reviewed at this link:
778 [OneDrive download link](#).

779 **C Authors declaration and Maintenance plan.**

780 **Author statement.** The first author of this paper declares that they bear all responsibility in case of
781 violation of rights, etc., and confirmation of the data license.

782 **Maintenance plan and dataset hosting information.** The authors of this work will be responsible
783 for the maintenance of this dataset. The benchmark has been hosted on one drive data-sharing
784 platform and all associated code-base is hosted on GitHub. Authors will maintain the dataset hosting
785 resources on a monthly basis.

786 **D How to use the dataset? Getting started with sample code files.**

787 For getting started with the CVRR-ES benchmark dataset, please refer to the code files provided in the
788 at the GitHub repository publically available at: [https://github.com/mbzuai-oryx/CVRR-Evaluation-](https://github.com/mbzuai-oryx/CVRR-Evaluation-Suite/)
789 [Suite/](#).

790 **Reproducing experimental results.** Instructions have been provided in the GitHub repository which
791 to reproduce the main experimental results of the main paper.