

Additional supplementary material for CVRR-ES.

A DATASET DOCUMENTATION AND INTENDED USES.

Motivation and Purpose of the Dataset. 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. In this work, we present the Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES), a novel benchmark dataset that comprehensively assesses the performance of Video-LMMs across 11 diverse real-world video dimensions. The evaluation results of our CVRR-ES dataset provide valuable insights for building the next generation of human-centric AI systems with advanced robustness and reasoning capabilities.

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

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

Collection Process. The authors of this work have collected the videos manually for the CVRR-ES benchmark. 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 (Goyal et al., 2017), CATER (Girdhar & Ramanan, 2020), Charades (Sigurdsson et al., 2016), ActivityNet (Caba Heilbron et al., 2015), HMDB51 (Kuehne et al., 2011), YFCC100M (Thomee et al., 2016). The remaining 40% of videos are collected from the internet.

Preprocessing/cleaning/labeling. The main filtration step was formulated for the cleaning and re-labeling the LLM-generated question-answer pairs. Specifically, a manual filtration step is employed, 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 QA pairs are shown in Tab. 4 in the Appendix.

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

B DISTRIBUTION OF CVRR-ES DATASET.

How to view the dataset? The final dataset files alongside code repository and instruction manual are anonymously hosted at <https://drive.google.com/drive/folders/1t2-DnLhJpchzKgW-2jWmVlu75wzp9L5W?usp=sharing>. Additionally, all instructions and code files to reproduce the experiments of the paper are present attached anonymous code repository.

How will the dataset be distributed? The dataset will be distributed to the public using GitHub and Onedrive platforms. We will publically release the code-base 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 CVRR-ES dataset are collected from public academic benchmarks (refer to main paper for more details) and crawled from YouTube and are for educational research use only. By using CVRR-ES, 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.

URL to Croissant metadata record documenting the dataset/benchmark available for viewing and downloading by the reviewers. Dataset files can be downloaded and reviewed at this link: Anonymous download link.

C AUTHORS DECLARATION AND MAINTENANCE PLAN.

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

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

D HOW TO USE THE DATASET? GETTING STARTED WITH SAMPLE CODE FILES.

For getting started with the CVRR-ES benchmark dataset, please refer to the code files attached with the supplementary material.

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