

Enhancing the Evaluation of Vision-Language Models for Video Understanding

Navya Rachakonda^{‡§}, George Awad[‡]

[‡]National Institute of Standards and Technology, Gaithersburg, MD 20899

[§]Northwest High School, Germantown, MD 20874

navya.rachakonda@nist.gov, navya6.rachakonda@gmail.com, george.awad@nist.gov

Abstract—Vision Language Models (VLMs) extend large language models (LLMs) using visual encoders and provide understanding about images and videos. Evaluating VLMs is necessary to improve the reliability of their outputs over time. In this work, we developed an enhanced question and answer dataset that will be used to evaluate VLMs. This was done by using a ground truth human annotated dataset, where a more diverse dataset was generated using the NVIDIA-Nemotron 3.

Keywords—Vision-Language Models, evaluation, videos

I. INTRODUCTION

Vision Language Models (VLMs) combine large language models (LLMs) and vision encoders to provide understanding about images and videos. They have a wide variety of applications such as robotics, video summarization, video captioning and disaster scene description.

For autonomous robotic platforms such as humanoids, quadrupeds, and drones operating in disaster environments, accurate visual scene description is a fundamental perception requirement before any human intervention[1,2]. The evaluation of VLMs is essential to ensure they generate useful information in the applications described.

The Video Question Answering (VQA)[3] task at the Text Retrieval Conference (TREC) assesses VLMs' multimedia comprehension by posing questions about video content. A model is provided with a short video clip and a natural language question, requiring it to answer the question based on the video content. These questions span categories like multimodality, quantitative, casual, temporal, person-object interaction, common sense, and multi-hop reasoning. The National Institute of Standards and Technology (NIST) employed five human annotators to watch individual videos and generate natural language questions about them, along with their respective correct answers and plausible incorrect answers. This created the ground truth (GT) human generated dataset for VQA. This GT dataset will then be used to evaluate these VLMs. There were ≈ 1300 unique videos, which were 30 to 120 seconds long, and they came from Youtube Shorts and segments from licensed movies. In this work, we expanded the dataset by generating more correct and incorrect answers for each question, increasing the diversity and coverage of possible answers when testing VLMs.

II. METHODS

A python-based pipeline was developed to generate three additional correct answers and three additional incorrect answers for each question. For example, for the question 'How many children get off the bus in this video?', the original GT

dataset may contain a correct answer of 'Zero' and an incorrect answer of '1'. Our pipeline generates additional correct answers, like 'None' or 'No kids got off', and additional incorrect answers, like '2' or '3'.

This was done using the following steps in Figure 1:

- **Answer Generation:** The NVIDIA-Nemotron 3 LLM was used to generate alternative correct and incorrect answers using original GT answers.
- **Prompt Engineering:** Prompts were iteratively refined with few-shot examples, role assignment, and chain-of-thought to improve answer quality.
- **Automated Validation:** Semantic checks and manual reviews were performed to identify any overlap and ensure response quality.

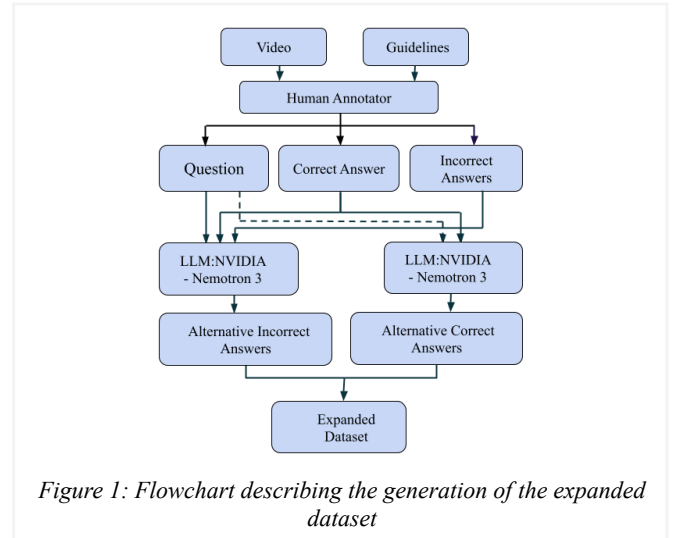


Figure 1: Flowchart describing the generation of the expanded dataset

III. VALIDATION

A quantitative validation of these generated answers was performed to assess the semantic consistency using metrics such as Semantic Textual Similarity (STS, using MPNet transformer[4]) and BERTScore[5]. These metrics were used to generate similarity scores between the original and generated answers.

For each question, the original correct answer was compared with the generated correct answers, where high semantic similarity was expected. The original correct answer was also compared with the generated incorrect answers, where a lower semantic similarity was expected. The resulting similarity scores were then analyzed to assess the effectiveness

of both the answer generation process and the prompt refinement process.

IV. RESULTS

The similarity scores demonstrated a significant improvement in semantic consistency of the generated correct answers following a prompt refinement. The prompt refinement incorporated an improved self-validation protocol. This protocol required the LLM to evaluate its generated responses for overlap with the original data and contextual inconsistencies, regenerating responses when necessary.

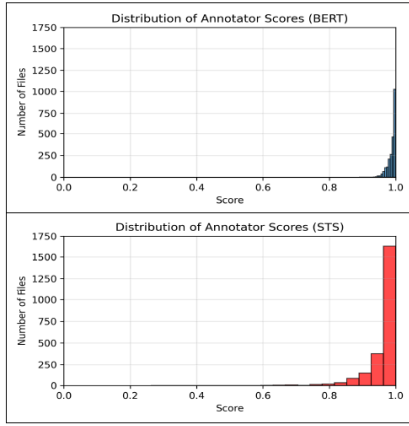


Figure 2: Histograms of refined generated correct answer scores vs original correct answer scores using BERT and STS

Before the prompt refinement, STS produced an average similarity score of 0.74 per human annotator, while BERTScore produced an average similarity score of 0.94 per human annotator. Following the prompt refinement, these scores increased to 0.96 for STS and 0.99 for BERTScore.

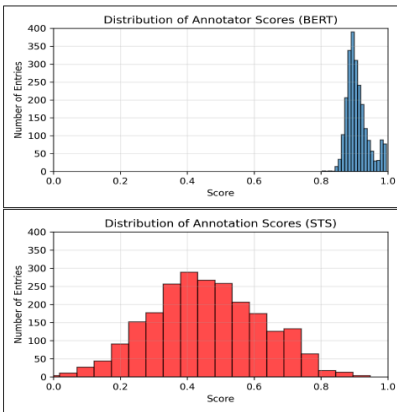


Figure 3: Histograms of generated incorrect answer scores vs original correct answer scores using BERT and STS

For the generated incorrect answers, STS produced an average similarity score of 0.46 when compared with the original correct answers, demonstrating greater semantic distinction. However, BERTScore produced an average similarity score of 0.91.

The relatively high BERTScore of 0.91 for the generated incorrect answers suggests that this metric might have

difficulty distinguishing between answers that may share contextual similarities but differ in correctness. This may contribute to an overestimation of similarity in cases where the generated answer is incorrect. This motivates future exploration for alternative metrics that can more effectively capture semantic differences between texts and distinguish between correct and incorrect answers.

V. SUMMARY AND CONCLUSION

We expanded the GT dataset to prepare for robust VLM evaluation, which we will conduct in future work using the wider diversity of correct and incorrect answers. The VLM evaluation will happen in two ways. First, we will assess the answer generation capabilities of VLMs by providing the videos and the human-generated questions to the models and evaluating their responses using metrics such as STS, BERTScore, and others. Second, we will assess VLMs' multiple-choice capabilities by providing the videos, human-generated questions, and multiple-choice answers and evaluating whether the models selected the correct answers.

Overall, the expanded GT dataset provides a broader set of valid and invalid responses for evaluating VLM performance. This will help establish a foundation for more robust assessment of multimedia understanding.

VI. DISCLAIMER

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government. This work was approved by NIST, research protection office, and IRB # ITL-2017-0061.

REFERENCES

- [1] G. Awad et al., "An overview on the evaluated video retrieval tasks at TRECVID 2022," 2023, <https://arxiv.org/abs/2306.13118v1>
- [2] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," 2023, <https://arxiv.org/abs/2307.15818v1>
- [3] G. Awad et al., "Video Question Answering (VQA) 2025 Track, TREC Proceedings" https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=961440
- [4] All-MPNET at SEMEVAL-2024 Task 1: Application of MPNET for evaluating semantic Textual relatedness. (2024). <https://aclanthology.org/2024.semeval-1.59.pdf>
- [5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2019, <https://arxiv.org/abs/1904.09675v3>