HD-EPIC VQA benchmark result using Qwen2.5VL and LLaVA-Video

Anonymous CVPR submission

Paper ID *****

Abstract

We report our result for HD-EPIC VQA benchmark using two open-source VLMs. We adopted both Qwen2.5VL and LLava-Video and tested 7B, 32B and 72B models. Only zero-shot inference were performed on the 26,550 questions provided without fine-tuning using any of the benchmark dataset and other egocentric dataset. The overall score averaged over all 30 categories shows that Qwen2.5VL is slightly higher than Llava-Video in this benchmark. which are 0.342 and 0.321 in accuracy, respectively.

1. Introduction

Vision language models (VLMs) have demonstrated strong ability on video understanding and reasoning tasks. In this report, we show results using two open-sourced VLMs on the HD-EPIC (A Highly-Detailed Egocentric Video Dataset) [2] video question answering benchmark with zero-shot setting. We first briefly give approach details on section 2 and then present our result on section 3.

2. Approach

We perform zero-shot inference on the provided 26,550 questions without fine tuning using the provided dataset and other egocentric dataset. We only used video files provided by the benchmark without including other modalities for inference. When testing Qwen2.5VL [1]-72B model, we have to limit the total pixels to 10240x28x28 because of GPU memory limitation.

We also tested model performance for different frame rates. The original 1408 x 1408 x 30 fps videos were processed by ffmpeg using the official script, generating 3 different frame rates as shown in Fig.3. We then tested the influence of frame rates on model performance for LLaVA-Video [3]. In this test, we set the maximum video length to 32 seconds.

3. Results

Figure 1 shows accuracy for each category for Qwen2.5VL 7B model, 32B and 72B combined (we didn't manage to run all 26,550 questions using only 32B or 72B model because the inference speed is very slow for our hardware setting. We therefore split questions into two parts and infer using 32B and 72B models separately.) and LLaVA-Video 7B, 72B models.

It demonstrates that even 7B model outperforms 32B and 72B models in some categories for both Qwen2.5VL and LLaVA-Video. Besides, Qwen2.5VL and LLaVA-Video have different strengths. The overall score (Fig.2) averaged over all 30 categories shows that Qwen2.5VL is slightly higher than LLava-Video in this benchmark. which are 0.342 and 0.321 in accuracy.

Figure 2 shows the average score over 30 question categories for LLavaVideo-7B models using different frame rates. It can be observed that increasing the frame rate also improves the model performance slightly.

4. Conclusion

In this report, we showed results for HD-EPIC VQA benchmark using two VLMs: Qwen2.5VL and LLaVA-Video. Overall, Qwen2.5VL performs slightly better than LLaVA-Video for the averaged score over all 30 categories, while each VLM has its strength on certain question categories. Additionally, for both VLMs, 7B model outperforms 32B and 72B models in some categories.

References

[1] Keqin Chen Xuejing Liu Jialin Wang Wenbin Ge Sibo Song Kai Dang et al Bai, Shuai. Qwen2. 5-vl technical report. In *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Ahmad Darkhalil Saptarshi Sinha Omar Emara Sam Pollard Kranti Parida Kaiting Liu et al Perrett, Toby. Hd-epic: A highly-detailed egocentric video dataset. In *arXiv preprint arXiv:2502.04144*, 2025. 1

[3] Wei Li Bo Li Zejun Ma Ziwei Liu Yuanhan Zhang, Jinming Wu and Chunyuan Li. Video instruction tuning with synthetic data. In *arXiv preprint arXiv:2410.02713*, 2024. 1

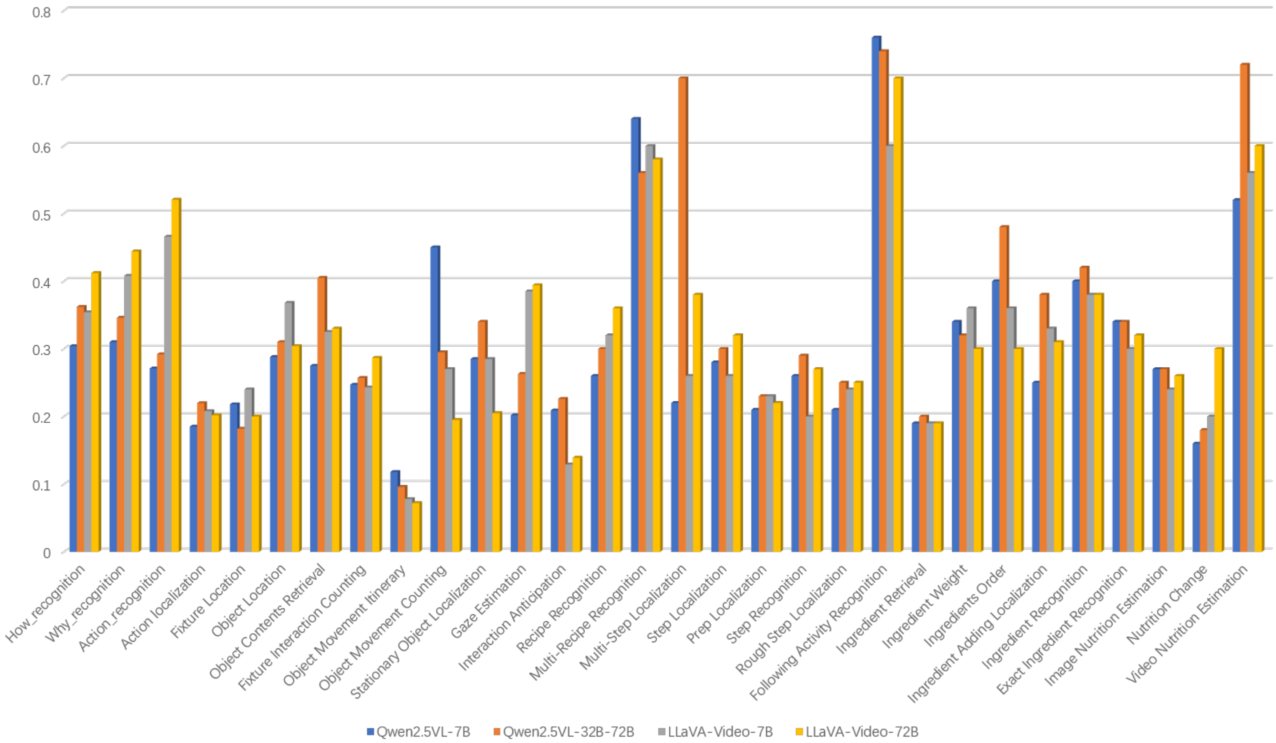
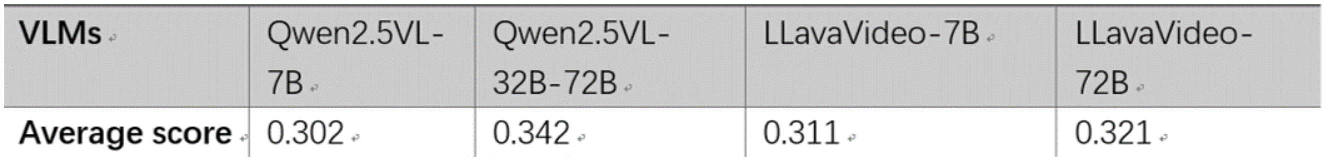
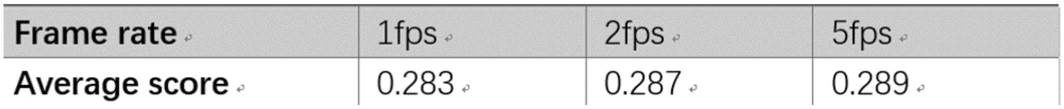


Figure 1. Model results per question category.



VLMs	Qwen2.5VL-7B	Qwen2.5VL-32B-72B	LLavaVideo-7B	LLavaVideo-72B
Average score	0.302	0.342	0.311	0.321

Figure 2. Average scores for different models.



Frame rate	1fps	2fps	5fps
Average score	0.283	0.287	0.289

Figure 3. The influence of different frame rates on the model performance using LLaVA-Video-7B.