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

Yupeng Zhang, Hong Yi
Smart vision group, Ricoh Software Research Center Beijing
3F East building, Genesis Beijing, 8 Xinyuan South Road, Chaoyang District, Beijing 100027, China
Yupeng.Zhang@cn.ricoh.com, Hong.Yi@cn.ricoh.com

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 7 categories shows that Llava-Video is slightly higher than Qwen2.5VL in this benchmark. which are 0.3234 and 0.3214 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 Table 2. We then tested the influence of frame rates on model performance of LLaVA-Video [3]. In this test, we set the maximum video length to 32 seconds.

VLMs	Average score (%)
Qwen2.5VL-7B	28.66
Qwen2.5VL-32B-72B	32.14
LLavaVideo-7B	30.19
LLavaVideo-72B	32.34

Table 1. Average scores for different models.

Frame rate	Average score (%)
1fps	27.66
2fps	28.01
5fps	28.25

Table 2. The influence of different frame rates on the model performance.

3. Results

Figure 1 shows the accuracy for each of the 30 subcategories of questions for Qwen2.5VL 7B model, 32B and 72B combined (we were unable to run all 26,550 questions using only the 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 shows that even 7B model outperforms 32B and 72B models in some categories for both Qwen2.5VL and LLaVA-Video. In addition, Qwen2.5VL and LLaVA-Video have different strengths. The overall score (Table 1) averaged over 7 categories (recipe, ingredient, nutrition, action, 3D, motion and gaze) shows that LLava-Video is slightly higher than Qwen2.5VL on this benchmark. The highest score is obtained by LLavaVideo-72B, which is 32.34 in percentage.

Table 2 shows the average score for seven categories for LLavaVideo-7B models using different frame rates. It can be seen that increasing the frame rate also slightly improves the model performance.

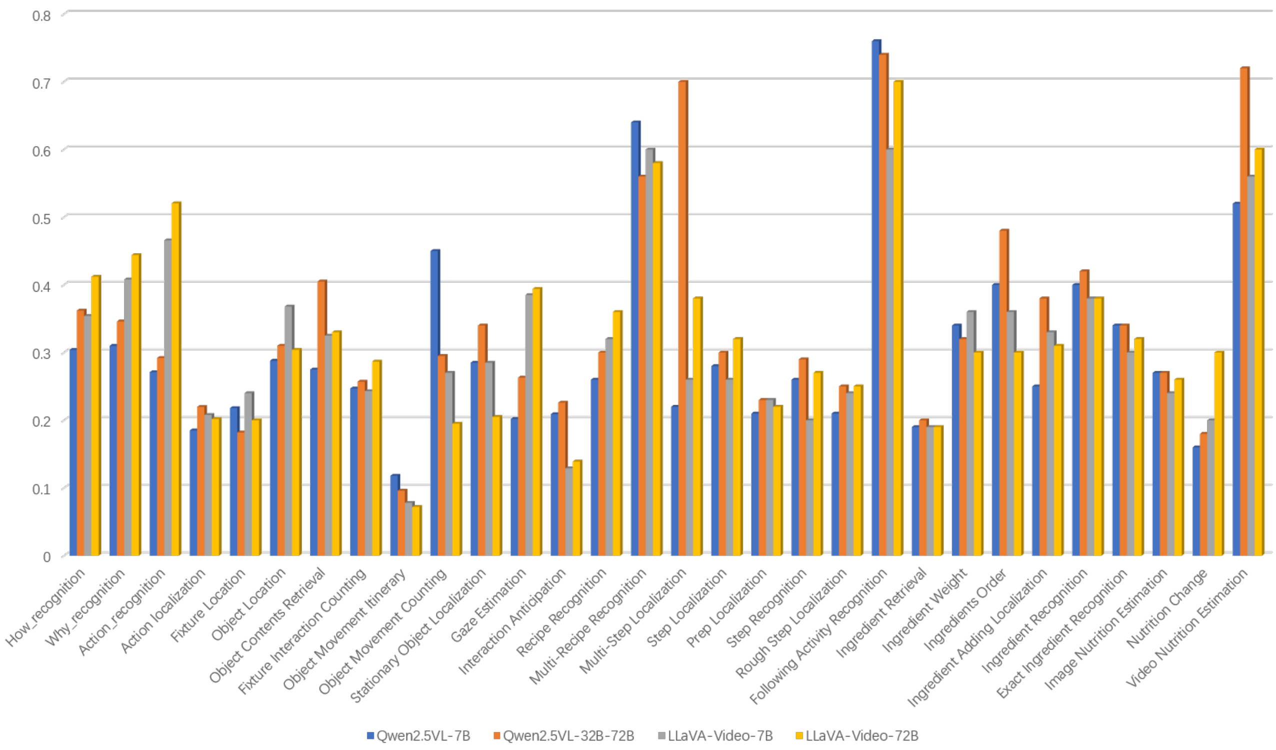

Figure 1. Model results per question category.

4. Conclusion

In this report, we showed results for HD-EPIC VQA benchmark using two VLMs: Qwen2.5VL and LLaVA-Video. Overall, LLaVA-Video performs slightly better than Qwen2.5VL for the averaged score for 7 categories of 26,550 questions, while each VLM has its strength in certain categories. Additionally, for both VLMs, 7B model outperforms 32B and 72B models in some categories.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2. 5-vl technical report. In *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *arXiv preprint arXiv:2502.04144*, 2025. 1

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