Simple Baselines for Interactive Video Retrieval with Questions and Answers (Supplementary Material)

Kaiqu Liang
Princeton University
kl2471@princeton.edu

Samuel Albanie
University of Cambridge
sma71@cam.ac.uk

In this supplementary material, we report results on additional datasets (Appendix A), analyze failure cases of our interactive system (Appendix B), provide additional results on the number of queries (Appendix C), and evaluate our system with user study (Appendix D).

A. Results on additional datasets

Method	ITA	R1 ↑	R5 ↑	R10 ↑	MdR ↓
CLIP4Clip [4]	×	21.6	41.8	49.8	11
BLIP [3]	×	21.3	38.2	45.3	16
BLIP + Auto-text	√	41.2	63.8	73.1	2
BLIP + Auto-text-vid	√	39.1	61.7	70.4	3
BLIP + Heuristic	√	**55.7**	**78.0**	**85.9**	**1**

Table 1: Text-to-video retrieval results on LSMDC dataset. ITA indicates whether it is an interactive retrieval system.

Method	ITA	R1 ↑	R5 ↑	R10 ↑	MdR ↓
CLIP4Clip [4]	×	43.4	70.2	80.6	2
BLIP [3]	×	41.5	67.1	75.3	2
BLIP + Auto-text	√	50.6	77.5	85.5	1
BLIP + Auto-text-vid	√	45.0	70.8	79.2	2
BLIP + Heuristic	√	**61.6**	**86.9**	**91.5**	**1**

Table 2: Text-to-video retrieval results on DiDeMo dataset. ITA indicates whether it is an interactive retrieval system.

Method	ITA	R1 ↑	R5 ↑	R10 ↑	MdR ↓
CLIP4Clip [4]	×	40.5	72.4	-	2
BLIP [3]	×	35.3	60.5	72.4	3
BLIP + Auto-text	√	40.5	65.0	75.0	2
BLIP + Auto-text-vid	√	35.4	59.3	69.9	3
BLIP + Heuristic	√	**44.8**	**70.4**	**79.9**	**2**

Table 3: Text-to-video retrieval results on ActivityNet. ITA indicates whether it is an interactive retrieval system.

We further evaluated our method using three additional widely recognized video retrieval datasets: LSMDC [5], DiDeMo [1], and ActivityNet [2]. For DiDeMo and ActivityNet, we set up the baseline using the same approach as in CLIP4Clip [4]. In our approach, we used the first sentence of the entire caption as the initial query and iteratively refined it through interaction. The results clearly demonstrate the effectiveness of our approach, as it consistently outperformed the baseline method across all three datasets. Particularly noteworthy is the fact that although the Heuristic question generator is manually designed, it exhibits remarkable robustness.

We noticed that the improvement on ActivityNet is not as pronounced as in the other datasets, which can be attributed to the limitations of BLIP [3]. BLIP's lack of pre-training on video datasets, as well as its inability to utilize temporal information, pose significant challenges in long-form video understanding within the zero-shot retrieval setting. Nevertheless, our approach still demonstrates a substantial improvement over the baseline, showcasing its effectiveness.

B. Analyze failure cases

It has been observed that the failures in the performance of VideoQA models can be attributed to the limitations of both the question and answer generators. To illustrate, in the first example depicted in Fig. 1, the VideoQA model fails to identify the object in the video (the object is not a toothbrush). In the second example, the question generated is misleading as the woman in the video is not talking on the phone. Furthermore, the generated questions are at times not meaningful, as the answer to the question is already present in the initial query. In such cases, the performance of our system remains unchanged or may even slightly deteriorate.

C. The influence of the number of queries (recall@5)

To supplement the main paper, we provide additional results illustrating how the number of queries influences re-

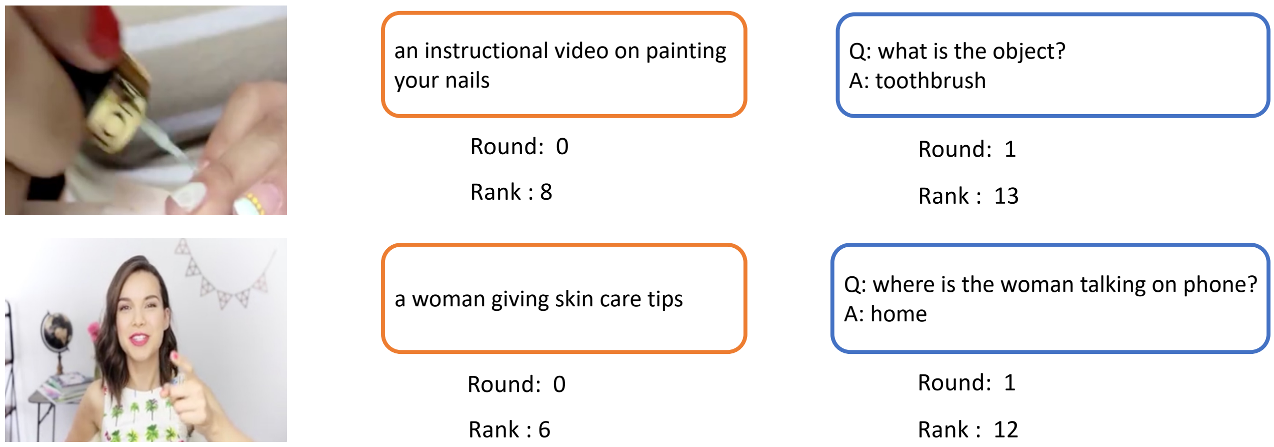
an instructional video on painting your nails
Round: 0
Rank : 8
Q: what is the object?
A: toothbrush
Round: 1
Rank : 13
a woman giving skin care tips
Round: 0
Rank : 6
Q: where is the woman talking on phone?
A: home
Round: 1
Rank : 12

Figure 1: **Qualitative results of the failure cases using our Q & A System on MSR-VTT.** The questions are generated by the Heuristic Question Generator. The initial queries of the target videos are shown in orange boxes. The questions and answers generated through each interaction are shown in blue boxes. Both the initial retrieval rank and the rank after the interaction are demonstrated.

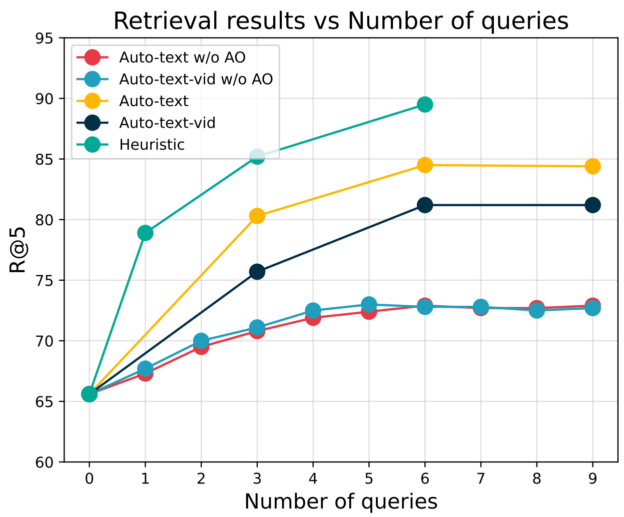

Figure 2: **Recall@5 vs number of queries on MSRVTT.** The performance of all approaches improves as the number of queries increases and then stabilizes as the number of queries grows beyond 6. The maximum number of queries for heuristic question generation is fixed to be 6.

trieval performance under a recall@5 metric (Fig. 2). We observe a similar trend to the recall@1 results reported in the main paper.

Method	VideoQA	CAP+LM	User
Heuristic	66.0	60.0	69.0
Auto-text	66.0	62.0	68.0
Auto-text-vid	62.0	56.0	66.0

(a) Recall@1

Method	VideoQA	CAP+LM	User
Heuristic	90.0	84.0	91.5
Auto-text	80.0	78.0	84.5
Auto-text-vid	82.0	76.0	84.0

(b) Recall@5

Method	VideoQA	CAP+LM	User
Heuristic	96.0	90.0	95.0
Auto-text	88.0	86.0	90.0
Auto-text-vid	88.0	88.0	91.0

(c) Recall@10

Table 1: **Comparing VideoQA answers to human answers on a selection of 50 randomly sampled videos in MSR-VTT**. Recall metrics are reported as percent. The retrieval results are averaged across 4 different users. We observe that VideoQA performs similarly to human answers.

D. User Study

We conducted a user study to confirm that our VideoQA model provides a reasonable simulation of human responses. Four volunteers (graduate students) were invited to participate in our interactive video retrieval experiments. For this study, we randomly selected 50 videos from the MSR-VTT dataset. Each participant watched the videos and provided short answers to the generated questions. All three question generation approaches (Heuristic, Auto-text, and Auto-text-vid) were considered. The final retrieval results are averaged across 4 different users.

As seen in Tab. 1, the VideoQA model obtains similar performance to the user, indicating that the VideoQA model provides a reasonable approximation. In addition, the CAP+LM model consistently performs worse than the VideoQA model. This result is expected since when a system generates a caption, it does not know what the question will be. As such, it is likely to miss relevant details. However, the VideoQA model (which incorporates the information from both videos and questions) addresses this issue.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[2] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[4] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[5] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.