RETRIEVAL-BASED VIDEO LANGUAGE MODEL FOR EFFICIENT LONG VIDEO QUESTION ANSWERING
Supplementary Material

Anonymous authors
Paper under double-blind review

A ABLATION STUDY ON INFLUENCE OF THE HYPERPARAMETER K

We study the influence of K. In general, when K is too small, it may lead to a loss of information necessary to answer the question. When K is too large, interference may be introduced, confusing the LLM. Table 5 shows the results of using different K. We found that as K gradually increases from 1 to 5, the average performance increases. When K increases from 5 to 7, the performance decreases. We found K=5 presents a good trade-off on most datasets, even though there are slight differences on different datasets. We leave the adaptive design of K as the future work.

Table 5: Ablation study on the influence of K, evaluated in terms of accuracy (%)/score. We use bold to mark the best performance and underline to mark the second-best performance.

Dataset	Video-ChatGPT	Ours(K=1)	Ours(K=3)	Ours(K=5)	Ours(K=7)
WildQA	58.00/3.30	57.45/3.18	60.58/3.31	**64.82/3.39**	63.44/**3.39**
QaEgo4D	29.74/2.43	32.42/2.41	32.04/2.42	32.51/**2.45**	**32.81**/2.42
lifeQA	33.87/2.55	37.63/2.62	38.44/2.62	**38.71**/2.61	37.63/**2.65**
Social-IQ 2.0	57.73/3.26	**63.92/3.44**	60.89/3.34	63.65/3.40	62.89/3.34
Average	44.84/2.89	47.86/2.91	47.99/2.92	**49.92/2.96**	49.19/2.95

B MORE VISUALIZATION RESULTS

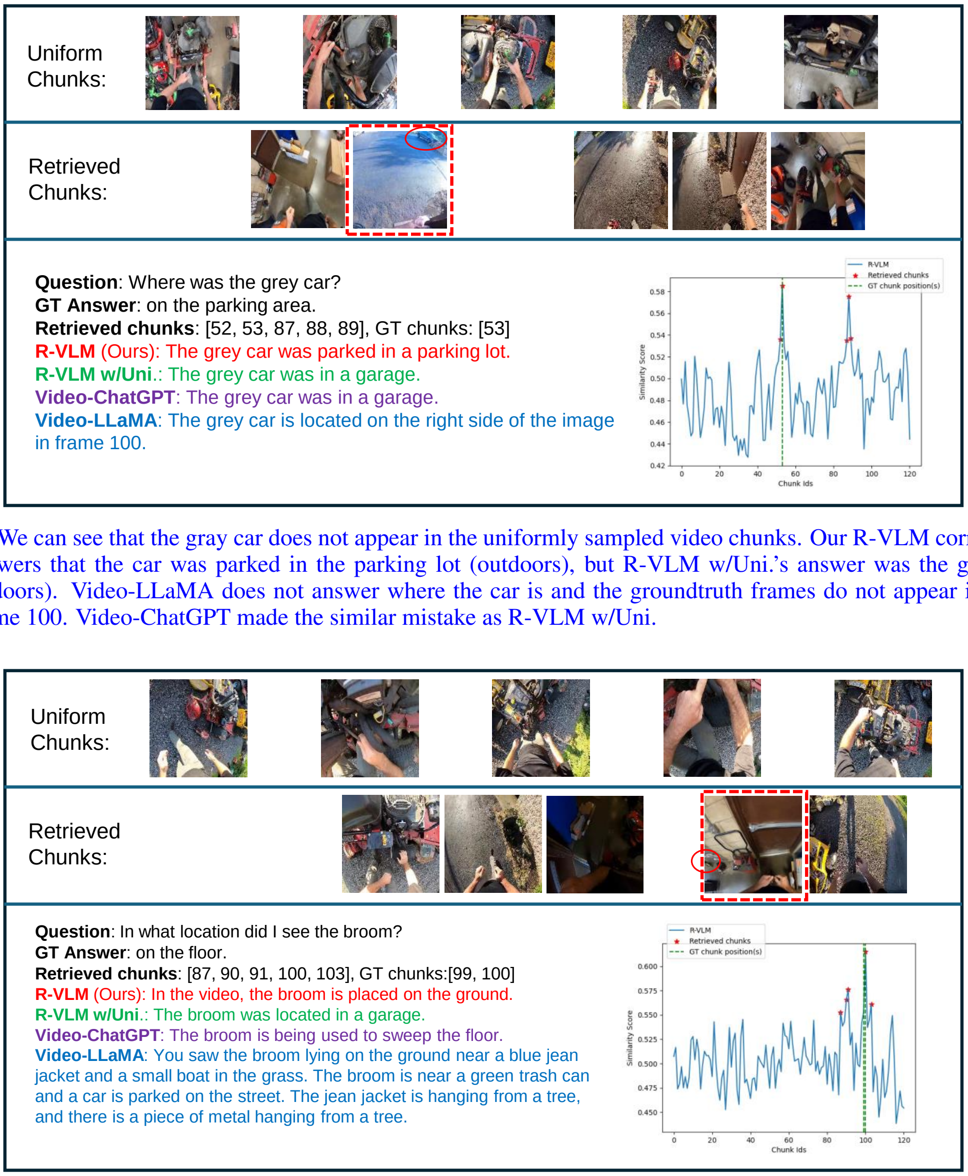
We visualize more examples from the QAEgo4D and WildQA datasets in Fig. 3 and Fig. 4, with the following information. 1) The first row shows the video chunk samples by uniformly selecting 5 video chunks. 2) The second row shows the retrieved 5 chunks (ordered by time order) in our *R-VLM*. We mark the groudtruth chunks by red box. 3) We show the learned similarity score curve based on which the top K chunks are selected. The horizontal axis represents the identity of chunk and the vertical axis denotes the similarity score of that chunk. The groundtruth chunks and our retrieved chunks are also marked. 4) The question and answers from different models: *R-VLM*, *R-VLM w/Uni.*, *Video-ChatGPT*, and *Video-LLaMA*, respectively.

We also show some failure examples in Fig. 5. A detailed analysis of the reasons for failure is given in the figure caption. There are two main cases of failure. One is that the retrieval does not select the correct video chunks. The other is that the retrieval correctly identified the correct video chunks, but the answer is wrong. For the later cases, we think more powerful vision feature extractor and LLMs would alleviate the problem.

C COMPUTATIONAL COMPLEXITY

The computational cost comes from two parts. The first part is to encode the video frames through the CLIP encoder and the spatial-temporal pooling to get chunks. The second part is the retrieval of K=5 chunks and put them to LLM for inference. The spatial-temporal pooing and retrieval is very fast and negligible. On a single A100, we tested 120 60s videos from Social-IQ 2.0 and calculated the average inference time cost for a video. For a single video, the first part for vision feature extraction takes an average of 0.14s (in parallel for 60 frames), and the second part takes an average of 2.42s. The total time is 2.56s. Actually, for an even longer video, the time consumption for the second part does not increase since the input number of vision tokens is fixed (i.e., 68×5=340) in our scheme, which is favored for long video or streaming video understanding. The GPU memory consumption is about 17GB. Note that the computational cost for the LLM is proportional to the number of tokens.

The FLOPs for LLM inference can be roughly estimated as 2PD, where P denotes the number of parameters (model size), and D denotes the number of tokens. The computational complexity of LLM is proportional to the number of tokens which consists of text tokens (question and answer) and vision tokens. The LLM model size P is 6.7B. On the training dataset, the average number of tokens for question and answers is 80, i.e., $D_{tex} = 80$. This varies on different datasets. For simplicity, we assume the number is the same for all the datasets. We denote the number of vision tokens as D_{vis}. The total number of tokens is $D = D_{tex} + D_{vis}$. For the four video datasets, WildQA, QaEgo4D, lifeQA, Social-IQ 2.0, the average number of vision chunks is 19, 122, 20, and 16, where each chunk has 68 tokens. Thanks to the retrieval, only $K = 5$ chunks ($D'_{vis} = 5 \times 68 = 340$ tokens) instead of all the chunks are needed as the input to LLM. Therefore, the computational cost (FLOPs) for LLM inference can be saved approximately $\frac{D_{vis} - D'_{vis}}{D_{tex} + D_{vis}}$, which are 69% (i.e., $(19 \times 68–5 \times 68)/(80 + 19 \times 68)$), 95%, 71%, and 64%, respectively.



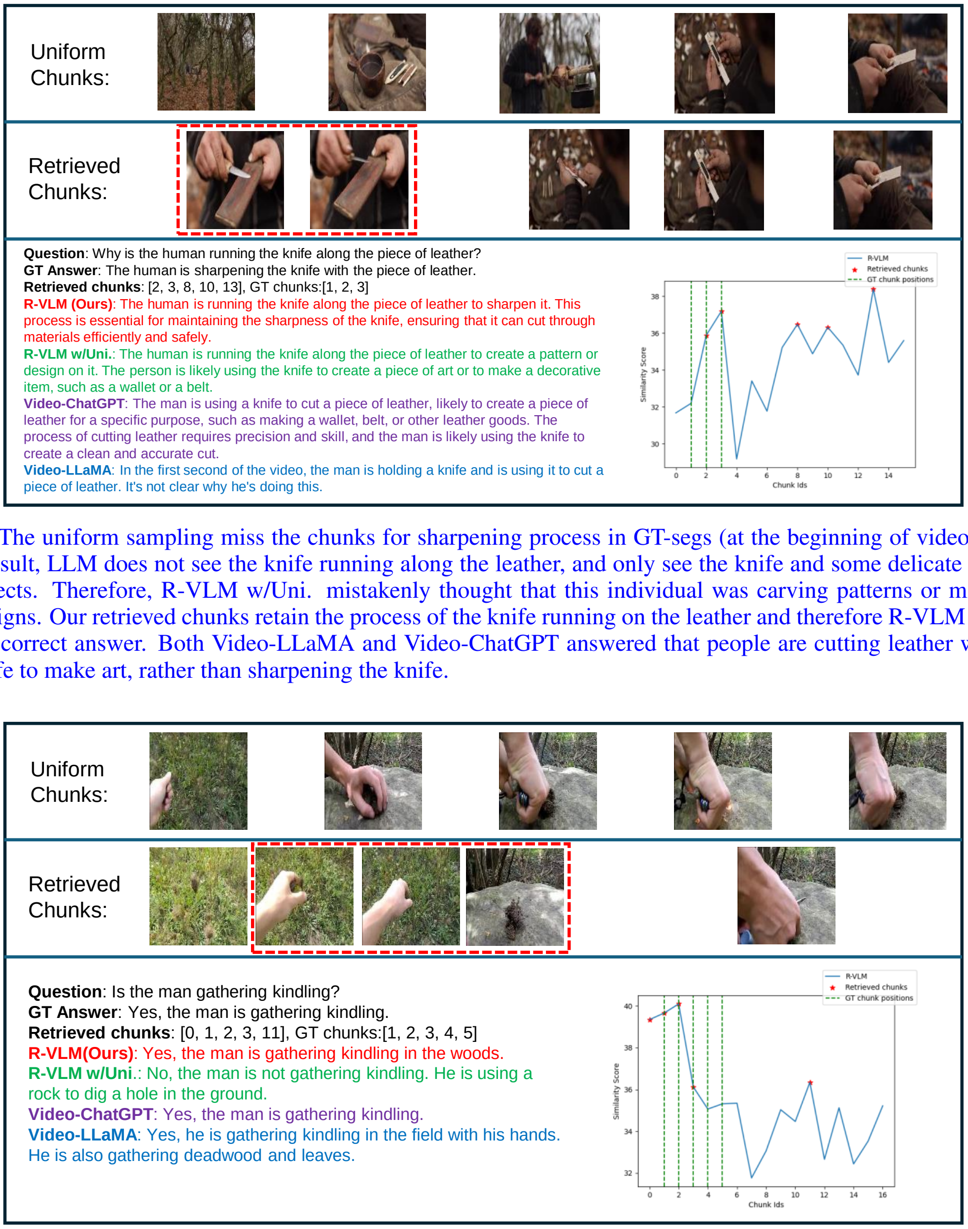
Question: Where was the grey car?
GT Answer: on the parking area.
Retrieved chunks: [52, 53, 87, 88, 89], GT chunks: [53]
R-VLM (Ours): The grey car was parked in a parking lot.
R-VLM w/Uni.: The grey car was in a garage.
Video-ChatGPT: The grey car was in a garage.
Video-LLaMA: The grey car is located on the right side of the image in frame 100.

(a) We can see that the gray car does not appear in the uniformly sampled video chunks. Our R-VLM correctly answers that the car was parked in the parking lot (outdoors), but R-VLM w/Uni.'s answer was the garage (indoors). Video-LLaMA does not answer where the car is and the groundtruth frames do not appear in the frame 100. Video-ChatGPT made the similar mistake as R-VLM w/Uni.

Question: In what location did I see the broom?
GT Answer: on the floor.
Retrieved chunks: [87, 90, 91, 100, 103], GT chunks:[99, 100]
R-VLM (Ours): In the video, the broom is placed on the ground.
R-VLM w/Uni.: The broom was located in a garage.
Video-ChatGPT: The broom is being used to sweep the floor.
Video-LLaMA: You saw the broom lying on the ground near a blue jean jacket and a small boat in the grass. The broom is near a green trash can and a car is parked on the street. The jean jacket is hanging from a tree, and there is a piece of metal hanging from a tree.

(b) The broom is small and is on the left in the red boxed image. Our R-VLM captures exactly where the broom is, i.e., on the ground. R-VLM w/Uni. does not capture the video chunks with broom and thus does not answer accurately. The answer of Video-ChatGPT is irrelevant to the question. The answer from Video-LLaMA is redundancy and tedious, where the mentioned blue jean jacket and boat actually do not appear in the video.

Figure 3: Visualization of video QA examples from QAEgo4D.



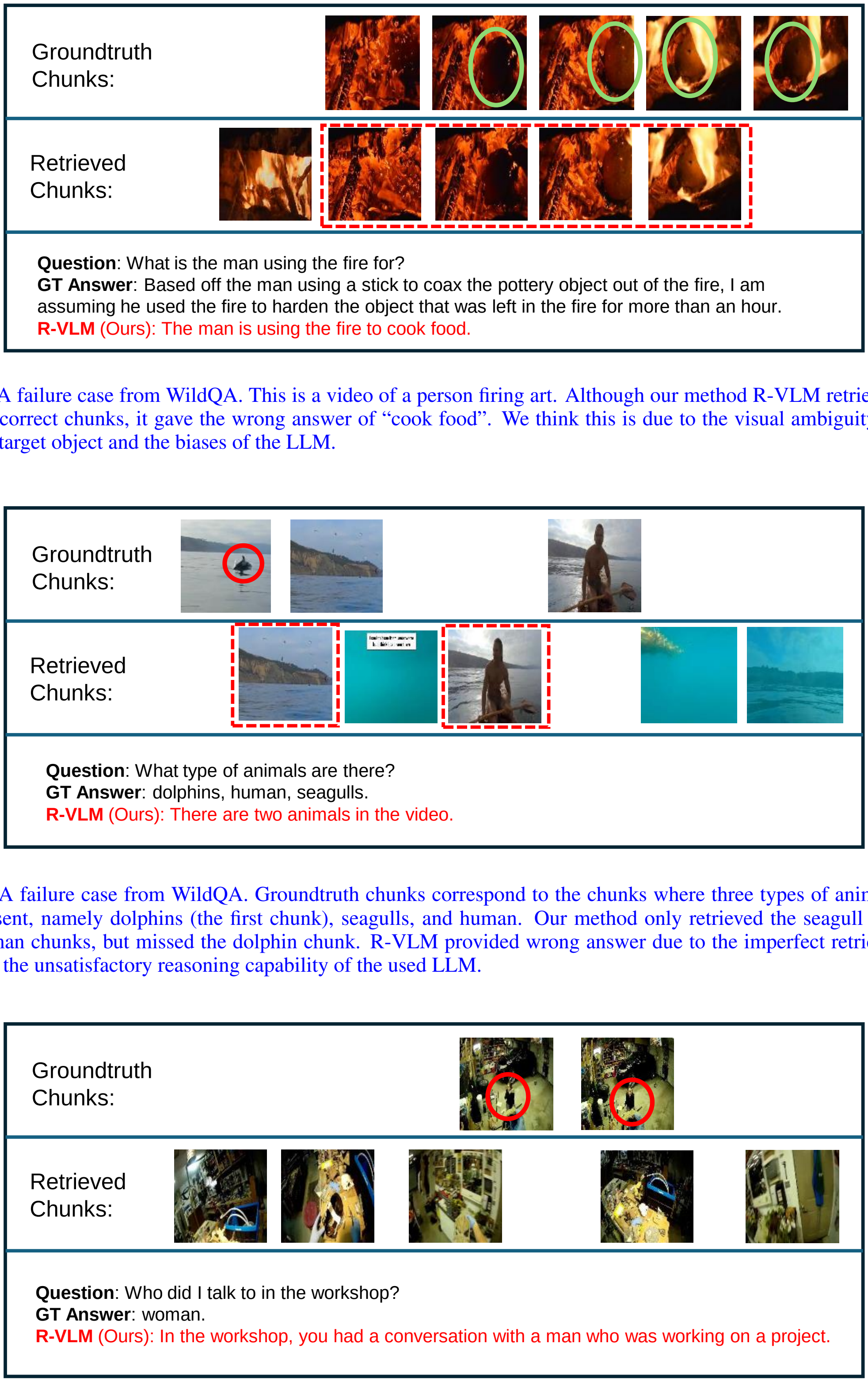
Question: Why is the human running the knife along the piece of leather?
GT Answer: The human is sharpening the knife with the piece of leather.
Retrieved chunks: [2, 3, 8, 10, 13], GT chunks:[1, 2, 3]
R-VLM (Ours): The human is running the knife along the piece of leather to sharpen it. This process is essential for maintaining the sharpness of the knife, ensuring that it can cut through materials efficiently and safely.
R-VLM w/Uni.: The human is running the knife along the piece of leather to create a pattern or design on it. The person is likely using the knife to create a piece of art or to make a decorative item, such as a wallet or a belt.
Video-ChatGPT: The man is using a knife to cut a piece of leather, likely to create a piece of leather for a specific purpose, such as making a wallet, belt, or other leather goods. The process of cutting leather requires precision and skill, and the man is likely using the knife to create a clean and accurate cut.
Video-LLaMA: In the first second of the video, the man is holding a knife and is using it to cut a piece of leather. It's not clear why he's doing this.

Question: Is the man gathering kindling?
GT Answer: Yes, the man is gathering kindling.
Retrieved chunks: [0, 1, 2, 3, 11], GT chunks:[1, 2, 3, 4, 5]
R-VLM(Ours): Yes, the man is gathering kindling in the woods.
R-VLM w/Uni.: No, the man is not gathering kindling. He is using a rock to dig a hole in the ground.
Video-ChatGPT: Yes, the man is gathering kindling.
Video-LLaMA: Yes, he is gathering kindling in the field with his hands. He is also gathering deadwood and leaves.

(a) The uniform sampling miss the chunks for sharpening process in GT-segs (at the beginning of video). As a result, LLM does not see the knife running along the leather, and only see the knife and some delicate small objects. Therefore, R-VLM w/Uni. mistakenly thought that this individual was carving patterns or making designs. Our retrieved chunks retain the process of the knife running on the leather and therefore R-VLM gives the correct answer. Both Video-LLaMA and Video-ChatGPT answered that people are cutting leather with a knife to make art, rather than sharpening the knife.

(b) In this video, collecting the kindling takes a short time, while placing the tinder on the stones takes a longer time. Uniform sampling makes LLM think that there is no process of collecting kindling and output the wrong answer of “digging a hole”. Our R-VLM identified the relevant chunks of “gathering” even though those chunks only take a small duration in the entire video, generating correct answer. Video-LLaMA’s prediction is not accurate since in fact the man did not gather deadwood and leaves in the video.

Figure 4: Visualization of video QA examples from WildQA.



Question: What is the man using the fire for?
GT Answer: Based off the man using a stick to coax the pottery object out of the fire, I am assuming he used the fire to harden the object that was left in the fire for more than an hour.
R-VLM (Ours): The man is using the fire to cook food.

(a) A failure case from WildQA. This is a video of a person firing art. Although our method R-VLM retrieved the correct chunks, it gave the wrong answer of "cook food". We think this is due to the visual ambiguity of the target object and the biases of the LLM.

Question: What type of animals are there?
GT Answer: dolphins, human, seagulls.
R-VLM (Ours): There are two animals in the video.

(b) A failure case from WildQA. Groundtruth chunks correspond to the chunks where three types of animals present, namely dolphins (the first chunk), seagulls, and human. Our method only retrieved the seagull and human chunks, but missed the dolphin chunk. R-VLM provided wrong answer due to the imperfect retrieval and the unsatisfactory reasoning capability of the used LLM.

Question: Who did I talk to in the workshop?
GT Answer: woman.
R-VLM (Ours): In the workshop, you had a conversation with a man who was working on a project.

(c) A failure case from QAEgo4D. Our method did not find the correct chunks. Therefore, large language model did not correctly answer the question and provided hallucinated answer.

Figure 5: Visualization of failure cases from WildQA and QAEgo4D.