The supplementary material provides:
- Section A: Ablation on different values of k in Top-k retrieval.
- Section B: Model Hallucination problem.
- Section C: Video LLM in text tasks.
- Section D: Video length robustness.
- Section E: Prompt details.
- Section F: Implementation details.
- Section G: Qualitative results.

A Top K Effect

In this section, we explore how performance through accuracy is affected by the value of k for top k neighbors for the retrieval design in Section 3 of the main paper. From Table 1, we can see that the Top 3 achieved the best results for the "Vision + subtitles" experiments. By employing the general model summary, we observed that the accuracy improved when incorporating information from various neighbors. However, when this information was excessively increased, such as including data from five neighbors, the accuracy declined due to the introduction of noise from numerous incorrect details unrelated to the question. This phenomenon is evident in the first four rows.

From row 5 to 8 we can see that the accuracy decreased by increasing the number of neighbours because the related information from the wrong clips distract the model. We observe the same behavior in the "Vision Only" and "Subtitle Only" experiments.

B Model Hallucinations

The model hallucinates in our case when the VideoLLM is asked questions unrelated to the video, so the videoLLM generates incorrect information which misguides the answer module to answer the right answer.

After retrieving the Top-k clips, our goal is to filter these clips to the single correct one. Theoretically, we could prompt each retrieved clip with the query and filter for which clip produces an answer. A common problem in generative models, we find that the model hallucinates and outputs an answer instead of stating it doesn't have the required information to answer. This issue particularly arises when the clips originate from the same episode. However, we do see that the videoLLM responding with its lack of information to answer the question if the clip is entirely unrelated to the question.

For instance, in the multi-choice questions in TVQA, if the top three retrieved clips are guaranteed that one of them is the correct clip and the other two clips are incorrect, when using the VideoLLM with the wrong clip it will choose a wrong choice and when feeding the other wrong one, it will choose another wrong choice, and when using the correct clip, it may choose the correct choice based on the correct video content or may choose the wrong choice. In both cases the answer module will see the context information has three choices and this

Table 1: Effect of the number of neighbors on TVQA.Where model summary is the summary generated by the video descriptor and the Q_related_info is the new summary that is related to the question

Model Variations	GPT-4 Accuracy (%)	GPT-4 Score
Vision + Subtitles		
Top 1 (Model Summary + Subtitles)	40.66	3.17
Top 2 (Model Summary + Subtitles)	40.89	3.20
Top 3 (Model Summary + Subtitles)	41.78	3.21
Top 5 (Model Summary + Subtitles)	40.12	3.01
Top 1 (Model Summary + Subtitles + Q_related_info)	29.00	2.75
Top 2 (Model Summary + Subtitles + Q_related_info)	28.12	2.71
Top 3 (Model Summary + Subtitles + Q_related_info)	27.72	2.69
Vision Only		
Top 1 (Model Summary)	26.97	2.77
Top 2 (Model Summary)	27.72	2.77
Top 3 (Model Summary)	28.61	2.78
Top 5 (Model Summary)	27.63	2.67
Top 1 (Model Summary + Q_related_Info)	27.83	2.62
Top 2 (Model Summary + Q_related_Info)	26.45	2.63
Top 3 (Model Summary + Q_related_Info)	26.59	2.61
Subtitles Only		
Top 1 (Subtitles)	40.23	3.15
Top 2 (Subtitles)	41.61	3.20
Top 3 (Subtitles)	41.80	3.22
Top 5 (Subtitles)	39.83	3.02

distracts it from answering correctly even if one of them is the correct answer as evidenced by the table 2. the accuracy dropped by around 14 % in the vision and subtitles and dropped by 2 % in the vision only.

Table 2: Effect of model hallucination.Where the model summary is the summary generated by the video descriptor and the Q_related_info is the new summary that is related to the question.

Model Variations	GPT-4 Accuracy (%)	GPT-4 Score
Vision + Subtitles		
Top 3 (Model Summary + Subtitles)	41.78	3.21
Top 3 (Model Summary + Subtitles + Q_related_info)	27.72	2.69
Vision Only		
Top 3 (Model Summary)	28.61	2.78
Top 3 (Model Summary + Q_related_Info)	26.59	2.61

C MiniGPT4-video in Text Tasks

Here, we will see how the fine-tuned version of Llama 2 (our MiniGPT4-video) performs compared to the original Llama2 in the text tasks. We used MiniGPT4-

video as an answer module in the Goldfish system. We can tell from the table 3 that MiniGPT4-video has lost some text skills during vision tasks fine-tuning, so we decided to use the original Llama to get the best performance.

Table 3: Ablation about answer module LLM

Top 3 (Model Summary + Subtitles)	GPT-4 Accuracy (%)	GPT-4 Score
Goldfish with MiniGPT4-video as answer module	35.07	2.93
Goldfish with original Llama2 as answer module	41.78	3.21

D Video Length Robustness.

To evaluate our framework's robustness with extended video lengths, we created three versions of the TVQA dataset by altering the aggregation window. This window compiles long videos from ground-truth short clips that include the answer to a question. Specifically, we combined 5, 10, and 20 clips to produce videos that average between 6, 12, and 24 minutes, respectively. Table 4 illustrates that our framework maintains its robustness regardless of video length, with both retrieval performance and overall accuracy remaining consistent even as video duration increases. These results, detailed in Table 4, are based on an analysis of 5% of the TVQA validation set.

Table 4: Ablation study about the video length impact on 5% of TVQA validation set.

Video Length	Retrieval Acc.	Overall Acc.
5-6 Min	60.2	40.8
10-12 Min	60.2	41.3
20-30 Min	60.2	40.8

E Prompts Details

Evaluation prompts. We followed the same evaluation setting in videochatgpt [2]. The {question}, {answer}, and {pred} correspond to the question, the ground truth answer, and the model prediction, respectively, in the prompt. The **System prompt** is as follows:

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:
INSTRUCTIONS:
- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

User prompt:
Please evaluate the following video-based question-answer pair:

Question: {question}
Correct Answer: {answer}
Predicted Answer: {pred}

Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where the value of 'pred' is a string of 'yes' or 'no' and the value of 'score' is an INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'pred': 'yes', 'score': 4.8}.

Summary prompts. Below is the summary prompt to obtain the vision summary of the clip:

Generate a description of this video. Pay close attention to the objects, actions, emotions portrayed in the video, providing a vivid description of key moments. Specify any visual cues or elements that stand out.

Extract the related information prompt : In the multi-choice questions, we added the choice *"I don't know"* as the fifth choice, and the {question} is a placeholder for the question itself in the prompt. The prompt is as follows:

From this video extract the related information to This multichioce question and provide an explaination for your answer and If you don't know the answer, say 'I DON'T KNOW' as option 5 because maybe the questoin is not related to the video content. the question is: {question} your answer:

F Implementation Details

Our models are trained with 4 A100 GPUs. The training process involved three distinct stages, with specific durations allocated to each. The initial stage focused on image-text training and spanned a period of two days. Subsequently, the second stage, dedicated to pre-training with video captions datasets, lasted one day, followed by the third stage, involving instruction tuning, which extended over three days. Throughout these stages, we maintained a batch size of 4 and utilized the AdamW optimizer in conjunction with a cosine learning rate scheduler, setting the learning rate to 1e-4.

Our visual backbone consisted of the EVA-CLIP V1 [3] architecture, with the frozen weights. Notably, we trained the linear projection layer and performed efficient fine-tuning of the language model using LoRA [1] (*Low-Rank Adaptation*). Specifically, we fine-tuned the W_q and W_v components with a rank (r) of 64 and a LoRA-alpha value equal 16. The entire model was trained with a consistent image resolution of 224×224 pixels, ensuring uniformity across all stages.

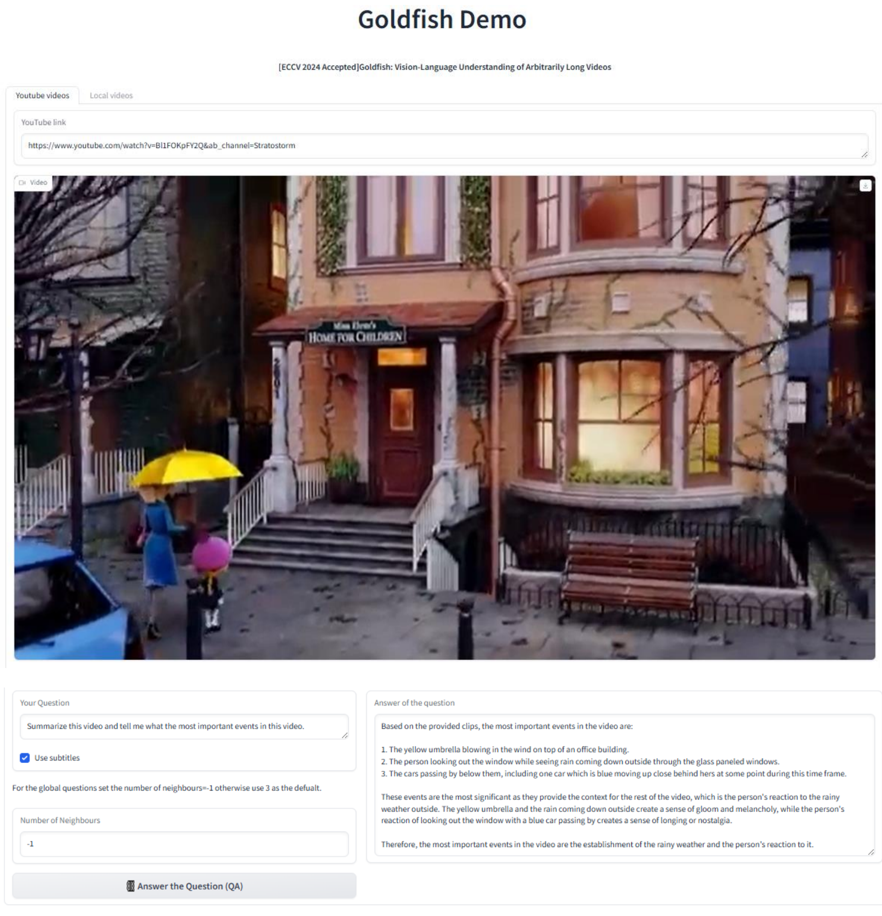

Fig. 1: Qualitative result of Goldfish model for long video understanding. [link].

G Qualitative Results

G.1 Long Video

Fig 1 and Fig 2 shows one example of the goldfish demo. Please refer to this link for more qualitative video demos.

G.2 Short Video

Figs. 3 and 4 demonstrate qualitative results of our model MiniGPT4-video on in-the-wild online videos.

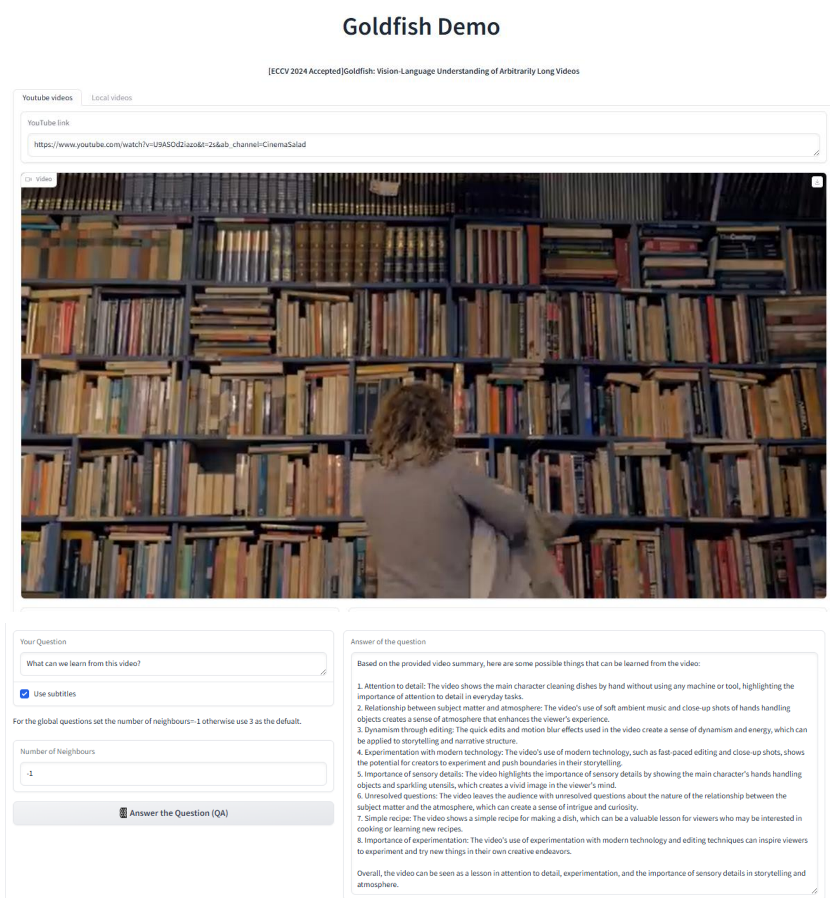

Fig. 2: Qualitative result of Goldfish model for long video understanding. [link].

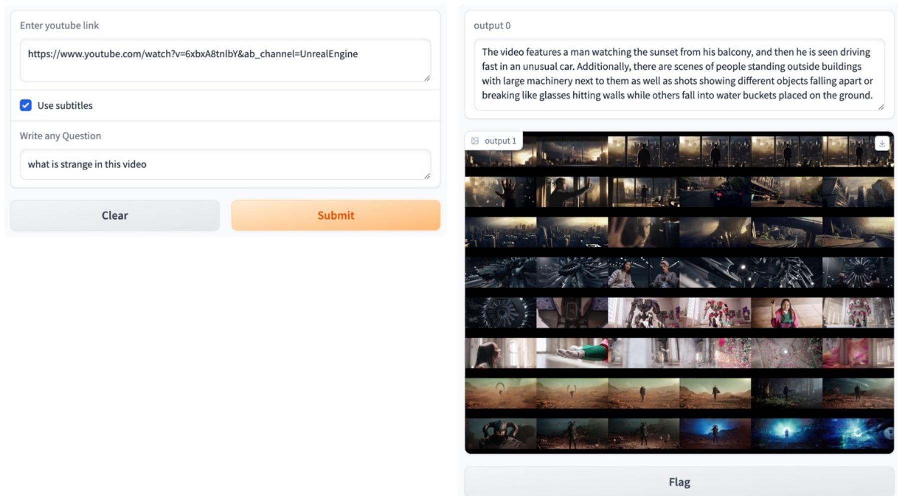

Fig. 3: Qualitative result of short video understanding with in-the-wild video [link].

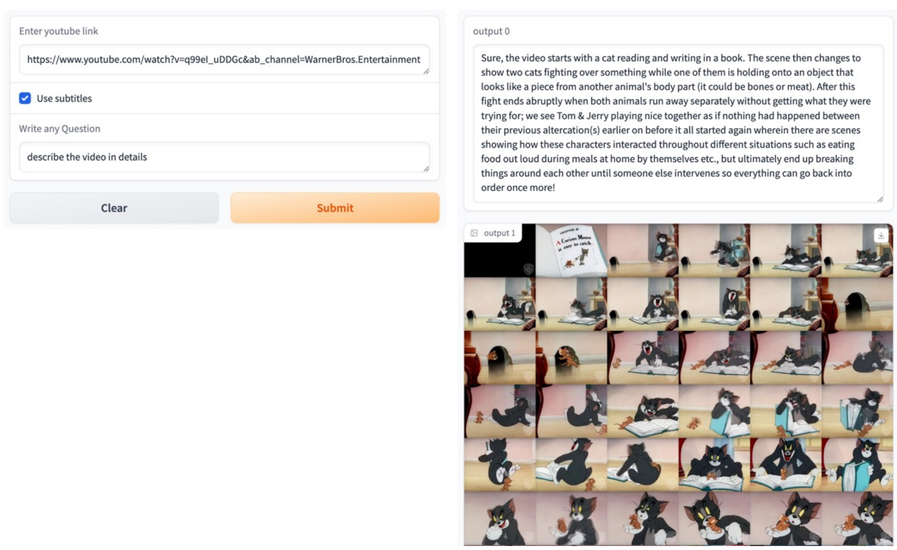

Fig. 4: Qualitative result of short video understanding with in-the-wild video [link].

Question: Generate a creative advertisement feature about this product that can be used on the product website.

VideoChatGPT response:

"Say goodbye to sunburn and hello to a glowing, healthy complexion with our new sunscreen product. With its lightweight formula and SPF 50+ protection, it's perfect for everyday use. Apply it daily to your face and neck for a long-lasting barrier against harmful UV rays. Don't let sun damage ruin your skin choose our sunscreen and enjoy the benefits of a healthy, radiant complexion. Get your glow on with our sunscreen today!"

VideoLLM (Ours)

Title: "Sunshine on Your Skin: Experience the Ultimate Moisturizing Journey"
[Scene: A beautiful beach with a person lying on the sand, enjoying the warmth of the sun on their skin. Suddenly, a bottle of moisturizer appears, and the person applies it to their skin, feeling the cool and refreshing sensation. The camera zooms in on the bottle, highlighting the product's name and key ingredients.]
[Scene: A close-up shot of the person's skin, now glowing with health and radiance. The camera then transitions to a text that appears on the screen, describing the product's benefits and unique features. The text is accompanied by a call-to-action, encouraging viewers to try the product for themselves.]
[Scene: A person is shown sitting on a couch, rubbing the moisturizer into their skin, feeling the soothing and nourishing sensation. The camera then cuts to a shot of the product bottle, with the text "Sunshine on Your Skin" appearing on the screen. The text is accompanied by a tagline that reads, "Experience the ultimate moisturizing journey."]
[Scene: The person is shown enjoying the benefits of the moisturizer throughout the day, with shots of them applying it to their skin in different settings, such as in the bathroom, at their desk, and outdoors. The camera captures their satisfaction and confidence, highlighting the product's effectiveness in providing long-lasting moisture and nourishment to the skin.]
[Scene: The final shot is a close-up of the product bottle, with the words "Sunshine on Your Skin" and "Experience the Ultimate Moisturizing Journey" appearing on the screen. The camera then fades to black, leaving viewers with a lasting impression of the product's benefits and the promise of a radiant, healthy-looking.complexion.]

Fig. 5: Short model qualitative results Ours vs VideoChatGPT [2].

References

1. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
2. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2023)
3. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)