From Trial to Triumph: Advancing Long Video Understanding via Visual Context Sample Scaling and Self-reward Alignment

Supplementary Material

1. Appendix

1.1. Implementation details

When reproducing the benchmark results, we follow Video-ollama2 to use a special prompt for the Egoschema dataset. Specifically, we enclose the option letters in brackets to prevent confusion with the option content. When adjusting the score weights on different datasets, we find that the confidence score weight α plays a vital role in local questions. Therefore, we set α to 2 for CG-Bench and LVBench, two datasets containing a high percentage of local questions. The value of β for different datasets ranges from 1 to 4. The GPU memory usage is about 40GB when inferencing 32 frames on 7B models.

1.2. Additional statistics

To further verify the effectiveness of the global and local questions, we provide detailed statistics on the Video-MME benchmark using Llava-Video. Fig. S1 left lane shows the percentage of global and local questions analyzed by Deepseek-R1 predictions. The right figure shows the percentage of short (shorter than 3 minutes), medium (4-15 minutes), and long video (longer than 15 minutes)questions that need a split decision. We find that the model tends to

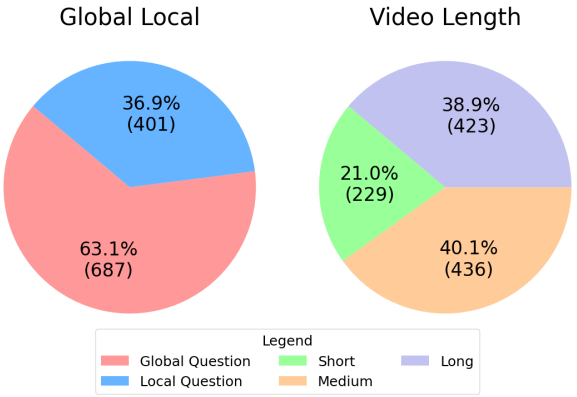

Figure S1. Global local questions and video length percentage in questions with divergent answers.

easily reach unanimous in short video questions, as short videos contain less redundant visual information. The accuracy of split decision in each type of video questions are reported in Tab. S1. Comparing with random select an option, our method consistently acquires higher accuracy in various types of questions. The improvements on global and local questions indicates the effectiveness of clue-based answering for global questions and temporal self-refocus for local questions. Moreover, our method is robust to different length of video.

Method	Global	Local	Short	Medium	Long
random	36.68%	36.15%	40.61%	33.94%	35.46%
Ours	**51.53%**	**52.12%**	**58.95%**	**51.15%**	**48.46%**

Table S1. Prediction selection accuracy compared with random selection.

1.3. Exploration on language model sampling

In the coverage test, we wish to test the importance of visual context. Therefore, all predictions follow the setting in every benchmark and video model that do not use language model sampling, *i.e.*, setting the temperature and top-p hyperparameter to 0. This is because enabling the sample will not promise the option with the largest possibility will always be picked as the final prediction. To further evaluate how the language model sampling affects the performance, we test the performance of majority voting and our method on the Video-MME benchmark using the Llava-Video model with various top-p and temperature in Fig. S2. The result shows that higher temperature and top-p leads

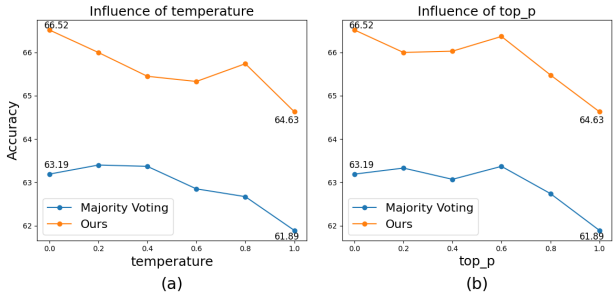

Figure S2. Influence of temperature and top-p.

to a lower performance of both methods. This is because the frequency of each option is disturbed when introducing randomness and may not reflect the real inherent inclinement of the video model. Moreover, with a higher top-p and temperature, we find that the model sometimes does not answer from the option sets, leading to potential hallucinations. However, our method consistently outperforms majority voting with an average improvement of $+3\%$, demonstrating the robustness towards randomness.

1.4. Prompt template

We provide the following detailed prompt templates.

Prompts for generating refocus questions:

Given a question, output a complementary question to check whether the current video clip meets the condition for the question. The complementary question can be answered by yes or no. For instance, given the question What is the standing man in the black shirt doing? The complementary question is: Is the video showing a standing man wearing black shirt? Now, output the complementary question for the following question: {question}. You only respond with the complementary question.

Answer the following question with the best option letter. {question} The best answer is:

Prompts for question categorization:

What type does the following question and options of a video belong to: {question}. You can choose between: 1. perception .(based on observing a scene of the video, for instance what are the man in black and the woman in pink doing?) 2. reasoning (questions like summarizing the video, for instance: What is the video about? What can we infer from the video?) 3. Counting (for instance How many cats appear in the video?) 4. Elimination. (for instance Which of the following item in not mentioned in the video?) 5. Ordering. (for instance Which of the following item in not mentioned in the video?). You only answer the question type.

Prompts for clue-based question answering (prompt for Deepseek-R1):

Given the information of a video consisting of different parts: {video clues}. Answer the best option of the question based on the information. The question is question. If no option perfectly matches the option, then select the most related one. You only response the option letter.

Prompts for global question narration instruction:

Transform a question into the information for answering the question. An example is the question "Who is the third billionaire who appeared in the video?" Considering to answer the question, we need to know each billionaire that appears in the video, so the question should be transformed into "the billionaire appeared". Now directly output the transformed result for the following question: {question}.

Transform a question into the information for answering the question. An example is the question "Which billionaire is not appeared in the video?" Considering to answer the question, we need to know each billionaire that appears in video, so the question should be transformed into "the billionaire appeared". Now directly output the transformed result for the following question: {question}.

Transform a question into the information for answering the question. An example is the question " How many red socks are above the fireplace at the end of this video?" Considering to answer the question, we need to know the number of red socks, so the question should be transformed into " the number of red socks". Another example is the question "In the first game, when the challenger scores for the first time, how many points has Bugs & Daffy scored?", to answer the question, we need to know the points of both Bugs & Daffy and the challenger, so that we can decide when the challenger scores for the first time, how many points has Bugs & Daffy scored. Therefore the transformed results are " the points of both Bugs & Daffy and the points of the challenger" Now directly output the transformed result for the following question: {question}.

Describe everything related to {transformed information} in the current video. Your response should not be longer than 250 words.