M-LLM Based Video Frame Selection for Efficient Video Understanding

Supplementary Material

A. Prompt for Pseudo Label Generation

Table 10 provides the prompt template for generating pseudo spatial labels. We uniformly sample $n = 128$ frames and use the prompt template to obtain a score for each frame independently. We use the logit to generate the word "True" or "False" after "Evaluation:" to compute the score using Equation 6. In a few cases, the M-LLM response may not follow the instruction and does not contain the text "Evaluation: True" or "Evaluation: False". We manually add 'Evaluation: True" to the end of the response and use the logit to generate the word "True" to compute the score.

The image is a video frame from a video. A question about the video is: {question} Evaluate whether the video frame provides useful information to answer this question about the video. First explain your reasoning. Then generate a Boolean evaluation of the frame's usefulness. For example: Evaluation: True

Table 10. Prompt template for spatial pseudo labels

Table 11 provides the prompt template for generating pseudo temporal labels. We first use the M-LLM to generate a concise caption for $n = 128$ uniformly sampled frames. Then we use the prompt in Table 11 to generate a list of frame indexes containing the most helpful frames.

I need to answer a question based on a long video. To do this, I have uniformly sampled 128 frames from the video, each with a corresponding caption. The question I need to answer is: {question} Below is the list of frames and their captions: Frame 1 : {caption1} Frame 2 : {caption2} . . . Frame 128 : {caption128} Please provide a list of 8 frames that would be most helpful for answering this question. Rule: ONLY provide a Python List without extra text.

Table 11. Prompt template for temporal pseudo labels

M-LLM	ANet-QA	NExT-QA
No pseudo-labels	53.5	62.4
LLaVA-NeXT 7B	53.9	62.8
Idefics2 8B	53.8	63.2
Qwen2 VL 7B	54.2	63.6

Table 12. Performance of LLaVA-NeXT-Video 7B on ActivityNet (ANet) and NExT QA with different spatial pseudo-labels

M-LLM	EgoSchema	LongVideoBench
Uniform 4 frames	45.8	45.3
$16 \rightarrow 4$	47.8	46.0
$32 \rightarrow 4$	48.2	48.9
$128 \rightarrow 4$	49.0	49.5

Table 13. Performance of selecting different number of frames on EgoSchema and LongVideoBench with LLaVA-NeXT-Video 34B as downstream video-LLM.

B. Additional Results

Pseudo Label Geneation with different M-LLM Qwen2-VL serves as the prompting M-LLM for spatial pseudo-labels generation as detailed in Section 3.3. We investigate the influence of alternative M-LLMs on video QA performance. Table 12 compares the performance of LLaVA-Next-Video 7B on ActivityNet and NExt-QA using frames selected based on spatial pseudo-labels generated by different prompting M-LLMs. A stronger M-LLM produces higher-quality pseudo-labels.

Number of frames before selection Existing frame selection methods [37, 57] typically sample $16 \sim 32$ frames from a video and then perform frame selection on these frames. In contrast, our method samples a significantly larger list of 128 frames prior to the frame selection process. We posit that a larger number of frames is essential for long video QA. To evaluate this, we assessed the video QA performance of LLaVA-NeXT-Video 34B taking 4 frames selected from different number of frames. Table 13 summarizes the results on the long-video QA benchmarks EgoSchema and LongVideoBench. The improvement on QA performance from $16 \rightarrow 4$ to $128 \rightarrow 4$ is significant, showing the necessity of have a large frame selection candidate pool.

C. More Visualization Results

Figure 5 and Figure 6 are the zoom-in for Figure 4. Figure 7 and Figure 8 are additional visualization results.

Four frames from the video using uniform sampling

Four frames sampled using the frame selector

Question about the video: What does the boy do after turning his head at the head?
Answer: wave at girl

Figure 5. One visualization example of the frame selection results.

Four frames from the video using uniform sampling

Four frames sampled using the frame selector

Question about the video: Why did the man with the cap move his hands at the start?
Answer: drink water

Figure 6. One visualization example of the frame selection results.

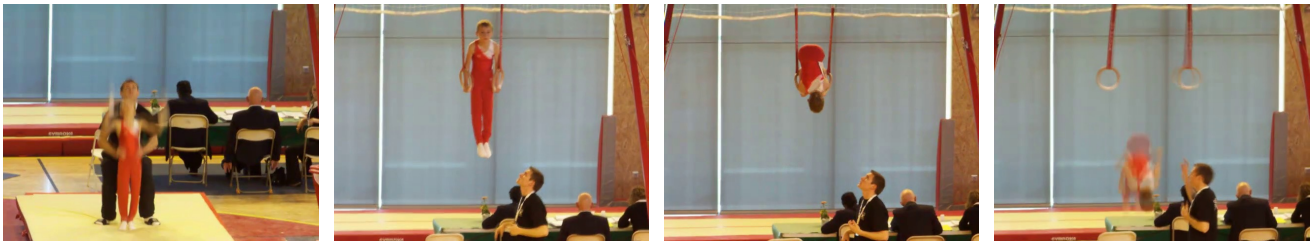
Four frames from the video using uniform sampling

Four frames sampled using the frame selector

Question about the video: What did the boy in brown do after he looked at the boy in red at the beginning of the video? Answer: point at the boy

Figure 7. One visualization example of the frame selection results.

Four frames from the video using uniform sampling



Four frames sampled using the frame selector

Question about the video: What did the man in black do after lifting the boy?
Answer: stand to the side

Figure 8. One visualization example of the frame selection results.