Efficient Frame Selection for Long Video Understanding via Reinforcement Learning

Supplementary Material

1. Selector Architecture

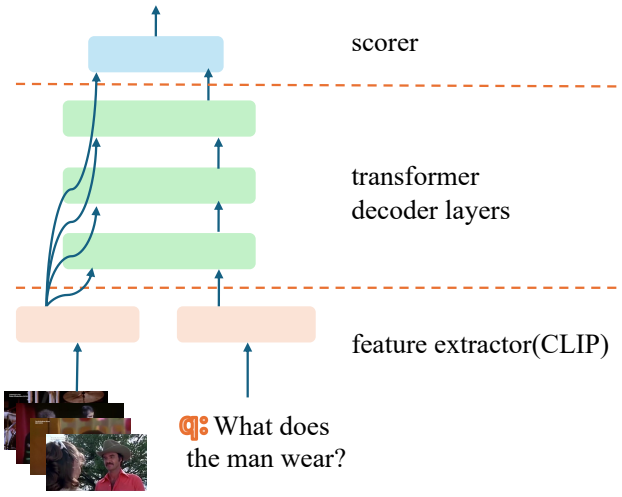

Figure 1. The overview of the selector architecture.

As shown in Fig. 1, the selector consists of three components: a feature extractor, a contextual feature model, and a scorer. The feature extractor is CLIP ViT-L/14, which we use to obtain high-quality embeddings.We adopt a contextual feature model consisting of three decoder layers. This module is designed to further capture interactions between the query and visual representations, enabling richer context aggregation and improved alignment. The scorer is implemented as a simple key–query (K–Q) attention module which outputs the final selection score for each frame.

As shown above, the proposed selector is compact and lightweight, introduces a small additional computational overhead.

2. Comparision with Existing Selection Methods on Shorter Videos

To further illustrate the performance of our selector on the shorter videos, we report both accuracy and inference time under different frame selection methods on EgoSchema. As shown in Table 1, dealing with videos with less temporal redundancy, we still achieve the best performance. Our method attains comparable inference time to CLIP-Top$K$, while outperforming the strong competitor AKS by 0.8 percentage points in accuracy.

The reported inference time covers only the frame selection stage (i.e., computing the selected frame indices per video) on a single L20 GPU.It excludes the downstream MLLM inference, which is identical across methods.

Table 1. Performance and frame selection time of Qwen2-VL-7B on EgoSchema with different frame selection methods.

Method	Accuracy(%)	Frame selection time(s)
Random	53.9	/
Uniform	56.5	/
CLIP-Top$K$	56.6	2.10
AKS	57.2	6.44
Ours	**58.0**	2.13

3. Ablation Study on the Weighting Factors of the Hierarchical Reward

Table 2. Performance of Qwen2-VL-7B on LongVideoBench and EgoSchema with different weighting factors of the hierarchical reward.

λ_{f}	λ_{comb}	LongVideoBench	EgoSchema
0.80	0.20	55.7	56.5
0.67	0.33	56.0	56.7
0.50	0.50	**56.5**	**58.0**
0.33	0.67	55.6	56.7
0.20	0.80	55.6	56.4

We study the impact of the hierarchical reward weighting by varying the frame-level weight λ_{f} and the combination-level weight λ_{comb} under five ratios: 4:1, 2:1, 1:1, 1:2 and 1:4. As shown in Table 2, the balanced setting (λ_{f}=0.50, λ_{comb}=0.50) achieves the best performance on both LongVideoBench and EgoSchema. These results suggest that a balanced mixture of frame-level and combination-level supervision provides the most effective guidance, while emphasizing only one source underutilizes useful information.

4. Detailed Analysis on MLVU: Open-Ended and Multiple-Choice

Unlike LongVideoBench, VideoMME and EgoSchema which focus on a single task format. MLVU is a heterogeneous benchmark that includes multiple video question-answering types (e.g., open-ended and multiple-choice). In

Table 3. Main performance comparison on MLVU across multiple-choice (MC) and open-ended (OE) tasks. Each method is evaluated under two selection strategies. Subscene is the QA subcategory within OE.

Method	Size	Selector	MLVU-mc	MLVU-oe	subscene
LLaVA-NeXT-Video	7B	✗	47.7	20.0	27.2
		✓	53.0(5.3↑)	20.1(0.1↑)	27.3(0.1↑)
	34B	✗	55.1	21.5	29.7
		✓	63.9(8.8↑)	23.3(1.8↑)	33.3(3.6↑)
InternVL2	8B	✗	47.0	14.6	17.9
		✓	54.4(7.4↑)	16.3(1.7↑)	21.3(3.4↑)
	40B	✗	50.0	15.1	19.1
		✓	58.2(8.2↑)	17.2(2.1↑)	23.4(4.3↑)
Idefics2	8B	✗	40.9	10.7	22.2
		✓	46.5(5.6↑)	10.4(0.3↓)	20.9(1.3↓)
VideoLLaMA3	7B	✗	63.4	**28.9**	30.0
		✓	**71.8**(8.4↑)	26.8(2.1↓)	32.3(2.3↑)
Qwen2-VL	7B	✗	58.7	22.6	30.6
		✓	66.3(7.6↑)	24.0(1.4↑)	**34.0**(3.4↑)

prior analyses, we reported aggregate scores that mix these types, which can obscure meaningful differences across task formats. To provide a clearer and more faithful assessment, we now present a fine-grained evaluation that separates open-ended and multiple-choice performance. This breakdown allows us to demonstrate the model's ability across diverse query types and highlight strengths that may be masked by aggregate reporting.

As shown in Table 3, across all model families and sizes, introducing the selector yields consistent and substantial gains on MC. While the aggregate OE column shows smaller or occasionally negative changes. This is largely driven by the "Summary" portion of OE, which is not representative for two reasons: our training data does not include video summarization task and each summary query is nearly identical ("summarize this video") providing little task-specific signal and thus having limited diagnostic value.To more faithfully assess OE question answering, we focus on the Subscene subset, where the selector generally improves performance. These results support two conclusions: our approach brings clear improvements on both MC tasks and OE tasks while OE is evaluated on its QA-relevant component; and the selector is an effective mechanism that consistently strengthens MC performance and generally benefits OE QA as well.

5. Fine-grained Performance on Multiple-Choice Questions

We further perform a comprehensive fine-grained performance comparison between our selector and the uniform sampling baseline on multiple-choice questions across various video benchmarks. Specifically, we evaluate both methods on the three datasets LongVideoBench, Video-MME and MLVU, which comprise diverse question categories. The performance on multiple-choice question answering is presented in a class-wise manner in Fig. 2, Fig. 3, and Fig. 4, respectively.

By adaptively selecting frames that are more temporally relevant to the given questions, our method achieves substantial performance improvement over uniform sampling in most of the question categories, particularly on perception and recognition tasks involving complex and fine-grained scene, object and event understanding.

6. Visualization of Selected Frames with Different Frame Selection Methods

To provide a more intuitive comparison among different selection strategies, we visualize the frames selected from the same video by each strategy, as shown in Fig. 5. In this example, the question requires understanding the little girl's expression after the performance. Our frame selector successfully captures the critical moment in the video and selects the corresponding frames, whereas the other selection strategies fail to preserve this critical information and consequently produce incorrect answers. In the Ellen Show example in Fig. 6, CLIP overly focuses on the host and the guest introducing redundant visual evidence while failing to capture what the question is actually asking. While our method selectively identifies keyframes that are effective for the query."

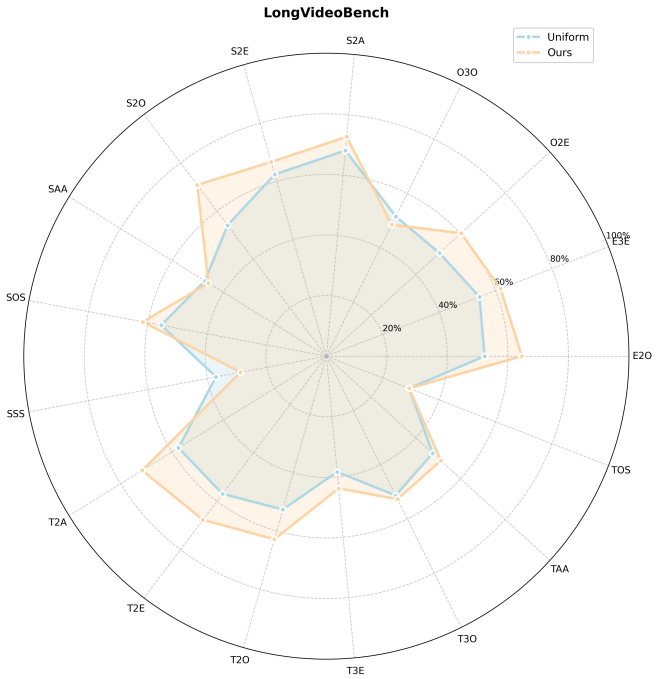

Figure 2. Performance comparison across different question categories on the LongVideoBench dataset.

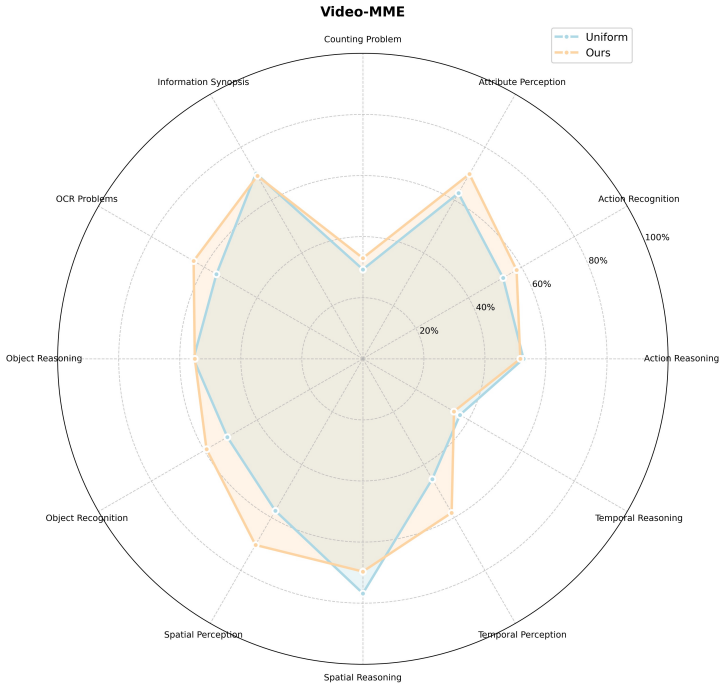

Figure 3. Performance comparison across different question categories on the VideoMME dataset.

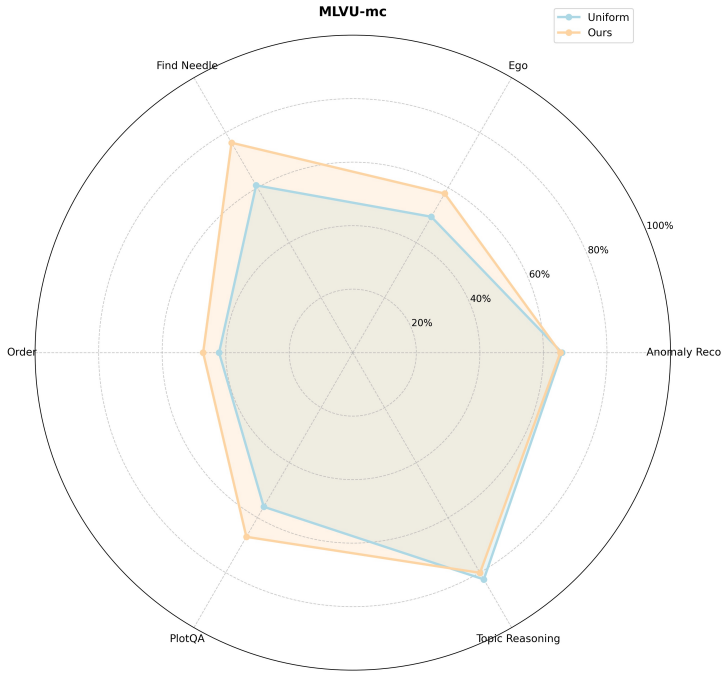

Figure 4. Performance comparison across different question categories on the MLVU dataset(only multiple choice questions).

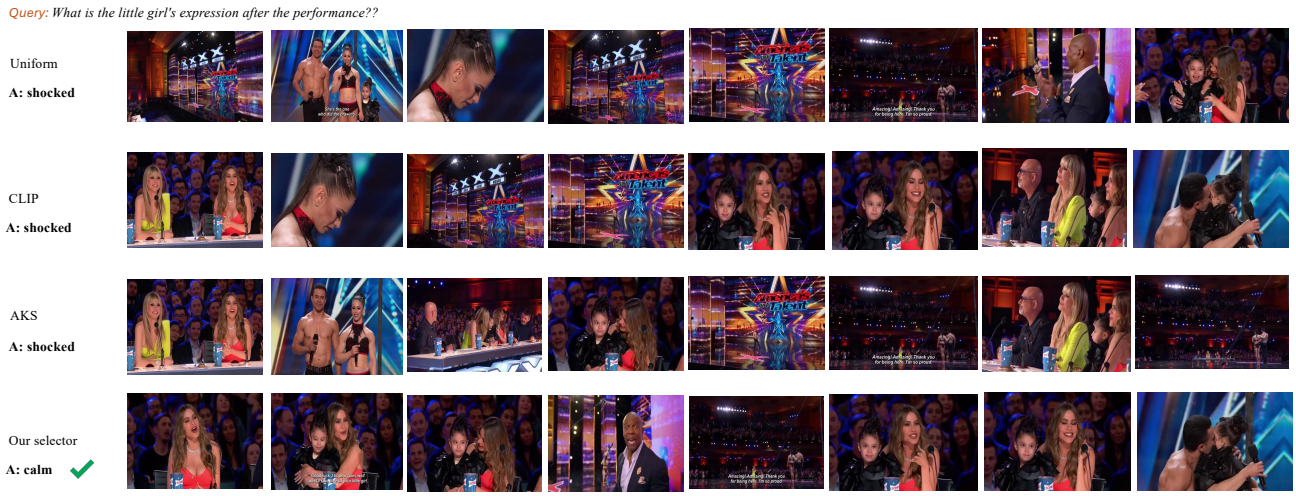

Figure 5. The qualitative comparison of different selection strategies. Our method answers the question correctly while other strategies fail.

Question: What type of flowers are placed on the table between the host and the guest in the video? Answer: D. Sunflowers

Figure 6. The qualitative comparison between our method and CLIP.