Free-MoRef: Instantly Multiplexing Context Perception Capabilities of Video-MLLMs within Single Inference

Supplementary Material

Table 1. Performance comparison with token compression method.

Method	MLVU	VideoMME		
		Medium	Long	Overall
128 frames(23296)	70.2	63.2	54.1	64.9
128 frames(11520) @Large Pooling	70.6	64.9	55.0	65.7
128 frames(23296) @Free-MoRef	**70.8**	**65.8**	**55.8**	**66.3**
256 frames(46592)	67.2	61.4	54.1	63.1
256 frames(23040) @Large Pooling	68.7	64.7	52.9	64.9
256 frames(46592) @Free-MoRef	**72.5**	**66.4**	**55.3**	**66.3**

1. Estimation of computational cost.

In video understanding tasks, visual tokens typically constitute over 99% of the total number of tokens. Moreover, within the reasoning process, the computational load during the prefilling stage is substantially higher than that during the decoding stage, particularly in video multi-choice tasks. Taking these factors into consideration, we use the amount of computation generated by calculating full-attention with vision-token in the perfilling stage as an estimate of the amount of computation for the whole response process.

The Qwen2-7B LLM contains 28 decoder layers. Denoting the input frames as F, the computing cost can be represented as $F \times F \times 28$. Applying Free-MoRef with N parallel references and fusion at layer L, the computing cost compared to the full-attention is calculated as follows:

$$\frac{F \times F \times 28}{(F/N)^2 \times N \times L + (F/N)^2 \times (28 - L)} \tag{1}$$

In our main experiments, when setting $F = 128, N = 2, L = 3$, Eq. 1 equals 27.6%, and the computational cost is 10.2% and 6.25% for $F = 256, N = 4, L = 6$ and $F = 512, N = 8, L = 12$ respectively.

2. Effects of each components.

Table 2. Effects of key components of Free-MoRef.

Multi-Reference Partition	MoRef Attention	Reference Fusion	Overall
✗	✗	✗	64.9
✗	✗	✓	63.9
✓	✗	✓	62.0
✓	✓	✗	65.8
✓	✓	✓	**66.3**

In Table 2, we perform ablation experiments on the key components of Free-MoRef. Directly applying Reference Fusion at the third layer without Multi-Reference Partition and MoRef-Attention is equivalent to dropping 50% of the vision tokens using the FastV method, which inevitably results in a performance decline. Building upon this baseline, applying Multi-Reference Partition to reconstruct the input vision sequence into two chunks and conducting inference with full attention separately leads to a further deterioration in performance. However, when MoRef-Attention is utilized to fuse the attention results across multiple reference, a significant improvement is observed. This clearly demonstrates that Free-MoRef enhances the contextual understanding capabilities of Video-MLLM primarily through the parallel reasoning of MoRef-Attention over Multi-References. Moreover, implementing Reference Fusion on the foundation of MoRef-Attention can further optimize the performance. This indicates that establishing connections among the vision references of different chunks could further help the overall understanding.

3. Comparison with Vision Compression Method.

Table 1 records the comparison between Free-MoRef and common vision compression methods. In the case of Large Pooling, it involves increasing the stride of spatial pooling within the connector, which leads to a reduction in the number of tokens per frame from 182 to 90. While vision compression can yield certain performance enhancements for 128-frame and 256-frame inputs, it remains far from comparable to Free-MoRef. As the number of input frames increases, the advantages conferred by the compression method become increasingly limited. In contrast, Free-MoRef attains better performance on longer-length inputs. This comparison highlights the distinct performance characteristics of Free-MoRef and traditional vision compres-

Table 3. Performance comparison on various task categories in MLVU. Tasks contain TR–Topic Reasoning, AR–Anomaly Recognition, NQA–Needle Question-Answering, ER–Ego Reasoning, PQA–Plot Question-Answering, AO–Action Order, and AC–Action Count. The best result is **bolded**, the second is underlined, and the worst is in red.

Context Length	**Holistic**		**Single Detail**			**Multi Detail**		**Avg**
	TR	AR	NQA	ER	PQA	AO	AC	
64 frames	86.0	72.0	76.3	**62.5**	76.4	63.7	43.2	70.3
128 frames	86.4	71.0	77.5	62.2	76.8	62.9	41.7	70.2
128 frames @Free-MoRef	85.6	71.5	77.2	62.2	76.4	64.1	**48.1**	70.8
256 frames @Free-MoRef	**88.6**	**72.5**	**78.3**	62.2	**80.1**	**66.0**	47.1	**72.5**

sion methods, underscoring the superiority of Free-MoRef in handling extended contexts.

4. Detailed analysis on MLVU benchmark.

Table 3 records the detailed performance on MLVU benchmark. When the number of input frames is extended to 256 frames, substantial performance improvements are observed across all tasks (except for Ego Reasoning task). Among these, the performance enhancements in Action Ordering and Action Counting problems are the most pronounced, followed by Needle Question Answering (NQA) and Plot Question Answering (PQA) problems. Notably, Ego Reasoning problems do not gain benefits from the increased number of frames. This is attributable to the fact that Ego Reasoning problems typically necessitate only a limited part of the entire video. For instance, consider the question *”Where was the insulated drink cup?”* Given the observation of the relevant fragment, adding more irrelevant frames will not provide effective references for the response to this question.

5. Attention map visualization.

Figure 1 presents the averaged attention map across different attention heads for each decoder layer during the reasoning process of LLaVA-Video on 64 frames. As depicted in the figure, distinct diagonal lines are observable in the attention maps of the first three decoder layers. These diagonal lines span across each vision token, suggesting that at the shallow layer, each vision-token contributes to the information aggregation. Commencing from the 4th layer, the weight assigned to the vision tokens starts to diminish. As the decoding process progresses to the deeper layers, an irregular weight distribution emerges among the vision-tokens. This phenomenon may signify that the model is engaged in extracting deep semantic features.

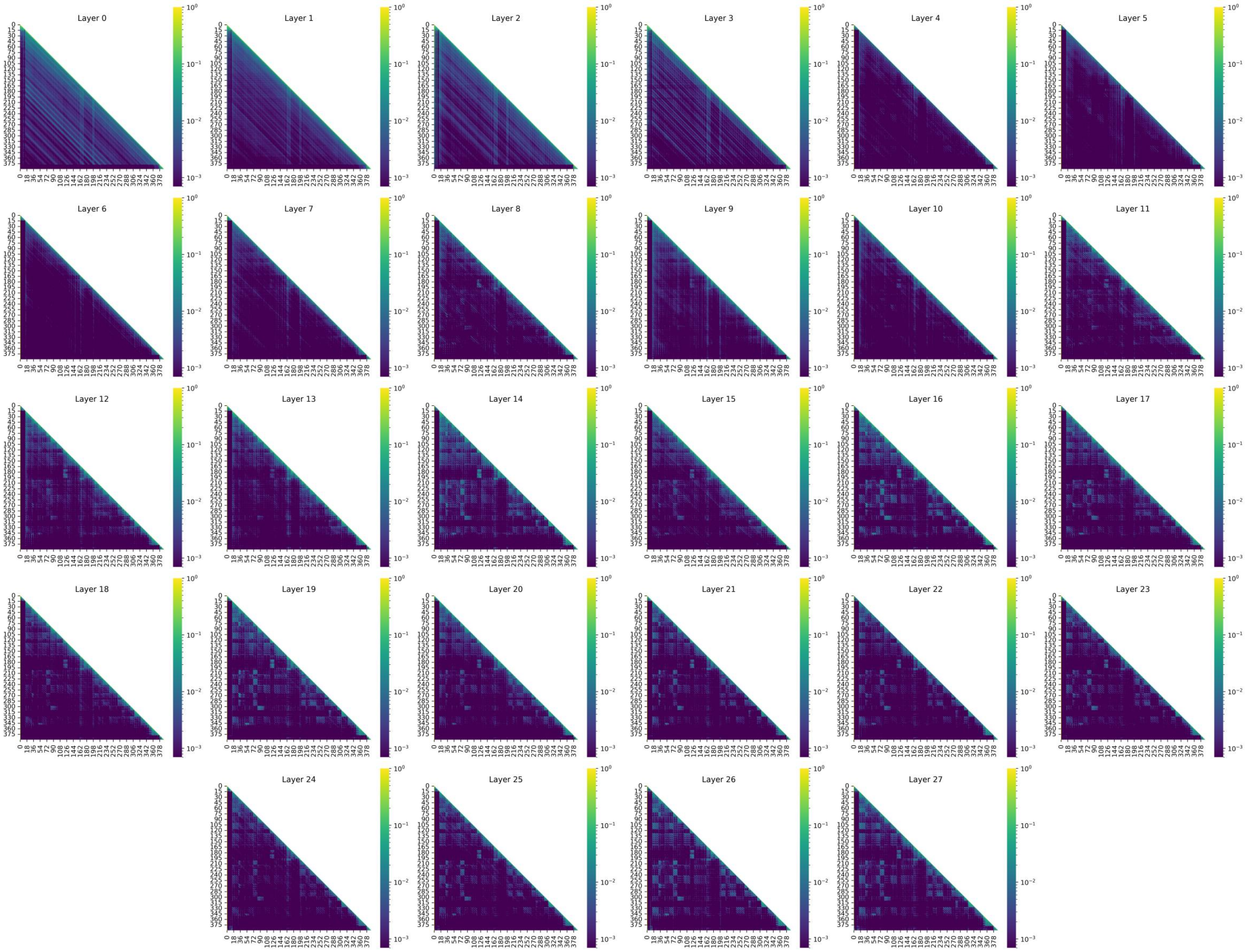

Figure 1. The averaged attention maps during reasoning 64 frames by LLaVA-Video.