

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

Kuo Wang^{1,2,4*} Quanlong Zheng^{3*} Junlin Xie⁵ Yanhao Zhang^{3†} Jinguo Luo⁶ Haonan Lu³

Liang Lin^{1,2,8} Fan Zhou^{1,4,7} Guanbin Li^{1,2,4,8‡}

¹Sun Yat-sen University, ²Peng Cheng Laboratory, ³OPPO AI Center, OPPO Inc., China

⁴Research Institute, Sun Yat-sen University, Shenzhen, China

⁵The Chinese University of Hong Kong, Shenzhen, China

⁶Harbin Institute of Technology, Shenzhen, China

⁷Shenzhen Key Laboratory of Digital Living Network and Content Service

⁸Guangdong Key Laboratory of Big Data Analysis and Processing

wangk229@mail2.sysu.edu.cn, {zhengquanlong, zhangyanhao, luhaonan}@oppo.com

junlinxie@link.cuhk.edu.cn, 23s153135@stu.hit.edu.cn

linliang@ieee.org, {isszf, liguanbin}@mail.sysu.edu.cn

Abstract

Video Multimodal Large Language Models (Video-MLLM) have achieved remarkable advancements in video understanding tasks. However, constrained by the context length limitation in the underlying LLMs, existing Video-MLLMs typically exhibit suboptimal performance on long video scenarios. To understand extended input frames, common solutions span token compression and streaming inference techniques, which sacrifice feature granularity or inference efficiency. Differently, to efficiently achieve comprehensive understanding of longer frame inputs, we draw ideas from MoE and propose a training-free approach **Free-MoRef**, which instantly multiplexes the context perception capabilities of Video-MLLMs within one inference pass. Specifically, Free-MoRef reconstructs the vision tokens into several short sequences as multi-references. Subsequently, we introduce MoRef-attention, which gathers clues from the multi-reference chunks in parallel to summarize unified query activations. After the shadow layers in LLMs, a reference fusion step is derived to compose a final mixed reasoning sequence with key tokens from parallel chunks, which compensates the cross-reference vision interactions that are neglected in MoRef-attention. By splitting and fusing the long vision token sequences, Free-MoRef achieves improved performance under much lower computing costs in reasoning multiplexed context length, demonstrating strong

efficiency and effectiveness. Experiments on VideoMME, MLVU, LongVideoBench show that Free-MoRef achieves full perception of $2\times$ to $8\times$ longer input frames without compression on a single A100 GPU while keeping instant responses, thereby bringing significant performance gains, even surpassing dedicatedly trained long-video-MLLMs. Codes are available at <https://github.com/wkfdb/Free-MoRef>

1. Introduction

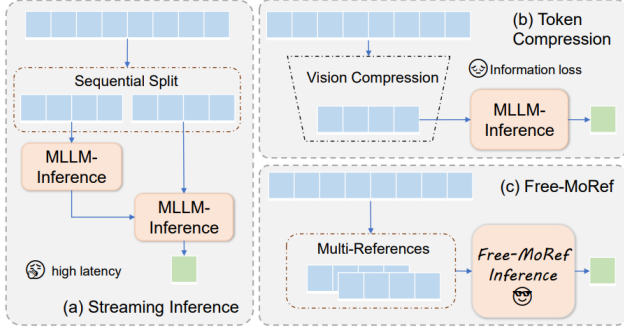
Large Language Models (LLMs) [2, 31, 40] have emerged as a revolutionary force towards general intelligence, marked by their universal capabilities in various language tasks. Through instruction tuning [20], Multimodal Large Language Models (MLLMs) further extend their exceptional reasoning ability to other modalities such as vision [14, 30, 32] and audio [24]. In recent studies, MLLMs have been extensively applied to the comprehension of video content [5, 16, 23, 34]. Notwithstanding the remarkable advances in video understanding, the context length restriction inherent in LLMs has emerged as a critical bottleneck, especially for long video understanding, where the abundant visual tokens readily surpass the threshold that ensures stability and consequently resulting in a decline in the effectiveness of these models.

To prevent over-length sequences that violate the context constraints of the reasoning LLM, existing MLLMs [30, 32, 43] typically impose a maximum limit on the length of vision tokens to maintain stable performance. (e.g. 64 frames

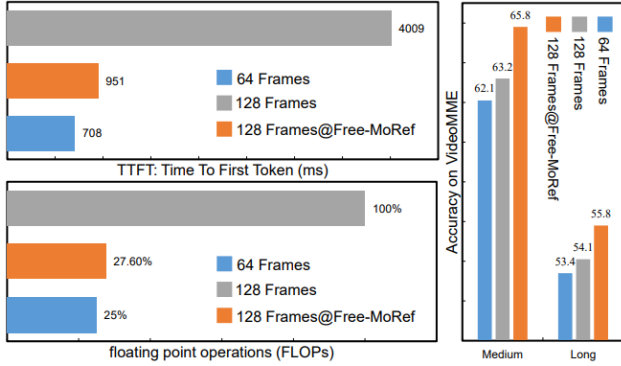
*Equally-contributed authors.

†Project Leader.

‡Corresponding author.



(a) Different designs to expand the context perception capability. Free-MoRef achieves both efficient and comprehensive understanding of the multiplexed vision inputs.



(b) Comparison of FLOPs, first token latency and overall QA accuracy in reasoning original and doubled input frames by LLaVA-Video [44].

Figure 1. Different inference designs and the advantages of Free-MoRef. In summary, Free-MoRef brings superior performance under much lower computing costs on longer vision contexts.

with 2×2 spatial pooling for LLaVA-Video [44] and 784 small rescaled frames for Qwen2-VL [32].) The tradeoff between resolution and number of input frames have to be made in video understanding models, which greatly restricts their effectiveness in fully exploiting the rich information contained within extended video sequences.

To expand the context perception capability under the context restriction of the foundation LLM in Video-MLLMs, common solutions primarily encompass token compression [18, 26, 28] and streaming inference technique [25, 33, 38]. However, both of these methods suffer from notable deficiencies. The Streaming Inference technique [38] achieves ultra-long context dependency by retaining and reusing the historical KV-CACHE, but the extra time cost is proportional to the context length benefit. For example, reasoning doubled contexts results in doubled latency. As an alternative, the token compression strategy can represent more information within a limited token length, thereby increasing the context within a single inference without exceeding the preset token length limit. However, longer context benefits lead to more severe informa-

tion loss. In light of these drawbacks, a crucial question emerges: *Is it possible to achieve longer context perception within a single inference while ensuring comprehensive understanding of the context?*

Motivated by this question, we have designed and implemented Free-MoRef, a training-free approach that instantly multiplexes the context throughput within one single inference pass, achieving full long-context understanding with flash efficiency. Inspired by the MoE paradigm [9], we abstract long visual tokens into multiple short sequences as multiple references, each of which encapsulates the overall information of the original long contexts. Subsequently, we further design the **Mixture of Reference** attention, which allows for the parallel querying of multiple references and the integration of the results into a unified activation within each decoding layer. This process could be considered as an expert solving problems according to different references and figuring out a final solution. As observed in FastV [4], after the shadow layers in LLM, the attention pattern would be more concentrated on query tokens. Leveraging this insight, we further extract key vision tokens in each chunk and mix them into a global reference for the remaining decoding layers, which compensates the neglected cross-reference vision interactions in the parallel reasoning. As illustrated in Figure 1b, by splitting and fusing the vision-tokens, Free-MoRef achieves instant comprehensive understanding of longer contexts with improved performance, demonstrating strong efficiency and effectiveness.

We apply Free-MoRef to LLaVA-Video-7B [44] and conduct experiments on several long video benchmarks, including VideoMME [12], MLVU [45], and LongVideoBench [37]. On a single A100 GPU, Free-MoRef can directly extend the context throughput from $2\times$ to $8\times$ with less than 27.6% of the FLOPs and negligible latency, while bringing superior performance on all the three above long video benchmarks. On the VideoMME benchmark, our method leads 3% to 5% performance gains on long and medium videos, reaching SOTA results, even surpassing dedicatedly trained long-video-MLLMs [28, 39, 42]. Notably, Free-MoRef supports Flash-Attention [10] and can also be integrated with streaming inference or token compression strategies. Despite training-free application, the MoRef-attention mechanism may also inspire the training design for long context scenarios.

2. Related Works

2.1. Video Large Language Models

Recent advancements in Video-MLLMs are mainly achieved by empowering Image-MLLM to comprehend video content. Here, large-scale video-text data is utilized to learn the vision features and temporal relations between the input frames. The structure of Video-MLLM typically

consists of a vision-encoder to tokenize the vision inputs, a foundation LLM for query-aware reasoning, and an intermediate connector to link the vision space and language space. In terms of the connector, several works [8, 16, 17] use the Q-Former [15] to merge visual and text features, where learnable tokens summarize the encoded patch embeddings. However, high compression rate usually results in lower performance in such designs. In other approaches [3, 19, 22, 43], patch embeddings are directly concatenated, which is more effective since it preserves more detailed features. However, the abundance of vision tokens poses a vital challenge in understanding longer videos.

2.2. Long Video Understanding

Through uniform sampling, existing Video-MLLMs can be directly applied to long video understanding tasks. However, it is evident that the limited input severely restricts the model’s performance. To tackle the challenge of long video understanding, existing research can be mainly categorized into three types: LLM context expansion [11, 29, 35, 39], token compression [18, 26, 28, 36], and streaming inference [25, 33]. Context extension methods aim to directly increase the context length limit of Video-MLLMs by conducting post-training with long-sequence data, thereby enhancing the model’s context perception ability. Although these methods are effective, they impose a substantial computational burden on long video understanding, which restricts their application in practical scenarios. Regarding vision token compression, training-free methods [4, 13, 27] prune visual tokens based on spatiotemporal redundancy, while training-based approaches [6, 28] introduce learnable summary tokens to achieve token compression. However, a low compression rate may not yield significant contextual improvement, while a high compression rate could result in substantial information loss. Recently, streaming inference techniques [38] have been applied to long video understanding [33]. This is achieved by invoking Video-MLLM multiple times to gradually comprehend the long context through identifying and reusing key historical KV-CACHE. Nevertheless, this process leads to exponential reasoning delays.

Our approach bears resemblance to the context extension method. However, the key distinction lies in the fact that our method is training-free, incurs low computational overhead, and can instantly achieve comprehensive perception of the exponentially increased context within less than 27.6% of the computational cost.

3. Method

In this section, we present the Free-MoRef method, which effectively extends the context perception capacity of Video-MLLMs with a high degree of flexibility. Notably, when dealing with extended input frames, Free-MoRef initially partitions the long vision tokens into parallel inference

chunks via multi-reference partitioning. Subsequently, it substitutes the self-attention layers of the LLM with MoRef attention. This allows for parallel reasoning over multiple references using the same query and aggregation of unified activations. At the mid-decoder layers, an optional Reference Fusion step is derived to combine the parallel chunks, thereby further enhancing the efficiency of the reasoning process. Without additional training, Free-MoRef overcomes the context length constraint of Video-MLLM during single inference and attains comprehensive perception of exponentially expanding context with minimal computational cost and achieves better performance. The overall architecture of Free-MoRef Inference is depicted in Figure 2, and the details of each component will be elucidated in the following subsections.

3.1. Multi-Reference Partition

In Video-MLLM, the reasoning LLM typically has a context length threshold to safeguard stable performance. For instance, the LLaVA-Video [44] model employs Qwen2 [1] as its underlying LLM, and the corresponding sequence length threshold is set at 32768 tokens. When the length of the inference sequence surpasses this threshold, it frequently results in performance deterioration or, more critically, Out-Of-Memory errors. Under these constraints, existing approaches commonly resort to vision token compression or streaming inference techniques to handle longer input frames. To minimize information loss while concurrently ensuring the efficiency of the reasoning process, we propose to partitioning the long visual token sequence into multiple parallel chunks, which serve as multi-references for comprehensive understanding.

To enhance flexibility, we initially divide the vision-token sequence into M units according to the temporal relationship. Subsequently, within each unit, we further temporally decompose it into N fragments. Here, both M and N are manually configured hyperparameters. Eventually, through the aggregation of fragments from diverse units, we are able to obtain N reference chunks. Each of these references can be considered as an abstraction of the extended video sequence. Notably, the larger the value of the hyperparameter M , the more pronounced the temporal intersection among the references. When setting $M = 1$, the N chunks will be temporally independent of each other. After completing the multi-reference partition, we assign identical system prompt and question to each vision sequence, thereby forming the final parallel inference chunks, which is designed for more comprehensive and efficient reasoning by leveraging the **Mixture-of-Reference (MoRef)** attention.

3.2. MoRef Attention

MoRef Attention is the key step in attaining comprehensive perception and parallel reasoning across multi-references.

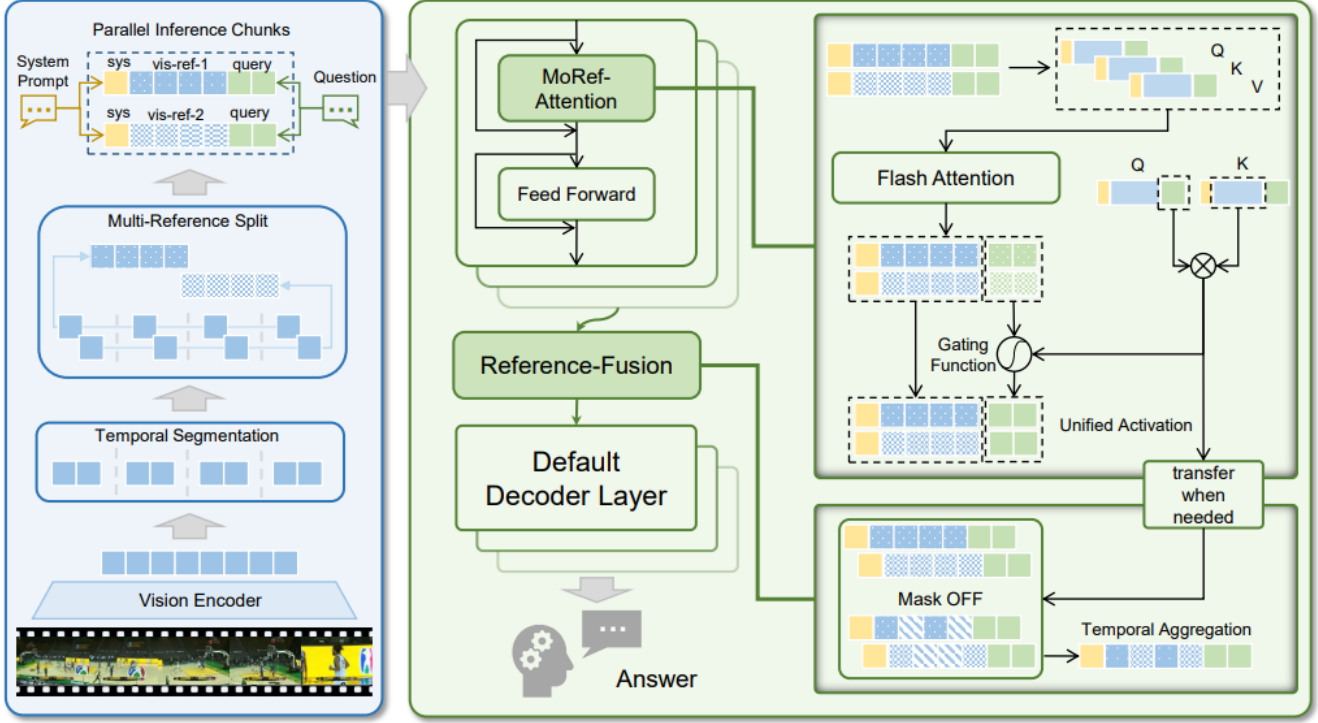


Figure 2. The framework of Free-MoRef Inference. For extended input frames, the initial step involves partitioning the vision tokens into multiple references and subsequently assigning identical system prompt and question to each of these references. To enable efficient comprehension of these multi-reference chunks, we design the MoRef attention mechanism, which concurrently extracts clues from multiple references to formulate responses to the posed question. At the middle layer of the decoder, a reference-fusion step is derived. This step serves to aggregate the parallel chunks into a unified global representation which not only further accelerates the reasoning process but also facilitates cross-chunk interactions, thereby enhancing the overall performance and effectiveness of the long context reasoning.

It concurrently queries distinct references in parallel using an identical question and combines multiple attention outcomes to summarize a unified response for updating the question tokens within each decoder layer.

Specifically, for the input parallel inference chunks, MoRef first constructs $Q, K, V \in \mathbb{R}^{N \times l \times D}$ (N is the chunk number, l is the sequence length and D is the embedding dimension), then executes flash-attention to obtain the initial attention results O , where $O = [O^{sys}, O^{vis}, O^{ques}]$. Here, O^{sys} , O^{vis} and O^{ques} respectively denote the attention results corresponding to the system prompt token, vision token, and question token. Owing to the unidirectional nature of causal attention, O^{sys} in different chunks would be exactly the same. In contrast, O^{vis} and O^{ques} yield divergent results due to the variance in vision-references. At this point, we maintain the variation of O^{vis} and aggregate O^{ques} across different chunks through the following function:

$$O^{fusion} = (\sum_{i=1}^N \omega_i \cdot O_i^{ques}).repeat(N), \sum_{i=1}^N \omega_i = 1 \quad (1)$$

In Eq. 1, O_i^{ques} represents the query result on each

reference and O^{fusion} is the unified summarization. By replacing O^{ques} in the initial attention results, the output of MoRef attention is constructed as $O^{MoRef} = [O^{sys}, O^{vis}, O^{fusion}]$.

The ω_i in Eq 1 indicates a gating function, which controls the information aggregation across different references. The gating function should be query-aware, as the key information required to answer the question may not be uniformly distributed among diverse references. In the training-free implementation, our objective is to seek the query-reference-correlation from the attention map. Since the flash-attention doesn't support output attention weight, we manually calculate the multi-model attention map between query and vision-reference:

$$A = softmax(Q^{ques} \times (K^{vis})^T) \quad (2)$$

Compared with the full-attention on the whole sequence, the cross-modal attention Eq. 2 introduces negligible computation. With respect to A , we set the gating weights of each reference chunk as:

$$\omega_i = \frac{max(A[i])}{\sum_{i=1}^N max(A[i])} \quad (3)$$

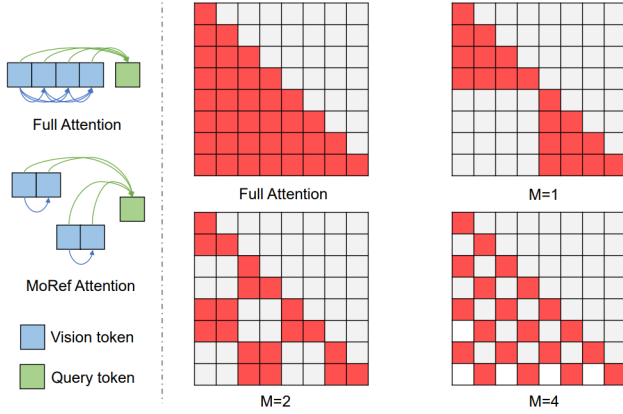


Figure 3. Visualization of attention patterns of full attention and MoRef attention. MoRef attention achieves the full-awareness of vision tokens, while the attention map among vision tokens are sparsified. The right part represents vision attention map of splitting 8 vision tokens into two chunks under different setting of temporal unit number (M).

By combining the attention results of the same query across diverse vision-references, all vision-tokens are effectively engaged in the updating of the query-token in each decoder layer. As shown in Figure 3, this integration strategy enables the full-context perception equivalent to that achieved by full attention. By partitioning the vision sequence into N non-overlapping chunks, the computational complexity is reduced by approximately a factor of $\frac{1}{N}$ compared with full attention. The number of temporal units M serves as a crucial parameter that significantly influences the formation of sparse attention patterns. For instance, consider a vision sequence of length 8 divided into $N = 2$ chunks. Figure 3 depicted the impact of different settings of M on the vision-attention-map.

3.3. Reference Fusion

MoRef attention efficiently enables the comprehensive perception of multi-references. Nevertheless, the division of multi-references disrupts the connectivity among vision-tokens across different chunks. To address this limitation, we design an additional Reference Fusion step. This step aims to achieve the integration of multiple reference chunks into a global one, thereby compensating for the lack of cross-chunk interaction within the deep decoder layer.

The implementation of Reference Fusion is grounded in an observation made by FastV [4]: vision-tokens contribute uniformly in the shallow decoder layers. In contrast, within the deep layers, the attention weights of the decoder layer would more concentrate on the query-token. We visualized the reasoning process of LLaVA-Video [44] and noted a similar phenomenon (as detailed in supplementary material). Leveraging this insight, we maintain parallel multi-

reference reasoning within the shallow layer. When the decoding process reaches a specific layer L , we perform the merging of multi-references based on the attention map A computed in Eq 2.

Specifically, $A \in \mathbb{R}^{N \times l_{ques} \times l_{vis}}$, where N represents the number of references, l_{ques} and l_{vis} denote the number of question token and vision token in each inference chunk. We compute the average of the attention map A along the l_{ques} dimension to construct the importance estimation matrix $E \in \mathbb{R}^{N \times l_{vis}}$, where each element E_{ij} in E quantifies the average contribution of the j -th vision token in the i -th chunk. Based on the estimation matrix E , we prune $1 - \frac{1}{N}$ of the less important vision-tokens within each inference chunk. Subsequently, we aggregate the remaining vision tokens into a global reference in accordance with their temporal relationships. System prompt tokens and question tokens are directly transferred from the local reference chunk to the global reference. For the following decoding process, only the global reference is used by the default decoder layers of the LLM.

Through the Reference fusion step, the pruning of non-crucial tokens further reduces the computational load, while the cross-chunk vision interaction that is lacking in shallow layers can be effectively compensated for, which results in optimized performance.

4. Experiments

4.1. Experimental Setup

Benchmarks VideoMME. [12] Video Multi-Modal Evaluation benchmark (VideoMME) consists of 900 videos with a total duration of 256 hours, covering a wide range of video types. The videos are associated with 2,700 manually labeled complex multiple-choice QA pairs across 30 subfields. According to video durations, VideoMME is partitioned into three subsets: short (< 2 minutes), medium ($4 \sim 15$ minutes), and long ($30 \sim 60$ minutes). **MLVU.** [45] Multi-Task Long Video Understanding Benchmark (MLVU) significantly expands the scope of durations with diverse types of videos and 7 different types of QA tasks. The video lengths range from 3 minutes to over 2 hours, with an average duration of 12 minutes. **LongVideoBench.** [37] LongVideoBench highlights referred reasoning questions, which are dependent on long frame inputs. It contains 17 finer-grained question categories on 10 different types of videos. The video duration covers 4 groups: 8-15 seconds, 15-60 seconds, 3-10 minutes, and 15-60 minutes.

Implementation Details We implement the Free-MoRef on the LLaVA-Video-7B [44] model. By default, LLaVA-Video-7B loads videos at a FPS=1, with a maximum of 64 frames, where each frame is represented by 182 tokens. To

Table 1. Performance of Free-MoRef@LLaVA-Video-7B under extended frame inputs. The red color indicates Out-Of-Memory error on a single A100 GPU. We managed the inference under the help of *accelerate* toolkit.

Context Length (Token Number)	FLOPs	MLVU	VideoMME			LongVideoBench		
			Medium	Long	Overall	600s	3600s	Overall
64 frames (11648)	100%	70.3	62.1	53.4	64.3	60.4	51.2	58.8
128 frames (23296)	400%	70.2	63.2	54.1	64.9	60.6	50.8	58.7
128 frames@Free-MoRef	110.4%	70.8	65.8	55.8	66.3	62.1	51.0	59.3
256 frames (46592)	1600%	67.2	61.4	54.1	63.1	57.2	48.5	56.7
256 frames@Free-MoRef	163.2%	72.5	66.4	55.3	66.3	62.1	51.2	59.3
512 frames (93184)	6400%	61.1	55.7	48.8	60.6	53.1	45.9	54.3
512 frames@Free-MoRef	400%	72.8	67.3	56.0	66.9	62.8	51.9	59.9

Table 2. Performance comparison on Long-Video Benchmarks: all models in this table are of the 7B $\sim$ 8B scale.

Method	MLVU	LVideoBench	VideoMME	
			Long	Overall
InternVL2 [30]	64.0	54.6	-	54.0
InternVL2.5 [7]	68.4	57.5	53.0	64.5
Qwen2-VL [32]	64.8	55.6	55.7	63.3
LLaVA-OneVision [14]	64.7	56.3	-	58.2
LLaVA-Video [44]	70.2	58.2	53.4	64.3
Kangaroo [21]	61.0	54.8	46.7	56.0
LongVILA [39]	-	57.1	47.0	60.1
LongVA [42]	56.3	-	46.2	52.6
Video-XL [28]	64.9	50.7	-	55.5
RETAKE [33]	69.8	-	56.2	63.9
LLaVA-Video@Free-MoRef	72.8	59.9	<u>56.0</u>	66.9

demonstrate the efficacy of Free-MoRef, we multiply the maximum number of frames to 128 (2x), 256 (4x), and 512 (8x) respectively. In our experimental setup, the vision token sequence is consistently partitioned into $M = 64$ temporal units. For the multiplexed inputs of 2x, 4x, and 8x, the number of parallel chunks is set as $N = 2, 4, 8$ respectively, and the reference fusion layer is configured as $L = 3, 6, 12$. We conduct the evaluation upon *lmms-eval* framework [41] and all experiments are executed on a single A100 GPU. Further details will be made accessible in our publicly-released code repository.

4.2. Main Results

Results on Multiplexed Context Understanding. We implemented the Free-MoRef method on the *llava-video-7B* model, which by default has a maximum input frame num-

ber of 64. In order to verify Free-MoRef’s ability to handle multiplexed contexts, we expanded the maximum number of input frames to 128, 256, and 512 for experiments. The experimental results are presented in Table 1.

When the number of input frames is doubled, the number of tokens required to encode the input frames amounts to 23,296, which is within the context length limit of Qwen2 (32,768). Under such circumstances, the performance of the model remains nearly unchanged on both the MLVU and LongVideoBench, while demonstrating a slight improvement on the VideoMME benchmark. After the application of Free-MoRef, the computational cost incurred during the inference process is reduced by $(1 - 110.4/400 = 72.4\%)$. Concurrently, both MLVU and LongVideoBench exhibit a 0.5% performance gain. The performance improvement on VideoMME is more pronounced, with a 2.6% improvement for medium-length videos and a 1.7% improvement for long videos.

When the number of input frames is quadrupled, the number of vision tokens surpasses 40,000. This length clearly exceeds the context length limit of Qwen2, and attempting to perform inference using a single A100 GPU will directly result in an Out-Of-Memory error. Notably, even without the assistance of the *accelerate* toolkit, our proposed method is capable of effectively reasoning on up to 512 frames using a single A100 GPU.

Leveraging the *accelerate* toolkit, we were enabled to conduct further comparative experiments. When the number of input frames reached 256, the model’s performance experienced a substantial decline across all benchmarks. However, upon the application of Free-MoRef, the degradation was effectively mitigated. In particular, on the MLVU dataset, which consists of relatively longer videos, the model’s performance was further enhanced to 72.5%. In terms of efficiency, the computational requirement for the model to infer a context of $4 \times$ length was merely 163.2% of the original. When the number of input frames is further in-

Table 3. Performance comparison on various task categories in VideoMME. Tasks contain Temporal Perception(TP), Spatial Perception(SP), Attribute Perception(AP), Action Recognition(ARec), Object Recognition(ORec), OCR Problems(OCR), Counting Problem(CP), Temporal Reasoning(TR), Spatial Reasoning(SR), Action Reasoning(AR), Object Reasoning(OR) and Information Synopsis(IS). The best result is **bolded**, the second is underlined, and the worst is in **red**.

Context Length	TP	SP	AP	ARec	ORec	OCR	CP	TR	SR	AR	OR	IS	Avg
64 frames	<u>74.5</u>	64.8	79.3	64.5	<u>71.5</u>	66.9	48.5	48.0	80.4	56.1	59.0	76.5	64.3
128 frames	70.9	59.3	79.3	<u>66.8</u>	70.6	<u>71.2</u>	47.4	48.0	80.4	54.7	<u>61.2</u>	78.9	<u>64.9</u>
128 frames @Free-MoRef	<u>74.5</u>	64.8	<u>78.8</u>	67.1	72.3	71.9	48.5	50.8	85.7	<u>58.2</u>	61.9	<u>79.6</u>	66.3
256 frames @Free-MoRef	80.0	<u>63.0</u>	77.9	67.1	72.3	70.5	<u>47.8</u>	<u>50.3</u>	<u>82.1</u>	60.0	61.9	80.8	66.3

creased by $8\times$, reaching 512 frames, the length of the vision token approaches nearly 100,000. In comparison to the scenario with 256-frame input, the performance of the baseline model experiences a substantial and further decline. Despite posing challenges to the baseline model, the extended context serves as a rich source of information and provides more abundant references for the reasoning process of Free-MoRef, as a result, Free-MoRef attains a further enhancement in performance. In summary, Free-MoRef enables the Video-MLLM to leverage the multiplexed frame inputs for more comprehensive understanding of long videos, thereby highlighting the robustness and efficiency of Free-MoRef in handling ultra-long context scenarios.

Comparison with other Models. We validated the efficacy of Free-MoRef by conducting comprehensive comparisons with other Video-MLLMs. These comparisons encompassed open-source MLLMs capable of video understanding, as well as specifically designed long-video understanding models. As depicted in Table 2, our proposed Free-MoRef method outperformed all the others, attaining the optimal results on the MLVU, LongVideoBench and VideoMME benchmarks.

The underlying reason for the SOTA performance lies in the fact that Free-MoRef enables an efficient and exhaustive understanding of ultra-long contexts. By simply expanding the input frames, Free-MoRef can achieve superior long-video understanding performance within a single inference. It is worth noting that our method is implemented in a training-free manner, which further confirms the potential of MoRef-attention. Its ability to fully understanding ultra-long contexts while maintaining a low computational burden offers significant inspiration for the development of future long-video understanding models, thereby highlighting the practical value and far-reaching implications of Mixture-of-Reference design in the field of long video understanding tasks.

5. Ablation & Analysis

In this section, we perform ablation experiments and in-depth analysis on the Free-MoRef method. Free-MoRef is principally associated with three hyperparameters: the number of sequential units M , the number of reference partitions N , and the specific decoder layer L for reference fusion. We conduct detailed ablations based on 128-frame inputs on the VideoMME benchmark to evaluate the impact of the hyperparameters. Please refer to supplementary materials for additional analysis.

5.1. Performance on different types of question.

Table 3 records the performance of varying context inputs across different types of questions. In general, compared with 64-frame input, the expansion of the context predominantly brings benefits in Information Synopsis, and diverse Recognition and Reasoning questions. By applying Free-MoRef for long-context inference, the performance of nearly all types of question on the VideoMME benchmark are enhanced, except for Attribute Perception task. Here is an example of an AP task. *Question-id 009-1: Which color of clothes is QuYuan wearing in the video?* This type of question only refers to a small clip in the video. Expanding the context introduces redundant unnecessary information in this case, thus impact the accuracy of answering such questions.

5.2. Impacts of various reference number N .

The Free-MoRef method partitions the video sequence into N identically sized parallel references. In general, the more parallel references there are, the more computational effort is saved during inference. In the context of a 128-frame input, we set $M = 128/N$ to keep the vision-tokens are split frame-by-frame into each reference chunk. The effects of the model under different values of N are presented in Table 4. As the number of reference chunk increases, both the computational load and the performance of Free-MoRef

Table 4. Performance of different setting of parallel chunk number N . $N = 1$ indicates default inference.

Chunk Number	FLOPs	Medium	Long	Overall
$N = 1$	100%	63.2	54.1	64.9
$N = 2$	27.6%	65.8	55.8	66.3
$N = 4$	25%	65.1	55.8	66.1
$N = 8$	23.6%	64.9	55.6	65.9

Table 5. Performance of different setting of temporal units M . TP indicates Temporal Perception task and SP represents Spatial Perception.

Temporal Units	Partition Units	TP	SP	Overall
$M = 1$	64 frames	70.9	68.5	66.3
$M = 4$	16 frames	72.7	66.7	66.0
$M = 32$	4 frames	74.5	66.7	66.0
$M = 64$	1 frames	74.5	64.8	66.3

exhibit a gradual decline. For best performance, we identically set $N = \text{input frame number} / 64$ for all the experiments.

5.3. Effects of different temporal units M .

The configuration of the temporal unit M determines the temporal intersection among different reference chunks. Its influence on the vision attention pattern is illustrated in Figure 3. Table 5 documents the effects of diverse values of M on various tasks. Overall, the performance remains relatively consistent. However, significant disparities are observed in Temporal Perception and Spatial Perception tasks. We provide examples of TP and SP questions for better explanation. *TP: In which part of the video does the red parrot appear? SP: What is the location of the scene being depicted in the video?* When $M = 1$ and $N = 2$, the middle part of the entire video serves as the tail and head for the first and second reference chunks, which leads to opposed interpretations of the temporal perception task, thereby reducing the performance for TP tasks. In the case of spatial perception problems, a detailed comprehension of a continuous video segment is essential. When $M = 64$, the specific video segments are separated to each reference chunk frame by frame, which reduces the feature density and thereby impacting the reasoning for SP questions.

5.4. Analysis of reference fusion layer L .

By discarding unimportant vision tokens in the middle layer and merging multi-reference chunks, further computational savings can be realized while the missing visual feature de-

Table 6. Performance of different Reference Fusion Layers L .

Context	Fusion	Medium	Long	Overall
128 frames	$\times$	64.7	55.4	65.8
drop rate	$L = 1$	64.4	54.4	65.4
50%	$L = 3$	65.8	55.8	66.3
256 frames	$\times$	64.6	54.8	65.5
drop rate	$L = 3$	64.4	56.3	66.0
75%	$L = 6$	66.4	55.3	66.3

pendencies between parallel chunks can be made up. However, it is crucial to note that premature execution of reference fusion may lead to a certain degree of information loss, which impairs the final performance. As depicted in Table 6, when conducting inference on a context of 128-frame length, performing fusion operations at the first layer yields sub-optimal performance. In the case of reasoning about a length of 256 frames, executing fusion operations at the third layer results in superior long video understanding performance compared to performing it at the sixth layer. This phenomenon indicates that the timely establishment of visual feature associations across parallel chunks is more conducive to longer video understanding. From a holistic perspective, performing reference fusion at the sixth layer can achieve relatively higher performance at the overall evaluation.

6. Conclusion

In this paper, we present Free-MoRef, a novel training-free approach that instantly multiplexes the context perception capacity of VideoLLM within a single inference pass. By partitioning long video inputs into multi-reference chunks, our proposed MoRef-attention concurrently extracts clues from multi-references and synthesize unified query responses, thus facilitates the understanding of long videos with extended input frames. Leveraging Free-MoRef, we successfully achieved a comprehensive understanding of 1024 frames using a 7B-VideoLLM on a single A100 GPU and brought substantial improvements across three long video understanding benchmarks.

Acknowledgment

This work is supported in part by the National Key R&D Program of China (2024YFB3908503), in part by the Guangxi Science and Technology Plan Project (GuikeAD23026034), in part by the National Natural Science Foundation of China (62322608), and in part by the Shenzhen Science and Technology Program (JCYJ20220530141211024).

References

- [1] Qwen2 technical report, 2024. [3](#)
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [1](#)
- [3] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. [3](#)
- [4] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. [2](#), [3](#), [5](#)
- [5] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2025. [1](#)
- [6] Shimin Chen, Yitian Yuan, Shaoxiang Chen, Zequn Jie, and Lin Ma. Fewer tokens and fewer videos: Extending video understanding abilities in large vision-language models. *arXiv preprint arXiv:2406.08024*, 2024. [3](#)
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. [6](#)
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. [3](#)
- [9] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jia Shi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. [2](#)
- [10] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. [2](#)
- [11] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. [3](#)
- [12] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. [2](#), [5](#)
- [13] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024. [3](#)
- [14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. [1](#), [6](#)
- [15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [3](#)
- [16] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. [1](#), [3](#)
- [17] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. [3](#)
- [18] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. [2](#), [3](#)
- [19] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. [3](#)
- [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. [1](#)
- [21] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoli Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. [6](#)
- [22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. [3](#)
- [23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. [1](#)
- [24] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. [1](#)
- [25] Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. Inf-mllm: Efficient streaming inference of multi-modal large language models on a single gpu. *arXiv preprint arXiv:2409.09086*, 2024. [2](#), [3](#)
- [26] Shuhuai Ren, Shihuo Chen, Shicheng Li, Xu Sun, and Lu Hou. Testa: Temporal-spatial token aggregation for long-form video-language understanding. *arXiv preprint arXiv:2310.19060*, 2023. [2](#), [3](#)

- [27] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 3
- [28] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 2, 3, 6
- [29] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3
- [30] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 1, 6
- [31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 6
- [33] Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*, 2024. 2, 3, 6
- [34] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1
- [35] Hongchen Wei and Zhenzhong Chen. Visual context window extension: A new perspective for long video understanding. *arXiv preprint arXiv:2409.20018*, 2024. 3
- [36] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 3
- [37] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 2, 5
- [38] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 2, 3
- [39] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2, 3, 6
- [40] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1
- [41] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 6
- [42] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2, 6
- [43] Y Zhang, B Li, H Liu, Y Lee, L Gui, D Fu, J Feng, Z Liu, and C Li. Llava-next: A strong zero-shot video understanding model. 2024. 1, 3
- [44] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2, 3, 5, 6
- [45] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2, 5