Design Choices for Extending the Context Length of Visual Language Models

Mukai Li[1] **Lei Li**[1] **Shansan Gong**[1] **Qi Liu**[1]
[1]The University of Hong Kong
kaikiaia3@gmail.com liuqi@cs.hku.hk

Abstract

Visual Language Models (VLMs) demonstrate impressive capabilities in processing multimodal inputs, yet applications such as visual agents, which require handling multiple images and high-resolution videos, demand enhanced long-range modeling. Moreover, existing open-source VLMs lack systematic exploration into extending their context length, and commercial models often provide limited details. To tackle this, we aim to establish an effective solution that enhances long context performance of VLMs while preserving their capacities in short context scenarios. Towards this goal, we make the best design choice through extensive experiment settings from data curation to context window extending and utilizing: (1) we analyze data sources and length distributions to construct ETVLM - a data recipe to balance the performance across scenarios; (2) we examine existing position extending methods, identify their limitations and propose M-RoPE++ as an enhanced approach; we also choose to solely instruction-tune the backbone with mixed-source data; (3) we discuss how to better utilize extended context windows and propose hybrid-resolution training. Built on the Qwen-VL series model, we propose GIRAFFE, which is effectively extended to 128K lengths. Evaluated on extensive long context VLM benchmarks such as VideoMME and Viusal Haystacks, our GIRAFFE achieves state-of-the-art performance among similarly sized open-source long VLMs and is competitive with commercial model GPT-4V.[1]

1 Introduction

Visual Language Models (VLMs) (OpenAI, 2023; Gemini Team, 2024) integrate visual and textual information, which are pivotal in understanding the multimodal world and excel in various applications, such as visual question answering and video understanding (Liu et al., 2023c; Li et al., 2022). However, more advanced scenarios involve multi-image and long video comprehension, which challenge the long-range modeling capabilities of VLMs. For instance, a 2K context length can only digest less than a few frames (Liu et al., 2023c,b; Li et al., 2023a), limiting the upper bound of long video understanding. Consequently, there is a pressing need for methods to extend the context window of VLMs and improve their performance in long context scenarios. This would benefit next-generation VLMs in performing long history visual agents or serving as world models (Liu et al., 2024a).

Recent efforts for longer context VLMs focus on extending base Large Language Models (LLMs), along with visual alignment or efficient architectures. LongVA (Zhang et al., 2024a) seeks to transfer long context ability from language models to vision by modifying position embeddings in the LLM backbone (PI, Chen et al. 2023b; NTK, LocalLLaMA 2023). LongVILA (Xue et al., 2024) and LongLLaVA (Wang et al., 2024b) accommodate longer sequences using multi-stage alignment and instruction tuning (Peng et al., 2023; Fu et al., 2024c) with additional infrastructure and architecture. Despite these initial explorations, they have not investigated the feasibility of directly extending the context window of existing VLMs or systematically explored the design space in the extending pipeline. To bridge this gap, we decompose the challenge of extending context windows of existing VLMs into three fundamental research questions: ***(1) How to effectively organize and curate training data? (2) How to efficiently train longer VLMs? (3) How to leverage the extended context window?***

In our work, our goal is to answer the three research questions and find a solution in practice. To validate our design choices, we implement thorough experiments based on Qwen-VL series model (Bai et al., 2023; Wang et al., 2024a) and conduct comprehensive evaluations on single image understanding, image interleave, and video

[1]https://github.com/kiaia/GIRAFFE

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33425–33438

tasks (§2.1). For data curation, we prepare a diverse dataset comprising long context instruction data, multimodal instruction data, multimodal interleave data, and video instruction data (§2.2). We analyze the impact of different data compositions, ratios, and lengths on model performance (§2.3) and find that (1) short multimodal instruction data is crucial for both extending long context capability and retaining short context performance; (2) a balanced data ratio contributes to balanced performance on downstream tasks. For the second research question on extending training, we examine the effective context length of previous position embedding extending alternatives such as PI and NTK, discovering that, akin to LLM studies (Gao et al., 2024a; An et al., 2024b), the effective length is shorter than the training length (§3.1). We propose M-RoPE++ (§3.2) to extend position embedding on spatial and temporal dimensions. Validation experiments reveal that our method achieves better downstream task performance and longer effective length under the same training length (§3.2). Different from LongVA (Zhang et al., 2024a) that first extend LLM base or LongLLaVA (Wang et al., 2024b) and LongVILA (Xue et al., 2024) that adopt multi-stage training with visual alignment and instruction tuning, we find that directly training VLMs by only updating LLM backbone's parameters achieves optimal results (§3.3). To figure out how to use long context well in VLM, the third research question, we examine the trade-off between single-frame resolution and frame numbers regarding task performance (§3.4). We consequently propose hybrid-resolution training, which further improves the utilization of a fixed context length (§3.5).

Based on our findings from the three research questions, we carefully select data recipes and training methods to extend Qwen-VL and Qwen2-VL to GIRAFFE-QwenVL and GIRAFFE with 128K length. Our final models are evaluated on both short context tasks such as single image understanding and long context tasks with multi-image and long videos. Experimental results demonstrate that our GIRAFFE achieves state-of-the-art performance among long VLMs and there is a significant improvement for our GIRAFFE-QwenVL compared with Qwen-VL base (§4.2). Summarized contributions:

1. We investigate different design choices to extend the context window of existing VLMs to 128K while maintaining comparable performance on short visual tasks.

2. Technically, M-RoPE++ and hybrid-resolution training methods are newly proposed by us to enhance model performance during training and inference.

3. On existing long VLM benchmarks, GIRAFFE achieves state-of-the-art performance among similar scale open-sourced long VLMs and is competitive to commercial models.

2 How to Curate Extending Data

Developing an effective recipe for extending the context window of VLMs is crucial. To systematically evaluate such recipes, we construct a comprehensive metric suite encompassing single-image, multi-image, and video tasks (§2.1), enabling a thorough assessment of model performance across diverse scenarios. This section focuses on the selection and preprocessing of training data (§2.2), with an emphasis on understanding how data compositions, ratios, and lengths influence the model's capabilities (§2.3).

2.1 Evaluation Tasks

We evaluate both long and short-context multimodal tasks, as it is essential for VLMs to sustain performance on short-context tasks after extended training. For short-context evaluation, we utilize widely adopted benchmarks such as single-image MME (Fu et al., 2023) and MMBench (Liu et al., 2024b), which capture the diverse capabilities of VLMs. For multi-image tasks, we incorporate Mantis-Eval (Jiang et al., 2024), QBench (Wu et al., 2024b), and BLINK (Fu et al., 2024b), in line with LLaVA-Interleave (Li et al., 2024a). Given the temporal nature of videos, which naturally represent long-context multimodal tasks, we evaluate on LongVideoBench (Wu et al., 2024a) and VideoMME (Fu et al., 2024a). Additionally, we include the Visual Haystack Single Needle Challenge (Wu et al., 2024c), which requires locating specific visual information within a long sequence of images, providing a robust measure of the model's effective context length.

2.2 Extending Data Curation

Data Composition To construct our extending training dataset, ETVLM, we incorporate four primary types of data with varying lengths: (i) Long-context instruction data, sourced primarily from

Categories	Task types	Data sources	%Part
Text	Long context instructions	LongAlign (Bai et al., 2024), LongAlpaca (Chen et al., 2023c)	20%
Image	Short visual instruction data	LLaVA-Instruct (Liu et al., 2023c), M3IT (Li et al., 2023b)	25%
	Image interleave data	MMDU (Liu et al., 2024c), Mantis (Jiang et al., 2024), ArXivQA-interleave*	25%
Video	Video QA Video Summary	ShareGPT4O (Chen et al., 2024), MLVU (Zhou et al., 2024), LLaVA-Video (Zhang et al., 2024b) ShareGPT4V (Chen et al., 2023a)	30%

Table 1: Overview of our ETVLM training dataset. This dataset encompasses a wide range of modalities and is concatenated to target context length. * indicates that we reconstruct this data by our own.

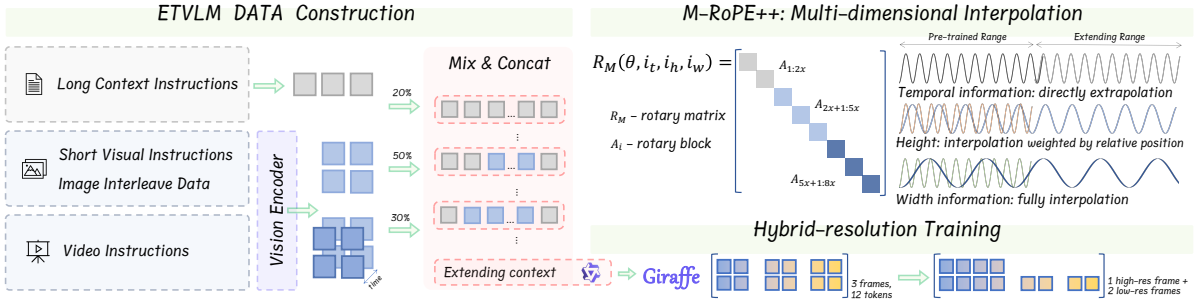

Figure 1: Pipeline of extending visual language models. We collect data from text, text-image pairs, and videos. We propose M-RoPE++ in extending training and hybrid-resolution inference to enhance the model performance.

LongAlign-10K (Bai et al., 2024) and LongAlpaca (Chen et al., 2023c), with typical lengths ranging from 10K to 100K tokens. (ii) Short multimodal instruction data, drawn mainly from LLaVA-Instruct (Liu et al., 2023c) and M3IT (Li et al., 2023b). While the original datasets are generally under 10K tokens, we concatenate samples to achieve lengths between 10K and 100K tokens. (iii) Interleaved multimodal pre-training data, comprising multiple images with typical lengths of 1K–10K tokens, sourced from MMDU (Liu et al., 2024c) and Mantis (Jiang et al., 2024). We also process interleaved image data from arXiv following the arXivQA protocol (Li et al., 2024c). (iv) Long multimodal instruction data, created by sampling multiple frames from video datasets, primarily sourced from ShareGPT4V (Chen et al., 2023a) and ShareGPT4O (Chen et al., 2024). To address the scarcity of long video instruction data, we sample videos longer than 5 minutes from MLVU (Zhou et al., 2024), ensuring MLVU is excluded from our test set to maintain fair evaluation. The data composition details are summarized in Table 1.

Data Processing All data are processed into a dialogue format consistent with ChatML style (OpenAI, 2024). Data are maintained in their original length and as concatenated multi-turn dialogues. For original-length text instruction data, we filter out special tokens. For short visual instruction and interleaved data, we adjust formatting and remove unnecessary symbols. Video data are sampled at 2 fps to reduce computational overhead. During data concatenation, we aim to match the target context length (e.g., 32K, 128K) as closely as possible without truncating content, ensuring a balance between efficiency and context preservation.

2.3 Data Recipe Exploration

We investigate the impact of different data ratio and data length on downstream task performance and provide recommendations for optimal data recipes. Using the same number of training tokens across all datasets, we conduct experiments with Qwen-VL (Bai et al., 2023) as the base model.

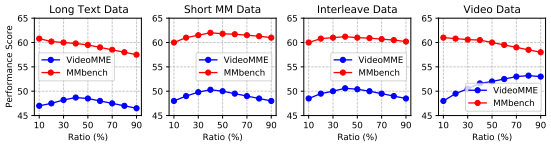

Figure 2: Performance of extending Qwen-VL with different data composition ratios.

Data Ratio To further investigate the impact of data composition on model performance, we conduct experiments by varying the proportion of a single data type from 10% to 90% while keeping the total training volume consistent. The results presented in Figure 2 reveal that increasing the proportion of long video data improves long video comprehension but compromises performance on

other tasks. Similarly, increasing the ratio of any specific data type predominantly enhances its associated downstream task performance. Based on these findings, we determine the final data composition strategy, as shown in Table 1, which modestly increases the proportion of video data while reducing the share of pure text data. This adjusted recipe achieves a well-balanced performance across diverse task types.

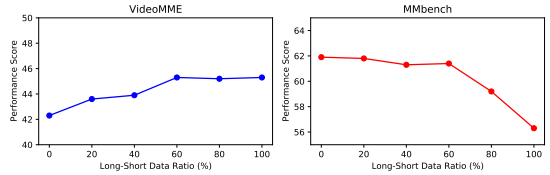

Figure 3: Performance on Qwen-VL trained with different composition ratio of long (>8K) and short data.

Data Length We categorize data into long data and short data based on whether their length exceeds 8K tokens. We investigate how different ratios of long and short data affect downstream performance on both long-context and short-context tasks. As shown in Figure 3, increasing the proportion of long data leads to improved performance on long-context tasks, with performance plateauing after the long data ratio reaches 60%. However, for short-context tasks, when the proportion of long data exceeds 60%, there is a notable decline in performance. Based on these observations, we adopt a 60% long data ratio for our extending training to achieve an optimal balance between long and short task performance.

Findings 1

Short multimodal instruction data is crucial for both extending long context capability and retaining short context performance. A balanced data ratio contributes to balanced performance on downstream tasks.

3 How to Extend Context Length

In this section, we test the effective length of existing length-extending methods, address their limitations (§3.1), and introduce our position embedding technique M-ROPE++ (§3.2). We find that for extending VLMs, it is sufficient to tune the LLM base of VLMs without requiring multi-stage training (§3.3). We propose hybrid-resolution training to further leverage the fixed context length (§3.5).

3.1 Effective Length of VLMs

To evaluate the effective context length of VLMs, we draw inspiration from recent studies on LLMs, which suggest that their effective lengths are often only about half of their training lengths (An et al., 2024b; Gao et al., 2024a). We adopt the single needle setting from Visual Haystack (Wu et al., 2024c), where models process varying numbers of input images and are tasked with identifying specific images and answering questions such as, "For the image with the anchor object, is there a target object?" This setup enables the assessment of performance across different context lengths, with random guessing yielding a 50% success rate. All tests are conducted using native image resolutions consistent with the original configuration. As shown in Figure 4, retrieval success rates decrease as the number of input images grows. We define an accuracy threshold of 60% to determine the effective length. The base Qwen2-VL model achieves effectiveness up to 15 images, corresponding to an effective length to approximately 10K tokens. After extending the training length to 128K tokens using existing length-extending methods like PI and NTK, the effective length increases to around 50 images, equivalent to approximately 40K tokens—still less than one-third of the training length. These findings highlight that the extended VLMs, similar to LLMs, exhibit the *falls short* phenomenon (An et al., 2024b), where effective length falls short of the training length. These findings highlight the need for a novel position-extending method to enhance the effective length of models.

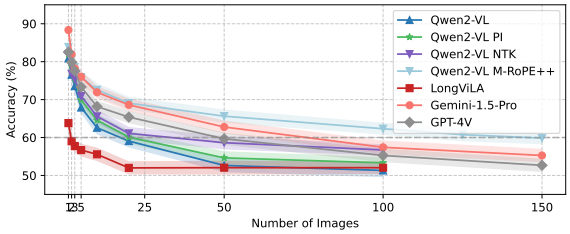

Figure 4: Results on visual haystack. The x-axis shows the number of input images, and the y-axis shows the retrieval success rate. The dashed line indicates the 60% threshold for effective length.

Findings2

The effective length in VLMs, including models that utilize existing position-extending methods, is smaller than the training length.

3.2 Position Extending on VLM

In this subsection, we briefly introduce M-RoPE, discuss potential issues associated with existing position extending methods, and then present our proposed M-RoPE++ along with experimental results validating its effectiveness.

M-RoPE Multimodal Rotary Position Embedding (M-RoPE) proposed in Qwen2-VL (Wang et al., 2024a) extends the RoPE (Su et al., 2024) to effectively model positional information with multi-dimensions. M-RoPE deconstructs the original rotary embedding into three components: temporal, height, and width. The formal definition of M-RoPE and RoPE can be found in Appendix B.

For a 16x-dimensional M-RoPE matrix, the dimensions are allocated in a 2:3:3 ratio for temporal, height, and width components respectively. This can be represented as:

$$R_M(\theta, i_t, i_h, i_w) = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{8x} \end{bmatrix}, \quad (1)$$

where each $A_i \in \mathbb{R}^{2\times 2}$ is a rotary block and i_t, i_w, i_h are position indices. θ represents the rotary base. The blocks are allocated as follows:

- A_1 to A_{2x} represent the temporal dimension;
- A_{2x+1} to A_{5x} represent the height dimension;
- A_{5x+1} to A_{8x} represent the width dimension.

Each rotary block A_i is defined as:

$$A_i = \begin{bmatrix} \cos(i_x\theta_d) & -\sin(i_x\theta_d) \\ \sin(i_x\theta_d) & \cos(i_x\theta_d) \end{bmatrix}, \quad (2)$$

where i_x represents i_t, i_h, or i_w depending on which dimension the block belongs to. The frequency basis θ is shared across all dimensions.

Position extending on M-RoPE In M-RoPE, the temporal index are allocated to the lower dimensions of the rotary embedding, which correspond to high-frequency information. Preserving this information is crucial for maintaining the model's ability to discern temporal order. Position extending methods such as position interpolation (PI; Chen et al. 2023b) or modifying the RoPE base (NTK; LocalLLaMA 2023) tend to compress high-frequency signals indiscriminately, **potentially confusing the model's perception of order of close-by frames**. Conversely, the height and width dimensions occupy higher-dimensional spaces in the rotary embedding, indicating that they may not have fully covered the rotational domain during pre-training. This necessitates the application of interpolation to these dimensions. To address this, we propose M-RoPE++ that applies extrapolation exclusively to the temporal index and apply interpolation on height and weight index.

M-RoPE++ We begin by defining key parameters following YaRN ((Peng et al., 2023) :

$$s = \frac{L'}{L_V}, \quad (3)$$

where s is the ratio between the extended context length L' and the original visual context length L_V.

We define λ_d as the wavelength of the RoPE embedding at the d-th hidden dimension:

$$\lambda_d = \frac{2\pi}{\theta_d} = 2\pi b^{\frac{2d}{|D|}}, \quad (4)$$

and introduce the ratio r:

$$r = \frac{L'}{\lambda}. \quad (5)$$

For M-RoPE, the index range is divided into three segments: temporal (t), height (h), and width (w). Temporal information is predominantly in high-frequency, which has been covered during pre-training stage. Therefore, we maintain extrapolation for this segment. For the height and width segments, where $\lambda > L'$, indicating insufficient rotational domain training, we employ interpolation to preserve their performance. This design is illustrated in Figure 1 right part.

We propose the following piecewise function to obtain the updated θ'_d for M-RoPE++:

$$\theta'_d = \begin{cases} \theta_d & \text{if } 0 < d \le 2x, \\ (\frac{1}{s} + (1-\frac{1}{s}) \cdot \frac{d - r_{5x}}{r_{2x} - r_{5x}}) \cdot \theta_d & \text{if } 2x < d \le 5x, \\ \frac{\theta_d}{s} & \text{if } 5x < d \le 8x. \end{cases} \quad (6)$$

Experiment Validation We conduct a comparative analysis of various methods for extending the

Method	VideoMME Long Score (Frames)					VH(Images)
	64	128	256	512	768	100
Direct extrapolation	52.5	54.3	56.0	55.4	55.6	51.3
PI training	52.1	54.6	56.7	56.0	55.1	57.8
NTK-aware	**53.8**	54.8	55.8	56.2	56.0	56.7
M-RoPE++	53.4	**55.9**	**57.5**	**58.5**	**58.5**	**61.3**

Table 2: Comparison of position embedding extension methods on VideoMME long video task and visual haystack on Qwen2-VL.

context length of VLMs, focusing on their performance on the VideoMME long context task and Single Needle Visual Haystacks in Table 2.

Our results demonstrate that M-RoPE++ consistently surpasses other methods, showing continued improvement as the number of frames increases in VideoMME Long tasks. This indicates that M-RoPE++ effectively captures long-range dependencies in video data. While direct extrapolation shows some potential for context extension, increasing the frame count without additional training does not lead to further performance gains. The PI method, due to significant interpolation of high-frequency information, exhibits slight performance degradation on shorter tasks. The NTK-aware approach achieves better results than the base model but still falls short of M-RoPE++ when handling higher frame counts, emphasizing the importance of preserving the original RoPE base in temporal dimensions. In the Visual Haystack test with 100 images, M-RoPE++ outperforms all baseline methods, demonstrating its ability to further enhance the effective length of VLMs. These findings highlight the effectiveness of M-RoPE++ in extending context length in VLMs.

Findings 3

The effective lengths achieved by existing position-extending methods remain insufficiently long. M-RoPE++ achieves better downstream task performance and longer effective length in the same training length.

3.3 Multi-Stage Training

We investigate whether multi-stage training strategies commonly used in VLM training are necessary for extending context length. Previous works on long-context VLMs, typically training from an LLM base, often employ multiple stages, including extending the text-based model's context length, multimodal alignment, and multimodal instruction tuning. For extending existing VLMs like Qwen2-VL, we explore three approaches: (1) train VLM with mixed instruction data while only updating LLM backbone, (2) extending the LLM base with additional pure text data (Wiki-103) followed by multimodal instruction data, like LongVA (Zhang et al., 2024a), and (3) multimodal alignment using image-text pairs (Sampled from LAION-5B) followed by instruction tuning (Xue et al., 2024; Wang et al., 2024b). As shown in Table 3, our

Training Strategy	MMBench	BLINK	VideoMME
One-stage MM Instruction	**82.8**	**54.6**	**58.5**
Two-stage Text Extending + MM Instruction	79.8	52.9	58.1
Two-stage MM Alignment + MM Instruction	80.5	51.2	57.8

Table 3: Comparison of different training strategies for extending Qwen2-VL context length.

experiments indicate that pre-extending the text-based model with pure text data provides no significant advantage. This is likely because training with long-context multimodal data already addresses diverse length distributions, rendering pure text extension redundant. Moreover, performing multimodal alignment before instruction tuning degrades performance on short-context tasks. This could be attributed to Qwen2-VL already undergoing instruction tuning before extending training; further tuning of MLP and ViT layers with alignment objectives may disrupt the model's learned distributions. With fixed training steps, this disruption negatively impacts short-context performance without yielding improvements for long-context multimodal tasks.

Findings 4

Directly train VLM with mixed instruction data while only updating LLM backbone's parameters achieves optimal results.

3.4 Trade-off in Fixed Context Length

When encoding videos with a fixed total number of visual tokens, there exists an inherent balance between the resolution of each frame and the number of frames included. To investigate this balance on video tasks, we test various combinations of frame counts and resolutions, adjusting one in response to changes in the other. Table 4 summarizes the results of GIRAFFE on VideoMME medium and long sub-tasks under these configurations, highlighting the impact of different frame-resolution trade-offs.

Frame Count	Image Token Count	VideoMME Medium	VideoMME Long
128	960	62.5	55.6
256	480	63.9	57.3
512	240	64.6	58.2
768	160	**64.8**	**58.5**
768	120	64.3	58.3
1024	120	64.7	**58.5**

Table 4: Performance of different frame counts and resolutions on VideoMME tasks for GIRAFFE.

From the perspective of frame count, performance on medium-length tasks tends to plateau at 512 frames, with little to no substantial improvement beyond this threshold. For longer tasks, however, increasing the frame count continues to yield performance gains, despite a corresponding reduction in the resolution of each frame. Notably, when the frame count is high but individual frame resolution is already low, further compression of resolution negatively impacts performance. These findings highlight the importance of a strategy that preserves high resolution for critical frames while accommodating longer sequences.

3.5 Hybrid-resolution Training

To address this, we propose hybrid-resolution training, inspired by SlowFast (Feichtenhofer et al., 2019), which reduces token usage while maintaining performance in long-form video understanding tasks. We partition the video frames into N groups, each containing L frames. For each group, we process the first frame using a high-resolution image that occupies m visual tokens. The subsequent $L-1$ frames within the group are processed that occupy $\frac{m}{s}$ tokens, where s is the compression ratio. This approach significantly reduces the token usage from $L*N*m$ tokens to $(1+\frac{L-1}{s})*N*m$ tokens. The high-resolution frames at the beginning of each group provide detailed visual information, while the low-resolution frames maintain temporal continuity and context at a reduced computational cost. This design is illustrated in Figure 1.

The results in Table 5 demonstrate the effectiveness of hybrid-resolution training. Comparing the first two rows, we observe that reducing the resolution of low-res frames using hybrid resolution only marginally affects downstream task performance while halving visual token usage. Furthermore, the bottom two rows reveal that under equivalent visual token constraints, hybrid-resolution inference enables increased resolution for high-res frames and successfully enhances downstream task performance. These findings suggest that hybrid-resolution inference offers a promising approach to optimize the trade-off between computational efficiency and model performance in long-form video understanding tasks. We use (L,m,s)=(4,240,3) by default for other evaluations.

Frames Count	(L,m,s)	Avg. Image Tokens	VideoMME Medium	VideoMME Long
512	(1,240,1)	240	64.2	57.9
512	(4,240,3)	120	64.0	57.6
1024	(1,120,1)	120	64.7	58.5
1024	(4,240,3)	120	**66.2**	**60.4**

Table 5: Performance comparison of hybrid-resolution training settings on VideoMME tasks.

Findings 5

Hybrid-resolution training can further improve the performance of VLM in a fixed context length.

4 Extended VLMs

In this section, we first present the experimental setup and the relevant models, followed by an analysis of their performance across various downstream tasks. For infrastructure and engineering details, please refer to Appendix E.

4.1 Models

We assess the following models: **Qwen-VL-Chat-7B** (Bai et al., 2023) A visual language model based on the Qwen language model, incorporating visual capabilities through cross-attention and learnable query embeddings. **VideoLLaVA-7B** (Lin et al., 2024) A video-language model that extends LLaVA to handle video inputs, capable of processing up to 8 frames. **VideoChat2-Mistral-7B** (Li et al., 2024b) An advanced VLM built on the Mistral-7B, designed to process up to 16 frames. **LongVA-7B** (Zhang et al., 2024a) A long context VLM based on Qwen-2 language model, utilizing a two-stage alignment process to handle up to 128 frames. **LongVILA-8B** (Xue et al., 2024) A long context VLM based on VILA language model, capable of processing up to 256 frames. **Qwen2-VL** (Wang et al., 2024a) A foundational VLM that employs dynamic image tokenization and M-RoPE, with pre-trained 16K context length. We select Qwen2-VL

Methods	Frames	VideoMME				Frames	LongVideoBench				Avg
		Short	Medium	Long	Overall		(8, 15)	(15, 60)	(180, 600)	(900, 3600)	
Close-source VLMs											
GPT-4V (turbo)	10	70.5	55.8	53.5	59.9	256	66.4	71.1	61.7	54.5	59.1
GPT-4o	384	80.0	70.3	65.3	71.9	256	71.6	**76.8**	**66.7**	**61.6**	**66.7**
Gemini-1.5-Pro	1/0.5fps	**81.7**	**74.3**	**67.4**	**75.0**	256	**68.3**	73.2	63.1	56.3	62.7
Open-source VLMs											
VideoLLaVA-7B	8	45.3	38.0	36.2	39.9	8	43.1	44.6	36.4	34.4	39.1
VideoChat2-Mistral-7B	16	48.3	37.0	33.2	39.5	16	49.3	49.3	39.0	37.5	39.3
VideoLLaMA2-7B	16	56.0	45.4	42.1	47.9	32	-	-	-	-	45.3
LLaVA-NeXT-Qwen2-7B	32	58.0	47.0	43.4	49.5	32	-	-	-	-	47.9
LongVA-7B	128	61.1	50.4	46.2	52.6	128	-	-	-	-	50.1
LongVILA-8B	256	61.8	49.7	39.7	50.5	256	-	-	-	-	48.7
Qwen-VL-Chat-7B	4	46.9	38.7	37.8	41.1	4	-	-	-	-	40.7
GIRAFFE-QwenVL	128	55.4	51.2	46.9	51.2	128	-	-	-	-	50.9
Qwen2-VL-7B	256	**71.2**	62.5	56.0	63.2	256	**67.8**	70.4	56.6	51.3	61.5
GIRAFFE	768	71.1	64.8	58.5	64.8	768	67.4	70.6	59.1	55.9	63.3
w/ Hybrid-res train&inf	1024	71.1	**66.2**	**60.5**	**65.9**	1024	67.4	**71.0**	**60.8**	**58.1**	**64.3**

Table 6: Performance comparison across VLMs on VideoMME and LongVideoBench tasks. We bold the best results for both close-source and open-source VLMs. We choose the best frames from our experiments in §3.4 and only use Hybrid-res inference on tasks above 512 frames.

(for GIRAFFE), Qwen-VL (for GIRAFFE-QwenVL) as the base model with the best extending training setting shown in §2 and §3.

4.2 Video Task Results

Our extended models, GIRAFFE-QwenVL and GIRAFFE, demonstrate substantial improvements in video understanding across various temporal scales while specifically maintaining competitive performance on short videos. Table 6 shows that GIRAFFE-QwenVL significantly outperforms its base model Qwen-VL-Chat, enabling better understanding of video content. Notably, GIRAFFE, based on an improved base model and capable of processing 1024 frames, achieves state-of-the-art performance among open-source models in both VideoMME and LongVideoBench, even surpassing GPT-4V in several categories. These results provide compelling evidence that our approach successfully extends the context window of VLMs, particularly benefiting long context video understanding tasks while reserving original short context capacities.

4.3 Image Task Results

The results from Table 7 demonstrate that our GIRAFFE maintains competitive performance on short-form multimodal tasks. This balanced capability can be attributed to our training strategy, which incorporates a mix of short instruction data alongside long context video inputs. Incorporating LLaVA-Instruct and M3IT in our training process ensures the model retains its capacity in single-image understanding. For multi-image task results, please refer to Appendix 5.

Model	MME_p	MME_c	MMBench(en)
GPT-4V	1590.5	573.2	82.8
Qwen-VL	1487.6	360.7	60.9
GIRAFFE-QwenVL	1489.7	372.9	61.5
Qwen2-VL	**1695.3**	1630.4	**82.8**
GIRAFFE	1692.9	**1635.4**	82.1

Table 7: VLM performance on the single-image scenario: MME and MMBench tasks. We bold the best results and underline the second best.

5 Multi Image Task Results

Model	Mantis-Eval	QBench	BLINK
LLaVA-v1.5-7B	31.3	49.3	37.1
GPT-4V	62.7	76.5	51.1
Qwen-VL	39.2	45.9	31.1
GIRAFFE-QwenVL	48.3	57.4	41.2
Qwen2-VL	63.4	**76.9**	53.3
GIRAFFE	**63.9**	76.8	**54.5**

Table 8: VLMs results on multi-image scenario: Mantis-Eval, QBench and BLINK. We bold the best results and underline the second best.

In the multi-image evaluation presented in Table 8, GIRAFFE-QwenVL exhibits substantial improvements, whereas GIRAFFE also demonstrates enhancements, validating the efficacy of our pipeline. In multi-image scenarios, context length is less critical than in long video tasks. Qwen-VL's

superior performance stems from capacities trained on the ETVLM dataset, compared to its initial 2K context length. In contrast, Qwen2-VL has already undergone substantial pre-training in 16K contexts. Additionally, Qwen2-VL benefits from a broader range of training data compared to Qwen-VL, rendering the incremental advantages from ETVLM data relatively modest.

6 Related Work

6.1 Long Context Language Models

The main solution for long context scenery addresses the out-of-distribution issue with position-embedding and enhancing model extrapolation capabilities. Training-free methods like streamingLLM (Xiao et al., 2024b), InfLLM (Xiao et al., 2024a) and ChunkLLaMA (An et al., 2024a) offer cost-effective ways to scale context window size. Additionally, further training using modified RoPE (Su et al., 2024) base frequency is introduced in NTK (LocalLLaMA, 2023), PI (Chen et al., 2023b) and YaRN (Peng et al., 2023), a effective practice adopted by models such as CodeLlama (Rozière et al., 2024) and LLaMA 3.1 (Team, 2024). Moreover, efforts have also been made on data curation for long context training (Bai et al., 2024; Gao et al., 2024b; Fu et al., 2024c). However, corresponding comprehensive studies on extending context for open-source VLMs remain limited.

6.2 Long Visual Language Models

For long context VLMs, recent LongVA (Zhang et al., 2024a) are first extending an LLM base model to 128K token lengths and then developing it into a VLM. Concurrent work LongVILA (Xue et al., 2024) also involves multi-stage training starting from an LLM backbone and employs an improved sequence parallel technique for efficient training, while LongLLaVA (Wang et al., 2024b) combines Mamba and Transformer blocks to reduce memory usage. In contrast, our model GIRAFFE optimizes various data recipes and position extending designs, establishing itself as the state-of-the-art among open-source long VLMs.

7 Conclusion and Future Work

We develop an effective solution to extend the context length of VLMs while preserving their performance on shorter contexts. Our comprehensive experiments led to the introduction of the ETVLM dataset for extended training and M-RoPE++ for improved position embedding learning. We use Hybrid-res training to better use long context window. Our extended model, GIRAFFE, achieves state-of-the-art performance for long context tasks. In the future, we aim to apply GIRAFFE to more complex scenarios, such as long-term history multimodal chats and visual agents in real-world applications.

Limitations

Our study has several limitations that warrant consideration. (i) Due to limited computational resources, we were unable to conduct a more comprehensive exploration of optimal data ratios through additional experiments. This limitation may have prevented us from determining a more precise and effective data composition for training. (ii) The current implementation of M-RoPE++ is restricted to models pre-trained with M-RoPE. Adapting this technique to other model architectures remains a subject for future investigation. (iii) Our evaluation primarily focused on question-answering tasks due to the scarcity of diverse long context video datasets. This constraint limits our ability to assess the model's performance in more realistic application scenarios, such as embodied agents or long-term visual agents. Addressing these limitations in future work could potentially yield more robust and generalizable long context visual language models.

Ethical Considerations

The ethical considerations for our study encompass several key aspects: (i) Data sourcing: All data utilized in our research was obtained from publicly shared sources, adhering strictly to their respective open-source licenses. (ii) Model development: Our further training on the Qwen model complies fully with Qwen's licensing agreements. (iii) Evaluation methodology: We exclusively employed automated evaluation tools for assessment, avoiding the need for human annotators. (iv) Potential misuse: While we have focused on benign applications, we acknowledge the potential for misuse of advanced visual language models and encourage ongoing discussions on responsible AI development and deployment.

References

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. 2024a. Training-free long-context scaling of large language models.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024b. Why does the effective context length of llms fall short?

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024. Longalign: A recipe for long context alignment of large language models.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation.

Yukang Chen, Shaozuo Yu, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023c. Long alpaca: Long-context instruction-following models. `https://github.com/dvlab-research/LongLoRA`.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. 2022. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024c. Data engineering for scaling language models to 128k context. In *Forty-first International Conference on Machine Learning*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024a. How to train long-context language models (effectively).

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024b. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. arXiv2405.01483.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024b. Videochat: Chat-centric video understanding.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024c. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023b. M^{3}IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, abs/2306.04387.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with ringattention. *arXiv preprint*.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *ArXiv*, abs/2310.01889.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player?

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024c. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*.

LocalLLaMA. 2023. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.

OpenAI. 2023. Gpt-4v(ision) system card. *OpenAI Research*.

OpenAI. 2024. Chatml documents.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code.

Jianlin Su. 2023. Extending llm context window beyond 2048 tokens.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer.

Llama Team. 2024. The llama 3 herd of models.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024b. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024b. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. 2024c. Visual haystacks: A vision-centric needle-in-a-haystack benchmark.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. 2024a. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. *arXiv*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024b. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2024. Longvila: Scaling long-context visual language models for long videos.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Zhilin Zhu. 2023. Ring flash attention. https://github.com/zhuzilin/ring-flash-attention.

A Data Composition

Table 1 shows the composition details of our ETVLM data.

B RoPE and M-RoPE

Attention is defined over C embeddings $X = [x_1, x_2, \ldots, x_C]^T \in \mathbb{R}^{C\times d}$ where d is the model dimension. Learned weight matrices $W_v \in \mathbb{R}^{d\times d_k}$, $W_q \in \mathbb{R}^{d\times d_k}$, and $W_k \in \mathbb{R}^{d\times d_k}$ are used to transform these inputs where d_k is the projected hidden dimension. The attention mechanism itself computes the attention matrix and applies it to produce a weighted sum of the value vectors:

$$\text{Attention}(Q, K, V) = AV = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (7)$$

Basic attention was originally defined with: $Q = XW_q$, $K = XW_k$, $V = XW_v$. However, this approach does not directly encode the relative position of keys and values.

Rotary Position Embeddings (RoPE) (Sun et al., 2022) encode positional information by applying a phase rotation to each element of the embedding vectors. Formally, we define a transformation f:

$$f_W(x_i, \theta) = R(\theta, i)W^T x_i \quad (8)$$

Here $x_i \in \mathbb{R}^{d_k}$ is an embedding for position i, W is a projection matrix, and $\theta \in \mathbb{R}^{d_k/2}$ is a frequency basis. The function is defined based on the rotary position matrix:

$$R(\theta, i) = \begin{bmatrix} \cos i\theta_1 & -\sin i\theta_1 & \cdots & 0 & 0 \\ \sin i\theta_1 & \cos i\theta_1 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & \cos i\theta_{d_k/2} & -\sin i\theta_{d_k/2} \\ 0 & 0 & \cdots & \sin i\theta_{d_k/2} & \cos i\theta_{d_k/2} \end{bmatrix} \quad (9)$$

Due to the arrangement of frequencies, this matrix has the property that $R(\theta, n-m) = R(\theta, m)^T R(\theta, n)$ by Ptolemy's identity. We redefine the query-key product between two positions m and n as,

$$q_m^T k_n = f_{W_q}(x_m, \theta)^T f_{W_k}(x_n, \theta) \quad (10)$$

Multimodal Rotary Position Embedding (M-RoPE) extends the concept of RoPE to effectively model positional information of multimodal inputs. M-RoPE deconstructs the original rotary embedding into three components: temporal, height, and width. For text inputs, these components utilize

identical position IDs, making M-RoPE functionally equivalent to 1D-RoPE. For image inputs, the temporal IDs remain constant, while distinct IDs are assigned to the height and width components based on the token's position in the image. For video inputs, the temporal ID increments for each frame, while the height and width components follow the same ID assignment pattern as images.

Formally, we define the M-RoPE transformation function f_M as:

$$f_M(x_i, \theta_t, \theta_w, \theta_h) = [R_t(\theta_t, i_t)W_t^T x_{it}; \\ R_w(\theta_w, i_w)W_w^T x_{iw}; \\ R_h(\theta_h, i_h)W_h^T x_{ih}] \quad (11)$$

where x_i is the embedding vector, θ_t, θ_w, θ_h are frequency bases, i_t, i_w, i_h are position indices, and W_t, W_w, W_h are projection matrices for temporal, width, and height dimensions respectively.

The query-key product for M-RoPE is then redefined as:

$$q_m^T k_n = f_M(x_m, \theta_t, \theta_w, \theta_h)^T f_M(x_n, \theta_t, \theta_w, \theta_h) \quad (12)$$

For a 16x-dimensional M-RoPE matrix, the dimensions are allocated in a 2:3:3 ratio for temporal, height, and width components respectively. This can be represented as:

$$R_M(\theta, i_t, i_h, i_w) = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{8x} \end{bmatrix} \quad (13)$$

where each $A_i \in \mathbb{R}^{2\times 2}$ is a rotary block. The blocks are allocated as follows:

- A_1 to A_{2x} represent the temporal dimension
- A_{2x+1} to A_{5x} represent the height dimension
- A_{5x+1} to A_{8x} represent the width dimension

Each rotary block A_i is defined as:

$$A_i = \begin{bmatrix} \cos(i_x\theta_d) & -\sin(i_x\theta_d) \\ \sin(i_x\theta_d) & \cos(i_x\theta_d) \end{bmatrix} \quad (14)$$

where i_x represents i_t, i_h, or i_w depending on which dimension the block belongs to. The frequency basis θ is shared across all dimensions.

This formulation allows M-RoPE to effectively model multimodal inputs while maintaining the rotary structure for each dimension.

C Impact of RoPE Base

We investigated the effect of different RoPE bases on the performance of Qwen-VL. Our findings indicate that the optimal performance was achieved by following the recommendations from Su's blog, specifically using a RoPE base of 500,000 for a context length of 128k. Increasing the base beyond this point did not yield significant improvements while keeping the default base of 10,000 resulted in a notable performance drop. Table 9 summarizes our results.

RoPE Base	VideoMME Long	VideoMME Avg	MME Sum	MMBench
10,000 (default)	39.5	41.1	1848.29	60.9
500,000 (optimal)	**43.2**	**51.2**	**1862.62**	**61.5**
1,000,000	43.1	51.1	1862.20	61.4

Table 9: Performance comparison of different RoPE bases across various benchmarks.

These results underscore the significance of meticulously adjusting the RoPE base when expanding the context window of visual language models. Our findings corroborate the conclusions presented in Su's blog (Su, 2023), which posits that for models with a context length of 128k, an optimal RoPE base of 4.9×10^6 is recommended. This value closely approximates our selected base of 5×10^5, which consistently demonstrates superior performance compared to the default configuration across all evaluated metrics.

Interestingly, further increasing the base beyond this point does not yield significant performance improvements. This observation is consistent with the approaches taken by models like LLaMA 2 and Qwen, which have opted for even larger base values. Such choices may provide additional flexibility for future extensions of model context lengths.

The effectiveness of the optimized RoPE base in capturing long-range dependencies in multimodal data underscores the critical role of position embedding strategies in enhancing the performance of extended visual language models.

D Progressive Extending

To ensure more stable training, we adopted a progressive extending strategy. For GIRAFFE-QwenVL, we set multiple incrementally increasing context lengths: 8K, 32K, 64K, and 128K. We concatenate and chunk ETVLM data according to these different context lengths. For GIRAFFE-QwenVL, we investigate the optimal RoPE base setting, as detailed in Appendix C. Following Su (2023), we

Method	MME_P	MME_e	VideoMME
Single-step (2k→128k)	1462.58	350.71	48.9
Progressive	**1487.58**	**360.71**	**51.2**

Table 10: Comparison of single-stage and progressive extension methods on Qwen-VL.

experiment with bases of 5×10^4, 1×10^6, 2.5×10^6, and 5×10^6. For GIRAFFE, we employ M-RoPE++, training up to 64K before extending to 128K. This approach allows the model to gradually adapt to longer sequences while maintaining performance on shorter contexts.

Ablation of progressive extending We conduct comparative experiments on Qwen-VL to evaluate two methods for extending the model's context length: a single-stage approach and a progressive multi-stage approach. Both methods are using the same number of training steps. The results are summarized in Table 10. Our experiments demonstrate that the progressive extending approach consistently outperforms the single-stage method across different evaluated tasks. This suggests that gradually increasing the context length during training allows the model to better adapt to longer sequences, resulting in improved performance on various tasks.

E Infrastructure and Engineering

We employ the NTK method for Qwen-VL and M-RoPE++ for GIRAFFE to extend the model's window length. Training long VLMs results in substantial memory demands, thus we employ several optimization strategies to perform training on such long sequences. These include FlashAttention-2 (Dao et al., 2022; Dao, 2024), Ring Attention (Liu et al., 2023a), ZERO (Rajbhandari et al., 2020) (including activation checkpointing, and parameter offload). To balance the load across 8 80G H100 GPUs, we shard the sequence in a zigzag way (Zhu, 2023). We use LoRA (Hu et al., 2022) to reduce the GPU memory usage to train longer VLMs. We train the model for an average of 80 H100 hours.