A Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

B Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. Code will be made publicly available to facilitate replication and verification after inspection. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We believe these measures will enable other researchers to reproduce our work and further advance the field.

C LLM Usage

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

D Appendix

D.1 Practical Considerations: MHRoPE vs. MRoPE-I

While both of our proposed methods are effective, we currently recommend MRoPE-I over MHRoPE for two primary reasons: its consistent (albeit slight) performance advantage and its greater implementation simplicity. We attribute MHRoPE's minor performance deficit to its head-level information partitioning, which prevents the integration of different positional axes within the self-attention mechanism. From an engineering perspective, MRoPE-I is also simpler, avoiding the complexities that MHRoPE introduces with distributed training paradigms like tensor parallelism. Nevertheless, MHRoPE's design offers a potentially more scalable architecture for future models that may need to accommodate a larger number of positional axes.

D.2 Derivation of the Attention Score Upper Bound in MRoPE

Here, we provide a formal derivation for the upper bound of the RoPE attention score. The RoPE dot product between a query $\boldsymbol{q}$ and a key $\boldsymbol{k}$ at a relative position $m-n$ can be expressed in complex form as:

$$(\boldsymbol{\mathcal{R}}_m \boldsymbol{q})^\top (\boldsymbol{\mathcal{R}}_n \boldsymbol{k}) = \mathrm{Re}\left[\sum_{i=0}^{d/2-1} (\boldsymbol{q}_{[2i:2i+1]} \cdot \boldsymbol{k}^*_{[2i:2i+1]}) e^{\mathrm{i}(m-n)\theta_i}\right] \tag{1}$$

where $\boldsymbol{v}^*$ denotes the complex conjugate of a 2D vector treated as a complex number, and $\cdot$ is the complex product.

To derive the upper bound, we analyze the magnitude of the summation term. To apply summation by parts, let us define a content-dependent sequence $h_i = \boldsymbol{q}_{[2i:2i+1]} \cdot \boldsymbol{k}^*_{[2i:2i+1]}$ and a position-dependent sequence of partial sums $S_j = \sum_{k=0}^{j-1} e^{\mathrm{i}(m-n)\theta_k}$. We also set the boundary conditions $S_0 = 0$ and $h_{d/2} = 0$. The standard summation by parts formula is $\sum_{i=a}^{b} u_i \Delta v_i = [u_i v_i]_a^{b+1} - \sum_{i=a}^{b} v_{i+1} \Delta u_i$. Applying this, the magnitude of the summation can be rewritten and bounded as follows:

$$
\begin{aligned}
\left| \sum_{i=0}^{d/2-1} h_i e^{\mathrm{i}(m-n)\theta_i} \right| &= \left| [h_i S_i]_0^{d/2} - \sum_{i=0}^{d/2-1} S_{i+1}(h_{i+1} - h_i) \right| \\
&= \left| (h_{d/2} S_{d/2} - h_0 S_0) - \sum_{i=0}^{d/2-1} S_{i+1}(h_{i+1} - h_i) \right| \\
&= \left| - \sum_{i=0}^{d/2-1} S_{i+1}(h_{i+1} - h_i) \right| \\
&\leq \sum_{i=0}^{d/2-1} |S_{i+1}||h_{i+1} - h_i| \\
&\leq \left(\max_{0 \leq i < d/2} |h_{i+1} - h_i| \right) \sum_{i=0}^{d/2-1} |S_{i+1}|
\end{aligned}
\tag{2}
$$

This final expression reveals that the upper bound is a product of two distinct components. $\max |h_{i+1} - h_i|$ is a content-dependent term that acts as a scaling factor based on the specific query and key vectors. The second, $\sum |S_{i+1}|$, is a purely position-dependent term whose value is determined only by the relative position $m - n$ and the fixed frequencies θ_i. Since the content-dependent term is independent of position, the long-range decay property of the attention score is governed primarily by this position-dependent term. Therefore, its average value, $\frac{1}{d/2} \sum_{i=1}^{d/2} |S_i|$, serves as a practical indicator to characterize how the upper bound attenuates with relative distance.

D.3 Compatibility with YaRN Extrapolation

As shown in Figure **??**, the interleaved frequency allocation of MRoPE-I makes it compatible with extrapolation algorithms like NTK-aware (**?**) and YaRN (**?**). Whereas standard MRoPE's partitioned spectrum complicates the application of a consistent frequency scaling boundary, our interleaved design provides a full spectrum across all positional axes, enabling a straightforward and symmetric application of these methods.

To validate this effect, we apply YaRN to both MRoPE and MRoPE-I under a 256K context window and evaluate their performance on LVBench and MLVU. As shown in Table 1, MRoPE-I with YaRN demonstrates substantially larger gains in long-video understanding compared to MRoPE.

Table 1: Performance comparison of MRoPE and MRoPE-I with and without YaRN under a 256K context.

Method	**LVBench**	**MLVU**
MRoPE	41.5	62.9
MRoPE + YaRN	41.2	63.3
MRoPE-I	42.0	63.2
MRoPE-I + YaRN	**43.6**	**64.1**

D.4 Long-Context Video Understanding

We further compare the performance of different methods on long video understanding, with context lengths ranging from 32K to 256K. As shown in Figure 1, apart from LVBench, we do not observe

clear performance improvements or degradation when extrapolating to longer sequences. The only exception is Vanilla RoPE, which suffers from a sharp performance drop at 128K/256K. We attribute this to excessively fast-growing position IDs, which lead to degraded extrapolation capability, which also discussed in other works (**??**).

Overall, methods such as VideoRoPE and HoPE, which allocate most low-frequency channels to the temporal axis, exhibit slightly better extrapolation ability in long video senario. However, when considering performance across images and grounding tasks, MHRoPE and MRoPE-I remain the most comprehensive and balanced designs.

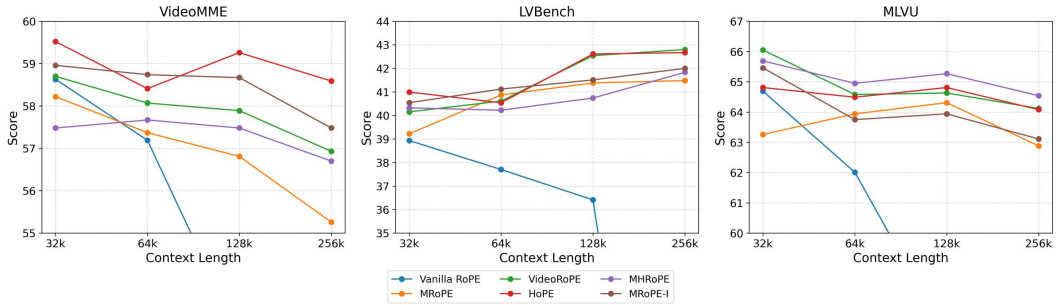

Figure 1: Video extrapolation performance. Models are trained with a context length of 32k (256 frames) and extrapolated to 64k (512 frames), 128k (1024 frames), and 256k (2048 frames).

D.5 More Ablation Results.

D.5.1 Enhanced visual attention in *spatial-reset*

To understand the mechanism driving the effectiveness of *spatial-reset*, we analyzed its impact on the model's attention patterns. As detailed in Table 2, we calculated the total attention scores on visual tokens using the DocVQA test set. Specifically, we extracted attention scores from layers 4, 12, 20, and 28, and averaged the scores across all attention heads and samples. The result demonstrates that MRoPE equipped with *spatial-reset* allocate more attention on visual content, particularly in deeper layers, confirming it's effectiveness in enhancing the model's visual focus.

Method	Layer 4	Layer 12	Layer 20	Layer 28
MHRoPE	40.31	21.76	32.05	19.00
w/o *spatial-reset*	35.99	19.68	22.02	9.93
MRoPE-I	37.48	15.68	28.08	23.23
w/o *spatial-reset*	31.22	17.66	16.02	11.69

Table 2: Average attention scores (%) on visual contents. The inputs are from DocVQA test set. And the scores are averaged between attention heads and samples.

D.5.2 Allocation Ratio of Frequency

We further investigate different allocation ratios under the interleave frequency strategy. The results are summarized in Table 3. The balanced allocation (t:h:w = 24:20:20) achieves the best overall performance. Increasing the proportion of channels assigned to the temporal axis reduces the available high-frequency capacity for spatial dimensions. This leads to a degradation in grounding ability and negatively impacts benchmarks involving spatial understanding in both images and videos.

D.5.3 Temporal Stride in Video Modeling

This section investigates the impact of different temporal strides between video frames. Specifically, we experiment with $\delta = 0.5, 1, 2$, as well as dynamic strides as used in V2PE and HoPE (with $\delta = 1$ applied during inference). The results are shown in Table 4.

Allocation Ratio	Image	Video	Grounding	Overall
0:32:32	66.42	51.01	76.02	64.48
12:26:26	66.30	51.93	75.77	64.67
24:20:20	66.65	52.36	75.85	**64.95**
32:16:16	64.07	51.15	74.65	63.29
48: 8: 8	65.06	51.17	72.87	63.03

Table 3: Ablation results of different frequency allocation ratios under interleave design.

Stride	MVBench	STAR	VideoMME	LVBench	MLVU	Charades	**Overall**
0.5	56.55	57.90	58.96	38.99	62.37	31.88	51.11
1	**57.05**	57.79	58.96	40.54	**65.46**	**34.36**	**52.36**
2	55.70	**58.13**	58.15	38.02	63.11	33.51	51.10
Dynamic	56.28	57.93	58.74	**41.12**	63.75	32.99	51.80

Table 4: Comparison of temporal stride settings for video benchmarks on MRoPE-I.

From the results, $\delta = 1$ achieves the best overall performance, while smaller ($\delta = 0.5$) or larger ($\delta = 2$) strides lead to performance drops. Incorporating the dynamic stride from V2PE does not shows a significant benefit.

D.6 Full Experiment Results on Qwen3-VL

We present the complete results of our main experiments on Qwen3-VL. Table 5 and Table 6 report the performance of **Qwen3-VL-4B-Instruct** and **Qwen3-VL-8B-Instruct**, respectively. Across both model scales, MHRoPE and MRoPE-I consistently outperform other multimodal positional encoding variants. The experimental settings are identical to those in the main experiments.

Table 5: Overall performance of multimodal RoPEs variants on various benchmarks, evaluated on Qwen3-VL-4B-Instruct. The highest score is shown in **bold**, while the second-highest score is underlined.

Types	**Benchmarks**	Vanilla RoPE	MRoPE	VideoRoPE	HoPE	CircleRoPE	MHRoPE	MRoPE-I
Image	MMMU	45.78	42.89	44.44	45.44	45.33	46.11	**46.33**
	MMBench$_{avg}$	68.77	71.26	69.20	70.79	71.35	**71.56**	71.20
	MMstar	43.67	42.87	41.87	43.53	43.73	43.73	**45.00**
	OCRBench	35.20	37.30	32.10	28.70	37.10	**38.60**	38.30
	AI2D	66.71	67.29	63.21	63.46	66.06	68.32	**69.29**
	RealworldQA	57.12	56.60	57.25	57.52	57.25	58.86	**59.78**
	DocVQA	46.59	43.13	25.27	27.25	44.60	46.77	**46.89**
	TextVQA	46.65	48.25	47.10	47.16	47.27	**48.60**	48.27
	InfoVQA	30.55	28.14	16.22	17.32	30.02	30.93	**32.46**
	ChartQA	39.64	38.80	39.37	39.52	40.68	39.76	**41.64**
	BLINK	36.33	36.80	34.44	34.46	37.52	37.22	**37.88**
Video	MVBench	54.30	53.95	53.83	54.53	53.13	**54.98**	54.30
	STAR	58.07	57.90	**58.77**	56.73	55.89	58.66	58.23
	MLVU	60.03	60.76	60.21	**61.72**	59.40	61.00	61.29
	VideoMME	50.93	50.41	50.19	**51.41**	50.44	51.30	50.70
	LVBench	36.93	36.35	36.09	37.02	35.54	**37.20**	36.99
	Charades-STA	39.16	38.20	40.06	40.08	40.98	**41.68**	41.44
Grounding	RefCOCO$_{val}$	22.51	25.53	19.58	19.74	20.38	26.70	**28.82**
	RefCOCO$_{testA}$	22.91	27.40	20.03	21.35	21.21	26.60	**28.91**
	RefCOCO$_{testB}$	23.20	25.14	21.90	20.27	20.88	28.13	**30.68**
	RefCOCO+$_{val}$	17.12	20.37	15.45	15.98	15.48	21.33	**22.61**
	RefCOCO+$_{testA}$	18.49	21.76	15.37	16.66	16.59	21.87	**23.85**
	RefCOCO+$_{testB}$	18.20	20.37	17.71	16.32	16.22	23.54	**25.81**
	RefCOCOg$_{val}$	22.32	24.90	20.34	20.45	20.10	26.80	**30.00**
	RefCOCOg$_{test}$	21.38	23.89	20.93	19.19	19.82	26.90	**29.52**
Overall	Image	47.00	46.67	42.77	43.20	47.36	48.22	**48.82**
	Video	49.90	49.60	49.86	50.25	49.23	**50.80**	50.49
	Grounding	20.77	23.67	18.91	18.74	18.84	25.23	**27.52**

Table 6: Overall performance of multimodal RoPEs variants on various benchmarks, evaluated on Qwen3-VL-8B-Instruct. The highest score is shown in **bold**, while the second-highest score is underlined.

Types	**Benchmarks**	Vanilla RoPE	MRoPE	VideoRoPE	HoPE	CircleRoPE	MHRoPE	MRoPE-I
Image	MMMU	51.18	50.11	51.78	51.89	50.67	53.33	**53.89**
	MMBench$_{avg}$	79.29	78.27	78.74	79.59	79.00	**80.78**	79.50
	MMstar	51.88	52.53	53.33	52.86	**54.40**	53.20	53.47
	OCRBench	73.20	72.90	67.70	61.70	73.60	**74.40**	73.90
	AI2D	78.90	77.56	78.34	77.85	77.59	78.22	**79.50**
	RealworldQA	63.27	64.05	62.48	62.48	61.96	**67.33**	65.22
	DocVQA	82.41	82.85	71.71	72.95	82.28	82.41	**83.67**
	TextVQA	63.15	66.05	61.54	62.33	64.94	67.20	**67.32**
	InfoVQA	**53.84**	49.60	39.15	41.69	51.06	52.22	52.62
	ChartQA	62.52	61.04	61.00	59.04	62.96	61.89	**63.44**
	BLINK	37.21	37.44	37.08	37.22	40.52	40.00	**40.45**
Video	MVBench	60.35	59.90	59.70	60.53	59.53	60.20	**60.70**
	STAR	61.65	62.29	61.21	62.19	61.51	62.33	**62.45**
	MLVU	66.01	67.34	68.31	**68.81**	66.61	67.20	67.52
	VideoMME	60.07	59.78	**61.48**	60.56	51.04	60.89	60.89
	LVBench	42.22	41.06	**42.93**	42.48	38.80	42.00	41.58
	Charades-STA	50.39	51.79	**52.90**	52.30	51.77	52.13	52.70
Grounding	RefCOCO$_{val}$	73.22	74.16	71.68	73.22	74.70	76.88	**77.72**
	RefCOCO$_{testA}$	76.45	77.11	73.77	74.50	76.77	79.41	**83.21**
	RefCOCO$_{testB}$	71.40	72.09	69.89	70.93	73.22	72.89	**76.45**
	RefCOCO+$_{val}$	65.77	65.94	62.05	66.33	66.37	67.99	**70.88**
	RefCOCO+$_{testA}$	71.17	72.25	68.08	72.89	72.41	75.08	**77.79**
	RefCOCO+$_{testB}$	59.97	60.46	59.26	60.58	61.16	63.18	**64.33**
	RefCOCOg$_{val}$	71.14	73.10	69.77	71.28	72.33	74.14	**76.47**
	RefCOCOg$_{test}$	71.57	72.68	69.45	71.82	72.67	74.65	**76.79**
Overall	Image	63.35	62.95	60.26	59.96	63.54	64.63	**64.82**
	Video	56.78	57.03	57.75	**57.81**	54.88	57.46	57.64
	Grounding	70.09	70.97	67.99	70.19	71.20	73.03	**75.46**