HoPE: Hybrid of Position Embedding for Long Context Vision-Language Models

Haoran Li[1] **Yingjie Qin**[2] **Baoyuan Ou**[2] **Lai Xu**[2] **Ruiwen Xu**[2]
[1]Carnegie Mellon University [2]Xiaohongshu Inc.
haoranl4@cs.cmu.edu

Abstract

Vision-Language Models (VLMs) have made significant progress in multimodal tasks. However, their performance often deteriorates in long-context scenarios, particularly long videos. While Rotary Position Embedding (RoPE) has been widely adopted for length generalization in Large Language Models (LLMs), extending vanilla RoPE to capture the intricate spatial-temporal dependencies in videos remains an unsolved challenge. Existing methods typically allocate different frequencies within RoPE to encode 3D positional information. However, these allocation strategies mainly rely on heuristics, lacking in-depth theoretical analysis. In this paper, we first study how different allocation strategies impact the long-context capabilities of VLMs. Our analysis reveals that current multimodal RoPEs fail to reliably capture semantic similarities over extended contexts. To address this issue, we propose HoPE, a Hybrid of Position Embedding designed to improve the long-context capabilities of VLMs. HoPE introduces a hybrid frequency allocation strategy for reliable semantic modeling over arbitrarily long contexts, and a dynamic temporal scaling mechanism to facilitate robust learning and flexible inference across diverse context lengths. Extensive experiments across four video benchmarks on long video understanding and retrieval tasks demonstrate that HoPE consistently outperforms existing methods, confirming its effectiveness. Our code is available at https://github.com/hrlics/HoPE.

1 Introduction

Vision-Language Models (VLMs) [1–5] have achieved remarkable success in multimodal tasks, including visual question answering [6–9], image captioning [10, 11], cross-modal retrieval [12, 13], and more [14–16]. However, VLMs suffer from significant performance degradation in long-context scenarios, particularly long videos [17–20]. In such settings, VLMs even struggle with simple tasks like object counting and temporal localization [21, 22], revealing a critical limitation in their ability to model extended spatial-temporal dependencies. This limitation substantially hinders their real-world deployment, where input length often exceeds the context window they have been pretrained on.

Rotary Position Embedding (RoPE) [23] has been widely adopted for length generalization in text-based LLMs [24–26]. Specifically, RoPE incorporates positional information by partitioning the query and key vectors into 2-dimensional pairs and rotating each pair at a unique frequency that decreases as the dimensional index increases. Despite its advantages, directly applying 1D RoPE fails to capture the intricate spatial-temporal dependencies in videos. Several methods have been proposed to extend 1D RoPE for multimodal inputs [2, 27, 28]. Among these, the most common approach is to allocate different frequencies to encode different positional components, as shown in the upper plots of Figure 1. For example, M-RoPE [2] allocates the *highest* frequencies for temporal modeling (t), and the remaining low frequencies for spatial modeling (x, y). In contrast, VideoRoPE [28] proposes to allocate the *lowest* frequencies for temporal dimensions (t), and further

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

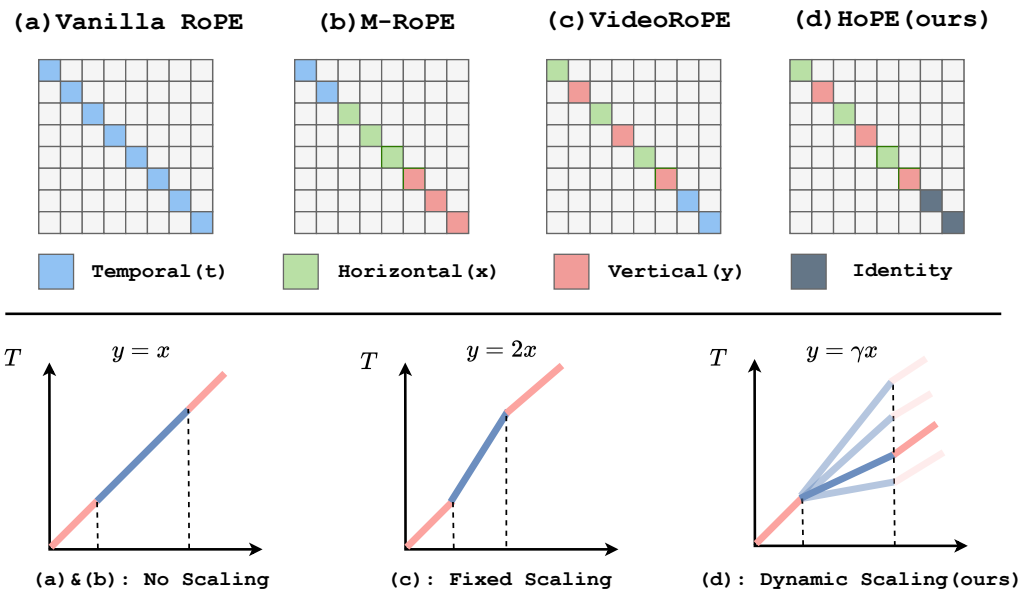


Figure 1: **Comparison of our HoPE and existing methods.** Upper plots illustrate the frequency allocation strategies in different RoPE variants. Here, frequency decreases along the diagonal. (d) HoPE sets the lowest frequencies to zero for reliable long-range semantic modeling. Lower plots demonstrate different temporal scaling mechanisms. (d) HoPE proposes dynamic and bidirectional scaling to learn temporal dynamics at multiple scales, facilitating robustness to various video speeds.

applies a fixed scaling factor to scale the temporal indices of visual tokens, as shown in the lower plots of Figure 1. Despite their improved performance, two significant challenges remain unsolved. Firstly, current methods mainly rely on heuristics rather than theoretical analysis to determine the frequency allocation strategy. Second, applying a fixed and unidirectional scaling factor for all videos is suboptimal in real-world scenarios, where videos proceed at different speeds and demonstrate significant variance in information densities.

To address these challenges, we begin with an in-depth theoretical analysis in Section 3, outlining the ideal properties that a multimodal RoPE should possess. Our analysis reveals that: (1) the flattening operation in vanilla RoPE inherently violates spatial-temporal locality prior, which is crucial in video modeling; (2) despite diverse frequency allocation strategies, existing multimodal RoPE variants fail to reliably capture semantic similarities over extended contexts; (3) temporal scaling of visual tokens should include both compression and expansion to accommodate varying video speeds in real-world scenarios. Guided by these insights, we propose HoPE, a **H**ybrid **o**f **P**osition **E**mbedding designed to improve the long-context capabilities of VLMs. As shown in Figure 1, HoPE first introduces a hybrid frequency allocation strategy to facilitate long-range semantic modeling. In this strategy, higher frequencies, which are more sensitive to positional differences and better at capturing local features, are allocated to spatial components (x, y) in an interleaved manner. Meanwhile, the lower frequencies are directly set to zero and allocated to temporal component (t) to enable reliable semantic modeling. Moreover, HoPE develops a dynamic temporal scaling mechanism for lengthy visual tokens. This mechanism not only enhances VLMs' robustness to various video speeds, which are common in real-world scenarios, but also offers flexible scaling during inference across diverse context lengths.

We summarize our contributions as follows:

- To our best knowledge, we provide the first theoretical analysis of how different frequency allocation strategies in multimodal RoPEs impact the long-context capabilities of VLMs, offering insights for the design and analysis of future multimodal RoPEs.
- Guided by our analysis, we propose HoPE, which consists of a hybrid frequency allocation strategy for reliable semantic modeling in long-context scenarios, and a dynamic temporal scaling mechanism for robust and flexible temporal comprehension.
- Extensive experiments on four video benchmarks demonstrate that HoPE consistently outperforms existing RoPE variants, achieving improvements of 22.23% and 8.35% on long video retrieval and long video understanding tasks, confirming its effectiveness.

2 Preliminaries

Rotary Position Embedding (RoPE). Current Transformer-based LLMs rely on Positional Encodings (PEs) to incorporate sequential information into the attention mechanism. Among various PEs, Rotary Position Embedding (RoPE) [23] has emerged as a dominant approach for long-context modeling in text-based LLMs [24–26]. The key to RoPE's success lies in its ability to encode *relative* position information through an *absolute* positional encoding approach, ensuring both effectiveness and efficiency. Consider query and key vectors with d dimensions (where d is an even number), RoPE partitions the dimensions into $d/2$ pairs, e.g., $\mathbf{q}_n = [\mathbf{q}_n^{(0)}; \mathbf{q}_n^{(1)}; \dots; \mathbf{q}_n^{(d/2-1)}]$. Each pair of dimensions is assigned a unique rotation angle $\theta_i = b^{-2i/d}, i \in \{0, 1, \dots, d/2-1\}$, where b is a pre-defined frequency base and set to $10,000$ by default [23]. This rotation can be achieved through a rotation matrix as follows:

$$r^{(i)} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}. \tag{1}$$

The overall rotation matrix $\mathbf{R}_n$ is constructed by concatenating each pair's rotation matrix along the diagonal to form a block-diagonal matrix, i.e., $\mathbf{R}_n = \mathrm{diag}(r^{(0)}, r^{(1)}, \dots, r^{(d/2-1)}) \in \mathbb{R}^{d\times d}$. Therefore, during attention computation, the attention score[1] $\mathbf{A}_{n,m}$ between the n-th query $\mathbf{q}_n$ and m-th key $\mathbf{k}_m$ is:

$$\mathbf{A}_{n,m} = (\mathbf{q}_n\mathbf{R}_n)(\mathbf{k}_m\mathbf{R}_m)^\top = \mathbf{q}_n\mathbf{R}_{n-m}\mathbf{k}_m^\top, \tag{2}$$

where the rotation matrix $\mathbf{R}_{n-m}$ can be formulated as:

$$\mathbf{R}_{n-m} = \begin{pmatrix} \cos\theta_0(n-m) & -\sin\theta_0(n-m) & \cdots & 0 & 0 \\ \sin\theta_0(n-m) & \cos\theta_0(n-m) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos\theta_{(d/2-1)}(n-m) & -\sin\theta_{(d/2-1)}(n-m) \\ 0 & 0 & \cdots & \sin\theta_{(d/2-1)}(n-m) & \cos\theta_{(d/2-1)}(n-m) \end{pmatrix}. \tag{3}$$

It can be observed that through pairwise attention computation, the final rotation matrix naturally incorporates the *relative* position information $(n-m)$ between the query-key pair.

No Positional Encoding (NoPE). Despite the popularity of RoPE, several works have pointed out that the *causal* attention mechanism in current decoder-only LLMs implicitly learns *absolute* positional information [30–32]. This motivates the development of No Positional Encoding (NoPE). Specifically, the *causal* attention mask enforces $A_{n,m} = 0$ for all $m > n$, ensuring that each token only attends to itself and previous tokens. Under this constraint, the attention score with NoPE is a simple dot product between the query vector $\mathbf{q}_n$ and the key vector $\mathbf{k}_m$, i.e., $\mathbf{A}_{n,m} = \mathbf{q}_n\mathbf{k}_m^\top$, providing no explicit positional information to Transformers.

3 Analysis

In this section, we conduct a comprehensive theoretical analysis of multimodal RoPE variants, aiming to answer the following questions: **(1)** Is vanilla RoPE enough for long-context VLMs? **(2)** How do different frequency allocation strategies impact semantic modeling in long-range multimodal contexts? **(3)** How should we assign the temporal indices for text and visual tokens?

3.1 Vanilla RoPE Fails in Spatial-Temporal Structure

Several recent VLMs [1, 5, 20, 33–35] still use vanilla RoPE for multimodal inputs. In their approach, each video frame is first encoded by a vision encoder (e.g., ViT [36]) and then flattened into a sequence of patch-level tokens. These visual tokens will be treated equally as text tokens for

[1]Here, we omit the softmax function and $1/\sqrt{d}$ scaling in standard Transformer [29] for simplicity.

positional encoding, with each token incorporating only 1D temporal information. We show in Proposition 3.1 that this approach, while easy to implement, distorts spatial-temporal localities and fundamentally limits VLMs' ability to model extended spatial-temporal dependencies.

Proposition 3.1 (1D RoPE violates spatial-temporal locality priors)**.** *Given any query* $\mathbf{q}$ *at position* (t, x, y) *and a relative distance of 1 in spatial or temporal dimensions, the flattening operation in 1D RoPE distorts the relative distance with a magnitude dependent on the frame resolution.*

We provide the proof in Appendix A.1. This mismatch between positional encoding and the 3D structure of videos creates distorted attention patterns, making it difficult for models to learn meaningful spatial-temporal relationships essential for video-related tasks.

Conclusion 1. Directly applying vanilla RoPE to multimodal long-context inputs inherently fails to capture their complex spatial-temporal dependencies.

3.2 Current Multimodal RoPEs Are Unreliable in Long-Range Semantic Modeling

To capture the spatial-temporal structure of multimodal inputs, a recent VLM, Qwen2-VL [2], has introduced a Multimodal Rotary Position Embedding (M-RoPE). Concretely, M-RoPE partitions the 128-dimensional rotary embedding into three distinct groups: the first 32 dimensions for temporal information (t), the subsequent 48 dimensions for horizontal spatial information (x), and the final 48 dimensions for vertical spatial information (y), i.e., $\mathbf{R}_{t,x,y} = \mathrm{diag}(\mathbf{R}_t, \mathbf{R}_x, \mathbf{R}_y)$. While this approach realizes a naive extension for RoPE, a fundamental question remains to be answered:

How do different frequency allocation strategies impact the performance of multimodal RoPE?

This question arises from the fact that in RoPE, different dimensions carry unique frequencies ($\theta_i = b^{-2i/d}$, $i \in \{0, 1, \ldots, d/2-1\}$), as shown in Equation 3. Therefore, different strategies exist for frequency allocation in multimodal RoPE. As shown in Figure 1, M-RoPE allocates the highest frequencies for t, intermediate frequencies for x, and the lowest frequencies for y. In contrast, VideoRoPE [28] proposes to assign the lowest frequencies to temporal modeling (t) and high frequencies to spatial dimensions (x, y). Their empirical justification stems from attention pattern analysis, which reveals that dimensions encoded with the lowest frequencies exhibit a more pronounced *attention sink* phenomenon [37], which has proven to be effective in long-context modeling. However, we argue that using the lowest frequencies for temporal modeling is still unreliable in capturing semantic similarities in extended multimodal contexts. Specifically, we first introduce semantic preference, a property where attention mechanisms should prioritize semantically similar tokens regardless of their relative distance, and formally define this concept in Definition 3.1.

Definition 3.1 (Semantic Preference)**.** For any query vector $\mathbf{q}$ and a semantically similar key vector $\mathbf{k}'$ that can be expressed as $\mathbf{k}' = \mathbf{q} + \delta$ where δ is a zero-mean perturbation, the attention score with RoPE should satisfy:

$$\mathbb{E}_{\mathbf{q},\mathbf{k},\delta}[\mathbf{q}\mathbf{R}_{\Delta t \Delta x \Delta y}\mathbf{k}'^{\top} - \mathbf{q}\mathbf{R}_{\Delta t \Delta x \Delta y}\mathbf{k}^{\top}] \geq 0,$$

where $\mathbf{k}$ is the key vector of a semantically unrelated token. This preference should hold regardless of the relative distance $(\Delta t, \Delta x, \Delta y)$ between the query and key.

Then, we show in Theorem 3.1 that ***all*** frequency allocation strategies of current multimodal RoPEs, including selecting the highest/lowest frequencies for temporal modeling, are unreliable in maintaining the semantic preference property over extended contexts. This limitation arises because, with sufficiently long contexts, even the lowest frequencies can produce arbitrary rotations, ultimately undermining semantic preference. We provide the proof in Appendix A.2.

Theorem 3.1. *Let* $X = [x_1, x_2, \ldots, x_L]$ *be an input sequence, and let RoPE use any fixed set of temporal frequencies (e.g., highest or lowest). Then there exists a critical length* L_c *such that for all* $L \geq L_c$*, the semantic preference property (Definition 3.1) is violated.*

Conclusion 2. There exist different frequency allocation strategies to extend vanilla RoPE to multimodal RoPE. However, we prove that none of these strategies can reliably maintain the semantic preference property over a sufficiently long context.

3.3 How to Assign Positional Index for Multimodal Inputs?

Currently, most VLMs [1–4, 19, 20] adopt the same temporal stride for video frames and text tokens, as shown in Figure 1. However, this approach overlooks the inherent difference in information densities between text and visual tokens. To address this issue, VideoRoPE [28] applies a fixed scaling factor (implemented as 2) to adjust the temporal indices of visual tokens, achieving better empirical performance. However, this rigid scaling approach lacks the flexibility needed for diverse real-world videos, which naturally vary in pace and information density. A more ideal approach would incorporate both temporal compression and expansion capabilities, allowing the model to learn multi-scale temporal relationships, thereby enabling more robust temporal modeling.

Conclusion 3. Temporal index scaling of visual tokens is crucial for balancing multimodal information, yet current methods lack flexibility and bidirectionality.

4 HoPE: Hybrid of Position Embedding for Long Context VLMs

To address the above challenges, we propose HoPE, a **H**ybrid **o**f **P**osition **E**mbedding designed to improve the long-context capability of VLMs. As illustrated in Figure 1 and Figure 2, HoPE first introduces a hybrid frequency allocation (HFA) strategy to better preserve the semantic preference property (Definition 3.1) in long-context modeling. Under this strategy, spatial information will be encoded with higher frequencies to capture local semantics, while the lowest frequencies will be set to zero (as in NoPE [30]) to facilitate long-range semantic modeling. Second, HoPE develops a dynamic temporal scaling (DTS) mechanism to enhance VLMs' robustness to various video speeds and enable flexible inference under diverse context lengths. We will detail these strategies as follows:

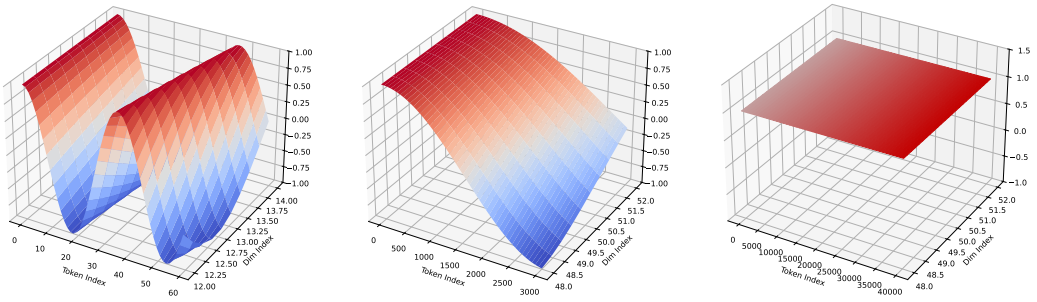
4.1 Hybrid Frequency Allocation Strategy

To extend vanilla RoPE to multimodal scenarios, a common approach is to allocate different frequencies to encode different positional components (t, x, y). For example, M-RoPE [2] assigns the highest frequencies for temporal modeling and lower frequencies for spatial encoding. In contrast, VideoRoPE [28] allocates the lowest frequencies for temporal modeling, achieving better empirical results. However, in Theorem 3.1, we theoretically prove that, despite using lower frequencies being more ideal for semantic modeling, none of these frequency allocation strategies can maintain the ideal semantic preference property (Definition 3.1) over extended contexts.

To provide a stronger theoretical guarantee for the semantic preference property, we propose a hybrid frequency allocation strategy. As shown in Figure 1, we encode spatial information (x, y) with high frequencies, as high frequencies are more sensitive to positional differences and thereby better at capturing local semantics [28, 38]. Following existing work [28], x and y are encoded in an interleaved manner to prevent biased spatial encoding. More importantly, unlike existing methods [23, 28, 2], we directly set the lowest frequencies to zero (as in NoPE [30]) to provide a stronger guarantee for the semantic preference property (Definition 3.1), as shown in Figure 2. Specifically, for $d = 128$, we interleave x and y positions in the first 96 dimensions of the rotation matrix and set the frequencies in the remaining 32 dimensions to zero, which corresponds to an identity matrix:

$$R_{\Delta x,\Delta y} = \mathrm{diag}\left(\begin{pmatrix} \cos\theta_0\Delta x & -\sin\theta_0\Delta x & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \sin\theta_0\Delta x & \cos\theta_0\Delta x & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\theta_1\Delta y & -\sin\theta_1\Delta y & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \sin\theta_1\Delta y & \cos\theta_1\Delta y & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{46}\Delta x & -\sin\theta_{46}\Delta x & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{46}\Delta x & \cos\theta_{46}\Delta x & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cos\theta_{47}\Delta y & -\sin\theta_{47}\Delta y \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \sin\theta_{47}\Delta y & \cos\theta_{47}\Delta y \end{pmatrix}, I_{32} \right)$$

We now provide a theoretical analysis of how this hybrid strategy helps the attention mechanism to capture long-range semantic similarities. Building on Definition 3.1 and Theorem 3.1, we first formalize the condition under which semantic preference is preserved in multimodal RoPE.



(a) High frequencies for temporal modeling in M-RoPE.

(b) Low frequencies for temporal modeling in VideoRoPE.

(c) Zero frequencies for temporal modeling in HoPE (ours).

Figure 2: **Multimodal RoPEs use different frequencies for temporal modeling.** M-RoPE uses the *highest* frequencies, which are suboptimal for long-context modeling. VideoRoPE utilizes the *lowest* frequencies for more stable semantic modeling. Our HoPE, employing *zero* frequencies for temporal modeling, establishes the upper bound of semantic modeling capabilities across all strategies.

In particular, Lemma 4.1 establishes a clear theoretical criterion for maintaining semantic preference with multimodal RoPE. It directly follows from our analysis in Theorem 3.1 and Appendix A.2, providing the theoretical foundation for our proposed method.

Lemma 4.1 (Necessary Condition for Semantic Preference)**.** *For a multimodal RoPE with rotation matrix* $\mathbf{R}_{t,x,y} = \mathrm{diag}(\mathbf{R}_t, \mathbf{R}_x, \mathbf{R}_y)$*, the semantic preference property (Definition 3.1) holds if, for all possible relative distances,*

$$\sum_{i \in i_t} 2\sigma^2 \cos(\Delta t \cdot \theta_i) + \sum_{i \in i_x} 2\sigma^2 \cos(\Delta x \cdot \theta_i) + \sum_{i \in i_y} 2\sigma^2 \cos(\Delta y \cdot \theta_i) \geq 0,$$

where σ^2 *is the variance of each component in the query/key vector,* i_t, i_x, i_y *are dimensions allocated to* t, x, y*, and* $\Delta t \in \{0, 1, \ldots, L-1\}$*,* $\Delta x \in \{0, 1, \ldots, H\}$*,* $\Delta y \in \{0, 1, \ldots, W\}$*.*

Based on this Lemma, we now prove how our hybrid frequency allocation strategy provides stronger guarantees for the semantic preference property over extended contexts. Specifically, HFA set $\theta_i = 0$ for all $i \in i_t$. Hence, the temporal terms in Lemma 4.1 reduce to $\sum_{i \in i_t} 2\sigma^2 \cdot 1$, noting that $\sum_{i \in i_t} 2\sigma^2 \cdot 1 \geq \sum_{i \in i_t} 2\sigma^2 \cos(\Delta t \cdot \theta_i)$ holds for any choice of temporal frequencies θ_i. Adding the identical spatial terms on both sides, we obtain:

$$\begin{aligned} &\sum_{i \in i_t} 2\sigma^2 \cdot 1 + \sum_{i \in i_x} 2\sigma^2 \cos(\Delta x \cdot \theta_i) + \sum_{i \in i_y} 2\sigma^2 \cos(\Delta y \cdot \theta_i) \\ &\geq \sum_{i \in i_t} 2\sigma^2 \cos(\Delta t \cdot \theta_i) + \sum_{i \in i_x} 2\sigma^2 \cos(\Delta x \cdot \theta_i) + \sum_{i \in i_y} 2\sigma^2 \cos(\Delta y \cdot \theta_i). \end{aligned} \tag{4}$$

This shows that our HFA strategy, by setting the lowest frequencies to zero, dominates any other choice of temporal frequencies and provides a stronger guarantee for preserving the semantic preference property under long-context scenarios, as in Theorem 4.1.

Theorem 4.1. *For multimodal position embeddings with dimensions allocated across temporal* (i_t)*, and spatial components* (i_x, i_y)*, setting* $\theta_i = 0$ *for all temporal dimensions* $i \in i_t$ *maximizes the semantic preference guarantee in Definition 3.1, compared to any alternative frequency allocation strategy, particularly under extended context lengths.*

Another interesting finding is that, if we set $|i_t| = d/4, |i_x| = |i_y| = d/8$ and $\theta_i = 0, i \in i_t$, for any context length t and spatial size x, y, semantic preference property invariably holds, as Lemma 4.1 reduces to $\sum_{i=0}^{d/8-1} 2\sigma^2(2 + \cos(\Delta x \cdot \theta_{2i}) + \cos(\Delta y \cdot \theta_{2i+1})) \geq 0$. However, the empirical results of this approach are inferior to our proposed HoPE, probably due to the decreased number of frequencies for spatial modeling. More discussions are provided in Appendix B.3.

4.2 Dynamic Temporal Scaling Mechanism

Considering the distinct information densities of text and visual tokens, HoPE introduces a dynamic temporal scaling mechanism that adjusts the temporal strides of visual inputs. Specifically, we first define a set of scaling factors, e.g., $\Gamma = \{0.5, 0.75, 1, 1.25, 1.5\}$, which includes both stretching ($\gamma > 1$) and compressing ($\gamma < 1$) operations. During training, the scaling factor γ is randomly selected from Γ and applied to each video. This allows the model to learn temporal relationships at multiple scales, making it more robust to variations in video speed, which are common in real-world scenarios. Consider a multimodal input $(text, video, text)$ of length L_t, L_v, and L_e, respectively. The position indices (t, x, y) for each token with our dynamic scaling factor γ are:

$$(t,x,y) = \begin{cases} (l,l,l), & 0 \le l < L_t \\ \begin{pmatrix} L_t + \gamma(l - L_t), \\ L_t + \gamma(l - L_t) + w - \frac{W}{2}, \\ L_t + \gamma(l - L_t) + h - \frac{H}{2} \end{pmatrix}, & L_t \le l < L_t + L_v \\ \begin{pmatrix} (\gamma - 1)L_v + l, \\ (\gamma - 1)L_v + l, \\ (\gamma - 1)L_v + l \end{pmatrix}, & L_t + L_v \le l < L_t + L_v + L_e \end{cases} . \tag{5}$$

Note that for visual tokens ($L_t \le l < L_t + L_v$), $l - L_t$ indicates the distance of the current frame from the start frame. During inference, scaling factors can be flexibly selected from the set to accommodate videos of different lengths. It is worth noting that unlike existing methods, which do not consider temporal scaling for visual tokens [1, 2, 4, 5] or just apply a fixed and unidirectional scaling factor for both training and testing [28], our methods not only help the model learn temporal relationships at multiple scales, but also offer flexibility during inference to accommodate various context lengths.

5 Experiment

In this section, we evaluate the performance of HoPE on four video benchmarks across long video understanding and long video retrieval tasks, aiming to validate its effectiveness in multimodal long-context modeling. Additionally, we conduct ablation studies to investigate the individual contribution of each strategy to overall performance and the interplay between task type, context length, and scaling factor selection.

5.1 Experimental Setups

Implementation Details. We utilize Qwen2-1.5B and Qwen2-7B [39] as the backbone models. By integrating these models with vision encoders from Qwen2-VL-2B/7B-Instruct [2], we obtain Qwen2-2B/7B-Video, respectively. During training, we adopt a batch size of 128, a learning rate of 1e-5(2B)/2e-5(7B) with a cosine scheduler. Following the instruction tuning settings in Qwen2-VL [2], we set the maximum video frames to 128 and the video sampling rate to 2. The training context length is set to 8k, with the entire training process taking approximately 304 GPU hours on machines equipped with H800-80GB GPUs. During evaluation, the minimum tokens per frame are set to 144.

Training Data. We train the models on a subset of LLaVA-Video-178k [40], which consists of 178k videos ranging from 0 to 3 minutes and 5M instruction samples, including captions, free-form, and multiple-choice question answering. Our selected subset includes 30k videos with durations under 2 minutes and 3k videos with durations between 2 and 3 minutes, resulting in roughly 300k pairs.

Baselines. We compare HoPE with the following RoPE variants: 1) vanilla RoPE [23], the standard approach in long-context LLMs, 2) M-RoPE [2], a famous RoPE extension in Qwen2-VL for multimodal inputs, 3) VideoRoPE [28], a specialized RoPE variant designed for video-related tasks.

Evaluation Benchmarks. We evaluate HoPE across four video benchmarks for long video understanding and long video retrieval tasks. For long video understanding, we utilize LongVideoBench [41], Video-MME [42], and MLVU [43], covering videos ranging from a few seconds to 2 hours. For long video retrieval, we employ V-NIAH (Visual Needle-In-A-Haystack) [17]. In this task, a

Table 1: **Performance comparison on long video understanding benchmarks.** The training context length of all methods is set to 8k, and we report the performance on 8k, 16k, 32k, and 64k to evaluate length generalization. The best results are **bold**, while the second best results are underlined.

Method	MLVU				LongVideoBench				Video-MME			
	8k	16k	32k	64k	8k	16k	32k	64k	8k	16k	32k	64k
Qwen2-2B-Video												
Vanilla RoPE	**55.10**	55.21	54.36	39.06	51.57	50.29	51.00	34.21	50.70	51.48	51.44	20.31
M-RoPE	53.26	53.69	54.73	40.63	50.81	52.26	51.30	44.74	51.44	51.22	51.52	23.44
VideoRoPE	54.75	55.19	54.00	42.19	52.17	52.02	51.31	36.84	50.89	50.52	50.56	15.63
HoPE (Ours)	54.89	**56.36**	**55.70**	**45.12**	**52.31**	**52.97**	**51.66**	**46.27**	**51.79**	**51.87**	**51.69**	**26.03**
Qwen2-7B-Video												
Vanilla RoPE	59.75	61.13	61.03	34.38	51.17	50.31	51.29	39.47	56.70	57.96	57.99	26.13
M-RoPE	59.70	61.68	62.46	46.88	52.27	53.29	53.49	50.00	56.81	57.77	58.37	23.43
VideoRoPE	60.40	61.82	62.51	45.31	52.89	53.13	53.82	47.37	57.51	59.00	59.13	26.52
HoPE (Ours)	**61.09**	**63.48**	**63.85**	**50.01**	**54.11**	**55.09**	**55.34**	**51.22**	**57.74**	**59.33**	**59.44**	**27.34**

"needle" image is randomly inserted into a "haystack" video, and the VLM is required to answer a question specifically about the embedded "needle" image. Following the protocol in V-NIAH [17], we utilize a haystack video with 1-hour duration (3,000 frames) and insert the needle image at 20% depth intervals (e.g., a frame depth of 0% would place the needle image at the very beginning of the video). For more detailed benchmark descriptions, please refer to Appendix B.1.

5.2 Results on Long Video Understanding

In this section, we provide a comprehensive comparison of HoPE and different RoPE variants in long video understanding. From Table 1, we observe that: **(1)** HoPE consistently outperforms all baselines across nearly all benchmarks, context lengths, and backbone sizes. Specifically, under the 7B model scale and 32k context lengths, HoPE surpasses vanilla RoPE by 2.82, 4.05, and 1.45 on MLVU, LongVideoBench, and Video-MME, respectively. This confirms its effectiveness and generalizability in multimodal long-context modeling. **(2)** The effectiveness of HoPE scales with backbone size. For instance, when the size of the backbone LLM increases from 2B to 7B, HoPE's performance gain on LongVideoBench (32k) significantly increases from 0.66 to 4.05 compared to vanilla RoPE. Notably, the performance gap between different methods on the 2B scale is less significant, probably due to the limited capabilities of the backbone LLM. **(3)** For context lengths under 64k, performance on Video-MME drops substantially, while the impact on MLVU and LongVideoBench is less pronounced. This suggests that extrapolating to extreme context lengths (e.g., up to 8x) remains highly challenging.

5.3 Results on Long Video Retrieval

We evaluate HoPE against other RoPE variants on V-NIAH [17] to demonstrate the superiority of our method in long video retrieval, where VLMs are required to identify specific frames in a video to answer the question. Figure 3 demonstrates that multimodal RoPEs significantly outperform vanilla RoPE, supporting our claim in Proposition 3.1 that the flattening operation in vanilla RoPE hinders spatial-temporal modeling. Furthermore, HoPE achieves better extrapolation than M-RoPE and VideoRoPE, confirming its effectiveness in multimodal long-context modeling. Quantitative results in Table 4 show that HoPE surpasses the best baseline by a significant margin of 22.23%.

5.4 Analysis

In this section, we first conduct ablation studies to analyze the effectiveness of each component in HoPE. We then present a comprehensive analysis exploring how different factors, including task type, context length, and the scaling factor of visual tokens, interact and impact model performance.

Figure 3: **Performance comparison on long video retrieval task (V-NIAH).** Here, each frame corresponds to 144 tokens. Cell colors indicate model accuracy (red: low, green: high), and the black dotted line marks the training context length (8k).

Ablation Studies. We conduct a series of ablation experiments to evaluate the impact of each component in HoPE and summarize the results in Table 2. According to the results, we observe that : (1) The 3D structure effectively improves the performance of vanilla RoPE in multimodal contexts, supporting our Proposition 3.1. (2) Based on the 3D structure, the hybrid frequency allocation (HFA) strategy further enhances long-range semantic modeling, achieving an average improvement of 1.69 across all context lengths. (3) The dynamic temporal scaling (DTS) mechanism facilitates VLMs' robustness to varying video speeds in real-world scenarios, yielding further performance gain. By combining the above strategies, our HoPE achieves the best overall performance across different context lengths in multimodal long-context modeling.

Table 2: **Ablation results on Video-MME from 8k to 64k.** Here, HFA: hybrid frequency allocation, DTS: dynamic temporal scaling.

Method	8k	16k	32k	64k
Vanilla RoPE	56.70	57.96	57.99	26.13
+ 3D structure	56.81	57.77	58.37	23.43
+ 3D + HFA	57.66	59.19	59.31	26.98
+ 3D + HFA + DTS	**57.74**	**59.33**	**59.44**	**27.34**

Impact of Test-time Scaling Factor Selection. We conduct further experiments to investigate how different scaling factors γ in our dynamic temporal scaling mechanism impact the performance of video-related tasks. We summarize the results on V-NIAH and LongVideoBench in Table 3. Our main observations are as follows: **(1)** *Long video retrieval generally prefers smaller scaling factors.* As shown in Table 3, when we utilize smaller scaling factors γ during inference, the performance on V-NIAH improves. We attribute this to the substantial length of 1-hour videos (3,000 frames), which far exceeds the training length (128 frames). In such cases, smaller scaling factors indirectly prevent the spatial position indices from becoming excessively large (see Equation 5), thereby providing a better guarantee for the semantic preference property. Therefore, we set $\gamma = 0.75$ for long video retrieval. **(2)** *Long video understanding generally benefits from larger scaling factors.* In contrast to retrieval, we find that long video understanding is relatively insensitive to the choice of scaling factor when the input context length is close to the training length. However, as the input length increases, employing larger scaling factors ($\gamma > 1$) results in better performance. We hypothesize that while smaller scaling factors help preserve the semantic preference property, larger

Table 3: **Ablation studies on test-time scaling factor selection.** We find that long video understanding generally benefits from larger scaling factors, while long video retrieval yields better results with smaller ones.

Scaling Factor γ	**LongVideoBench**				**V-NIAH**
	8k	16k	32k	64k	
0.50	**54.48**	54.29	54.36	52.63	60.89
0.75	54.36	54.97	54.72	52.63	**63.56**
1.00	54.11	54.48	54.97	52.63	62.67
1.25	54.11	54.84	**55.70**	52.63	62.67
1.50	54.11	**55.09**	55.34	51.22	61.78

scaling factors are beneficial for maintaining spatial details (also see Equation 5), which are crucial for complex understanding tasks. This introduces a natural tradeoff between semantic preference and spatial detail preservation. Compared to long video retrieval (3,000 frames, roughly 432k tokens), where extended temporal distances can significantly degrade semantic preference, in long video understanding tasks with context lengths of 16k–32k, the negative impact on semantic preference is relatively small. At the same time, the positive effect of larger scaling factors on capturing spatial details outweighs the semantic preference loss, making larger scaling factors overall more effective for complex video understanding. In our experiments, we set $\gamma = 1.5$ for long video understanding.

6 Related Work

Position Embedding in LLMs. Rotary Position Embedding (RoPE) [23] has become a common choice for position embedding in modern LLMs [24–26, 44]. As discussed in Section 2, RoPE achieves this success through rotating query and key vectors, encoding *relative* position information through an *absolute* positional encoding approach. Despite its success, several works have pointed out that No Position Embedding (NoPE) still works for decoder-only LLMs, arguing that the causal attention mechanism implicitly learns *absolute* position information [30, 31, 38]. These works even suggest that NoPE outperforms RoPE in out-of-distribution (OOD) scenarios. However, this observation remains unexplored in multimodal settings, where positional encoding strategies may have different implications for cross-modal interactions. Based on Lemma 4.1, we find that incorporating NoPE's zero frequency strategy indeed improves the length generalization of multimodal RoPE.

Multimodal Position Embedding in VLMs. In VLMs [1–5], images are first processed by vision encoders and then flattened into 1D tokens. Several early models [1, 4, 5] rely on vanilla RoPE for positional encoding, which distorts spatial-temporal locality (see Section 3) and limits VLMs' long-context capability. Recently, Qwen2-VL [2] introduced M-RoPE, which extends 1D RoPE to multimodal settings by assigning distinct frequency ranges to different positional components. Specifically, M-RoPE allocates the *highest* frequencies to the temporal component t, while distributing the lower frequencies sequentially to the spatial components x and y. Conversely, VideoRoPE [28] allocates the *lowest* frequencies to t to capture long-range dependencies, achieving stronger length generalization. However, these allocation strategies mainly rely on heuristics, lacking in-depth theoretical analysis. In contrast, our work theoretically analyzes how different frequency allocation strategies impact the performance of multimodal RoPE. By zeroing out low frequencies for temporal modeling, our proposed HoPE provides the strongest theoretical guarantee for long-range semantic modeling. HoPE's strength is further enhanced by its dynamic temporal scaling of visual tokens, which enables robust temporal learning during training and flexible scaling during inference. By integrating these advantages, HoPE achieves state-of-the-art performance in long video understanding and retrieval tasks, making it well-suited for long context VLMs.

7 Conclusion

This paper theoretically analyzes the limitations of current multimodal RoPE variants. Our analysis reveals that: (1) vanilla RoPE inherently fails in spatial-temporal modeling; (2) keeping all frequencies in multimodal RoPE is unreliable in capturing long-range semantic similarities; (3) temporal scaling of lengthy visual tokens should include both compression and expansion to accommodate various video speeds. Consequently, we introduce HoPE, a hybrid of position embedding designed to enhance the long-context capabilities of VLMs. HoPE proposes a hybrid frequency allocation strategy to facilitate long-range semantic modeling, and a dynamic temporal scaling mechanism to enhance VLMs' robustness to varying video speeds in real-world scenarios. Experimental results on long video understanding and long video retrieval tasks demonstrate that HoPE consistently outperforms existing methods across diverse context lengths and backbone sizes, confirming its effectiveness.

Limitations. While HoPE's performance gains scale from 2B to 7B backbones, our work does not use larger models or training data. We observe that the performance of all methods degrades significantly at 64k, though HoPE remains the most robust. While these resource-constrained evaluations are essential for uncovering genuine algorithmic benefits of multimodal RoPE, we note that training with more data, particularly long-context data, could further improve length generalization.

References

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[2] 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.

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[4] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[5] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[7] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, 2024.

[9] Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *arXiv preprint arXiv:2506.00555*, 2025.

[10] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023.

[11] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3136–3146, 2023.

[12] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11851–11861, 2024.

[13] Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305, 2024.

[14] Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alexander G Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. *ICLR*, 2022.

[15] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024.

[16] Feng Wang, Yaodong Yu, Guoyizhe Wei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Scaling laws in patchification: An image is worth 50,176 tokens and more. *ICML*, 2025.

[17] 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.

[18] 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, 2024.

[19] 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.

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[21] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024.

[22] Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tuny Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

[23] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[25] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[26] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[27] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.

[28] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? In *International Conference on Machine Learning*, 2025.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, 2022.

[31] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[32] Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuan-Jing Huang, and Xiaoling Wang. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14024–14040, 2024.

[33] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023.

[34] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[35] Haoran Li, Junqi Liu, Zexian Wang, Shiyuan Luo, Xiaowei Jia, and Huaxiu Yao. LITE: Modeling environmental ecosystems with multimodal large language models. In *First Conference on Language Modeling*, 2024.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

[38] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025.

[39] 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.

[40] 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.

[41] 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, 2024.

[42] Chaoyou Fu, Yuhan 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.

[43] 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.

[44] Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 87386–87410. Curran Associates, Inc., 2024.

[45] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10853–10862, 2022.

[46] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Flowar: Scale-wise autoregressive image generation meets flow matching. In *ICML*, 2025.

[47] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *ICCV*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are summarized in Figure 1. Section 3 provides theoretical analysis and Section 4 offers detailed methodology.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, please see Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the proofs in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setups are provided in Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use fully public datasets in our experiments and the details are provided in Section 5.1 and Appendix B.1. Code will be released in camera-ready version.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: In our experiments, the variance between different runs is negligible. Additionally, our training and evaluation pipelines are deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work focuses on improving the long-context capabilities of Vision-Language Models, which uses public datasets and is purely academic. We believe it has no direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not pose safety risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credited them in appropriate ways and followed their licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not include crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Proofs

In this section, we provide detailed proofs for the theoretical statements presented in this paper.

A.1 Vanilla RoPE Fails in Spatial-Temporal Structure

Proposition 3.1. *Given any query* $\mathbf{q}$ *at position* (t, x, y) *and a relative distance of 1 in spatial or temporal dimensions, the flattening operation in 1D RoPE distorts the relative distance with a magnitude dependent on the frame resolution.*

Proof. Consider a video of shape $T \times H \times W$, where each token at position (t, x, y) is flattened by

$$f(t, x, y) = tHW + xW + y.$$

Now consider two types of local neighbors:

1. **Spatial neighbors within the same frame**:

Let (t, x, y) and $(t, x + 1, y)$ be adjacent in the spatial dimension. Then,

$$|f(t, x + 1, y) - f(t, x, y)| = |((x + 1)W + y) - (xW + y)| = W. \tag{6}$$

Note that a relative distance of 1 in x becomes W after flattening, which grows linearly with the frame width.

2. **Temporal neighbors at the same spatial position**: Let (t, x, y) and $(t + 1, x, y)$ be adjacent in time. Then,

$$|f(t + 1, x, y) - f(t, x, y)| = |(t + 1)HW + xW + y - (tHW + xW + y)| = HW. \tag{7}$$

For a 1-frame shift in time, the index difference becomes HW, which grows with spatial resolution.

In both cases, spatially or temporally adjacent tokens are mapped to indices with significant differences. Since vanilla RoPE incorporates positional information based on these 1D index differences, such flattening leads to distorted spatial-temporal relationships. □

A.2 Semantic Preference Property

We now prove that the frequency allocation strategies in current multimodal RoPEs are unreliable in capturing semantic similarities over extended contexts, as defined in Definition 3.1.

Definition 3.1. **(**Semantic Preference**).** For any query vector $\mathbf{q}$ and a semantically similar key vector $\mathbf{k}'$ that can be expressed as $\mathbf{k}' = \mathbf{q} + \delta$ where δ is a zero-mean perturbation, the attention score with RoPE should satisfy:

$$\mathbb{E}_{\mathbf{q},\mathbf{k},\delta}[\mathbf{q}\mathbf{R}_{\Delta t \Delta x \Delta y}\mathbf{k}' - \mathbf{q}\mathbf{R}_{\Delta t \Delta x \Delta y}\mathbf{k}] \geq 0, \tag{8}$$

where $\mathbf{k}$ is the key vector of a semantically unrelated token. This preference should hold regardless of the relative distance $(\Delta t, \Delta x, \Delta y)$ between query-key pairs.

Firstly, we use Lemma A.1 to show why using lower frequencies for temporal modeling is more ideal in multimodal RoPE. Intuitively, larger rotation angles (frequencies) are more likely to produce negative cosine similarity values between semantically related tokens under long-context scenarios.

Lemma A.1. *Let* Δt *be drawn uniformly from* $\{0, 1, \ldots, L - 1\}$*, and define*

$$P_{\neg}(\theta) \;=\; \frac{1}{L}\,\big|\{\Delta : \cos(\theta\,\Delta t) < 0\}\big|.$$

Then for any $L > 1$*:*

1. *If* $0 < \theta < \frac{\pi}{2(L-1)}$*, then* $P_{\neg}(\theta) = 0$*.*
2. *For* $\theta \geq \frac{\pi}{2(L-1)}$*,* $P_{\neg}(\theta)$ *is non-decreasing in* θ*.*
3. $\lim_{\theta \to \infty} P_{\neg}(\theta) = \frac{1}{2}$*.*

Proof of Lemma A.1. 1. **No negative region for small θ.** If $0 < \theta < \frac{\pi}{2(L-1)}$, then for every $\Delta t \in \{0, \dots, L-1\}$ we have

$$0 \;\le\; \theta\,\Delta t \;<\; \theta\,(L-1) \;<\; \pi/2,$$

so $\cos(\theta\,\Delta t) > 0$. Hence $P_\neg(\theta) = 0$.

2. **Monotonicity once the first zero enters.** As soon as $\theta \ge \frac{\pi}{2(L-1)}$, the point satisfying $\theta\,\Delta t = \pi/2$ lies in $\{0, \dots, L-1\}$. Each further increase in θ extends the interval of length $\theta(L-1)$, adding more half-periods of cosine. Each added half-period contains exactly one "negative" region of length π. Therefore the count $\big|\{\Delta : \cos(\theta\Delta) < 0\}\big|$ (and hence $P_\neg(\theta)$) can only stay the same or increase, up to $O(1/L)$ rounding errors on the discrete grid.

3. **Limit to one half for large θ.** For large θ, the values $\{\theta\,\Delta\}_{\Delta=0}^{L-1}$ become equidistributed mod 2π. Since the negative region $\{\; x \mod 2\pi \mid \cos x < 0\}$ has total length π over each 2π-cycle, one finds

$$\lim_{\theta\to\infty} P_\neg(\theta) \;=\; \frac{\pi}{2\pi} \;=\; \tfrac{1}{2}.$$

□

We can now prove the frequency allocation strategies in current multimodal RoPE cannot reliably maintain the semantic preference property, i.e., semantically similar tokens should receive higher attention than semantically unrelated pairs.

Theorem 3.1. *Let $X = [x_1, x_2, \dots, x_L]$ be an input sequence, and let RoPE use any fixed set of temporal frequencies (e.g., highest or lowest). Then there exists a critical length L_c such that for all $L \ge L_c$, the semantic preference property (Definition 3.1) is violated.*

Proof. We first recall the definition of multimodal RoPE, where the rotation matrix is partitioned to encode different dimensions:

$$\mathbf{R}_{t,x,y} = \mathrm{diag}(\mathbf{R}_t, \mathbf{R}_x, \mathbf{R}_y),$$

where $\mathbf{R}_t$, $\mathbf{R}_x$, and $\mathbf{R}_y$ are rotation matrices applied to temporal, horizontal spatial, and vertical spatial dimensions, respectively, with each dimension carrying a frequency of $\theta_i = b^{-2i/d}, i \in \{0, \dots, d/2-1\}$. Note that the (t, x, y) ordering is purely notational and does not constrain the actual dimension allocation strategy.

Assume that each component of the query vector $\mathbf{q}$ is independently and identically distributed with mean μ and variance σ^2. We denote key vector that is semantically similar to $\mathbf{q}$ as $\mathbf{k}' = \mathbf{q} + \delta$, where δ is a zero-mean perturbation. The semantically unrelated key vector $\mathbf{k}$ is independently drawn with the same distribution as $\mathbf{q}$. Let $\Delta t, \Delta x, \Delta y$ denote relative temporal and spatial distances between the query and each key. According to Definition 3.1, the semantic preference property requires that:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{q},\mathbf{k},\delta}[\mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{k}'^{\top} - \mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{k}^{\top}] \\
&= \mathbb{E}_{\mathbf{q},\mathbf{k},\delta}[\mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}(\mathbf{q}+\delta)^{\top} - \mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{k}^{\top}] \\
&= \mathbb{E}_{\mathbf{q}}[\mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{q}^{\top}] - \mathbb{E}_{\mathbf{q},\mathbf{k}}[\mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{k}^{\top}] \\
&= \mathbb{E}_{\mathbf{q}}[\mathbf{q}\mathbf{R}_{\Delta t,\Delta x,\Delta y}\mathbf{q}^{\top}] - \mu^2\mathbf{R}_{\Delta t,\Delta x,\Delta y} \\
&= \sum_{i\in i_t} 2(\mu^2+\sigma^2)\cos(\Delta t)\theta_i + \sum_{i\in i_x} 2(\mu^2+\sigma^2)\cos(\Delta x)\theta_i + \sum_{i\in i_y} 2(\mu^2+\sigma^2)\cos(\Delta y)\theta_i - \qquad (9) \\
&\sum_{i\in i_t} 2\mu^2\cos(\Delta t)\theta_i + \sum_{i\in i_x} 2\mu^2\cos(\Delta x)\theta_i + \sum_{i\in i_y} 2\mu^2\cos(\Delta y)\theta_i \\
&= \sum_{i\in i_t} 2\sigma^2\cos(\Delta t\cdot\theta_i) + \sum_{i\in i_x} 2\sigma^2\cos(\Delta x\cdot\theta_i) + \sum_{i\in i_y} 2\sigma^2\cos(\Delta y\cdot\theta_i) \ge 0,
\end{aligned}$$

where i_t, i_x, i_y denote dimensions allocated to encode temporal (t), horizontal spatial (x), and vertical spatial (y) information. To satisfy the semantic preference property (Definition 3.1), the expected attention between a query and its semantically similar key should remain higher than that

for an unrelated key, regardless of their relative distance. This implies the following condition must hold universally:

$$\sum_{i\in i_t} 2\sigma^2 \cos(\Delta t \cdot \theta_i) + \sum_{i\in i_x} 2\sigma^2 \cos(\Delta x \cdot \theta_i) + \sum_{i\in i_y} 2\sigma^2 \cos(\Delta y \cdot \theta_i) \geq 0, \\ \Delta t \in \{0, 1, \ldots, L-1\}, \Delta x \in \{0, 1, \ldots, H\}, \Delta y \in \{0, 1, \ldots, W\}. \tag{10}$$

Now consider a long-context scenario, where $L \gg H, W$, we can now theoretically prove that why VideoRoPE [28] (using lowest frequencies for t) is better than M-RoPE (using highest frequencies for t) in maintaining semantic preference property in long contexts. Simply, by Lemma A.1, we show that when the context length L is sufficiently large, the probability that $cos(\Delta t \cdot \theta_i)$ leads to negative values becomes higher when θ_i becomes larger. Therefore, lower frequencies, which rotate less, are less likely to violate the semantic preference property.

However, despite using the lowest frequencies, VideoRoPE still fails to guarantee that the semantic preference property holds for all context lengths (Equation 10). Let VideoRoPE allocate only the smallest frequency to the temporal dimensions, instead of $|i_t|$ smallest frequencies:

$$\theta_{\min} \;=\; b^{-2(\frac{d}{2}-1)/d},$$

so that in Equation (10) the temporal sum reduces to:

$$2\sigma^2 \; |i_t| \cos(\Delta t \cdot \theta_{\min}).$$

Here, under the reasonable assumption that semantically related tokens co-occur in nearby spatial positions across frames, the spatial sums in Equation 10 remains non-negative. Thus we only consider the temporal sum in Equation 10:

$$2\sigma^2 |i_t| \cos(\Delta t \cdot \theta_{\min}).$$

Now pick any context length L so large that there exists

$$\Delta t \;\in\; \{0, 1, \ldots, L-1\} \quad \text{with} \quad \Delta t\, \theta_{\min} \;\in\; \left(\tfrac{\pi}{2},\, \tfrac{3\pi}{2}\right).$$

Such a Δt indeed exists as soon as $\theta_{\min}\,(L-1) > \frac{\pi}{2}$, i.e. for any

$$L \;>\; L_c = \frac{\pi}{2\,\theta_{\min}} + 1.$$

For that choice of Δt, we have

$$\cos(\Delta t \;\cdot \theta_{\min}) < 0,$$

and hence the left-hand side of Equation (10) becomes

$$2\sigma^2 \; |i_t| \cos(\Delta t \;\cdot \theta_{\min}) < 0.$$

This single counterexample $(\Delta t, \Delta x, \Delta y)$ violates the semantic preference condition, since no further temporal frequencies are available to "rescue" the sum. Therefore, despite using the lowest frequencies for temporal modeling, VideoRoPE still fails to guarantee the semantic preference property. In conclusion, all frequency allocation strategies in current multimodal RoPEs fail to maintain the semantic preference property in Definition 3.1, completing the proof. □

B Further Experimental Details

In this section, we provide further details of our experiments, including benchmark descriptions, experimental settings, and further results.

B.1 Detailed Benchmark Description

There is a growing interest in video generation and understanding [45–47, 34, 19], given their broad applications in content creation and analysis. In this subsection, we provide detailed descriptions of the video benchmarks we used in the experiments, i.e., LongVideoBench [41], Video-MME [42], MLVU [43], and V-NIAH [17].

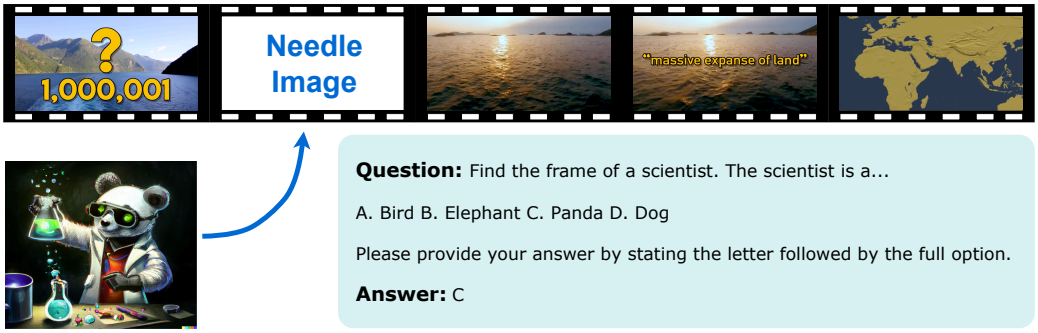


Figure 4: Illustration of V-NIAH, which consists of a randomly inserted needle image, a haystack video, and a specific question related to the needle.

- **LongVideoBench** is a comprehensive benchmark for evaluating Vision-Language Models on long video understanding tasks. Unlike traditional video benchmarks that focus on short clips under one minute, this dataset features videos ranging from 8 seconds to 1 hour across diverse sources, including everyday life, movies, knowledge, and news. The benchmark encompasses 17 fine-grained question categories organized into two levels: perception and relation. In our experiment, questions that are free from subtitles are retained.
- **Video-MME** is a full-spectrum evaluation benchmark of Vision-Language Models in video analysis, spanning 6 primary visual domains with 30 subfields to ensure generalizability. It features temporal diversity by incorporating both short- (<2 minutes), medium- (4-15 minutes), and long-term videos (30-60 minutes), ranging from 11 seconds to 1 hour.
- **MLVU** is a high-quality benchmark designed to evaluate the video understanding capabilities of Vision-Language Models. The temporal duration of videos within MLVU spans from 3 minutes to 2 hours, covering genres such as movies, life records, and egocentric videos. In our experiment, we evaluate all methods on the following multiple-choice tasks: Action Count, Action Order, Topic Reasoning, Ego Reasoning, Needle QA, Plot QA, and Anomaly Recognition.
- **V-NIAH** is a challenging benchmark designed to evaluate VLMs' ability to identify specific frames within long videos. In this task, a "needle" image is inserted into a "haystack" video, and the VLMs are required to answer specific questions about this "needle" image, as shown in Figure 4. Following the settings in V-NIAH [17], we utilize a haystack video with 1-hour duration (3,000 frames). The needle image is inserted at 20% depth intervals (e.g., a frame depth of 0% would place the needle image at the very beginning of the video.)

Table 4: Quantitative performance of different RoPE variants on V-NIAH. Here, we report the average accuracy across different context lengths and frame depths.

	Vanilla RoPE	M-RoPE	VideoRoPE	HoPE (ours)
V-NIAH	21.00	47.11	<u>52.00</u>	**63.56**

B.2 Quantitative Results on V-NIAH

Here, we provide the quantitative results of different RoPE variants on long video retrieval task in Table 4. It can be observed that our HoPE demonstrates a 22.23% improvement compared to the best baseline, justifying its effectiveness in multimodal long-context modeling.

B.3 Ideal Condition for Semantic Preference

As discussed after Theorem 4.1, the semantic preference property (Definition 3.1) invariably holds for any context length t and spatial size x, y when we set $|i_t| = d/4, |i_x| = |i_y| = d/8$ and $\theta_i = 0, i \in i_t$, since Lemma 4.1 reduces to:

$$\sum_{i=0}^{d/8-1} 2\sigma^2(2 + \cos(\Delta x \cdot \theta_{2i}) + \cos(\Delta y \cdot \theta_{2i+1})) \geq 0.$$

In our original HoPE implementation, the frequencies allocated to t, x, y are $16, 24, 24$, respectively, with the lowest 16 frequencies for t set to zero. For this proposed variant (HoPE-X), we redistribute these allocations to $32, 16, 16$ for t, x, y, respectively, while setting the lowest 32 frequencies for t to zero. To evaluate the comparative effectiveness of these configurations, we conduct further experiments on LongVideoBench.

Table 5: Performance comparison between HoPE-X and HoPE.

Method	LongVideoBench			
	8k	16k	32k	64k
HoPE-X	52.68	52.73	53.01	46.32
HoPE	54.11	55.09	55.34	51.22

Table 5 demonstrates that HoPE consistently outperforms HoPE-X across diverse context lengths. We deduce that the inferior performance of HoPE-X is due to its decreased dimensions allocated for spatial modeling. While this configuration helps to maintain the semantic preference property, it negatively impacts HoPE-X's ability to model local features. Therefore, it is necessary to keep adequate dimensions for spatial modeling in multimodal RoPE.