VideoRoPE++: Towards Better Video Rotary Position Embedding

Xilin Wei, Xiaoran Liu, Yuhang Zang, Shengyuan Ding, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, Dahua Lin

***Abstract*—While Rotary Position Embedding (RoPE) and its variants are widely adopted for their long-context capabilities, the extension of the 1D RoPE to video, with its complex spatio-temporal structure, remains an open challenge. This work first introduces a comprehensive analysis that identifies five key characteristics essential for the effective adaptation of RoPE to video, which have not been fully considered in prior work. As part of our analysis, to reveal the limitations of current position embedding designs, we introduce a challenging V-RULER benchmark. The Needle Retrieval under Distractor (NRD) subtask of V-RULER highlights the challenges posed by periodic distractors, demonstrating that previous RoPE variants, lacking appropriate temporal dimension allocation, are easily misled by such distractors. Based on our analysis, we introduce VideoRoPE++, with a *3D structure* designed to preserve spatio-temporal relationships. VideoRoPE++ features *low-frequency temporal allocation* to mitigate periodic oscillations, a *diagonal layout* to maintain spatial symmetry, *adjustable temporal spacing* to decouple temporal and spatial indexing. To improve extrapolation performance, VideoRoPE++ integrates our proposed *YaRN-V*, which interpolates along the low-frequency temporal dimension while preserving the spatial positional structure. VideoRoPE++ consistently surpasses previous RoPE variants, across diverse downstream tasks such as long video retrieval, video understanding, and video hallucination. Our code is available at https://github.com/Wiselnn570/VideoRoPE.**

***Index Terms*—Video Large Language Models, Rotary Position Embedding, Long-context Modeling, Video Understanding**

I. Introduction

Rotary Position Embedding (RoPE) [1] helps Transformer models understand word order by assigning each token a unique positional 'marker' calculated using a mathematical rotation matrix. RoPE has advantages in long-context understanding [2], and continues to be a default choice in leading Large Language Models (LLMs) like the LLaMA [3]–[5] and Qwen [6], [7] series.

The original RoPE implementation (Vanilla RoPE) [1] is designed for sequential 1D data like text. However, recent Video Large Language Models (Video LLMs) [8]–[15] process video, which has a more complex spatio and temporal structure. As shown in Tab. I, although several RoPE-based approaches [16], [17] have been proposed to support video inputs, these variants exhibit limitations and do not fully satisfy the following five key characteristics:

TABLE I: Comparison between different RoPE variants for Video Large Language Models (Video LLMs).

	2D/3D Structure	Frequency Allocation	Spatial Symmetry	Temporal Index Scaling	Extrapolation Capability
Vanilla RoPE [1]	✗	✗	✗	✗	✗
TAD-RoPE [16]	✗	✗	✗	✓	✗
RoPE-Tie [18]	✓	✗	✓	✗	✗
M-RoPE [17]	✓	✗	✗	✗	✓
M-RoPE++ [19]	✓	✗	✗	✗	✓
VideoRoPE++ (Ours)	✓	✓	✓	✓	✓

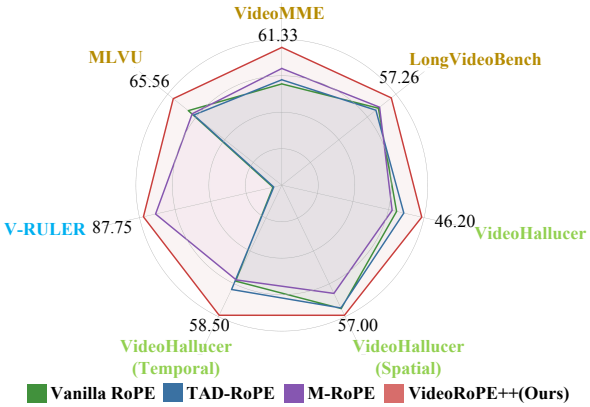


Fig. 1: VideoRoPE++ outperforms RoPE variants on benchmarks.

(1) 2D/3D Structure. Some existing Video LLMs direct flatten the video frame into 1D embeddings and apply the 1D structure RoPE [1], [16]. These solutions fail to capture video data's inherent 2D or 3D (temporal (t), horizontal (x), and vertical (y)) structure, thus hindering explicit spatial and temporal representation.

(2) Frequency Allocation. Previous approaches such as M-RoPE used in Qwen2-VL [17] employ 3D structure, dividing feature dimensions into distinct subsets for (t, x, y) encoding, respectively. How to determine the optimal allocation of these dimension subsets, and their associated frequencies[1] are not well studied. Some previous work allocates the lower dimensions corresponding to the high frequency to represent the t. However, the temporal dimension t is significantly tortured by periodic oscillation, and distant positions may have the same embeddings.

To verify this point, we propose the V-RULER benchmark

Xilin Wei, Xiaoran Liu, Shengyuan Ding, Xipeng Qiu are with the School of Computer Science at Fudan University. Xiaoran Liu, Xipeng Qiu, Qipeng Guo, Jiaqi Wang are with Shanghai Innovation Institute.

Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Qipeng Guo, Jiaqi Wang are with Shanghai AI Laboratory.

Dahua Lin is with The Chinese University of Hong Kong.

Corresponding author: Yuhang Zang (zangyuhang@pjlab.org.cn), Qipeng Guo (guoqipeng@pjlab.org.cn), Jiaqi Wang (wangjiaqi@pjlab.org.cn).

[1]In RoPE, frequencies are determined by $\beta^{-2n/d}$, where β is a constant, n is the dimension index, d is the total number of dimensions. Thus, choosing which dimensions represent t, x, and y directly determines the frequencies used for each.

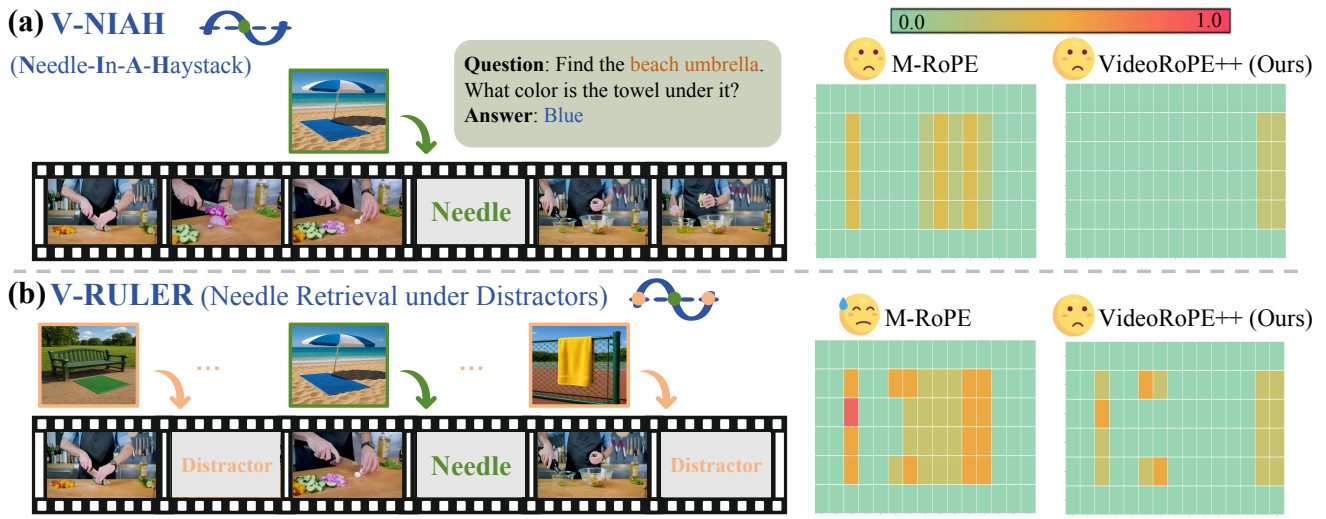


Fig. 2: **Left:** To highlight the importance of frequency allocation, we build upon V-NIAH (**a**) and introduce a more challenging setting—V-RULER. One of its key tasks, Needle Retrieval under Distractors (NRD), is illustrated in (**b**), where visually similar frames are inserted as distractors to increase task difficulty. **Right:** Compared to M-RoPE, our VideoRoPE++ is more robust in retrieval and is less affected by distractors. See Fig. 2 in the Experiments section for details on the horizontal and vertical axes.

to reveal the limitations of current position embedding designs, particularly in how they allocate frequency components. Based on the previous long-video retrieval task V-NIAH (Visual Needle-In-A-Haystack) [12], we insert several similar images that do not affect the question's answer before and after the needle image as distractor [20], [21], forming a new setting, Needle Retrieval under Distractors (NRD). NRD is one of the core tasks in our proposed V-RULER benchmark. As shown in Fig. 2, we find that the previous M-RoPE is misled by distractors, showing a significant performance decline from V-NIAH to NRD. Our observation demonstrates that the periodic oscillation reduces Video LLMs' robustness.

(3) Spatial Symmetry. The distance between the end of the precedent textual input and the start of visual input equals the distance between the end of visual input and the start of subsequent textual input [22]. Such a symmetry ensures that the visual input receives equal contextual influence from both the preceding and subsequent textual information.

(4) Temporal Index Scaling. Spatial and temporal dimensions often exhibit different granularities (e.g., a unit change in x/y differs from a unit change in t) [16]. Employing varying index intervals in positional encoding allows for dimension-specific encoding, capturing diverse scales and enhancing efficiency.

(5) Extrapolation Capability. One challenge for positional encoding in video LLMs is the capacity to extrapolate beyond the training context length. Existing RoPE-based methods, including M-RoPE [17], are typically trained within a limited positional range (e.g., 32k tokens). However, during inference, especially in long-context video understanding, the model often encounters positions far beyond this range. Due to the nature of RoPE's exponential frequency formulation, unseen indices lead to out-of-distribution behaviors, resulting in severe attention misalignment and degraded performance.

Driven by our analysis, we present a new video position embedding strategy, **VideoRoPE++**, which can simultaneously satisfy the five properties in Tab. I. Specifically, we use a 3D structure to model spatiotemporal information, allocating higher dimensions (lower frequencies), to the temporal axis (**L**ow-frequency **T**emporal **A**llocation, **LTA**) to prioritize temporal modeling. The right panel of Fig. 2 demonstrates that our LTA allocation mitigates oscillations and exhibits robustness to distractors in the V-RULER task. We further employ a **D**iagonal **L**ayout (**DL**) design to ensure spatial symmetry and preserve the relative positioning between visual and text tokens. Regarding temporal index scaling, we propose **A**djustable **T**emporal **S**pacing (**ATS**), where a hyperparameter controls the relative temporal spacing of adjacent visual tokens. To extend the applicability of positional encoding beyond the training range, we propose an extrapolation method, **YaRN-V** that performs frequency interpolation exclusively along the low-frequency temporal axis. YaRN-V maintains temporal consistency across extended contexts while avoiding distortion in spatial encoding. Spatial information remains stable and periodic due to the use of high-frequency components, which already span a full positional cycle during training, allowing for direct extrapolation without additional adjustment and ensuring reliable generalization capability. In summary, our proposed position encoding scheme demonstrates favorable characteristics for modeling video data, yielding a robust and effective representation of positional information.

A preliminary version of this work (VideoRoPE [23]) was accepted as an Oral presentation at ICML 2025. In this extended version, we provide valuable materials including a challenging benchmark V-RULER, an extrapolation method YaRN-V, exhaustive experiments, additional ablation studies, and detailed implementations that are summarized as follows:

- To reveal the limitations of current position embedding designs, we propose a more challenging and broad benchmark, V-RULER, to assess fine-grained temporal localization, entity tracking, and robustness to distractors. The results show that existing Video LLMs are highly sensitive to frequency-based distractors, which impair

their ability to perform accurate temporal localization and semantic reasoning.
- We introduce a new extrapolation method, YaRN-V, which improves temporal generalization beyond the training range. YaRN-V consistently outperforms prior extrapolation methods such as YaRN [24] and M-RoPE++ [19], achieving robust performance even under extreme context lengths and validating its superior extrapolation capability.
- We conduct comprehensive experiments on more benchmarks, including general video understanding, video captioning, and streaming video understanding. We also compare our method against strong baselines at both 3B and 7B scales to evaluate scalability.
- We conduct more ablation studies by varying the extrapolation factor to analyze the sensitivity of model performance to its value. We also validate the method across additional language models to verify the generality.
- On top of the conference version, we provide more analysis of existing RoPE-based methods, existing extrapolation methods, and additional experimental results.

II. Related Work

Rotary Position Embedding (RoPE). RoPE [1] is a widely used technique in large language models (LLMs) to encode positional information via complex rotations in embedding space. Compared to absolute or relative positional encodings, RoPE provides a smooth inductive bias for both near and far-token interactions. In RoPE, sinusoidal embeddings are generated by applying trigonometric functions with exponentially decaying frequencies across dimensions [24], [25], where lower-indexed dimensions correspond to higher frequencies. The efficiency and simplicity of RoPE have led to its adoption in many mainstream models such as LLaMA [4], Qwen [6], InternLM [26], [27], and Gemma [28].

Extending RoPE to Multi-modal data. Adapting RoPE from text to visual or multi-modal sequences [29]–[33] introduces new challenges, especially regarding spatial and temporal structure [34]. Some works, like TAD-RoPE [16], apply 1D RoPE across flattened image/video tokens, treating all modalities as a single token stream. However, such methods neglect spatial symmetry and misrepresent temporal structure. To address this, 2D and 3D RoPE variants have emerged. RoPE-Tie [18] integrates positional embeddings across spatial axes while maintaining alignment with textual context. M-RoPE [17], used in Qwen2-VL, generalizes RoPE to (t, x, y) axes, but suffers from performance drops under distractors, as seen in long video retrieval [23]. This work presents a comprehensive analysis of the important characteristics essential for extending RoPE to video and proposes VideoRoPE++ according to our analysis.

Extrapolation Methods for Rotary Position Embedding. Several *training-free* methods have been proposed to improve the extrapolation capability of RoPE by adjusting the frequency spectrum. Some approaches [25], [35] introduce frequency scaling based on theoretical analyses such as Neural Tangent Kernel (NTK) theory. Others, like YaRN [24], apply length-aware rescaling to adapt positional encoding for longer contexts. M-RoPE++ [19] modifies the spatial frequencies (x, y) through hybrid interpolation but keeps the temporal axis (t) unchanged, which restricts support for extended temporal inputs. Different from previous approaches, our **YaRN-V** applies extrapolation solely along the temporal axis while keeping the spatial frequency structure fixed. Focusing interpolation on the temporal axis enhances generalization across longer time spans and avoids distortion in the spatial domain, resulting in more stable and effective position encoding for long-context video tasks.

Video Large Language Models. Video LLMs build upon the success of image-based vision-language models (VLMs) [17], [36]–[39] to video scenarios, which require handling temporal dependencies [40]–[43] and long-form video understanding [12]–[14]. Various studies extend Video LLMs' capabilities to longer content, such as caption summarization [44], [45], reduce the number of video tokens [46]–[49], streaming-based processing [50], [51], memory-augmented models [52], [53], and hierarchical representations [54]. Orthogonal to the previous approaches, this work studies video rotary position embedding with the extrapolation method, which is an important component of video LLMs and is beneficial to the long-context understanding ability of video LLMs.

III. Analysis

3D Structure. The vanilla RoPE defines a matrix $\boldsymbol{A}_{t_1,t_2}$ that represents the relative positional encoding between two positions t_1 and t_2 in a 1D sequence:

$$\boldsymbol{A}_{t_1,t_2} = \left(\boldsymbol{q}_{t_1}\boldsymbol{R}_{t_1}\right)\left(\boldsymbol{k}_{t_2}\boldsymbol{R}_{t_2}\right)^{\top} = \boldsymbol{q}_{t_1}\boldsymbol{R}_{\Delta t}\boldsymbol{k}_{t_2}^{\top}, \tag{1}$$

where $\Delta t = t_1 - t_2$, the symbols $\boldsymbol{q}_{t_1}$ and $\boldsymbol{k}_{t_2}$ are the query and key vectors at positions t_1 and t_2. The *relative rotation matrix* $\boldsymbol{R}_{\Delta t}$ is defined as $\boldsymbol{R}_{\Delta t} = \exp(\Delta t i \theta_n)$, while i is the imaginary unit, $\theta_n = \beta^{-2n/d}$ is the frequency of rotation applied to a specific n-th pair of d dimensions ($n = 0, \ldots, d/2-1$), and β is the frequency base parameter. The vanilla RoPE uses $d = 128$, thus $n = 0, \ldots, 63$. Consequently, the $\boldsymbol{A}_{t_1,t_2}$ in Eq. (1) can be extended as:

$$\begin{pmatrix} q^{(0)} \\ q^{(1)} \\ \vdots \\ q^{(126)} \\ q^{(127)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_0\Delta t & -\sin\theta_0\Delta t & \cdots & 0 & 0 \\ \sin\theta_0\Delta t & \cos\theta_0\Delta t & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos\theta_{63}\Delta t & \sin\theta_{63}\Delta t \\ 0 & 0 & \cdots & \sin\theta_{63}\Delta t & \cos\theta_{63}\Delta t \end{pmatrix} \begin{pmatrix} k^{(0)} \\ k^{(1)} \\ \vdots \\ k^{(126)} \\ k^{(127)} \end{pmatrix} \tag{2}$$

While the vanilla RoPE operates on 1D sequences, it can also be applied to higher-dimensional input by flattening the input into a 1-D sequence. However, the flattening process discards crucial neighborhood information, increases the sequence length, and hinders the capture of long-range dependencies. Therefore, preserving the inherent 3D structure is essential when adapting RoPE for video data. Some recent RoPE-variants (e.g., M-RoPE in Qwen2-VL [17]) incorporate the 3D structure. The corresponding relative matrix $\boldsymbol{A}_{(t_1,x_1,y_1)}$ is computed as:

$$\boldsymbol{A}_{(t_1,x_1,y_1),(t_2,x_2,y_2)} = \boldsymbol{q}_{(t_1,x_1,y_1)}\boldsymbol{R}_{\Delta t,\Delta x,\Delta y}\boldsymbol{k}_{(t_2,x_2,y_2)}^{\top}, \tag{3}$$

where $\Delta t = t_1 - t_2$, $\Delta x = x_1 - x_2$, and $\Delta y = y_1 - y_2$. M-RoPE divides the $d = 128$ feature dimensions into 3 groups: the first 32 for temporal positions (t), the middle 48 for horizontal positions (x), and the last 48 for vertical positions (y). As

shown in Eq (4), $\boldsymbol{A}_{(t_1,x_1,y_1),(t_2,x_2,y_2)}$ in M-RoPE is extended as:

$$
\begin{aligned}
&\underbrace{\begin{pmatrix} q^{(0)} \\ q^{(1)} \\ q^{(2)} \\ q^{(3)} \\ \vdots \\ q^{(30)} \\ q^{(31)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_0\Delta t & -\sin\theta_0\Delta t & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_0\Delta t & \cos\theta_0\Delta t & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_1\Delta t & -\sin\theta_1\Delta t & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_1\Delta t & \cos\theta_1\Delta t & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{15}\Delta t & -\sin\theta_{15}\Delta t \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{15}\Delta t & \cos\theta_{15}\Delta t \end{pmatrix} \begin{pmatrix} k^{(0)} \\ k^{(1)} \\ k^{(2)} \\ k^{(3)} \\ \vdots \\ k^{(30)} \\ k^{(31)} \end{pmatrix}}_{\text{modeling temporal dependency with higher frequency}} \\
&+ \underbrace{\begin{pmatrix} q^{(32)} \\ q^{(33)} \\ q^{(34)} \\ q^{(35)} \\ \vdots \\ q^{(78)} \\ q^{(79)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_{16}\Delta x & -\sin\theta_{16}\Delta x & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_{16}\Delta x & \cos\theta_{16}\Delta x & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_{17}\Delta x & -\sin\theta_{17}\Delta x & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_{17}\Delta x & \cos\theta_{17}\Delta x & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{39}\Delta x & -\sin\theta_{39}\Delta x \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{39}\Delta x & \cos\theta_{39}\Delta x \end{pmatrix} \begin{pmatrix} k^{(32)} \\ k^{(33)} \\ k^{(34)} \\ k^{(35)} \\ \vdots \\ k^{(78)} \\ k^{(79)} \end{pmatrix}}_{\text{modeling horizontal dependency with intermediate frequency}} \\
&+ \underbrace{\begin{pmatrix} q^{(80)} \\ q^{(81)} \\ q^{(82)} \\ q^{(83)} \\ \vdots \\ q^{(126)} \\ q^{(127)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_{40}\Delta y & -\sin\theta_{40}\Delta y & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_{40}\Delta y & \cos\theta_{40}\Delta y & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_{41}\Delta y & -\sin\theta_{41}\Delta y & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_{41}\Delta y & \cos\theta_{41}\Delta y & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{63}\Delta y & -\sin\theta_{63}\Delta y \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{63}\Delta y & \cos\theta_{63}\Delta y \end{pmatrix} \begin{pmatrix} k^{(80)} \\ k^{(81)} \\ k^{(82)} \\ k^{(83)} \\ \vdots \\ k^{(126)} \\ k^{(127)} \end{pmatrix}}_{\text{modeling vertical dependency with lower frequency}}
\end{aligned}
\tag{4}
$$

Frequency Allocation. Incorporating 3D structure raises the question of how to allocate the temporal (t), horizontal (x), and vertical (y) components within the d dimensions. Note that different allocation strategies are not equivalent in the rotation frequency $\theta_n = \beta^{-2n/d}$. As shown in Eq. (4), M-RoPE assigns higher frequencies (corresponding to lower dimensions) to the temporal dimension (t).

To highlight the importance of frequency allocation, we introduce the **Needle Retrieval under Distractors** (NRD) subtask from our proposed benchmark **V-RULER** (will be detailed in Sec. IV-B). The NRD subtask builds upon V-NIAH [12], a benchmark originally designed to evaluate visual long-context understanding. However, prior studies [20], [21] have shown that such retrieval tasks often reflect only superficial comprehension. To address this, our subtask incorporates semantically similar distractors, sourced via Google Image Search [55] or Flux [56], in order to reduce the likelihood of correct predictions by chance and to better test the model's ability to perform fine-grained temporal localization. These distractors are designed to be unambiguous to the question in Fig. 2.

As shown in Fig. 2, M-RoPE exhibits a clear performance drop from V-NIAH to V-RULER (Needle Retrieval under Distractors). To investigate this decline, we follow previous works [25], [57], [58] to visualize the attention scores in Fig. 3. We decompose the attention scores into their corresponding temporal (t), horizontal (x), and vertical (y) components for visualization.

Fig. 3 reveals unusual M-RoPE's attention patterns, despite locating the needle image, it fails to answer the multi-choice question. According to M-RoPE's attention, the needle is located primarily through vertical positional information, rather than temporal features. Thus, the temporal dimension fails to capture long-range semantic dependencies, focusing on local relationships. Conversely, the spatial dimensions capture long-range rather than local semantic information. Lastly, the horizontal and vertical dimensions display distinct characteristics,

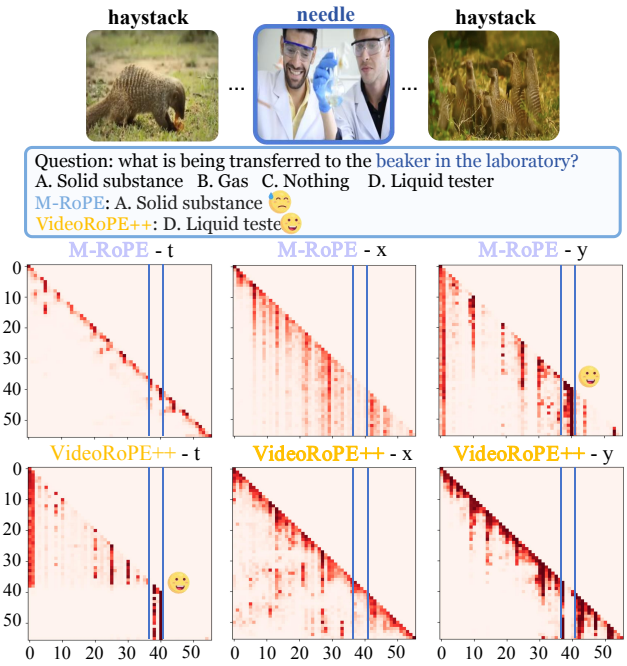


Fig. 3: Attention-based frequential allocation analysis. **Middle**: M-RoPE's temporal dimension (t) is limited to local information, resulting in a diagonal layout. **Bottom**: VideoRoPE++ effectively retrieves the needle using the temporal dimension. The x and y coordinates represent the video frame number, e.g., 50 for 50 frames.

with the vertical dimension exhibiting phenomena reminiscent of attention sinks [57]. These suggest the performance decline primarily results from sub-optimal frequency allocation designs of M-RoPE.

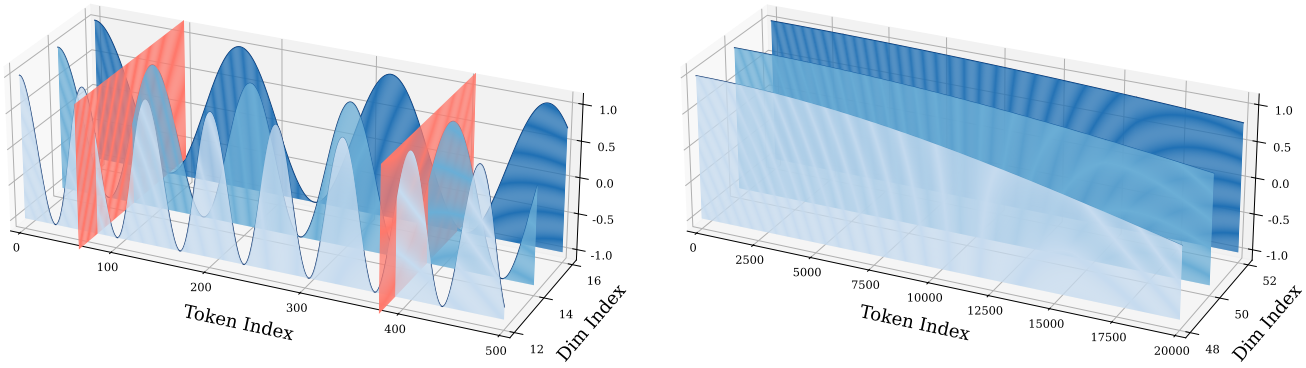
Spatial Symmetry. Given the text tokens T and the visual tokens T_v, spatial symmetry [22] claims that the distance between the end of the preceding textual input (T_{pre}) and the beginning of the visual input (T_v^{start}) is equal to the distance between the end of the visual input (T_v^{end}) and the beginning of the subsequent textual input (T_{sub}):

$$T_v^{\text{start}} - T_{\text{pre}} = T_{\text{sub}} - T_v^{\text{end}}. \tag{5}$$

The spatial symmetrical structure can potentially simplify the learning process and reduce bias toward input order. However, existing 3D RoPE variants such as M-RoPE do not meet the spatial symmetry, we will elaborate related discussion in Fig. 6.

Temporal Index Scaling. The frame index in video and the token index in text are inherently different [22], [59]. Recognizing this difference, methods like TAD-RoPE, a 1D RoPE adaptation for Video LLMs, introduce distinct step offsets for image and text token indices: γ for image tokens and $\gamma+1$ for text tokens. Consequently, an ideal RoPE design for video data should permit scaling of the temporal index to meet the inherent difference between the frame index and the text index.

Extrapolation Capability. In multimodal inputs, the temporal position index increases rapidly due to token concatenation across modalities. Prior RoPE designs have largely overlooked this aspect, leading to failure on our proposed Lengthy Multimodal Stack subtask in the V-RULER benchmark (see



(a) Temporal Frequency Allocation in M-RoPE (b) Temporal Frequency Allocation in VideoRoPE++ (ours)

Fig. 4: **(a)** M-RoPE [17] models temporal dependencies using the *first* 16 rotary angles, which exhibit higher frequencies and more pronounced oscillations. **(b)** In contrast, VideoRoPE++ models temporal dependencies using the *last* 16 rotary angles, characterized by significantly wider, monotonic intervals. Our frequency allocation effectively mitigates the misleading influence of distractors in the NRD subtask of V-RULER.

Tab. VII for details). An ideal RoPE design should be capable of handling such rapid growth in cross-modal positional indices and generalizing to unseen positions beyond the training range.

IV. VideoRoPE++

A. Design of Video Rotary Position Embedding

Based on some previous research and the above analysis, we claim that a good RoPE design for Video LLMs, especially for long videos, should satisfy five requirements: 3D structure, Appreciate Frequency Allocation, Spatial Symmetry, Temporal Index Scaling, and Extrapolation Capability. The first requirement has been solved by RoPE-Tie [18] and the subsequent M-RoPE [17]. To solve the last four requirements and mitigate the performance decline observed in V-RULER (especially the NRD subtask), we propose our VideoRoPE++, comprising the following four key components. (1) Low-frequency Temporal Allocation; (2) Diagonal Layout; (3) Adjustable Temporal Spacing, and (4) Extrapolation Capability with Yarn-V.

Low-frequency Temporal Allocation (LTA). As shown in Eq. (2), the vanilla RoPE [1] uses all dimensions to model the 1D position information. And as indicated in Eq. (4), M-RoPE [17] uses different dimensions to model temporal, horizontal, and vertical dimensions sequentially. However, previous frequency allocation strategies are suboptimal because different RoPE dimensions capture dependencies at varying ranges. As shown in Fig. 3, an interesting observation is that the local attention branch (as reported in [60]) corresponds to lower dimensions, while the global branch (or attention sink, as in [57]) corresponds to higher dimensions. To sum up, lower dimensions (higher frequency, shorter monotonic intervals, larger θ_n) tend to capture relative distances and local semantics [58], [61], while higher dimensions (lower frequency, wider monotonic intervals, smaller θ_n) capture longer-range dependencies [58].

Based on our analysis, VideoRoPE++ uses higher dimensions for temporal features in longer contexts and lower dimensions for spatial features, which are limited by resolution and have a fixed range. To avoid the gap between horizontal and vertical positions, we interleave the dimensions responsible for these spatial features. The dimension distribution for VideoRoPE++ is shown in Eq. (6):

$$
\begin{aligned}
&\underbrace{\begin{pmatrix} q^{(96)} \\ q^{(97)} \\ q^{(98)} \\ q^{(99)} \\ \vdots \\ q^{(126)} \\ q^{(127)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_{48}\Delta t & -\sin\theta_{48}\Delta t & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_{48}\Delta t & \cos\theta_{48}\Delta t & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_{49}\Delta t & -\sin\theta_{49}\Delta t & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_{49}\Delta t & \cos\theta_{49}\Delta t & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{63}\Delta t & -\sin\theta_{63}\Delta t \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{63}\Delta t & \cos\theta_{63}\Delta t \end{pmatrix} \begin{pmatrix} k^{(96)} \\ k^{(97)} \\ k^{(98)} \\ k^{(99)} \\ \vdots \\ k^{(126)} \\ k^{(127)} \end{pmatrix}}_{\text{modeling temporal dependency with lower frequency}} \\
&+ \underbrace{\begin{pmatrix} q^{(0)} \\ q^{(1)} \\ q^{(4)} \\ q^{(5)} \\ \vdots \\ q^{(92)} \\ q^{(93)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_{0}\Delta x & -\sin\theta_{0}\Delta x & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_{0}\Delta x & \cos\theta_{0}\Delta x & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_{2}\Delta x & -\sin\theta_{2}\Delta x & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_{2}\Delta x & \cos\theta_{2}\Delta x & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{46}\Delta x & -\sin\theta_{46}\Delta x \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{46}\Delta x & \cos\theta_{46}\Delta x \end{pmatrix} \begin{pmatrix} k^{(0)} \\ k^{(1)} \\ k^{(4)} \\ k^{(5)} \\ \vdots \\ k^{(92)} \\ k^{(93)} \end{pmatrix}}_{\text{modeling horizontal dependency with interleaved high frequency}} \\
&+ \underbrace{\begin{pmatrix} q^{(2)} \\ q^{(3)} \\ q^{(6)} \\ q^{(7)} \\ \vdots \\ q^{(94)} \\ q^{(95)} \end{pmatrix}^{\top} \begin{pmatrix} \cos\theta_{1}\Delta y & -\sin\theta_{1}\Delta y & 0 & 0 & \cdots & 0 & 0 \\ \sin\theta_{1}\Delta y & \cos\theta_{1}\Delta y & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\theta_{3}\Delta y & -\sin\theta_{3}\Delta y & \cdots & 0 & 0 \\ 0 & 0 & \sin\theta_{3}\Delta y & \cos\theta_{3}\Delta y & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\theta_{47}\Delta y & -\sin\theta_{47}\Delta y \\ 0 & 0 & 0 & 0 & \cdots & \sin\theta_{47}\Delta y & \cos\theta_{47}\Delta y \end{pmatrix} \begin{pmatrix} k^{(2)} \\ k^{(3)} \\ k^{(6)} \\ k^{(7)} \\ \vdots \\ k^{(94)} \\ k^{(95)} \end{pmatrix}}_{\text{modeling vertical dependency with interleaved high frequency}}
\end{aligned}
\tag{6}
$$

The horizontal position x and vertical position y are interleaved to occupy the lower dimensions, followed by temporal t, which occupies the higher dimensions. We keep the same allocation number for x, y, and t as M-RoPE for a fair comparison, with values of 48, 48, and 32, respectively. The advantages of this distribution are evident in Fig. 4. For a RoPE-based LLM with a 128-dimensional head (64 rotary angles θ_n), we visualize the function of $\cos\theta_n t$ for 3 dimensions using parallel blue planes.

As shown in Fig. 4 (**a**), M-RoPE's temporal position embeddings are significantly distorted by periodic oscillations [61], leading to identical embeddings for distant positions. For instance, considering the last three rotary angles, the temporal embeddings are severely affected by these oscillations due to their short monotonic intervals (and even shorter intervals in lower dimensions). This periodicity creates "hash collisions" (red planes), where distant positions share near-identical embeddings, making the model susceptible to distrac-

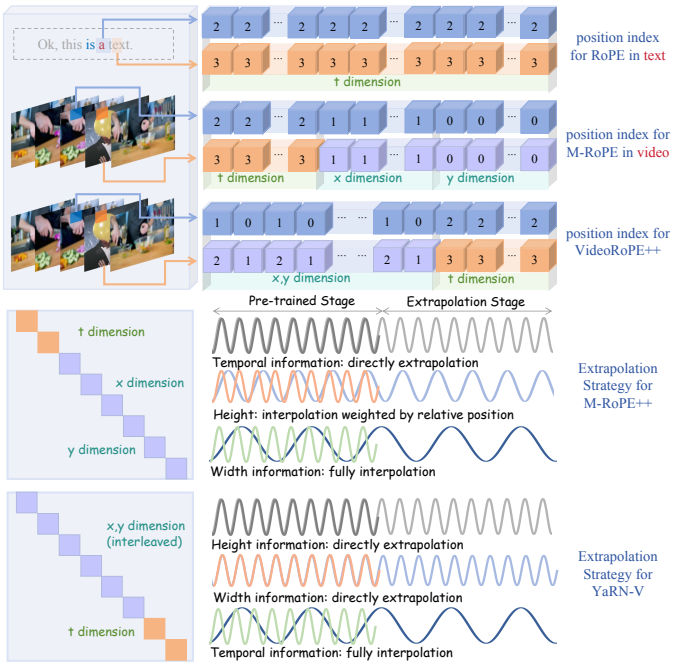


Fig. 5: The upper part shows that VideoRoPE++ adopts a 3D position indexing structure and a frequency allocation strategy that maintains a coherent spatiotemporal encoding design. The lower part illustrates the spectral behavior of different extrapolation strategies: while M-RoPE++ performs extrapolation or interpolation along spatial axes, VideoRoPE++ applies frequency interpolation exclusively along the temporal axis.

tor influence. Fortunately, our VideoRoPE++ (Fig. 4 (**b**)) is free from oscillation and Hash collision in temporal modeling. The relationship between periodicity, monotonicity, and temporal modeling is visualized in Fig 4.

Diagonal Layout. Fig. 6 provides a visual comparison of spatial symmetry in positional encodings. For vanilla RoPE (Fig. 6a), no spatial relation is considered and the index for every dimension increases directly. While M-RoPE (Fig. 6b), incorporates spatial information within each frame, it introduces two significant discontinuities between textual and visual tokens. This arises from M-RoPE's placement strategy, if the first visual token is at $(0, 0)$, the last token in each frame will always be placed at $(W-1, H-1)$, creating a stack in the bottom-left corner. Furthermore, like vanilla RoPE, M-RoPE's indices increase with input length across all dimensions.

To address these limitations, VideoRoPE++ arranges the entire input along the diagonal, see Fig. 6c. The central patch's 3D position for each video frame is (t, t, t), with other patches offset in all directions. Our **Diagonal Layout** has two advantages: (1) our design preserves the relative positions of visual tokens and ensures approximate equidistance from the image corners to the center, preventing text tokens from being overly close to any corner. (2) It maintains the indexing pattern of vanilla RoPE (Fig. 5), as the position index increment between corresponding spatial locations in adjacent frames mirrors that of adjacent textual tokens.

Adjustable Temporal Spacing. To scale the temporal index, we introduce a scaling factor δ to better align temporal information between visual and textual tokens.

Suppose the symbol τ denotes the token index, for the starting text ($0 \leq \tau < T_s$), the temporal, horizontal, and vertical indices are simply set to the raw token index τ. For the video input ($T_s \leq \tau < T_s + T_v$), The difference $\tau - T_s$ represents the index of the current frame relative to the start of the video, which is then scaled by δ to control the space in the temporal dimension. For the ending text ($T_s+T_v \leq \tau < T_s+T_v+T_e$), the temporal, horizontal, and vertical index are the same, creating a linear progression.

According to our adjustable temporal spacing design, for a multi-modal input that consists of a text with T_s tokens, a following video with T_v frame with $W \times H$ patches in each frame, and an ending text with T_e tokens, the position indices (t, x, y) of VideoRoPE++ for τ-th textual token or (τ, w, h)-th visual token are defined as Eq. (7):

$$(t,x,y) = \begin{cases} (\tau, \tau, \tau) & \text{if } 0 \leq \tau < T_s \\ \begin{pmatrix} T_s + \delta(\tau - T_s), \\ T_s + \delta(\tau - T_s) + w - \frac{W}{2}, \\ T_s + \delta(\tau - T_s) + h - \frac{H}{2} \end{pmatrix} & \text{if } T_s \leq \tau < T_s + T_v \\ \begin{pmatrix} \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v \end{pmatrix} & \text{if } T_s + T_v \leq \tau < T_s + T_v + T_e \end{cases}, \tag{7}$$

where w and h represent the horizontal and vertical indices of the visual patch within the frame, respectively.

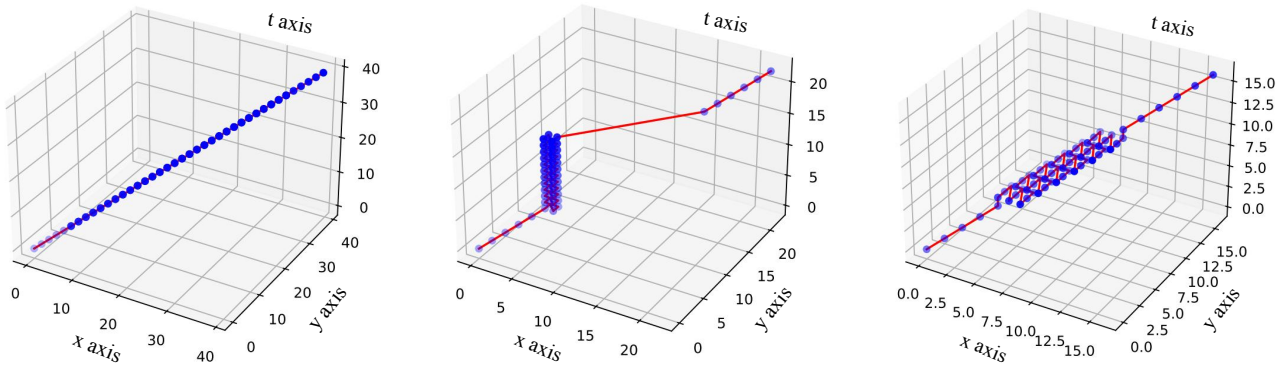
In summary, the parameter δ in our adjustable temporal spacing allows for a flexible and consistent way to encode the relative positions of text and video tokens.

Extrapolation Capability with YaRN-V. Video understanding poses specific challenges for position encoding due to the differing properties of spatial and temporal dimensions. Spatial information, such as textures and edges, is typically local and repetitive, whereas temporal information spans longer, variable ranges and demands extended contextual coverage. To address this asymmetry, we propose **YaRN-V**, an extrapolation method that applies frequency interpolation exclusively along the **temporal dimension** (Δt), while leaving the spatial dimensions (Δx, Δy) unchanged. This selective design preserves spatial structure and improves temporal generalization in long-context video modeling.

The effectiveness of YaRN-V is supported by the distinct spectral behavior of space and time. Spatial dimensions operate in higher frequency bands, where position indices complete a full cycle within the observed training range. This periodicity allows the model to generalize to unseen spatial positions without interpolation. In contrast, the temporal dimension lies in a lower frequency regime, where the model does not observe a full cycle during training. Therefore, interpolation along the temporal axis becomes both necessary and sufficient to support extrapolation in extended video contexts.

Let β denote the base frequency used in rotary embeddings (e.g., $\beta = 10000$), and d be the rotary embedding dimensionality. The frequency used for the n-th pair of dimensions in vanilla RoPE is:

$$\theta_n = \beta^{-2n/d}, \quad \text{where } n = 0, 1, \ldots, \frac{d}{2} - 1. \tag{8}$$



(a) 3D visualization for Vanilla RoPE. (b) 3D visualization for M-RoPE. (c) 3D visualization for VideoRoPE++.

Fig. 6: The 3D visualization for different position embedding. **(a)** The vanilla 1D RoPE [1] does not incorporate spatial modeling. **(b)** M-RoPE [17], which has the 3D structure, introduces a discrepancy in index growth for visual tokens across frames, with some indices remaining constant. **(c)** In contrast, our VideoRoPE++ achieves the desired balance, maintaining the consistent index growth pattern of vanilla RoPE while simultaneously incorporating spatial modeling.

For temporal extrapolation, we define a scaled base $\tilde{\beta}$ as:

$$\tilde{\beta} = \beta \cdot \left(\frac{T'}{T}\right)^{\frac{d}{d-2}}, \tag{9}$$

where T is the maximum sequence length during pre-training, and T' is the target extrapolated length.

The temporal rotary frequencies are then computed as:

$$\theta_n^{(t)} = \tilde{\beta}^{-2n/d}, \quad \text{for temporal dimensions only.} \tag{10}$$

In contrast, spatial frequencies $\theta_n^{(x)}$ and $\theta_n^{(y)}$ remain unchanged:

$$\theta_n^{(x)} = \theta_n^{(y)} = \beta^{-2n/d}. \tag{11}$$

Thus, the composite rotation matrix for VideoRoPE++ becomes:

$$\boldsymbol{R}^{\text{VideoRoPE++}}_{\Delta t, \Delta x, \Delta y} = \text{diag}\left(\boldsymbol{R}^{(\tilde{\beta})}_{\Delta t}, \boldsymbol{R}^{(\beta)}_{\Delta x}, \boldsymbol{R}^{(\beta)}_{\Delta y}\right). \tag{12}$$

By applying frequency scaling exclusively to the temporal axis, VideoRoPE++ preserves the spatial encoding structure and uses natural periodicity of high-frequency spatial dimensions. This selective strategy avoids the distortions introduced by global scaling and leads to significantly more robust extrapolation in long-context video modeling.

B. V-RULER: A Long-Context Retrieval Benchmark with Difficult Tasks

To rigorously evaluate the long-context capabilities of Video LLMs, we propose **V-RULER**, a benchmark suite built upon and significantly extending on V-NIAH-D [23]. As shown in Fig. 7, V-RULER introduces five task types, each simulating realistic and cognitively challenging scenarios that stress-test different facets of long video retrieval.

Multi-Key, Multi-Value (MKMV) [20] is designed to test the model's ability to resolve fine-grained temporal and semantic associations across multiple entities. In this task, several visual subjects appear at different timestamps, each associated with multiple actions or roles. For example, the balloon-headed person in Fig. 7 performs distinct activities such as cleaning, jumping, and working at different points in the video. The model must correctly identify all relevant actions and link them to the same entity across time. In another case, the frames where shoes are being tied involve multiple characters—such as the player, boy, and goat—that the model must recognize and differentiate. These settings require the model to track entity identity, interpret varying behavior, and resolve multiple semantic targets in a long video stream.

Needle Retrieval under Distractors (NRD) [23] is adapted from the V-NIAH-D setting [23] and forms one of the key tasks in the V-RULER benchmark. The model receives a question that refers to a specific but implicit target frame in the video, such as "What color is the towel under it?". The goal is to locate the correct frame (referred to as the needle) and generate the answer based only on its content.

To increase difficulty, the needle is surrounded by distractor frames that are visually similar but semantically irrelevant. These distractors may share background, character identity, or posture with the needle frame, but do not contain the information needed to answer the question. This setup forces the model to rely on fine-grained visual cues for accurate localization, rather than approximate similarity.

The data is constructed using GPT-4o to produce image sequences with controlled variation. In these sequences, some entities appear repeatedly with different actions, while others differ in appearance but occur within consistent backgrounds. These controlled variations are inserted at multiple locations in the video, and questions are designed to require disambiguation along both the identity and semantic axes.

This task evaluates the model's ability to perform precise temporal localization, resolve entity consistency over time, and distinguish between relevant and irrelevant visual evidence in densely populated input sequences.

Lengthy Multimodal Stack is designed to assess how well a model can isolate relevant video information when presented with mixed-modality input. In this task, the question depends entirely on the video content, but the input sequence also includes a large amount of unrelated text, such as noisy or misleading subtitles. As shown in Fig. 7, the video stream contains the required visual signal (e.g., a football match), while the accompanying long transcript does not help and

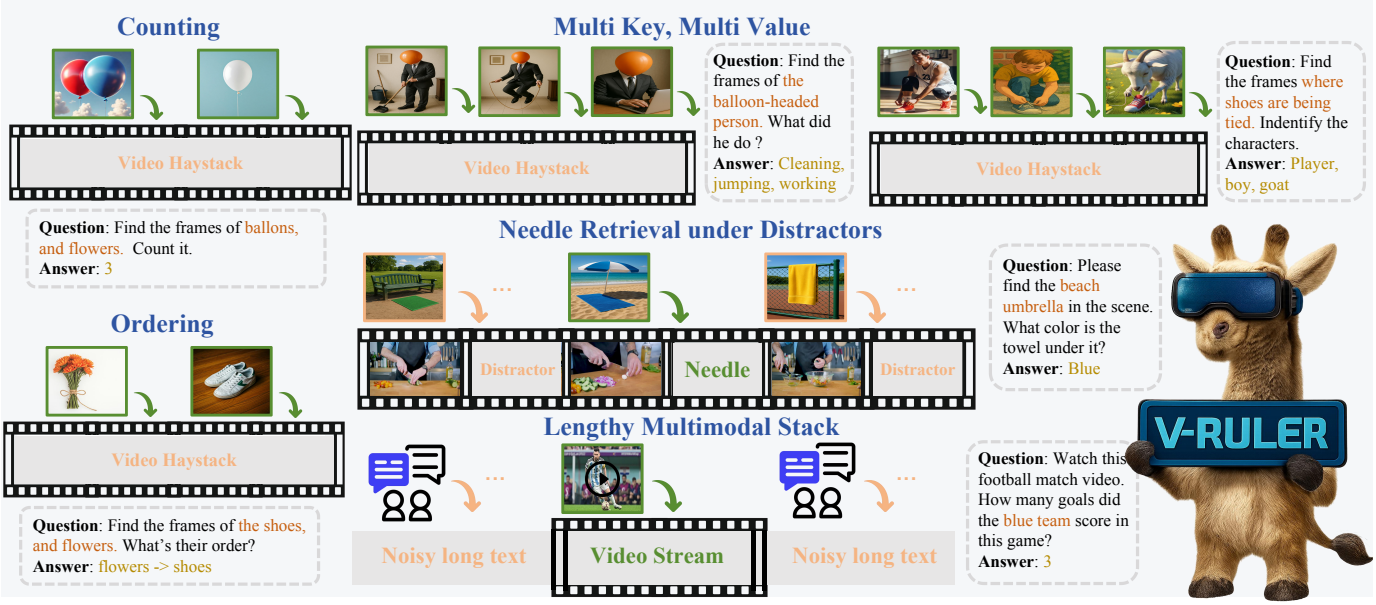


Fig. 7: **Overview of V-RULER benchmark**. Each task type is designed to evaluate a distinct aspect of video haystack retrieval ability, including fine-grained action disambiguation, semantic abstraction, long-range counting, temporal ordering, modality focus, and robustness to visual distractors.

contains unrelated information. The model must identify that the question is grounded in the video and focus on the correct source of information. This task evaluates the model's ability to select the appropriate modality and maintain performance when irrelevant inputs from other modalities are present in a long sequence.

Counting [62], [63] evaluates a model's ability to detect and aggregate repeated visual events across extended video input. In this task, a small number of frames contain the target object or action (e.g., balloons in Fig. 7) and are sparsely distributed within a long video haystack. The model is asked to count how many such target objects exist. The main difficulty comes from the sparsity of the signal: relevant frames are far apart, and no strong positional patterns are available. The model must remember where the target object appears throughout the video and count all occurrences, even if they are far apart and not grouped together.

Ordering [62], [63] tests a different aspect of temporal reasoning: identifying the relative order of visual events. A few frames showing distinct entities (e.g., flowers and shoes in Fig. 7) are inserted into a longer sequence filled with unrelated content. The model must determine the correct order in which the target frames appear. While earlier benchmarks focused on short clips with clear transitions, our implementation supports longer input sequences and increases the temporal range that the model can process. This reduces the reliability of shallow position encoding and requires the model to resolve global temporal structure based on frame-level alignment alone.

Together, V-RULER provides a comprehensive benchmark for evaluating the long-context generalization, temporal reasoning, semantic abstraction, and cross-modal robustness of Video LLMs. Its design encourages the development of models capable of long-context understanding.

V. Experiments

A. Experimental Setup

Training Data. We train VideoRoPE++ using a curated subset of the LLaVA-Video-178k dataset [64], which comprises 178k videos and approximately 5 million question-answer (QA) pairs drawn from diverse sources such as HD-VILA [65], Kinetics [66], and ActivityNet [67]. To ensure a balance between training efficiency and long-context understanding, we sample 136k videos with durations under 2 minutes and 18k videos ranging from 2 to 3 minutes. This results in a training corpus containing roughly 1.3 million QA pairs.

Implementation Details. Using the aforementioned video training data, we fine-tune different models that use different positional encoding strategies, such as the Vanilla RoPE [1], Time-Aware Dual RoPE (TAD-RoPE) [16], M-RoPE [17], and our VideoRoPE++. All models are initialized with the Vision Transformer from Qwen2-VL-7B or Qwen2.5-VL-7B, paired respectively with the language backbone (using Vanilla RoPE) from Qwen2-7B or Qwen2.5-7B [6], [7]. Our fine-tuning incorporates our VideoRoPE++ to process the spatiotemporal nature of the video data effectively. We adopt Qwen2-VL's fine-tuning settings, processing each video at 2 fps with a maximum of 128 frames and dynamically adjusting the image resolution to maintain a consistent token count. However, to prevent memory overflow, we use a context window of 8192 tokens.

We fine-tune all models using a batch size of 128, a cosine learning rate schedule with a peak learning rate of 1×10^{-5}, and a warm-up ratio of 0.01. The entire training process consumes a total of 704 NVIDIA A100 GPU hours.

For evaluation, videos are sampled at 2 frames per second, and each frame is encoded into a minimum of 144 image tokens. To enable inference with extremely long sequences

TABLE II: **Comparison of different RoPE methods on LongVidionBench, MLVU, and Video-MME**. The benchmarks evaluate performance across four context lengths: 8k, 16k, 32k, and 64k, where **8k** represents context within the training range, and others represent context outside the training range. Our VideoRoPE++ outperforms other RoPE variants across all three benchmarks. The best results are marked in **bold**, and the second-best results are underlined.

Method	Backbone	LongVideoBench				MLVU				Video-MME			
		8k	16k	32k	64k	8k	16k	32k	64k	8k	16k	32k	64k
Vanilla RoPE [1]	Qwen2 [6]	54.97	54.87	54.56	54.04	63.31	65.79	65.93	62.02	60.67	60.00	61.33	58.33
TAD-RoPE [16]	Qwen2 [6]	54.14	55.08	53.94	53.42	63.67	65.28	65.28	60.73	60.33	61.33	62.00	58.67
M-RoPE [17]	Qwen2 [6]	53.42	52.80	53.11	54.35	60.41	60.68	61.56	61.10	60.67	59.67	61.00	59.67
M-RoPE [17]	Qwen2.5 [7]	**61.22**	60.05	59.33	58.71	**69.19**	70.70	**69.51**	68.09	**64.33**	**64.33**	**62.33**	60.00
VideoRoPE++ (Ours)	Qwen2 [6]	54.46	55.29	57.15	57.26	65.19	66.29	66.02	65.56	61.33	61.00	61.67	61.33
VideoRoPE++ (Ours)	Qwen2.5 [7]	59.85	**62.03**	**59.54**	**59.12**	68.74	**70.72**	69.06	**68.64**	63.33	**64.33**	**62.33**	**61.67**

(exceeding 32k tokens), we adopt the vLLM framework [68], which provides efficient memory and throughput optimizations for large-scale generative models.

Evaluation Benchmarks. We evaluate our approach across eight video benchmarks spanning six core categories: long video understanding, long video retrieval, video hallucination, short general video understanding, video captioning, and streaming video understanding.

For *long video understanding*, we adopt three representative benchmarks: (1) **LongVideoBench** [69] focuses on reasoning questions that require access to extended frame sequences, which cannot be answered by a single frame or a few sparse snapshots. The videos range from 8 seconds to 1 hour in duration. To ensure fair evaluation, we retain only questions that do not rely on subtitle cues. (2) **MLVU** [62] is a comprehensive benchmark designed to evaluate multimodal LLMs on videos ranging from 3 minutes to 2 hours. It comprises nine diverse evaluation tasks; in our analysis, we focus on seven multiple-choice tasks: Topic Reasoning, Anomaly Recognition, Needle QA, Ego Reasoning, Plot QA, Action Order, and Action Count. (3) **Video-MME** [70] provides a high-quality evaluation suite that spans six major visual domains and 30 subfields. Covering a wide temporal range—from short 11-second clips to long videos up to 1 hour—it serves as a rigorous testbed for general long-video comprehension.

For *long video retrieval*, we evaluate on our proposed **V-RULER**, an extension of V-NIAH-D [23] designed to assess model performance under extreme long-context scenarios with greater task diversity and realism.

For all non-Lengthy Multimodal Stack tasks, we adopt the LongVA [12] configuration: a target "needle" image is inserted at a random position within a 3,000-frame haystack, where each frame is encoded into 144 visual tokens. The needle corresponds to a visual query that is unrelated to the surrounding video content. Frames are inserted at fixed intervals (0.2 in normalized depth), with evaluations beginning at 100 frames and increasing in 200-frame increments up to 3,000.

For the Lengthy Multimodal Stack task, we specifically evaluate the model's ability to handle extremely long multimodal inputs. We construct input sequences containing both video and irrelevant textual noise, where only the visual stream contains the answer. The total visual token length is fixed at 32k, sampled from video clips containing 224k to 288k visual tokens in total. This high token count isolates the effect of long-context inference.

To rigorously evaluate extrapolation capabilities, we stress-test various RoPE-based strategies—including M-RoPE and VideoRoPE—that encode position in 3D structures. In such designs, even with over 100k tokens, the temporal position indices often remain shallow (e.g.,$\tilde{1}$k) due to compact spatiotemporal layout. However, inserting long sequences of irrelevant text causes rapid positional growth due to diagonal placement of textual tokens. By intentionally injecting such noise, we ensure that all RoPE variants encounter positional indices far beyond their pretraining range.

We validate this extrapolation setting using a compact **Qwen2.5-3B** backbone, selected for its ability to process extremely long input sequences without causing GPU memory overflow. This lightweight yet capable model enables extrapolation testing under tight hardware constraints. Inference is performed using 2× A800 GPUs within the vLLM framework. This setup enables meaningful comparison of how different RoPE strategies generalize under extreme positional shifts in a multimodal, noise-rich environment, while maintaining practical computational feasibility.

For *video hallucination* evaluation, we adopt **VideoHallucer** [71], a benchmark designed to assess a model's ability to accurately answer both factual and hallucinated questions about video content. VideoHallucer categorizes hallucinations into two main types—*intrinsic* and *extrinsic*—and further breaks them down into subtypes for fine-grained analysis, including object-relation, temporal, and semantic detail hallucinations (intrinsic), as well as extrinsic factual and extrinsic non-factual hallucinations. This framework enables a comprehensive evaluation of the model's robustness against various forms of hallucinated information.

For *video captioning*, we evaluate on the **Video Detailed Captions (VDC)** [45] benchmark, which is designed to assess the ability of multimodal models to generate comprehensive, coherent, and detailed textual descriptions of videos. Over 87% of the videos in this benchmark have durations between 10 and 30 seconds. The captions include detailed information such as background elements, main subjects, and camera movements.

For *streaming video understanding*, we conduct evaluation on the **StreamingBench** [72] benchmark. The goal is to assess the performance of multimodal large models in real-time visual analysis. The benchmark consists of three tasks: (1) real-time visual understanding, (2) omni-source understanding,

and (3) contextual understanding. Among these, only the first task involves streaming video inputs and requires real-time processing. Therefore, we limit our evaluation to the real-time visual understanding task.

For *short general video understanding*, we evaluate on **MVBench** [73], which comprises 20 challenging video tasks that cannot be effectively solved using single-frame inputs. These tasks require models to demonstrate a wide range of temporal skills—from low-level perception to high-level cognition—across videos ranging from 5 to 35 seconds in length, with balanced difficulty levels across questions.

B. Results on Long Video Understanding

As shown in Tab.II, we compare our VideoRoPE++ with existing RoPE variants, including vanilla RoPE [1], TAD-RoPE [16], and M-RoPE [17], across three long-context video understanding benchmarks. VideoRoPE++ consistently surpasses all baseline methods on these benchmarks, with stable and measurable improvements. Specifically, under the Qwen2 backbone, VideoRoPE++ yields gains of up to 2.91, 4.46, and 1.66 points over the M-RoPE baseline at the 64k context length on LongVideoBench, MLVU, and Video-MME, respectively. When using the stronger Qwen2.5 backbone, VideoRoPE++ achieves best-case results that are either higher than or on par with those of M-RoPE. These results indicate that VideoRoPE++ can more effectively model long-range dependencies and sustain performance across diverse and demanding video understanding tasks.

C. Results on Long Video Retrieval

Fig. 8 presents the average retrieval performance of V-RULER when combined with VideoRoPE++ and several RoPE variants. The values are aggregated across all four sub-tasks to show overall trends. In Fig. 8 (1) and (2), both Vanilla RoPE and TAD-RoPE show clear degradation as the context length increases. Their failure to maintain temporal alignment leads to a significant drop in overall accuracy.

By contrast, Fig. 8 (3) and (4) show that M-RoPE and VideoRoPE++ produce higher and more stable average performance across the same range. Among them, VideoRoPE++ consistently achieves the best scores, indicating stronger ability to handle long-range retrieval.

Tab. III complements the figure by providing detailed results for each individual sub-task in the V-RULER benchmark: Multi-Key Multi-Value (MKMV), Needle Retrieval under Distractors (NRD), Counting, and Ordering. These sub-tasks assess different aspects of long video retrieval, including semantic matching, temporal localization, and distractor resistance.

Vanilla RoPE and TAD-RoPE perform poorly across all sub-tasks. For example, their scores on MKMV are 32.26 and 27.46, while Counting yields 29.33 and 36.00, respectively. The average accuracy for both methods remains below 30, reflecting their limited capacity to encode position information in long sequences.

M-RoPE performs strongly across all tasks, with an average score of 82.24. Building on that, VideoRoPE++ improves further, achieving the best results on every sub-task: 88.53 on MKMV, 95.78 on NRD, 82.67 on Counting, and 84.00 on Ordering. The overall average reaches 87.75. These results confirm that VideoRoPE++ effectively supports long-range retrieval by preserving semantic relations and temporal structure.

TABLE III: **Performance comparison on the four V-RULER subtasks.** Here, **MKMV** refers to Multi-Key, Multi-Value, and **NRD** denotes Needle Retrieval under Distractors.

Method	MKMV	NRD	Counting	Ordering	Avg.
Vanilla RoPE [1]	32.26	23.78	29.33	30.67	29.01
TAD-RoPE [16]	27.46	23.11	36.00	27.99	28.64
M-RoPE [17]	<u>84.53</u>	<u>89.78</u>	<u>81.33</u>	<u>73.33</u>	<u>82.24</u>
VideoRoPE++	**88.53**	**95.78**	**82.67**	**84.00**	**87.75**

TABLE IV: **Performance comparison of different RoPEs on VideoHallucer**, evaluated at context lengths of 8k, 16k, 32k, and 64k. The maximum result for each RoPE variant across these context lengths is displayed, with bold for the top result and underlined for the second-highest. 'OR' = Object-Relation, 'T' = Temporal, 'SD' = Semantic Detail, 'F' = Factual, 'NF' = Non-factual.

Method	OR	T	SD	F	NF	Avg.
Vanilla RoPE [1]	<u>51.5</u>	30.0	<u>48.0</u>	8.0	43.0	36.1
TAD-RoPE [16]	51.0	<u>37.0</u>	<u>48.0</u>	11.5	<u>47.5</u>	39.0
M-RoPE [17]	39.0	29.0	43.5	<u>12.5</u>	<u>47.5</u>	34.3
VideoRoPE++	**57.0**	**58.5**	**50.5**	**15.0**	**50.0**	**46.2**

D. Results on Video Hallucination

As highlighted in Tab. IV, VideoRoPE++ significantly surpasses current RoPE methods on the VideoHallucer benchmark. In particular, for the Temporal Hallucination task, VideoRoPE++ demonstrates a substantial performance improvement of 29.5%, indicating its enhanced capability to accurately capture and process temporal dependencies. This improvement suggests that VideoRoPE++ is better equipped to handle dynamic video sequences, where the understanding of time-based relationships is critical. Similarly, for the Spatial Hallucination task, specifically the Object-Relation Hallucination subtask, VideoRoPE++ achieves an impressive 18.0% improvement over existing methods, highlighting its ability to better discern complex spatial interactions. These results underscore VideoRoPE++'s robustness in solving video hallucination and potential for real-world video analysis.

E. Results on Other Video Tasks

To evaluate the generalization ability of our model beyond the primary benchmarks, we assess its performance on three representative video understanding tasks: VDC for video captioning, MVBench for short-form general video understanding, and StreamingBench (RTVU) for real-time video analysis. Table V reports a comparison across multiple RoPE-based models, including our VideoRoPE++.

VideoRoPE++ achieves the best or tied-best results on all three benchmarks. On the VDC captioning task, both VideoRoPE++ and M-RoPE reach the top score of 44.0, indicating strong ability in generating detailed and coherent

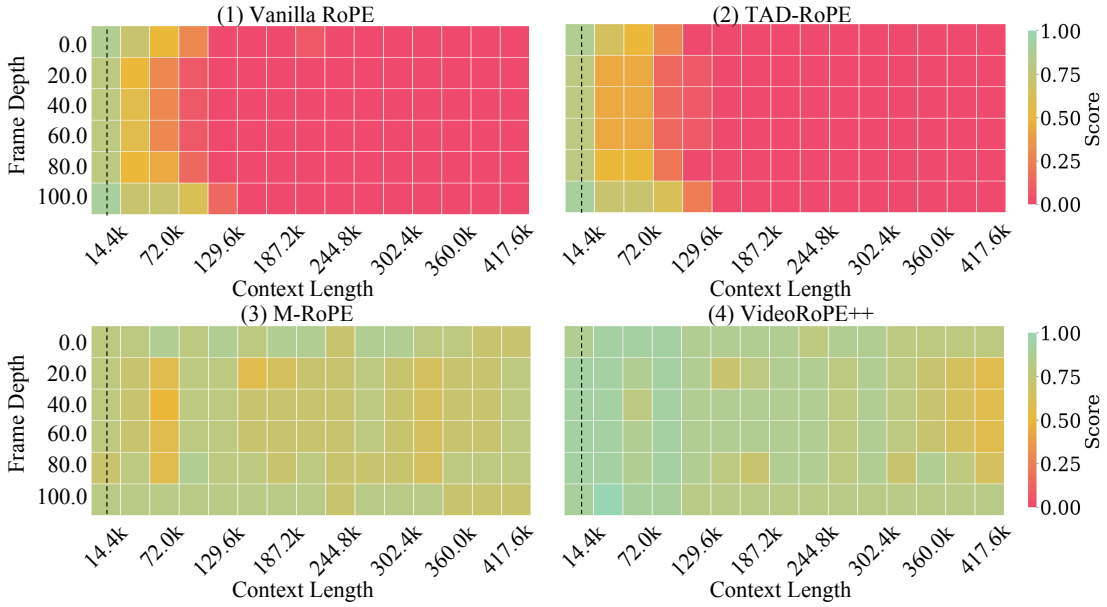


Fig. 8: **The average performance of V-RULER sub-tasks—MKMV, NRD, Counting, and Ordering—evaluated across various RoPE variants.** The black dashed line marks the division between the training context (left side) and the extrapolated region (right side).

TABLE V: **Comparison across three representative benchmarks**: **MVBench** for general short video understanding, **VDC** for video captioning, and **StreamingBench (RTVU)** for real-time video-streaming understanding tasks. Our **VideoRoPE++** achieves the best or competitive performance in all settings.

Model	**VDC** (Caption)	**MVBench** (General)	**StreamingBench** (RTVU)
Auroracap-7B	38.2	–	–
Vanilla RoPE	43.0	67.1	75.0
TAD-RoPE	43.8	66.9	75.8
M-RoPE	**44.0**	67.8	76.2
VideoRoPE++	**44.0**	**68.4**	**77.1**

video descriptions. On MVBench, which focuses on general video understanding across 20 diverse tasks, VideoRoPE++ achieves a score of 68.4, surpassing all other baselines. On StreamingBench (RTVU), which requires real-time comprehension of streaming video, VideoRoPE++ obtains a leading score of 77.1, outperforming M-RoPE (76.2), TAD-RoPE (75.8), and Vanilla RoPE(75.0).

These results confirm that the improvements introduced in VideoRoPE++ support stronger temporal reasoning and more stable inference in both offline and streaming video scenarios.

F. Comparison with State-of-the-Art Models

To further validate the effectiveness of our proposed VideoRoPE++, we compare it against several recent state-of-the-art VideoLLMs, including LLaVA OneVision [74], LongVU [75], Apollo [76], and LLaVA-Video [64], at both 3B and 7B scales. As shown in Table VI, our **VideoRoPE++-3B** outperforms all 3B-scale baselines on the LongVideoBench, MLVU, and VideoMME benchmarks, achieving scores of 54.7, 62.6, and 58.3, respectively.

On 7B-scale models, our **VideoRoPE++-7B** also achieves competitive or superior performance compared to larger models:

TABLE VI: **Comparison with state-of-the-art VideoLLMs at both 3B and 7B scales on three long video understanding benchmarks.** Our VideoRoPE++ achieves the best or competitive results while using significantly fewer fine-tuning samples.

Model	**LongVideoBench**	**MLVU**	**VideoMME**
Models with 3B scale			
VILA1.5-3B	42.9	44.4	42.2
Phi-3.5-Vision-4.2B	–	–	50.8
LongVU-3.2B	–	55.9	51.5
VideoRoPE++-3B	**54.7**	**62.6**	**58.3**
Models with 7B scale			
LLaVA OneVision-7B	56.3	64.7	58.2
LongVU-7B	–	65.4	60.6
Apollo-7B	58.5	**70.9**	61.3
LLaVA-Video-7B	58.2	70.8	63.3
VideoRoPE++-7B	**62.0**	70.7	**64.4**

62.0 on LongVideoBench, 70.7 on MLVU, and 64.4 on VideoMME, outperforming LLaVA OneVision-7B, Apollo-7B, and LLaVA-Video-7B on most metrics.

Notably, our method uses only 0.33 million fine-tuning samples—significantly fewer than the 8.8 million used by LLaVA OneVision, 3.75 million by LongVU, 3.2 million by Apollo, and 2.7 million by LLaVA-Video. These results highlight the efficiency and generalization benefits of our RoPE-based design, enabling strong long-video understanding performance even under low-resource supervision.

G. Results on Different RoPE Extrapolation Strategies

In this experiment, we compare several extrapolation strategies by applying them to inputs with position indices that exceed the training range. The evaluation is conducted on the Lengthy Multimodal Stack subtask in the V-RULER benchmark, as summarized in Table VII and visualized in Fig. 9.

The default strategy, which applies no extrapolation, fails entirely under this condition, indicating that position encoding without any adaptation does not generalize to longer sequences. Among the baseline methods, NTK-Aware and MRoPE++ achieve scores of 67.66 and 62.30, respectively. While these approaches preserve some generalization, their performance remains limited.

YaRN achieves a score of 68.33, slightly outperforming NTK-Aware. Our method, YaRN-V, reaches 81.33, which is the highest among all evaluated strategies. Compared to the strongest baseline, YaRN-V provides an improvement of 13.0 points. This gain reflects a more stable handling of long-range position indices, especially in the presence of mixed-modality distractors, as required by the Lengthy Multimodal Stack task. The experimental results suggest that YaRN-V is better suited for video LLMs under extended input lengths, where maintaining temporal alignment and avoiding degradation from positional overflow are essential.

TABLE VII: **Comparison of different extrapolation strategies.** on the Lengthy Multimodal Stack subtask of V-RULER.

Extrapolation Strategy	Lengthy Multimodal Stack
Default	fail
NTK-Aware [35]	67.66
MRoPE++ [19]	62.30
YaRN [24]	68.33
YaRN-V	**81.33**

H. Ablation Studies

Ablation Studies on Module Design. We conduct ablation studies on the components introduced in Section VIII, evaluating their effects on LongVideoBench and MLVU across 8k to 64k context lengths. Results are shown in Table VIII.

Starting from the baseline M-RoPE [17], the model scores 54.35 and 61.10 at 64k on LongVideoBench and MLVU, respectively. Adding the Diagonal Layout (DL) alone brings no consistent improvement. When combined with Low-frequency Temporal Allocation (LTA), performance increases across all lengths, reaching 57.06 and 63.26 at 64k. The full model, including DL, LTA, and Adjustable Temporal Spacing (ATS), achieves the best results: 57.26 on LongVideoBench and 65.56 on MLVU at 64k. The gains are consistent across other context lengths as well.

These results confirm that each module contributes to performance, with LTA and ATS providing the most noticeable improvements in long-context settings.

Ablation Studies on the Scaling Factor δ in ATS. To further examine the effect of the scaling factor δ in the Adjustable Temporal Spacing (ATS) module, we perform controlled experiments varying δ in a fixed architecture. The ATS module is designed to adjust the spacing of temporal position indices, which may affect the model's ability to align semantic and sequential information, particularly in long-context video-language settings.

We evaluate performance across three representative benchmarks—**LongVideoBench**, **MLVU**, and **VideoMME**—by sweeping δ from 0.5 to 3.0. Results are shown in Table IX. As δ increases, performance improves steadily and reaches a maximum when $\delta = 2$, where the average score across tasks reaches **60.92**. Further increases beyond this point do not yield additional gains and may introduce instability.

TABLE VIII: **Ablation study about different modules of VideoRoPE++.**

Method	LongVideoBench				MLVU			
	8k	16k	32k	64k	8k	16k	32k	64k
Baseline	53.42	52.80	53.11	54.35	60.41	60.68	61.56	61.10
+ DL	52.17	52.07	53.31	53.63	62.06	63.03	62.52	62.75
+ DL & LTA	**54.46**	**55.49**	54.66	55.60	63.35	64.09	64.00	63.26
+ DL & LTA & ATS	**54.46**	55.29	**57.15**	**57.26**	**65.19**	**66.29**	**66.02**	**65.56**

TABLE IX: **Performance under different scaling factors δ across multiple benchmarks.**

δ	LongVideoBench	MLVU	VideoMME	Avg
0.5	50.83	59.87	58.33	56.34
1.0	54.11	63.54	59.67	59.11
2.0	55.50	65.59	61.67	**60.92**
3.0	53.83	63.38	60.33	59.18

Ablation Studies on the Extrapolation Factor. We conduct an ablation study to examine the impact of the extrapolation factor on the performance of YaRN-V in the Lengthy Multimodal Stack task of V-RULER, as shown in Table X. We vary the extrapolation factor from 1.0 to 8.0 while keeping other configurations fixed. When the factor is 1.0, the model fails to produce valid outputs, indicating that a minimal extrapolation range is insufficient for generalization. Increasing the factor to 2.0 yields a substantial performance improvement, reaching 78.00 accuracy. The performance peaks at a factor of 4.0, achieving 81.33, which is also the default configuration recommended by Qwen2 [6]. However, further increasing the factor to 8.0 results in degraded performance (60.66), suggesting that overly aggressive extrapolation may lead to positional distortion and reduced robustness. These results indicate that moderate extrapolation, particularly with the Qwen2-default factor of 4.0, provides the best trade-off between range extension and model stability.

Ablation Studies on x, y Allocation. To analyze the impact of different spatial allocation strategies in the positional encoding scheme, we perform controlled ablation studies using **VideoRoPE++**. Specifically, we compare a sequential layout—where all x coordinates are placed before all y coordinates (e.g., $x, x, \ldots, y, y, \ldots$)—against an interleaved layout, where x and y positions alternate (e.g., $x, y, x, y, \ldots$).

Results are shown in Table XI. The interleaved allocation consistently yields higher scores across two benchmarks at all context lengths. For instance, on LongVideoBench at 64k, the interleaved design reaches 57.26 compared to 54.77 for the sequential version. A similar trend is observed on MLVU, where interleaved allocation improves the 64k score from 63.08 to 65.56. We attribute these gains to the structural balance introduced by the interleaved pattern. By alternating x and y dimensions, the model receives more uniform spatial information at each step, which helps maintain local coherence. In contrast, the sequential layout creates a separation between dimensions, which may weaken spatial continuity and increase

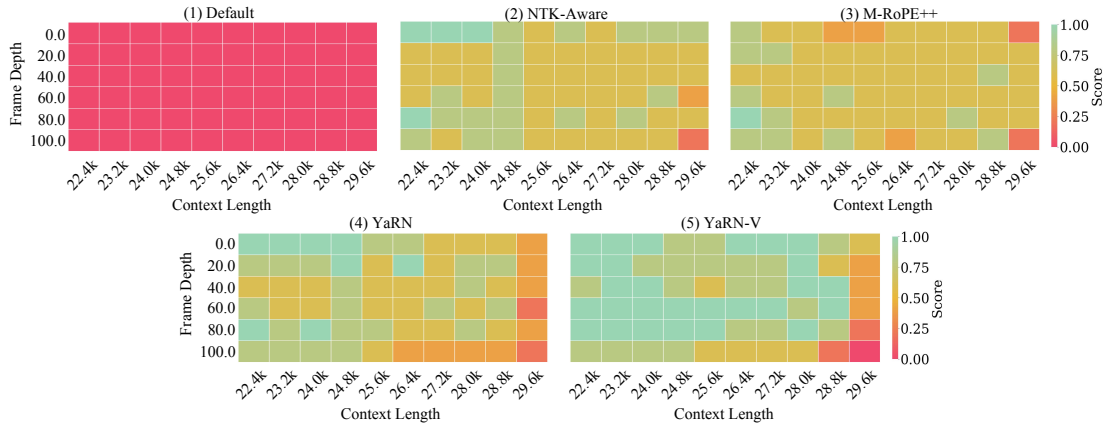


Fig. 9: **Visualization of results under different extrapolation strategies on the Lengthy Multimodal Stack task.** Without any extrapolation method, VideoRoPE++ fails when the position indices go far beyond the training range.

TABLE X: **Effect of Extrapolation Factor on VideoRoPE++ Performance in Lengthy Multimodal Stack task.**

Extrapolation Factor on VideoRoPE++	Lengthy Multimodal Stack
1.0	fail
2.0	78.00
4.0	**81.33**
8.0	60.66

TABLE XI: **Ablation Study on x, y Allocation. Sequential** represents the sequential allocation of x and y, following the pattern $x, x, x, \ldots, y, y, y, \ldots$ (similar to M-RoPE [17]). **Interleaved** represents the interleaved allocation, following the pattern $x, y, x, y, \ldots$ (similar to [77]).

Method	LongVideoBench				MLVU			
	8k	16k	32k	64k	8k	16k	32k	64k
Sequential	53.73	53.52	54.97	54.77	62.75	63.31	62.75	63.08
Interleaved	**54.46**	**55.29**	**57.15**	**57.26**	**65.19**	**66.29**	**66.02**	**65.56**

difficulty in learning position-sensitive patterns.

Ablation Studies on Frequency Allocation Strategies. We compare three frequency allocation strategies for rotary position embeddings: (1) the *M-RoPE* approach, which encodes temporal positions with high-frequency components and arranges dimensions in a sequential format such as `[t...x...y...]`; (2) an interleaved scheme, e.g., `[t t x y x y x y]`, that mixes temporal and spatial dimensions evenly across the embedding space; and (3) our proposed *VideoRoPE++* design, which allocates the high-frequency spectrum to spatial coordinates and applies low-frequency encoding to the temporal axis in a format like `[x y...t...]`.

We evaluate all three strategies on the *LongVideoBench* benchmark under different context lengths. This benchmark includes diverse video types, ranging from rapid temporal transitions to static or slowly changing content, making it well-suited for assessing temporal modeling fidelity. As shown in Table XII, our low-frequency temporal allocation consistently yields stronger performance than the interleaved scheme, which suggests that modeling temporal variation with low-frequency signals allows more stable generalization across a wide range of temporal granularities. In contrast, interleaving dilutes both temporal and spatial axes across frequencies, which may weaken structure in either domain.

TABLE XII: **Comparison of different frequency allocation strategies under various context lengths.**

Context	[t...x...y...]	[t t x y x y]	[xy...t...] (Ours)
16k	60.05	59.95	**62.03**
32k	59.33	58.40	**59.54**
64k	58.71	57.73	**59.12**
Avg	59.36	59.06	**60.14**

VI. Conclusion

This paper identifies five key criteria for effective positional encoding: 2D/3D structure, frequency allocation, spatial symmetry, temporal index scaling, and extrapolation capability. To reveal the limitations of current position embedding designs, we propose a challenging benchmark, V-RULER. As part of our analysis, through the Needle Retrieval under Distractors sub-task of V-RULER, we demonstrate that previous RoPE variants are vulnerable to distractors because of a lack of proper temporal allocation. As a result, we propose VideoRoPE++ that uses a 3D structure for spatiotemporal coherence, low-frequency temporal allocation to reduce oscillations, a diagonal layout for spatial symmetry, adjustable temporal spacing, and YaRN-V for extrapolation. VideoRoPE++ outperforms previous RoPE variants in various tasks like long video retrieval, video understanding, and video hallucination.

VII. *Acknowledgments

This work was supported by the National Key R&D Program of China 2022ZD0161600, Shanghai Artificial Intelligence Laboratory, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Dahua Lin is a PI of CPII under the InnoHK.

References

[1] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, 2024. 1, 3, 5, 7, 8, 9, 10

[2] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, and M. Yang, "LongRoPE: Extending llm context window beyond 2 million tokens," *arXiv preprint arXiv:2402.13753*, 2024. 1

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 1

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "LLaMA 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. 1, 3

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024. 1

[6] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024. 1, 3, 8, 9, 12

[7] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024. 1, 8, 9

[8] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "VideoChat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023. 1

[9] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning united visual representation by alignment before projection," in *EMNLP*, 2024. 1

[10] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, L. Yuan, Y. Qiao, D. Lin, F. Zhao, and J. Wang, "Sharegpt4video: Improving video understanding and generation with better captions," in *NeurIPS*, 2024. 1

[11] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *ACL*, 2024. 1

[12] P. Zhang, K. Zhang, B. Li, G. Zeng, J. Yang, Y. Zhang, Z. Wang, H. Tan, C. Li, and Z. Liu, "Long context transfer from language to vision," *arXiv preprint arXiv:2406.16852*, 2024. 1, 2, 3, 4, 9

[13] X. Wang, D. Song, S. Chen, C. Zhang, and B. Wang, "LongLLaVA: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture," *arXiv preprint arXiv:2409.02889*, 2024. 1, 3

[14] Y. Chen, F. Xue, D. Li, Q. Hu, L. Zhu, X. Li, Y. Fang, H. Tang, S. Yang, Z. Liu, Y. He, H. Yin, P. Molchanov, J. Kautz, L. Fan, Y. Zhu, Y. Lu, and S. Han, "Longvila: Scaling long-context visual language models for long videos," in *ICLR*, 2025. 1, 3

[15] P. Zhang, X. Dong, Y. Cao, Y. Zang, R. Qian, X. Wei, L. Chen, Y. Li, J. Niu, S. Ding, Q. Guo, H. Duan, X. Chen, H. Lv, Z. Nie, M. Zhang, B. Wang, W. Zhang, X. Zhang, J. Ge, W. Li, J. Li, Z. Tu, C. He, X. Zhang, K. Chen, Y. Qiao, D. Lin, and J. Wang, "Internlm-xcomposer2.5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions," *arXiv preprint arXiv:2412.09596*, 2024. 1

[16] M. Gao, J. Liu, M. Li, J. Xie, Q. Liu, B. Zhao, X. Chen, and H. Xiong, "TC-LLaVA: Rethinking the transfer from image to video understanding with temporal considerations," *arXiv preprint arXiv:2409.03206*, 2024. 1, 2, 3, 8, 9, 10

[17] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3, 5, 7, 8, 9, 10, 12, 13

[18] J. Su. (2024, March) Transformer upgrade path: 17. insights into multimodal positional encoding. [Online]. Available: https://spaces.ac.cn/archives/10040 1, 3, 5

[19] M. Li, L. Li, S. Gong, and Q. Liu, "GIRAFFE: Design choices for extending the context length of visual language models," *arXiv preprint arXiv:2412.12735*, 2024. 1, 3, 12

[20] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg, "RULER: What's the real context size of your long-context language models?" in *COLM*, 2024. 2, 4, 7

[21] T. Yuan, X. Ning, D. Zhou, Z. Yang, S. Li, M. Zhuang, Z. Tan, Z. Yao, D. Lin, B. Li *et al.*, "Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k," *arXiv preprint arXiv:2402.05136*, 2024. 2, 4

[22] J. Su. (2024, Sep) A brief discussion on multimodal thinking: 3. positional encoding. [Online]. Available: https://spaces.ac.cn/archives/10352 2, 4

[23] X. Wei, X. Liu, Y. Zang, X. Dong, P. Zhang, Y. Cao, J. Tong, H. Duan, Q. Guo, J. Wang *et al.*, "VideoRoPE: What makes for good video rotary position embedding?" in *ICML*, 2025. 2, 3, 7, 9

[24] B. Peng, J. Quesnelle, H. Fan, and E. Shippole, "YaRN: Efficient context window extension of large language models," in *ICLR*, 2024. [Online]. Available: https://openreview.net/forum?id=wHBfxhZu1u 3, 12

[25] X. Liu, H. Yan, S. Zhang, C. An, X. Qiu, and D. Lin, "Scaling laws of rope-based extrapolation," in *ICLR*, 2024. 3, 4

[26] Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu *et al.*, "Internlm2 technical report," *arXiv preprint arXiv:2403.17297*, 2024. 3

[27] InternLM. (2025, January) Internlm3-8b. [Online]. Available: https://huggingface.co/internlm/internlm3-8b-instruct 3

[28] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024. 3

[29] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *TPAMI*, 2024. 3

[30] Z. Liang, Y. Xu, Y. Hong, P. Shang, Q. Wang, Q. Fu, and K. Liu, "A survey of multimodel large language models," *TPAMI*, 2024. 3

[31] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Cap4Video++: Enhancing video understanding with auxiliary captions," *TPAMI*, 2024. 3

[32] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Learning to answer visual questions from web videos," *TPAMI*, 2022. 3

[33] C. Liang, W. Wang, T. Zhou, J. Miao, Y. Luo, and Y. Yang, "Local-global context aware transformer for language-guided video," *TPAMI*, 2023. 3

[34] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *TPAMI*, 2019. 3

[35] u/bubblematt, "Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation," June 2023, accessed on Reddit. [Online]. Available: https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/ 3, 12

[36] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023. 3

[37] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li *et al.*, "Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd," in *NeurIPS*, 2024. 3

[38] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025. 3

[39] Y. Zang, X. Dong, P. Zhang, Y. Cao, Z. Liu, S. Ding, S. Wu, Y. Ma, H. Duan, W. Zhang *et al.*, "InternLM-XComposer2. 5-Reward: A simple yet effective multi-modal reward model," in *Findings of ACL*, 2025. 3

[40] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021. 3

[41] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *CVPR*, 2021. 3

[42] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021. 3

[43] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "Vtimellm: Empower llm to grasp video moments," in *CVPR*, 2024. 3

[44] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu *et al.*, "Mm-vid: Advancing video understanding with gpt-4v (ision)," *arXiv preprint arXiv:2310.19773*, 2023. 3

[45] W. Chai, E. Song, Y. Du, C. Meng, V. Madhavan, O. Bar-Tal, J.-N. Hwang, S. Xie, and C. D. Manning, "AuroraCap: Efficient, performant video detailed captioning and a new benchmark," in *ICLR*, 2025. 3, 9

[46] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *ECCV*, 2024. 3

[47] P. Jin, R. Takanobu, C. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *CVPR*, 2024. 3

[48] L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng, "PLLaVA: Parameter-free llava extension from images to videos for video dense captioning," *arXiv preprint arXiv:2404.16994*, 2024. 3

[49] S. Zhang, Q. Fang, Z. Yang, and Y. Feng, "LLaVA-Mini: Efficient image and video large multimodal models with one vision token," *arXiv preprint arXiv:2501.03895*, 2025. 3

[50] R. Qian, S. Ding, X. Dong, P. Zhang, Y. Zang, Y. Cao, D. Lin, and J. Wang, "Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction," in *CVPR*, 2025. 3

[51] Y. Li, J. Niu, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian, P. Zhang, Y. Zang, Y. Cao, C. He, and J. Wang, "Ovo-bench: How far is your video-llms from real-world online video understanding?" in *CVPR*, 2025. 3

[52] R. Qian, X. Dong, P. Zhang, Y. Zang, S. Ding, D. Lin, and J. Wang, "Streaming long video understanding with large language models," in *NeurIPS*, 2024. 3

[53] S. Ding, R. Qian, X. Dong, P. Zhang, Y. Zang, Y. Cao, Y. Guo, D. Lin, and J. Wang, "SAM2Long: Enhancing sam 2 for long video segmentation with a training-free memory tree," *arXiv preprint arXiv:2410.16268*, 2024. 3

[54] Z. Wang, S. Yu, E. Stengel-Eskin, J. Yoon, F. Cheng, G. Bertasius, and M. Bansal, "VideoTree: Adaptive tree-based video representation for llm reasoning on long videos," in *CVPR*, 2024. 3

[55] Google, "Google image search," 2025, accessed: 2025-01-12. [Online]. Available: https://images.google.com 4

[56] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2023. 4

[57] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks," in *ICLR*, 2024. 4, 5

[58] F. Barbero, A. Vitvitskyi, C. Perivolaropoulos, R. Pascanu, and P. Veličković, "Round and round we go! what makes rotary positional encodings useful?" in *ICLR*, 2025. 4, 5

[59] L. Li, Y. Liu, L. Yao, P. Zhang, C. An, L. Wang, X. Sun, L. Kong, and Q. Liu, "Temporal reasoning transfer from text to video," in *ICLR*, 2025. 4

[60] C. Han, Q. Wang, H. Peng, W. Xiong, Y. Chen, H. Ji, and S. Wang, "LM-Infinite: Zero-shot extreme length generalization for large language models," in *ACL*, 2024. 5

[61] X. Men, M. Xu, B. Wang, Q. Zhang, H. Lin, X. Han, and W. Chen, "Base of rope bounds context length," *arXiv preprint arXiv:2405.14591*, 2024. 5

[62] J. Zhou, Y. Shu, B. Zhao, B. Wu, S. Xiao, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu, "MLVU: A comprehensive benchmark for multi-task long video understanding," in *CVPR*, 2025. 8, 9

[63] Z. Zhao, H. Lu, Y. Huo, Y. Du, T. Yue, L. Guo, B. Wang, W. Chen, and J. Liu, "Needle in a video haystack: A scalable synthetic framework for benchmarking video mllms," in *ICLR*, 2025. 8

[64] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024. 8, 11

[65] H. Xue, T. Hang, Y. Zeng, Y. Sun, B. Liu, H. Yang, J. Fu, and B. Guo, "Advancing high-resolution video-language representation with large-scale video transcriptions," in *CVPR*, 2022. 8

[66] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 8

[67] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015. 8

[68] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *ACM SIGOPS*, 2023. 9

[69] H. Wu, D. Li, B. Chen, and J. Li, "LongVideoBench: A benchmark for long-context interleaved video-language understanding," 2024. 9

[70] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," in *CVPR*, 2025. 9

[71] Y. Wang, Y. Wang, D. Zhao, C. Xie, and Z. Zheng, "Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models," *arxiv*, 2024. 9

[72] J. Lin, Z. Fang, C. Chen, Z. Wan, F. Luo, P. Li, Y. Liu, and M. Sun, "Streamingbench: Assessing the gap for mllms to achieve streaming video understanding," *arXiv preprint arXiv:2411.03628*, 2024. 9

[73] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024. 10

[74] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "LLaVA-OneVision: Easy visual task transfer," *TMLR*, 2024. 11

[75] X. Shen, Y. Xiong, C. Zhao, L. Wu, J. Chen, C. Zhu, Z. Liu, F. Xiao, B. Varadarajan, F. Bordes *et al.*, "Longvu: Spatiotemporal adaptive compression for long video-language understanding," *arXiv preprint arXiv:2410.17434*, 2024. 11

[76] O. Zohar, X. Wang, Y. Dubois, N. Mehta, T. Xiao, P. Hansen-Estruch, L. Yu, X. Wang, F. Juefei-Xu, N. Zhang, S. Yeung-Levy, and X. Xia, "Apollo: An exploration of video understanding in large multimodal models," *arXiv preprint arXiv:2412.10360*, 2024. 11

[77] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet *et al.*, "Pixtral 12b," *arXiv preprint arXiv:2410.07073*, 2024. 13