VRoPE: Rotary Position Embedding for Video Large Language Models

Anonymous ACL submission

Abstract

Rotary Position Embedding (RoPE) has shown strong performance in text-based Large Language Models (LLMs), but extending it to video remains a challenge due to the intricate spatiotemporal structure of video frames. Existing adaptations, such as RoPE-3D, attempt to encode spatial and temporal dimensions separately but suffer from two major limitations: positional bias in attention distribution and disruptions in video-text transitions. To overcome these issues, we propose Video Rotary Position Embedding (VRoPE), a novel positional encoding method tailored for Video-LLMs. Our approach restructures positional indices to preserve spatial coherence and ensure a smooth transition between video and text tokens. Additionally, we introduce a more balanced encoding strategy that mitigates attention biases, ensuring a more uniform distribution of spatial focus. Extensive experiments on Vicuna and Qwen2 across different model scales demonstrate that VRoPE consistently outperforms previous RoPE variants, achieving significant improvements in video understanding, temporal reasoning, and retrieval tasks.

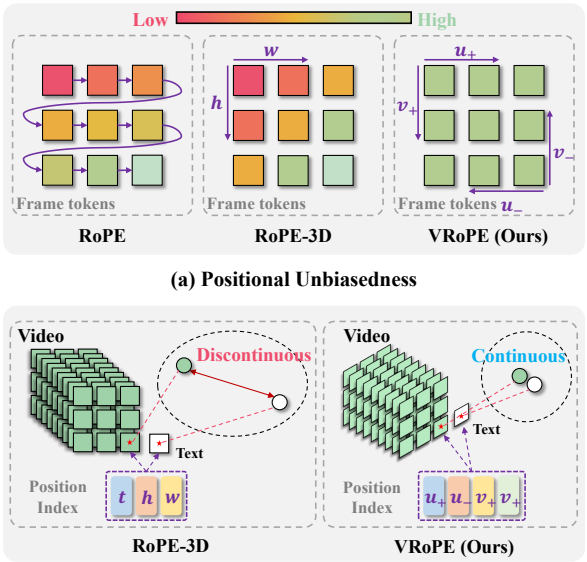

Figure 1: Comparison of RoPE, RoPE-3D, and our VRoPE in video positional encoding. (a) Positional Unbiasedness: RoPE and RoPE-3D exhibit spatial biased attention, particularly towards later tokens or specific frame regions, while VRoPE ensures more uniform attention. (b) Seamless Video-Text Transition: RoPE-3D causes a discontinuity when transitioning from video to text tokens, which VRoPE smooths for better cross-modal dependency modeling.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved remarkable progress (Touvron et al., 2023; Bai et al., 2023). Building on the success of LLMs, Video Large Language Models (Video-LLMs) (Maaz et al., 2023; Li et al., 2024d; Jin et al., 2024) have emerged as a powerful paradigm for video-language understanding. These models typically integrate LLMs with pre-trained vision encoders, enabling the joint modeling of video and textual information. However, a fundamental challenge in Video-LLMs lies in effectively modeling positional relationships within video sequences.

In LLMs, positional encoding plays a crucial role in enabling models to capture order-dependent patterns, as self-attention mechanisms themselves are inherently permutation-invariant. Among various positional encoding schemes, Rotary Position Embedding (RoPE) (Su et al., 2024) has gained widespread adoption due to its ability to encode relative position relationships. RoPE enables efficient long-range dependencies, making it highly effective in text-based models. However, when applied directly to video data, vanilla RoPE—where video tokens are treated as a simple sequence akin to text—fails to account for the complex spatiotemporal structure inherent in video frames, leading to suboptimal representations. Despite its critical role, an effective video-specific positional encoding

strategy remains an open challenge.

To optimally encode positional relationships in Video-LLMs, we identify three key properties that an ideal video positional encoding method should satisfy:

(1) Spatiotemporal Structure Modeling. Unlike text, where positional relationships are strictly one-dimensional, video frames exhibit both spatial (width, height) and temporal (frame index) dimensions. An effective encoding must reflect this inherent structure to facilitate accurate modeling of spatiotemporal dependencies. Recent approaches (Wang et al., 2024) have attempted to extend RoPE for video structure. These methods, which we refer to as RoPE-3D, split the feature channels into 3 parts to separately encode frame, width, and height positions.

(2) Positional Unbiasedness. A critical yet often overlooked aspect of positional encoding is its impact on attention distribution. As illustrated in Figure 1 (a), RoPE, by design, applies a long-term decay over increasing positional indices, inadvertently introducing a bias that amplifies attention toward later tokens. This issue persists in RoPE-3D, where spatial positions within video frames are unevenly weighted, causing attention to be disproportionately focused on certain areas—typically the bottom-right regions of frames—while suppressing others, which is shown in Figure 1 (a). Such biases distort spatial contextual modeling and can lead to suboptimal video comprehension. An effective video positional encoding should mitigate these biases to ensure uniform attention across the entire frame.

(3) Seamless Video-Text Transition. For effective multimodal understanding, an ideal positional encoding should ensure a seamless transition between video and text tokens. However, as demonstrated in Figure 1 (b), RoPE-3D introduces a discontinuity when transitioning from video to text tokens, as the positional indices of text tokens are arbitrarily offset by the maximum position index of the video sequence (determined by the largest of frame count, width, and height, which often vary significantly). This artificial "jump" in the positional encoding space disrupts the smooth flow of information between modalities, making it harder for the model to establish meaningful cross-modal dependencies.

Based on the above principles, we propose Video Rotary Position Embedding (VRoPE), a novel positional encoding method specifically designed for Video-LLMs. Our approach consists of two key components to satisfy those principles. *(1) Cross-Modal Continuity Rotation*: We introduce a spatial rotation transformation that redistributes RoPE-3D's positional indices while preserving local spatial structures. This transformation not only maintains spatial coherence within video frames but also ensures a smooth transition between video and text tokens, mitigating discontinuities in the positional encoding space. *(2) Symmetric Bias Mitigation*: To counteract the attention bias present in RoPE-based encodings, we design a symmetric positional representation that encodes each spatial coordinate with both positive and negative components. By distributing attention more uniformly across spatial locations, this method prevents positional distortions and improves overall video understanding.

Overall, VRoPE effectively enhances Video-LLMs by preserving spatiotemporal structure, mitigating attention bias, and ensuring smooth video-text transitions. We conduct extensive experiments on Vicuna (Chiang et al., 2023) and Qwen2 (Yang et al., 2024) across different model scales (1.5B and 7B parameters). Our results demonstrate significant performance improvements over RoPE, RoPE-3D, and other positional encoding variants on multiple video benchmarks, covering general video understanding, temporal reasoning, long video comprehension, and video retrieval. These findings establish VRoPE as a robust and efficient positional encoding method tailored for Video-LLMs.

2 Related Work

2.1 Video Large Language Models

Recent advancements in Video-LLMs have significantly enhanced video processing by integrating multiple modalities and employing instruction fine-tuning. Notable innovations include Video-ChatGPT (Maaz et al., 2023), which introduced video instruction tuning for text generation, and VideoChat (Li et al., 2023) and VideoChat2 (Li et al., 2024b), which improved modality alignment via cross-attention and multi-stage bootstrapping etc. Other models, such as Chat-UniVi (Jin et al., 2024) and LLaMA-VID (Li et al., 2024d), focus on efficient video representations through techniques like token compression and dual-token methods that separate context and content. Additionally, PLLaVA (Xu et al., 2024) explores the use of image-pretrained LLaVA models for video tasks, utilizing simple spatial pooling techniques.

2.2 Multimodal Position Embedding

Most Video-LLMs inherit the default design from LLMs by using Rotary Position Embedding (RoPE) (Su et al., 2024) for positional encoding. RoPE encodes relative distance information as absolute position embeddings, offering key advantages like no additional training parameters and improved performance in various tasks (Su et al., 2024). It is widely used in modern LLMs due to its ability to extrapolate context length, extending a model's window size without the need for expensive retraining. However, RoPE's 1D design, effective for text, overlooks the spatiotemporal structure of video data, limiting its suitability for Video-LLMs. To address this, several approaches have adapted RoPE for video. For instance, RoPE-2D (Agrawal et al., 2024; Wang et al., 2024) extends the encoding to capture spatial relationships in video frames, while RoPE-3D (Wang et al., 2024) divides the channel dimension into three groups to better represent the spatiotemporal dimensions.

However, these approaches still face issues like Positional Attention Bias and Cross-Modal Positional Discontinuity, which are discussed in Section 3. Our VRoPE method addresses these limitations, offering more accurate and robust positional encoding tailored for Video-LLMs.

3 Motivation

3.1 Preliminary: Rotary Position Embedding

Rotary Positional Embedding (RoPE) is a widely adopted method in LLMs that encodes absolute positional information while preserving relative positional relationships. This property makes RoPE particularly effective for self-attention mechanisms, as it allows models to capture the relative distance between tokens in a computationally efficient manner. Given a token embedding $\mathbf{x}$ at position index m, RoPE applies a complex-valued rotation operation, formulated as:

$$\text{RoPE}(\mathbf{x}, m) = \mathbf{x}e^{\mathrm{i}m\boldsymbol{\theta}} \tag{1}$$

where i is the imaginary unit, and the frequency encoding vector $\boldsymbol{\theta}$ is defined as:

$$\boldsymbol{\theta}_j = \text{base}^{\frac{-2j}{d}} \tag{2}$$

where base is a hyperparameter, d is the feature dimension, and $j = [0, 1, ..., d/2 - 1]$ denotes the index of each feature channel.

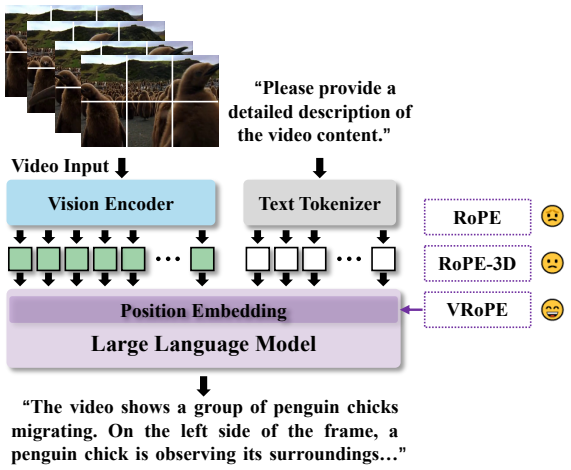

Figure 2: Overall architecture of a typical Video-LLM. The input videos are first encoded by vision encoders, followed by a connector layer. The resulting video features are then flattened and concatenated with text embedding tokens to form a multimodal input sequence for the LLM. In this work, our improvements primarily target the positional embedding component of the LLM to enhance its video understanding capability.

In the self-attention mechanism, RoPE transforms absolute position embeddings into relative ones. The attention score between m-th query $\mathbf{q}_m$ and n-th key $\mathbf{k}_n$ is

$$\mathbf{A}_{(m,n)} = \Re\left[\mathbf{q}_m \cdot \mathbf{k}_n^* e^{\mathrm{i}(m-n)\boldsymbol{\theta}}\right] \tag{3}$$

where $\Re[\cdot]$ denotes the real part, and * represents the complex conjugate.

While RoPE excels in sequential text modeling, its direct application to video-text interleaved sequences poses challenges due to the complex spatiotemporal relationships inherent in video frames.

3.2 RoPE for Video-LLMs

In Video-LLMs, video frames are typically processed by vision encoders (e.g., ViTs (Alexey, 2020) or CNNs (He et al., 2016)) and transformed into a sequence of visual tokens. These visual tokens are then concatenated with text tokens and fed into an LLM backbone. The overall architecture of typical Video-LLMs is denoted in Figure 2.

In most existing approaches, video tokens are treated as a simple 1D sequence, with position indices assigned in an increasing order, similar to text. However, this naive approach, referred to as RoPE, overlooks the inherent spatiotemporal structure of video data. Flattening video frames this way disrupts spatiotemporal structure and leads to inefficient token usage. Unlike text, video tokens

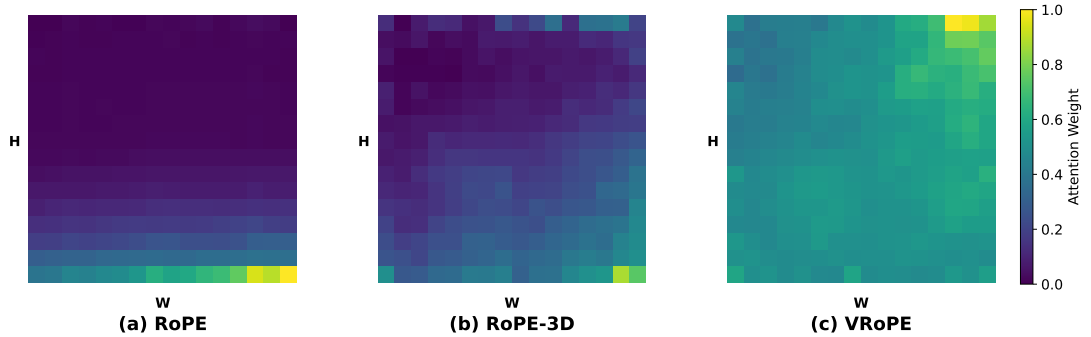

Figure 3: Attention weight visualization of RoPE, RoPE-3D, and VRoPE. We compute average text-to-video frame attention weights on VideoMME (Fu et al., 2024) benchmark (lighter color indicates higher attention). (a) RoPE exhibits row-wise attention decay within frames. (b) RoPE-3D shows a similar decay from the bottom-right to the top-left, introducing positional bias that skews attention toward spatially closer frame tokens. (c) VRoPE mitigates this bias, leading to a more balanced attention distribution.

carry less dense semantic information, and their excessive sequence length can weaken contextual dependencies, making long-range understanding harder.

3.3 RoPE-3D for Video-LLMs

Recent approaches, such as M-RoPE in Qwen2-VL(Wang et al., 2024), have proposed RoPE-3D as an extension of RoPE for video structure preserving. RoPE-3D intuitively partitions the feature dimensions to separately encode spatial (width, height) and temporal (frame index) positions. Given a video token with coordinates (w, h, t), RoPE-3D computes:

$$\text{RoPE-3D}_j(\mathbf{x}, w, h, t) = \begin{cases} \text{RoPE}_j(\mathbf{x}, w), j \in D_w \\ \text{RoPE}_j(\mathbf{x}, h), j \in D_h \\ \text{RoPE}_j(\mathbf{x}, t), j \in D_t \end{cases} \quad (4)$$

where where D_w, D_h, D_t denote feature partitions assigned to width, height, and temporal axes, respectively. For text tokens, the encoding remains consistent with the original RoPE by setting $w = h = t = m$, ensuring that:

$$\text{RoPE-3D}_j(\mathbf{x}, m, m, m) \equiv \text{RoPE}_j(\mathbf{x}, m) \quad (5)$$

This design explicitly models spatial and temporal positions while preserving text token behavior. However, RoPE-3D still exhibits two key limitations, which we elaborate on below.

3.4 Problem Analysis

While RoPE-3D introduces a promising design by partitioning the feature dimensions to encode spatial (width, height) and temporal (frame index) positions separately, two critical issues persist when handling video–text data.

(1) Positional Attention Bias. As is demonstrated in Figure 3 (a), RoPE naturally applies a long-term decay over increasing positional indices, which amplify attention toward later positions. Unfortunately, we find that this issue persists in RoPE-3D, where the decay leads to *an uneven distribution of focus across spatial positions in video frames*. As is shown in Figure 3 (b), notably, tokens in the bottom-right of each frame receive disproportionately higher attention, while those in the top-left are increasingly suppressed. This imbalance can distort spatial contextual modeling by weakening dependencies on earlier tokens, which in turn affects the model's understanding of the video.

(2) Cross-Modal Positional Discontinuity. RoPE-3D introduces separate positional encodings for spatial (width, height) and temporal (frame index) dimensions. However, when video tokens are concatenated with subsequent text tokens, their positional indices do not follow a smooth transition. Instead, text tokens inherit positional indices that are arbitrarily offset by the maximum position value across spatial (W, H) and temporal dimensions T, i.e., $\max(W, H, T)$. This results in an artificial "jump" in the positional encoding space when transitioning from video to text tokens. *The discontinuity creates an abrupt and non-uniform gap between the final video token and the subsequent text token.* The magnitude of this gap depends on video dimensions rather than being a fixed offset, making it inconsistent across different video-text samples. Such a discrepancy can degrade the model's ability to establish seamless contextual dependencies across modalities. This issue is particularly problematic

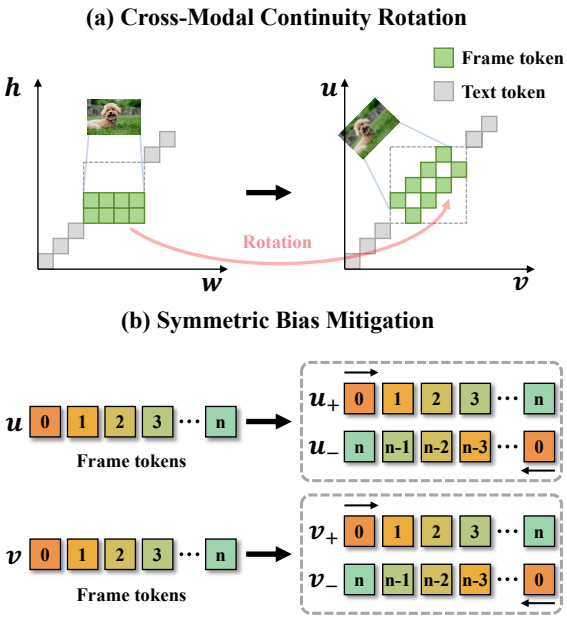

Figure 4: Method illustration of VRoPE.

in long videos, as the increasing frame count T exacerbates the positional gap.

4 Method: VRoPE

In this section, we introduce Video Rotary Position Embedding (VRoPE), a novel positional encoding method tailored for Video-LLMs. Our approach addresses the inherent limitations of RoPE-3D, including spatiotemporal attention bias and cross-modal positional discontinuity, by leveraging a combination of Cross-Modal Continuity Rotation and Symmetric Bias Mitigation. The overall framework of VRoPE is illustrated in Figure 4.

4.1 Cross-Modal Continuity Rotation

To mitigate the cross-modal positional discontinuity introduced when concatenating video and text tokens, we propose a spatial rotation transformation that redistributes positional indices while preserving the inherent spatial structure. Specifically, we transform the width and height coordinates into a rotated space as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} w + h \\ w - h + H - 1 \end{pmatrix} \tag{6}$$

where the term $H - 1$ ensures that v remains non-negative. As is shown in Figure 4 (a), this transformation provides two key benefits: (1) it preserves the local spatial relationships of tokens within the frame, and (2) it ensures a smooth transition from video to text tokens by aligning their positional indices along diagonal axes in the (u, v) space. Consequently, the model naturally reduces abrupt jumps in positional encoding, maintaining consistent contextual dependencies across modalities.

4.2 Symmetric Bias Mitigation

While Cross-Modal Continuity Rotation alleviates discontinuity, the issue of spatial attention bias persists. To address this, we introduce Symmetric Bias Mitigation, which ensures a more uniform attention distribution across spatial dimensions.

As is shown in Figure 4 (b), instead of using scalar positional values, we represent each spatial position u, v as a symmetric vector to capture bidirectional positional dependencies, i.e., $u = (u_+, u_-)$ and $v = (v_+, v_-)$. To achieve this, we evenly partition the feature dimensions into four equal parts, assigning each part to a different positional encoding term. For a video of size (W, H, T) with an initial position index p_{start}, its first frame's position indices are computed as follows:

$$\begin{pmatrix} u_+ \\ u_- \\ v_+ \\ v_- \end{pmatrix} = \begin{pmatrix} u \\ -u \\ v \\ -v \end{pmatrix} + \begin{pmatrix} 0 \\ bias \\ 0 \\ bias \end{pmatrix} + p_{start} \tag{7}$$

where $bias = H + W - 2$ ensures non-negative values. After processing the first frame, the position index p_{start} is updated to $p_{start} + H + W - 1$, and subsequent frames follow the same encoding scheme.

This encoding strategy ensures that the decay pattern is counterbalanced across positive and negative position pairs. Consequently, the distance between frame tokens at different spatial positions—and between frame tokens and subsequent text tokens—remains more consistent. As is shown in Figure 3 (c), this alleviates the skewed focus on specific regions and improves overall video understanding.

Finally, our VRoPE computes the positional encoding as:

$$\begin{aligned} &\text{VRoPE}_j(\mathbf{x}, u_+, u_-, v_+, v_-) \\ &= \begin{cases} \text{RoPE}_j(\mathbf{x}, u_+)_{j=4k} \\ \text{RoPE}_j(\mathbf{x}, u_-)_{j=4k+1} \\ \text{RoPE}_j(\mathbf{x}, v_+)_{j=4k+2} \\ \text{RoPE}_j(\mathbf{x}, v_-)_{j=4k+3} \end{cases} \end{aligned} \tag{8}$$

where $k \in \{0, 1, 2, ...\}$. For text tokens, we retain the original RoPE encoding structure (Eq. 5) to ensure compatibility with LLMs.

4.3 Discussion

In addition to preserving the video structure, ensuring a continuous transition between video and text tokens, and mitigating spatial attention bias, VRoPE also exhibits the following characteristics.

Implicit Temporal Modeling. While VRoPE does not directly encode frame indices, it implicitly captures temporal cues by arranging frame tokens in an ascending sequence and applying symmetric positional transformations. The temporal correlations among the frames are implicitly captured through the opposite direction design, ensuring the model differentiate temporal progressions.

Dimensional Adaptability. A key advantage of VRoPE is its ability to degenerate into lower-dimensional embeddings without altering its fundamental structure. Unlike methods that assign separate feature channels for each coordinate, VRoPE employs linear combinations of the original coordinates, allowing any dimension set to 1 to seamlessly adapt into lower-dimension form. For instance, when $H = 1$, the encoded positions simplify to $(w, -w, w, -w)$, effectively reducing to a 1D form—unlike previous methods that rely on separate encodings, such as $(w, 0)$. This property is particularly beneficial for adapting pre-trained model's positional encodings from images (2D) or videos (3D) to data of varying dimensions without disrupting the original encoding scheme. Consequently, models can transfer more effectively across modalities while preserving consistent positional behavior.

5 Experiments

5.1 Experimental Setup

Implementation Details. We apply our proposed VRoPE to Video-LLM architectures with three widely used LLM backbones: Vicuna-7B, Qwen2-1.5B, and Qwen2-7B, the resulting models are denoted as Video-Vicuna-7B, Video-Qwen2-1.5B, and Video-Qwen2-7B. For the vision encoder, we leverage Eva-CLIP (Sun et al., 2023), and connect the Vision Encoder to the LLM using a Multi-Layer Perceptron (MLP) (Tolstikhin et al., 2021). For video input, the frames are tokenized using a 2×2 pooling kernel with a stride of 2, i.e., each frame has 64 tokens as input. Training follows a two-stage paradigm: in the pre-training stage, only the MLP connector is trained, while in the instruction-tuning stage, both the MLP and LLM backbones are fine-tuned, with the Vision Encoder frozen throughout. During pre-training, we use a batch size of 256 and a learning rate of 1e-3, while for instruction-tuning, we reduce the batch size to 128 and set the learning rate to 2e-5. A warm-up ratio of 0.03 is used, followed by cosine learning rate decay after the linear warm-up phase. The training was conducted on 8 Nvidia A800 GPUs.

Training Data. For Vicuna-7B, we pre-train the model on the LLaVA-558K dataset (Liu et al., 2024a) with WebVid samples (Bain et al., 2021) and fine-tune it on the LLaVA-mix665K (Liu et al., 2024a) dataset augmented with VideoChatGPT data (Maaz et al., 2023). For the Qwen2 LLM series, we pre-train the models on a randomly sampled 1M caption dataset, which includes LLaVA-558K, WebVid, DenseFusion-1M (Li et al., 2024c), VALOR (Liu et al., 2024b), and CC3M (Changpinyo et al., 2021). The models are then fine-tuned on a combination of LLaVA-mix665K, VideoChatGPT, OneVision (Li et al., 2024a), and LLaVA-Video-178K (Zhang et al., 2024). We use a 224×224 resolution for both image and video inputs.

Evaluation Benchmarks. We evaluated VRoPE across diverse video benchmarks, covering *general video understanding* (Video-MME (Fu et al., 2024)), *video temporal understanding* (MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024c)), *long video understanding* (MLVU (Zhou et al., 2024), LongVideoBench (Wu et al., 2025), EgoSchema (Mangalam et al., 2024)), and *long video retrieval* (Video-NIAH (Zhao et al., 2024)) to validate its effectiveness.

5.2 Main Results

We evaluate the performance of RoPE, RoPE-3D, and our proposed VRoPE across six video understanding benchmarks, with the number of input frames fixed to 16. As shown in Table 1, VRoPE consistently outperforms both RoPE and RoPE-3D, achieving the highest average scores across all tasks and backbones.

For instance, in the Video-Vicuna-7B row, VRoPE achieves an average score of 44.48, surpassing RoPE by 1.13 points. Similarly, when evaluated with larger backbones like Qwen2-1.5B and Qwen2-7B, VRoPE demonstrates consistent improvements across all benchmarks. Notably, it outperforms RoPE and RoPE-3D by significant margins on tasks such as Video-MME (a 3.4-point

Table 1: Performance comparison of RoPE variants on video benchmarks across different LLM backbones. Results across tasks, including general video understanding (Video-MME), video temporal understanding (MVBench, TempCompass), and long video understanding (MLVU, LongVideoBench, EgoSchema).

Method	Video-MME (w/o subs)	MLVU @M-Avg	MVBench	LongVideoBench @Val	TempCompass @Multi-Choice	EgoSchema @Test	Avg.
Video-Vicuna-7B							
w/ RoPE	38.5	47.00	43.90	41.66	53.10	35.92	43.35
w/ RoPE-3D	38.0 (↓0.5)	46.30 (↓0.7)	44.55 (↑0.65)	40.16 (↓1.5)	54.94 (↑1.84)	39.79 (↑3.87)	43.96 (↑0.61)
w/ VRoPE	38.9 (↑0.4)	47.37 (↑0.37)	45.18 (↑1.28)	40.69 (↓0.97)	54.05 (↑0.95)	40.71 (↑4.79)	44.48 (↑1.13)
Video-Qwen2-1.5B							
w/ RoPE	39.0	51.15	51.15	46.63	56.96	48.50	48.90
w/ RoPE-3D	39.3 (↑0.3)	51.19 (↑0.04)	50.45 (↓0.70)	48.01 (↑1.38)	57.97 (↑1.01)	49.00 (↑0.50)	49.32 (↑0.42)
w/ VRoPE	42.4 (↑3.4)	51.76 (↑0.61)	50.78 (↓0.37)	47.79 (↑1.16)	57.15 (↑0.19)	49.90 (↑1.40)	49.96 (↑1.06)
Video-Qwen2-7B							
w/ RoPE	50.1	54.87	54.33	49.36	63.73	57.14	54.92
w/ RoPE-3D	49.5 (↓0.6)	56.06 (↑1.19)	54.23 (↓0.1)	49.55 (↑0.19)	64.49 (↑0.76)	58.80 (↑1.66)	55.44 (↑0.52)
w/ VRoPE	50.6 (↑0.5)	57.81 (↑2.94)	54.70 (↑0.37)	50.48 (↑1.12)	65.88 (↑2.15)	58.60 (↑1.46)	56.35 (↑1.43)

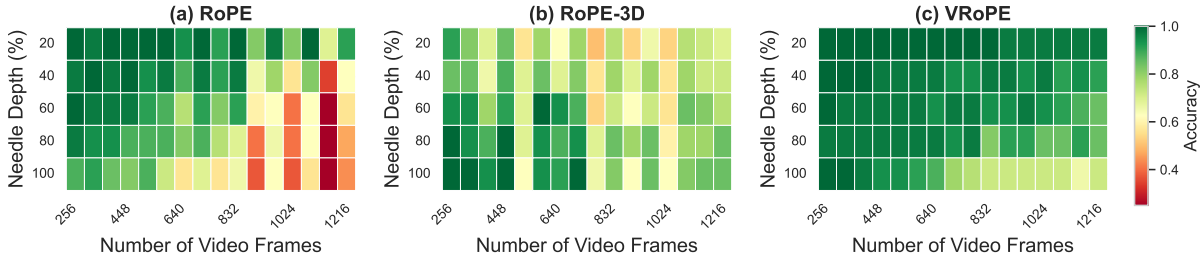

Figure 5: Visualization of long video retrieval results on Video-NIAH (Zhao et al., 2024). Our VRoPE consistently achieves high accuracy across varying background lengths and needle depths, showing strong retrieval capability in long videos.

increase for Qwen2-1.5B) and MLVU (a 2.94-point increase for Qwen2-7B).

These results highlight the superior adaptability of VRoPE across different LLM types and parameter sizes. Importantly, VRoPE introduces no new learnable parameters and does not increase computational complexity, making it a cost-free performance enhancement for Video-LLMs.

5.3 Results on Long Video Retrieval

We compare our method with RoPE (Su et al., 2024) and RoPE-3D (Wang et al., 2024) on the long video retrieval task to evaluate the model's generalization ability with longer video inputs. Following the setup in Video-NIAH (Zhao et al., 2024), we conduct Video Needle-In-A-Haystack (V-NIAH) experiments, where a target "needle" frame is inserted into a sequence of background frames, with the total frame count varying between 256 and 1216.

As shown in Figure 5, the retrieval accuracy of RoPE drops significantly when the number of input frames exceeds 832. While RoPE-3D demonstrates better frame extrapolation capabilities compared to RoPE, VRoPE outperforms both approaches by a considerable margin. The quantitative results, presented in Table 3, further evidence this finding. Specifically, VRoPE achieves an accuracy that is 32.19 points higher than RoPE and 14.22 points higher than RoPE-3D when the number of input frames increases to 1024-1216. Notably, these results are obtained even though the input frame count in this range is dozens of times greater than the maximum number seen during training. This demonstrates the exceptional extrapolation ability of VRoPE.

5.4 Ablation Studies

Comparison of RoPE Variants. We conduct experiments to assess the impact of three key properties: Spatiotemporal Structure Modeling (S.S.M.), Positional Unbiasedness (P.U.), and Seamless Video-Text Transition (S.V.T.), as discussed in Section 1. The results, summarized in Table 2, highlight the importance of these properties.

We first compare RoPE-2D (Agrawal et al., 2024) and RoPE-3D (Wang et al., 2024) with the baseline RoPE (Su et al., 2024) method. *RoPE-2D*

Table 2: We assess various RoPE designs to validate the necessity of the three desired properties: Spatiotemporal Structure Modeling (S.S.M), Positional Unbiasedness (P.U.), and Seamless Video-Text Transition (S.V.T.). The results indicate that the model attains optimal performance when all properties are fully incorporated.

Method	S.S.M.	P.U.	S.V.T.	Video-MME	EgoSchema	LongVideoBench	Avg.
RoPE	✘	✘	✔	39.0	48.50	46.63	44.71
RoPE-2D	✔	✘	✘	43.2 (↑4.2)	47.60 (↓0.90)	46.33 (↓0.30)	45.71 (↑1.00)
RoPE-3D	✔	✘	✘	39.3 (↑0.3)	49.00 (↑0.50)	48.01 (↑1.38)	45.44 (↑0.73)
RoPE-Share	✘	✔	✔	39.7 (↑0.7)	48.66 (↑0.16)	45.10 (↓1.53)	44.49 (↓0.22)
RoPE-Compact	✔	✘	✔	38.1 (↓0.9)	50.77 (↑2.27)	45.96 (↓0.67)	44.94 (↑0.23)
VRoPE	✔	✔	✔	42.4 (↑3.4)	49.90 (↑1.40)	47.79 (↑1.16)	46.70 (↑1.99)

Table 3: Average retrieval accuracy across different input frame length intervals on Video-NIAH (Zhao et al., 2024). Compared to RoPE, the performance advantage of VRoPE becomes more pronounced at longer video lengths.

Method	256-512	512-768	768-1024	1024-1216
RoPE	94.84	87.03	73.28	54.84
RoPE-3D	88.90	80.94	69.69	72.81
VRoPE	**98.28**	**95.16**	**90.31**	**87.03**

Table 4: Ablation study on VRoPE components. We evaluate the impact of Cross-Modal Continuity Rotation (Continuity) and Symmetric Bias Mitigation (Symmetric). The model achieves the best performance when both components are applied together.

Continuity	Symmetric	Video-MME	LongVideoBench
✘	✘	39.0	46.63
✔	✘	42.3	46.30
✘	✔	41.3	47.27
✔	✔	**42.4**	**47.79**

encodes only the spatial coordinates (w, h) of each frame, arranged sequentially. While it resolves the cross-modal positional discontinuity of RoPE-3D, it still suffers from positional bias. Both RoPE-2D and RoPE-3D show improvements over RoPE, demonstrating the benefits of spatiotemporal structure modeling.

Next, we evaluate two additional variants, *RoPE-Share* and *RoPE-Compact*, to further ablate the impact of S.S.M. and P.U. RoPE-Share uses identical positional embeddings within each frame, arranged sequentially. While it resolves positional bias and ensures continuity, it neglects the spatial structure of the frames, leading to a performance drop compared to RoPE. RoPE-Compact is a extention of RoPE-3D that addresses positional discontinuity by encoding subsequent text tokens with $(W + 1, H + 1, T + 1)$, but it deviates from text compatibility requirements, which slightly limits its performance. In contrast, our proposed method (VRoPE) incorporates all three properties, achieving a 1.99-point improvement over the RoPE baseline, surpassing all other variants.

Ablation on VRoPE Components. We conduct ablation experiments to evaluate the individual contributions of the Cross-Modal Continuity Rotation and Symmetric Bias Mitigation components. The results, presented in Table 4, reveal that when applied separately, each method produces mixed effects. Specifically, Cross-Modal Continuity Rotation improves performance on Video-MME, indicating its effectiveness in enhancing smooth translation for general video understanding. Symmetric Bias Mitigation shows a significant gain on LongVideoBench, indicating its effectiveness in reducing bias in long video tasks. When combined in VRoPE, the two components work synergistically, resulting in more consistent performance.

6 Conclusion

In conclusion, we propose VRoPE, a dedicated positional encoding strategy for Video-LLMs that balances spatiotemporal structure, mitigates attention bias, and ensures a smooth transition between video and text tokens. Extensive experiments on different model scales validate its superior performance in video understanding, temporal reasoning, and retrieval tasks. We believe VRoPE can serve as a useful building block for future Video-LLMs, enabling better video-language understanding.

7 Limitations

While VRoPE demonstrates strong performance, there are some limitations. Due to computational resource constraints, our experiments were limited to models with 1.5B and 7B parameters.

Larger-scale models could potentially yield further performance gains. Additionally, although VRoPE is adaptable across different dimensions, its extension to other modalities (e.g., audio, 3D point clouds, Electroencephalography (EEG)) and higher-dimensional data (e.g., 4D spatiotemporal or medical imaging data) remains an area for future research and validation.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. 2024c. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024d. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024b. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024c. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2025. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

An Yang, Baosong Yang, and Binyuan Hui et al. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. 2024. Needle in a video haystack: A scalable synthetic framework for benchmarking video mllms. *arXiv preprint arXiv:2406.09367*.

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