RIFLEx: A Free Lunch for Length Extrapolation in Video Diffusion Transformers

Min Zhao[1] **Guande He**[2] **Yixiao Chen**[1 3] **Hongzhou Zhu**[1 3] **Chongxuan Li**[4 5 6] **Jun Zhu**[1 3 7]

Abstract

Recent advancements in video generation have enabled models to synthesize high-quality, minute-long videos. However, generating even longer videos with temporal coherence remains a major challenge and existing length extrapolation methods lead to *temporal repetition* or *motion deceleration*. In this work, we systematically analyze the role of frequency components in positional embeddings and identify an *intrinsic frequency* that primarily governs extrapolation behavior. Based on this insight, we propose **RIFLEx**, a *minimal yet effective* approach that reduces the intrinsic frequency to suppress repetition while preserving motion consistency, without requiring any additional modifications. RIFLEx offers a true *free lunch*—achieving high-quality $2\times$ extrapolation on state-of-the-art video diffusion transformers in a completely *training-free* manner. Moreover, it enhances quality and enables $3\times$ extrapolation by minimal fine-tuning without long videos. Project page and codes: https://riflex-video.github.io/.

1. Introduction

Recent advances in video generation (Brooks et al., 2024; Bao et al., 2024; Yang et al., 2024; Kong et al., 2024; Zhao et al., 2023; Team, 2024b; Zhao et al., 2024) enable models to synthesize minute-long video sequences with high fidelity and coherence. A key factor behind this progress is the emergence of diffusion transformers (Peebles & Xie,

2023; Bao et al., 2023), which combines the scalability of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) with the expressive power of transformers (Vaswani, 2017).

Despite these advancements, generating longer videos with high-quality and temporal coherence remains a fundamental challenge. Due to computational constraints and the sheer scale of training data, existing models are typically trained with a fixed maximum sequence length, limiting their ability to extend content. Consequently, there is increasing interest in length extrapolation techniques that enable models to *generate new and temporally coherent content that evolves smoothly over time without training on longer videos*.

However, existing extrapolation strategies (Chen et al., 2023b; bloc97, 2023; Peng et al., 2023; Lu et al., 2024b; Zhuo et al., 2024), originally developed for text and image generation, fail when applied to video length extrapolation. Our experiments show that these methods exhibit distinct failure patterns: *temporal repetition* and *slow motion*. These limitations suggest a fundamental gap in understanding how positional encodings influence video extrapolation.

To address this, we isolate individual frequency components by zeroing out others and fine-tuning the target video model. We find that high frequencies capture short-term dependencies and induce temporal repetition, while low frequencies encode long-term dependencies but lead to motion deceleration. Surprisingly, we identify a consistent *intrinsic frequency component* across different videos from the same model, which primarily dictates repetition patterns among all components during extrapolation.

Building on this insight, we propose **R**educing **I**ntrinsic **F**requency for **L**ength **Ex**trapolation (**RIFLEx**), a *minimal yet effective* solution that lowers the intrinsic frequency to ensure it remains within a single cycle after extrapolation. Without any other modification, it suppresses temporal repetition while preserving motion consistency. As a byproduct, RIFLEx provides a principled explanation for the failure modes of existing approaches and offers insights that naturally extend to spatial extrapolation of images.

Extensive experiments on state-of-the-art video diffusion transformers, including CogVideoX-5B (Yang et al., 2024)

[1]Dept. of Comp. Sci. & Tech., BNRist Center, THU-Bosch ML Center, Tsinghua University. [2]The University of Texas at Austin. [3]ShengShu. [4]Gaoling School of Artificial Intelligence Renmin University of China Beijing, China. [5]Beijing Key Laboratory of Research on Large Models and Intelligent Governance. [6]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE. [7]Pazhou Laboratory (Huangpu). Correspondence to: Chongxuan Li <chongxuanli@ruc.edu.cn>, Jun Zhu <dcszj@tsinghua.edu.cn>.

(a) $2\times$ temporal extrapolation from 129 to 261 frames.

(b) From 480×720 to 960×1440. (c) $2\times$ temporal and spatial extrapolation from $480 \times 720 \times 49$ to $960 \times 1440 \times 97$.

Figure 1. **Visualization of RIFLEx for $2\times$ temporal, spatial, and combined extrapolation.** Our base models are (a) Hunyuan-Video (Kong et al., 2024) and (b-c) CogVideoX-5B (Yang et al., 2024), where we do not use any videos longer or larger than those used for pre-training. More demos and all the prompts used in this paper are listed in Appendix B and the supplementary materials, respectively.

and HunyuanVideo (Kong et al., 2024), validate the effectiveness of RIFLEx (see Fig. 1). Remarkably, for $2\times$ extrapolation, RIFLEx enables high-quality and natural video generation in a completely *training-free* manner. When fine-tuning is applied with only 20,000 original-length videos—requiring just 1/50,000 of the pre-training computation—sample quality is further improved, and the effectiveness of RIFLEx extends to $3\times$ extrapolation. Moreover, RIFLEx can also be applied in the spatial domain simultaneously to extend both video duration and spatial resolution.

Our key contributions are summarized as follows:

- We provide a *comprehensive understanding* of video length extrapolation by analyzing the failure modes of existing methods and revealing the role of individual frequency components in positional embeddings.

- We propose **RIFLEx**, a *minimal yet effective* solution that mitigates repetition by properly reducing the intrinsic frequency, without any additional modifications.

- RIFLEx offers a true *free lunch*—achieving high-quality $2\times$ extrapolation on state-of-the-art video diffu-

sion transformers in a completely *training-free* manner. Moreover, it enhances quality and enables $3\times$ extrapolation by minimal fine-tuning without long videos.

2. Background

2.1. Video Generation with Diffusion Transformers

Given a data distribution p_{data}, diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) progressively perturb the clean data $\boldsymbol{x}_0 \sim p_{\text{data}}$ with a transition kernel $q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I})$, i.e., $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$, where $t \in [0, T]$, α_t, σ_t are pre-defined noise schedule, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is Gaussian noise. Under proper designs of α_t, σ_t, the distribution of $\boldsymbol{x}_T$ is tractable, e.g., a standard Gaussian.

A generative model is obtained by reversing this process from $t = T$ to 0, whose dynamic is characterized by the score function $\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t)$. The score function is usually parameterized by a neural network $\boldsymbol{s_\theta}(\boldsymbol{x}_t, t)$ and learned with the denoising score matching (Vincent, 2011): $\mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}[w(t)\|\boldsymbol{s_\theta}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2],$

where $w(t)$ is a weighting function. The de facto approach for modeling video data via diffusion models is to first encode the video data into sequences of latent space, then perform diffusion modeling with transformer-based neural network (Peebles & Xie, 2023; Bao et al., 2023).

2.2. Position Embedding in Diffusion Transformers

A position embedding is a fixed or learnable vector-valued function $\boldsymbol{f}$ that maps a n-axes position vector $\boldsymbol{p} \in \mathbb{N}_+^n$ to some representation space. This position information can be incorporated into transformers through various mechanisms, such as through additive (Vaswani, 2017; Raffel et al., 2020; Press et al., 2021) or multiplicative (Su et al., 2021) operations with other input or hidden embeddings.

Rotary Position Embedding (RoPE) (Su et al., 2021) has emerged as the predominant method in transformers. RoPE encodes relative positional information by interacting with two absolute position embeddings within the attention mechanism. Specifically, for a sequence indexed by a single axis (i.e. $n = 1$), given an input $\boldsymbol{x} \in \mathbb{R}^d$ with position $p \in \mathbb{N}_+$, RoPE maps it to an absolute-position encoded embedding on $\mathbb{R}^{d'}$ with $d' \le d$, i.e.,

$$\boldsymbol{f}^{\text{RoPE}}(\boldsymbol{x}, p, \boldsymbol{\theta})_j = \begin{bmatrix} \cos p\theta_j & -\sin p\theta_j \\ \sin p\theta_j & \cos p\theta_j \end{bmatrix} \begin{bmatrix} x_{2j} \\ x_{2j+1} \end{bmatrix}. \quad (1)$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d'/2}$ with $\theta_j = b^{-2(j-1)/d'}$ for $j = 1, \ldots, d'/2$. Here, $\boldsymbol{\theta}$ represents the frequencies for all dimensions of the RoPE embedding and b is a hyperparameter that adjusts the base frequency. It can be verified that the dot product between two RoPE-embedded vectors encodes the relative positional information between them. In practice, RoPE is applied to the query and key vectors before the dot product operation in the attention mechanism, and thus the result attention matrix encodes the relative positional information.

RoPE with Multiple Axes. RoPE can be extended to multi-axes position vector $\boldsymbol{p} \in \mathbb{N}_+^n$ for $n > 1$. One popular practice is to encode each axis independently. For example, consider a video input $\boldsymbol{x} \in \mathbb{R}^d$ with three-dimensional coordinates (t, h, w), there are three axis-specific parameters $\boldsymbol{\theta}^t, \boldsymbol{\theta}^h, \boldsymbol{\theta}^w$. Single-axis RoPE, as defined in Eqn. (1), is then applied separately along the feature dimension with these three parameters. The final multi-axes RoPE is obtained by concatenating the three single-axis RoPE embeddings.

2.3. Length Extrapolation with RoPE

In this section, we briefly recap the techniques for length extrapolation with RoPE adopted in text and image.

The most straightforward approach, **Position Extrapolation (PE)**, extends the input sequence length without modifying the positional encoding, which purely relies on the

generalization ability of the network and the positional encoding. Whereas **Position Interpolation (PI)** (Chen et al., 2023b) uniformly down-scales all frequencies in RoPE embedding to match the training sequence length. In specific, the new RoPE frequencies are calculated as $\boldsymbol{\theta}^{\text{PI}} = \boldsymbol{\theta}/s$, where $s = L'/L$, and L, L' is the sequence length for training and inference, respectively.

A key limitation of both PE and PI is their reliance on training at the target sequence length, otherwise, the performance degrades drastically. To address this, **NTK-Aware Scaled RoPE (NTK)** (bloc97, 2023) combines the ideas of both position extrapolation and interpolation. Specifically, NTK adjusts the base frequency b for all dimensions as:

$$\theta_j^{\text{NTK}} = (\lambda b)^{-2(j-1)/d'}, \lambda = s^{d'/(d'-2)}, j = 1, \ldots, d'/2, \quad (2)$$

where $s = L'/L$. NTK effectively applies PE for high frequencies and PI for low frequencies, enabling training-free extrapolation. (bloc97, 2023; Zhuo et al., 2024).

YaRN (Peng et al., 2023) introduces a fine-grained base frequency adjustment strategy. It first categorizes all frequencies into three groups based on *the number of cycles elapsed over the training length*, defined as $r_j = (2\pi)^{-1} L\theta_j$. Given two pre-determined thresholds α, β with $r_{d'/2} \le \alpha < \beta \le r_1$, YaRN adjusts the RoPE frequencies as:

$$\theta_j^{\text{YaRN}} = \gamma(r_j)\theta_j + (1 - \gamma(r_j))\frac{\theta_j}{s}, \ j = 1, \ldots, d'/2,$$

$$\gamma(r_j) = \begin{cases} 1, & \text{if } r_j > \beta, \\ 0, & \text{if } r_j < \alpha, \\ \frac{r_j - \alpha}{\beta - \alpha}, & \text{otherwise}, \end{cases} \quad (3)$$

In practice, YaRN exhibits better training-free extrapolation performance compared to NTK and can achieve great performance with a relatively small fine-tuning budget on target length (Peng et al., 2023).

Length Extrapolation in Image Diffusion Transformers. Image diffusion transformers have two key characteristics related to RoPE: (1) image data is represented as a sequence with height and width axes, and (2) an iterative diffusion sampling procedure. These characteristics inform specific length extrapolation techniques for image diffusion models.

For multi-axes sequence, RoPE is independently applied to each axis, allowing length extrapolation techniques like NTK and YaRN to be used separately on height and width, termed Vision NTK and Vision YaRN (Lu et al., 2024b). For sampling, different RoPE adjustments can be employed at various diffusion timesteps. For instance, **Time-aware Scaled RoPE (TASR)** (Zhuo et al., 2024) leverages PI at large timesteps to preserve global structure while using NTK at smaller timesteps to enhance visual quality.

	2× **length extrapolation**	2× **spatial extrapolation**
Normal length	Video of 49 frames	Image of 1K resolution
PE	(a) Temporal repetition	(d) Spatial repetition
PI	(b) Slower motion	(e) Blurred details
NTK	(c) Temporal repetition	(f) Spatial repetition

Figure 2. **Visualization of existing methods for 2× extrapolation in video and image generation.** The base models CogVideoX-5B (Yang et al., 2024) and Lumina-Next (Zhuo et al., 2024) are trained to sample videos of up to 49 frames and images of up to 1K resolution, respectively. Existing methods lead to *temporal repetition* or *slower motion* in video extrapolation and *spatial repetition* or *blurred content* in image extrapolation, respectively. Please refer to Appendix C for more results and details.

3. Method

Our goal is to understand and solve the video length extrapolation problem thoroughly. We first highlight the intriguing failure patterns of existing methods, analyze the role of different frequency components in positional embeddings, and identify an *intrinsic frequency*. Based on this, we derive a minimal solution that enables length extrapolation: *reducing the intrinsic frequency*. As byproducts, our method not only provides a principled explanation for the failure of existing approaches in video extrapolation but also offers insights applicable to spatial extrapolation in images.

3.1. Failure Patterns of Existing Methods

Although the term "extrapolation" is widely used across different domains, its role in video generation is fundamentally different from text and images. In video generation, the objective is to create *new and temporally coherent content that evolves smoothly over time*. In contrast, text extrapolation primarily extends the context window, while image extrapolation typically involves adding high-resolution details rather than generating meaningful new content.

As a result, extrapolation strategies developed for text and images fail in video length extrapolation, exhibiting intriguing failure patterns, as illustrated in Fig. 2. To better understand these patterns, we also conduct the counterparts on image spatial extrapolation, revealing parallels to video.

PE, which directly extends positional encoding beyond the training range, leads to *temporal repetition*, causing videos to loop instead of progressing naturally (Fig. 2a). A similar phenomenon occurs in image generation, where *spatial repetition* occurs instead of generating new content.

Conversely, PI (Chen et al., 2023b), which compresses positional encodings within the training range, leads to *slow motion* by stretching frames over time (Fig. 2b). While this approach preserves structural coherence, it lacks temporal novelty. In image generation, this results in *blurred details* rather than new content (Fig. 2e).

As shown in Fig. 2c, NTK (bloc97, 2023) also induces *temporal repetition*, failing to generate meaningful motion progression. In image generation, it causes *spatial repetition* (Fig. 2f). While other methods (Peng et al., 2023; Lu et al., 2024b; Zhuo et al., 2024) differ from NTK in implementation, they invariably suffer from one or both of these two failure patterns: either motion deceleration or content repetition (see Appendix C for further analysis).

Beyond revealing these limitations, our findings provide an intuitive understanding of how positional embeddings fundamentally shape temporal motion, motivating our in-depth frequency component analysis in the next section.
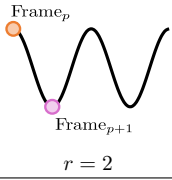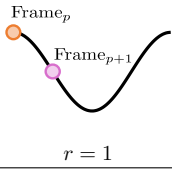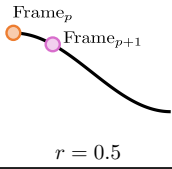
RoPE components	Dynamic & repetition behavior in training length	Repeat times under $2\times$ extrapolation
Frame$_p$ ... Frame$_{p+1}$ ... $r = 2$	Rapid changes accompanied by short-range repetitions	4 times
Frame$_p$ ... Frame$_{p+1}$ ... $r = 1$	Regular dynamics and no repetition	2 times
Frame$_p$ ... Frame$_{p+1}$ ... $r = 0.5$	Slow motion and no repetition	No repetition

Figure 3. **Visualization of frequency components and their roles in video generation.** High frequencies capture rapid movements and short-term dependencies, inducing temporal repetition, while low frequencies encode long-term dependencies with slow motion.

3.2. Frequency Component Analysis in RoPE

We begin by analyzing the role of individual frequency components in RoPE (Su et al., 2021). We follow the notation in Sec. 2.2 but focus on the time axis and omit the subscript t for simplicity. We isolate specific frequency components by zeroing out others and fine-tune the target model (Yang et al., 2024) on its training length to adapt to the modified RoPE. Two key insights emerge from this analysis.

First, different frequency components θ_j capture temporal dependencies at varying scales, dictated by their *periods*:

$$N_j = \frac{2\pi}{\theta_j}, \quad (4)$$

where j denotes the frequency index in RoPE. As illustrated in Fig. 3, when the frame interval exceeds N_j, the periodic nature of the cosine function forces positional encodings—and consequently, generated video content—to repeat. Given a training length L, the number of temporal repetitions can be quantified as:

$$r_j = \frac{L}{N_j} = \frac{L\theta_j}{2\pi}. \quad (5)$$

As shown in Fig. 3, when a high-frequency component has $r_j = 2$, the video completes two cycles within the training length and four cycles during $2\times$ extrapolation. In contrast, a low-frequency component with $r_j = 0.5$ remains within a single cycle even when extrapolated.

Second, frequency components influence the perceived motion speed in video generation. This effect correlates to the *rate of change* in positional encoding between consecutive (e.g., p-th and $(p+1)$-th frames:

$$\Delta_j = \cos((p+1)\theta_j) - \cos(p\theta_j). \quad (6)$$

Higher frequencies (i.e., larger θ_j) typically result in larger Δ_j, making the model more sensitive to rapid movements. Conversely, lower-frequency components induce minimal positional encoding shifts between adjacent frames, favoring slow-motion dynamics, aligning with results in Fig. 3.

Given that each component has its own period N_j, a key question arises: **which frequency primarily dictates the observed repetition pattern** in length extrapolation?

We define the *intrinsic frequency component* as the one whose period N_j is closest to the first observed repetition frame N in a video, as it determines the overall behavior:

$$k = \arg\min_j |N_j - N|, \quad (7)$$

where j denotes the frequency index in RoPE. Surprisingly, this intrinsic frequency remains consistent across different videos generated by the same model, despite slight variations in N. For instance, k is 2 for CogVideoX-5B (Yang et al., 2024) and 4 for HunyuanVideo (Kong et al., 2024) respectively, as detailed in Appendix E.

In the rare case where a model exhibits inconsistent intrinsic frequencies across videos, we suggest treating all such
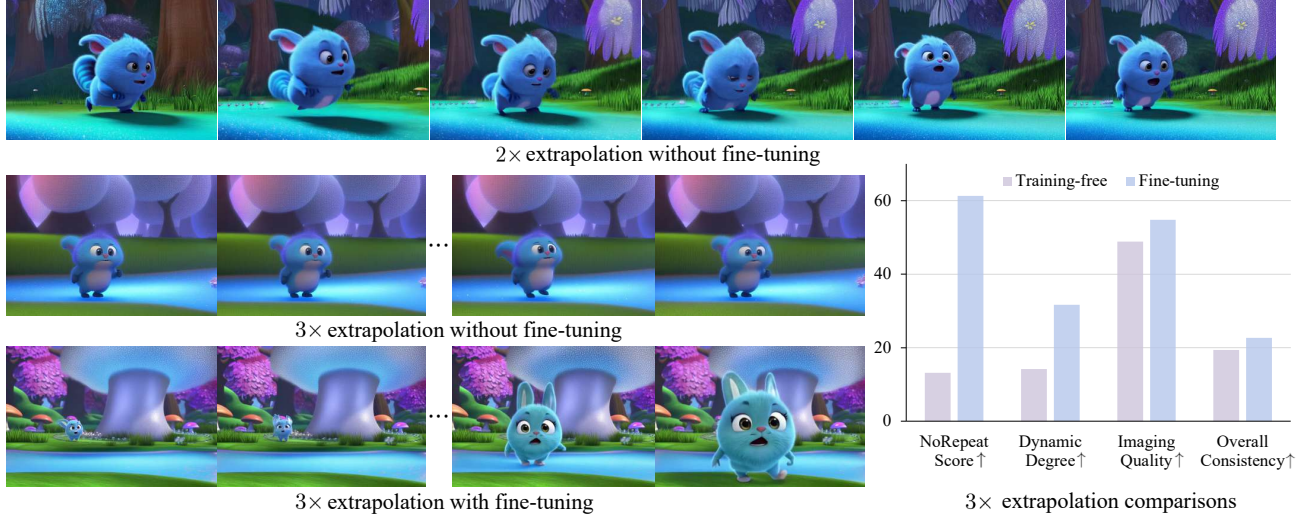
Figure 4. **Exploring the necessity of fine-tuning.** For 2× extrapolation, RIFLEx generates high-quality videos without fine-tuning. For 3× extrapolation, due to the large intrinsic frequency shift, fine-tuning is required to improve dynamic effects and visual quality.

frequencies as intrinsic. Our preliminary experiments further validate this assumption, showing that incorporating all lower-frequency components into our method maintains strong performance, as discussed in Appendix E.

3.3. Reducing Intrinsic Frequency: A Minimal Solution

Consider a video diffusion transformer trained on sequences of length L. We aim to generate videos of length sL via extrapolation by a factor of s[1]. Based on previous findings, we propose a natural and minimal solution: **R**educing **I**ntrinsic **F**requency for **L**ength **Ex**trapolation (**RIFLEx**). RIFLEx lowers the intrinsic frequency so that it remains within a single period after extrapolation:

$$N'_k \geq Ls \Rightarrow \theta'_k \leq \frac{2\pi}{Ls}. \tag{8}$$

By setting $\theta'_k = \frac{2\pi}{Ls}$, we achieve a minimal modification. Ablation studies on other frequencies (Appendix E) confirm that modifying only the intrinsic frequency is sufficient: adjusting higher-frequency components disrupt fast motion while altering lower frequencies has negligible impact. We present RIFLEx formally in Algorithm 1.

We further investigate **whether fine-tuning is necessary** for RIFLEx. Surprisingly, for 2× extrapolation, RIFLEx can generate high-quality videos in a *training-free* manner, as shown in Fig. 4. Fine-tuning with only 20,000 original-length videos and 1/50,000 of the pre-training computation further enhances dynamic quality and visual quality.

For 3× extrapolation, the intrinsic frequency shift becomes too large, leading to a notable training-testing mismatch. This occurs because the position embeddings used during in-

---

[1]We assume s is sufficiently large such that $N_k < Ls$. Otherwise, it is trivial to generate long videos by PE.

Algorithm 1 RIFLEx

Require: The extrapolation factor s, frequencies θ_j in the RoPE, the first observed repetition frame N

1: **for** $j = 1$ to $\dfrac{d'}{2}$ **do**
2:     Compute the period of each θ_j (Eqn. (4))
3: **end for**
4: Identify the intrinsic frequency component k (Eqn. (7))
5: Modify $\theta_k = \frac{2\pi}{Ls}$

---

ference deviate slightly from those seen during training due to modified frequencies. While this discrepancy does not undermine the conclusion about our non-repetition condition in Eqn. (8), it may affect visual quality since the model lacks explicit training on these specific position embeddings. Nevertheless, the fine-tuning process still succeeds, as shown in Fig. 4.

3.4. Principled Explanation for Existing Methods

Our findings provide a principled explanation for the failure patterns observed in Section 3.1. The repetition observed in PE and NTK (bloc97, 2023; Lu et al., 2024b) stems from their intrinsic frequency component violating the non-repetition condition in Eqn. (8). As a result, the generated video content loops instead of progressing naturally. PI (Chen et al., 2023b) and YaRN (Peng et al., 2023) cause slow motion by interpolating high-frequency components, which are crucial for fast motion. Divided by s in such methods, these components cannot generate rapid movements. TASR (Zhuo et al., 2024) combines both approaches mentioned above, resulting in a mixture of temporal repetition and motion slowdown. See Appendix C for more details and experiments.

Table 1. **Quantitative results in length extrapolation.** The red-marked areas in the NoRepeat Score and Dynamic Degree indicate severe issues with repetition and slow motion, making other metrics meaningless. In the user study, the ratio for no extrapolation represents the proportion of users who prefer the samples of the training length over RIFLEx. The others are the corresponding ranks among all methods.

Method	Automatic Metrics↑				User Study↓		
	NoRepeat Score	Dynamic Degree	Imaging Quality	Overall Consistency	Motion Quality	Visual Quality	Overall Aspects
CogVideoX-5B with 2× extrapolation, training-free							
No extrapolation	-	67.5	64.4	25.8	66.4%	76.0%	70.2%
PE	46.6	<u>58.6</u>	55.0	<u>22.9</u>	<u>2.1</u>	<u>1.6</u>	2.4
NTK	43.4	58.3	<u>55.3</u>	<u>22.9</u>	<u>2.1</u>	1.8	<u>2.1</u>
PI	<u>59.0</u>	5.0	44.3	19.2	3.7	4.1	3.8
TASR	10.8	26.9	50.5	21.5	3.3	3.8	3.6
YaRN	**59.4**	5.6	44.6	19.3	3.6	4.2	3.7
RIFLEx (**ours**)	54.2	**59.4**	**56.9**	**23.5**	**1.4**	**1.5**	**1.1**
CogVideoX-5B with 2× extrapolation, fine-tuning							
No extrapolation	-	65.6	62.7	25.8	61.8%	66.0%	65.0%
PE	13.2	50.6	56.6	24.2	1.8	1.8	1.7
RIFLEx (**ours**)	**61.3**	**54.7**	**60.4**	**25.0**	**1.2**	**1.2**	**1.3**
HunyuanVideo with 2× extrapolation, training-free							
No extrapolation	-	63.0	65.9	19.6	62.8%	62.0%	61.6%
PE	36.0	**63.0**	64.3	**19.1**	2.3	<u>1.2</u>	2.4
NTK	81.0	55.0	**65.3**	18.9	**1.5**	1.4	<u>1.6</u>
PI	**86.0**	11.0	57.4	18.9	4.3	2.8	3.8
TASR	<u>85.0</u>	18.0	61.3	<u>19.0</u>	4.2	2.2	3.4
YaRN	**86.0**	15.0	58.2	18.8	3.9	2.7	3.7
RIFLEx (**ours**)	72.0	<u>57.0</u>	<u>65.2</u>	<u>19.0</u>	<u>1.6</u>	**1.1**	**1.4**
HunyuanVideo with 2.3× extrapolation, training-free							
NTK	20.0	46.0	**65.5**	**18.3**	1.7	1.6	1.7
RIFLEx (**ours**)	**54.0**	**51.0**	65.0	18.1	**1.3**	**1.4**	**1.3**
HunyuanVideo with 2× extrapolation, fine-tuning							
No extrapolation	-	79.0	71.6	18.8	62.6%	51.2%	56.0%
PE	40.0	74.0	71.6	**18.7**	1.9	1.6	1.8
RIFLEx (**ours**)	**89.0**	**82.0**	**72.0**	18.1	**1.1**	**1.4**	**1.2**

4. Experiments

4.1. Setup

We describe the dataset and evaluation setup below, with implementation details in Tab. 3 (see Appendix D).

Datasets. We use a private dataset of 20,000 videos for fine-tuning. For CogVideoX-5B, We adopt the VBench (Huang et al., 2024) prompts to ensure consistency with prior work (Yang et al., 2024). Due to the high computational cost of HunyuanVideo (Kong et al., 2024), we evaluate it

using 100 diverse prompts across multiple categories.

Evaluation metrics. Following prior work (Huang et al., 2024; Yang et al., 2024), we assess video generation using Imaging Quality, Dynamic Degree, and Subject Consistency, measuring visual quality, motion magnitude, and temporal consistency, respectively. Additionally, we introduce the NoRepeat Score, where a higher score indicates less repetition (detailed in Appendix D). We also conduct a user study with 10 participants, evaluating visual quality, motion quality, and overall preference. Motion quality reflects both repetition and slow motion. Users rank their preferences
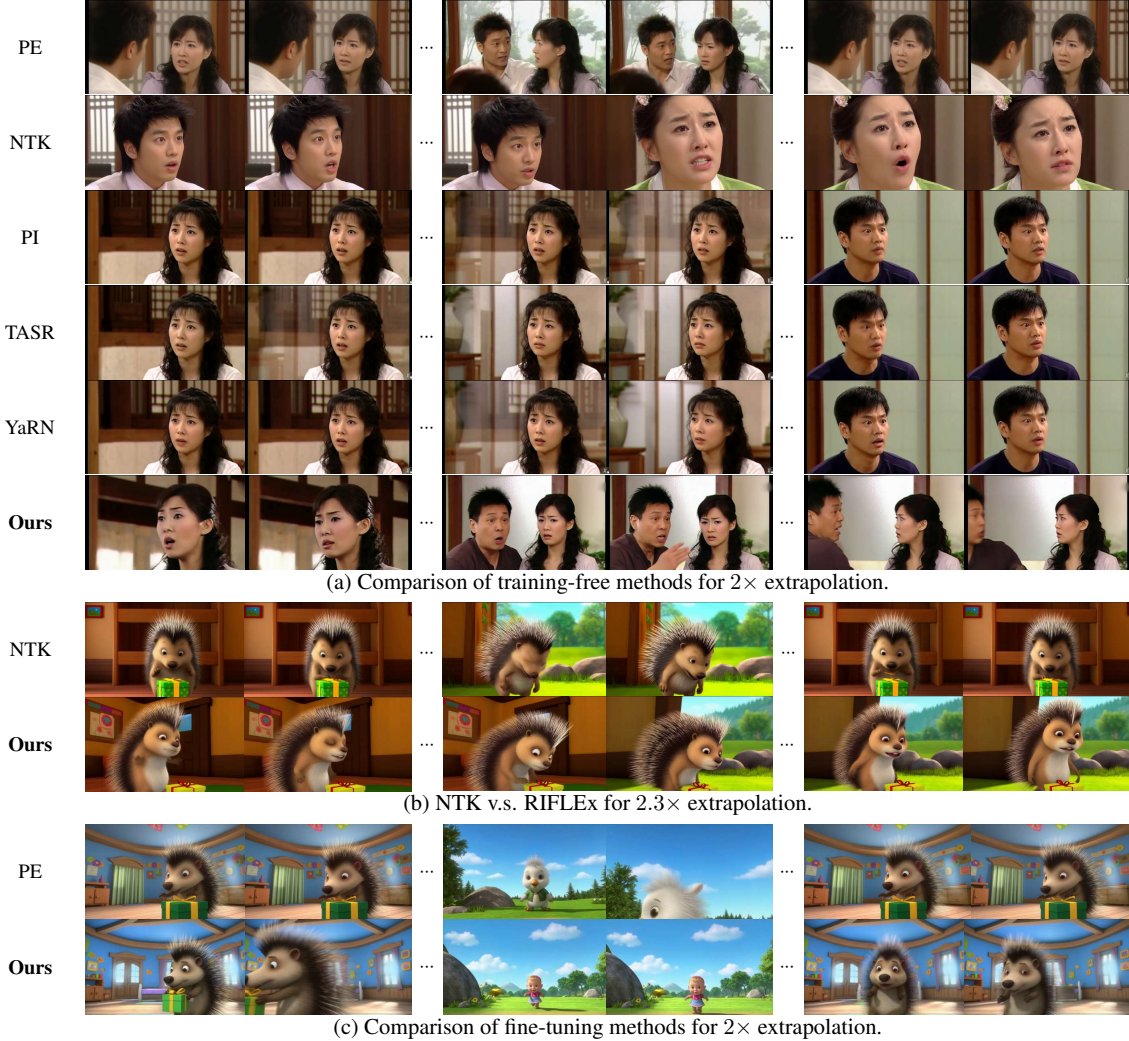
(a) Comparison of training-free methods for $2\times$ extrapolation.

(b) NTK v.s. RIFLEx for $2.3\times$ extrapolation.

(c) Comparison of fine-tuning methods for $2\times$ extrapolation.

Figure 5. **Visualization results of length extrapolation based on HunyuanVideo.** We achieve better video quality by effectively addressing issues of slow motion and repetition. Notably, while the NTK in HunyuanVideo incidentally avoids repetition at $2\times$ extrapolation, it still encounters significant repetition at longer extrapolations, such as $2.3\times$ extrapolation.

among all extrapolation methods, allowing for ties. We also perform pairwise comparisons between the results of normal samples and RIFLEx. See more details in Appendix D.

4.2. Performance Comparison

Results. Quantitative results are summarized in Tab. 1. Our approach achieves superior overall performance, generating new temporal content without compromising other aspects of video quality. For example, in CogVideoX-5B, PI and YaRN suffer from slow motion, leading to lower *Dynamic Degree*, while PE and NTK experience repetition issues, resulting in lower *NoRepeat Score*. By effectively addressing both challenges, our method significantly enhances motion quality and ranks highest in user studies across all methods.

Notably, NTK coincidentally performs well for Hunyuan-

Video at $2\times$ extrapolation, but our analysis attributes this to an unintended intrinsic frequency reduction that happens to satisfy the non-repetition condition in Eqn. (8), rather than its intended mechanism. This is evident as NTK fails on CogVideo-X and HunyuanVideo with $2.3\times$ extrapolation, reflected in its low *NoRepeat Score* in Tab. 1.

Qualitative results are shown in Fig. 5 for HunyuanVideo, with additional comparisons for CogVideoX-5B in Appendix F. Fig. 5 aligns with the quantitative findings, demonstrating our method's ability to effectively mitigate slow motion and repetition, thereby improving overall video quality.

Additionally, a minimal fine-tuning procedure requiring just 1/50,000 of the pre-training computation on short videos improves the *Dynamic Degree*, *Imaging Quality*, and *NoRepeat Score*. Finally, leveraging the strong HunyuanVideo base model, our approach achieves performance close to

that of the training length—with 56.0% and 61.6% of users preferring the training length over our method.

Maximum extent of extrapolation. Empirically, RIFLEx supports up to $3\times$ extrapolation, beyond which quality degrades significantly (e.g., at $4\times$ extrapolation, see Fig. 9 in Appendix). This may occur because excessive frequency reduction diminishes the effectiveness of RoPE, resulting in minimal encoding changes over the training length.

Extension to other extrapolation types. Video diffusion transformers typically apply 1D RoPE (see Sec. 2.2) independently to both spatial and temporal dimensions. This shared mechanism leads to analogous extrapolation challenges in both domains. Consequently, our method naturally extends to spatial extrapolation and joint temporal-spatial extrapolation, offering a unified framework for extrapolation in diffusion transformers. As shown in Fig. 1b and Fig. 1c, adjusting the intrinsic frequency for the corresponding dimensions enables resolution extrapolation and joint spatial-temporal extension. Additional demos and implementation details are provided in Appendix B and Appendix D.

5. Conclusion and Discussion

We provide a comprehensive understanding of video length extrapolation by analyzing the role of frequency components in RoPE. Building on these insights, we propose RIFLEx, a minimal yet effective solution that prevents repetition by reducing intrinsic frequency. RIFLEx achieves high-quality $2\times$ extrapolation on SOTA video diffusion transformers in a training-free manner and enables $3\times$ extrapolation by minimal fine-tuning without long videos.

In this paper, we primarily adopt an empirical approach—visual inspection—for intrinsic frequency identification when adapting the pre-trained video diffusion transformer. While this approach is effective for adaptation, establishing a theoretical foundation for intrinsic frequency identification is crucial. Achieving this would require fundamental research into how intrinsic frequencies emerge during the pre-training process, potentially analysis from a training-from-scratch perspective. What's more, as discussed in the main text, the $3\times$ limitation stems from diminished ability to discriminate sequential positions due to excessive frequency reduction. To further extend beyond this, it is promising to investigate the mechanism of positional encoding during training, specifically tailored for extrapolation.

Acknowledgements

Impact Statement

This paper presents work aimed at advancing the field of video generation. It is crucial to use this technology responsibly to prevent negative social impacts, such as the creation of misleading fake videos.

References

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22669–22679, 2023.

Bao, F., Xiang, C., Yue, G., He, G., Zhu, H., Zheng, K., Zhao, M., Liu, S., Wang, Y., and Zhu, J. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233, 2024.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., and Rombach, R. Stable video diffusion: Scaling latent video diffusion models to large datasets. NONE, 2023.

bloc97. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024.

Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., and Shan, Y. Videocrafter1: Open diffusion models for high-quality video generation, 2023a.

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., and Shan, Y. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

Chen, J., Long, F., An, J., Qiu, Z., Yao, T., Luo, J., and Mei, T. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. arXiv preprint arXiv:2501.09019, 2025.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595, 2023b.

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

Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., and Parikh, D. Long video generation with time-agnostic vqgan and time-sensitive transformer. In European Conference on Computer Vision, pp. 102–118. Springer, 2022.

He, Y., Yang, T., Zhang, Y., Shan, Y., and Chen, Q. Latent video diffusion models for high-fidelity long video generation. 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv: 2210.02303, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868, 2022.

Hu, J., Ai, Q., Guo, D., Zhou, Q., Sun, X., Zhang, Q., and Luo, C. Long-context extrapolation via periodic extension, 2024.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807–21818, 2024.

Kim, J., Kang, J., Choi, J., and Han, B. FIFO-diffusion: Generating infinite videos from text without training. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=uikhNa4wam.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.-C., et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

Li, Z., Hu, S., Liu, S., Zhou, L., Choi, J., Meng, L., Guo, X., Li, J., Ling, H., and Wei, F. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. arXiv preprint arXiv: 2410.20502, 2024.

Liang, J., Wu, C., Hu, X., Gan, Z., Wang, J., Wang, L., Liu, Z., Fang, Y., and Duan, N. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. Advances in Neural Information Processing Systems, 35:15420–15432, 2022.

Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al. Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131, 2024.

Lin, H., Zala, A., Cho, J., and Bansal, M. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. arXiv preprint arXiv: 2309.15091, 2023.

Lu, Y., Liang, Y., Zhu, L., and Yang, Y. Freelong: Training-free long video generation with spectralblend temporal attention. arXiv preprint arXiv:2407.19918, 2024a.

Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., and Bai, L. Fit: Flexible vision transformer for diffusion model. International Conference on Machine Learning., 2024b.

NVIDIA, :, Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., Dworakowski, D., Fan, J., Fenzi, M., Ferroni, F., Fidler, S., Fox, D., Ge, S., Ge, Y., Gu, J., Gururani, S., He, E., Huang, J., Huffman, J., Jannaty, P., Jin, J., Kim, S. W., Klár, G., Lam, G., Lan, S., Leal-Taixe, L., Li, A., Li, Z., Lin, C.-H., Lin, T.-Y., Ling, H., Liu, M.-Y., Liu, X., Luo, A., Ma, Q., Mao, H., Mo, K., Mousavian, A., Nah, S., Niverty, S., Page, D., Paschalidou, D., Patel, Z., Pavao, L., Ramezanali, M., Reda, F., Ren, X., Sabavat, V. R. N., Schmerling, E., Shi, S., Stefaniak, B., Tang, S., Tchapmi, L., Tredak, P., Tseng, W.-C., Varghese, J., Wang, H., Wang, H., Wang, H., Wang, T.-C., Wei, F., Wei, X., Wu, J. Z., Xu, J., Yang, W., Yen-Chen, L., Zeng, X., Zeng, Y., Zhang, J., Zhang, Q., Zhang, Y., Zhao, Q., and Zolkowski, A. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv: 2501.03575, 2025.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. International Conference on Learning Representations., 2023.

Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.-Y., Chuang, C.-Y., Yan, D., Choudhary, D., Wang, D., Sethi, G., Pang, G., Ma, H., Misra, I., Hou, J., Wang, J., Jagadeesh, K., Li, K., Zhang, L., Singh, M., Williamson, M., Le, M., Yu, M., Singh, M. K., Zhang, P., Vajda, P., Duval, Q., Girdhar, R., Sumbaly, R., Rambhatla, S. S., Tsai, S., Azadi, S., Datta, S., Chen, S., Bell, S., Ramaswamy, S., Sheynin, S., Bhattacharya, S., Motwani, S., Xu, T., Li, T., Hou, T., Hsu, W.-N., Yin, X., Dai, X., Taigman, Y., Luo, Y., Liu, Y.-C., Wu, Y.-C., Zhao, Y., Kirstain, Y., He, Z., He, Z., Pumarola, A., Thabet, A., Sanakoyeu, A., Mallya, A., Guo, B., Araya, B., Kerr, B., Wood, C., Liu, C., Peng, C., Vengertsev, D., Schonfeld, E., Blanchard, E., Juefei-Xu, F., Nord, F., Liang, J., Hoffman, J., Kohler, J., Fire, K., Sivakumar, K., Chen, L., Yu, L., Gao, L., Georgopoulos, M., Moritz, R., Sampson, S. K., Li, S., Parmeggiani, S., Fine, S., Fowler, T., Petrovic, V., and Du, Y. Movie gen: A cast of media foundation models. arXiv preprint arXiv: 2410.13720, 2024.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409, 2021.

Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., and Liu, Z. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21 (140):1–67, 2020.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

Su, J., Lu, Y., Pan, S., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding, 2021.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14398–14409, 2024.

Team, F. Fastvideo: A lightweight framework for accelerating large video diffusion models., December 2024a. URL https://github.com/hao-ai-lab/FastVideo.

Team, G. Mochi 1. https://github.com/genmoai/models, 2024b.

Vaswani, A. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

Vincent, P. A connection between score matching and denoising autoencoders. Neural computation, 23(7):1661–1674, 2011.

Wang, F.-Y., Chen, W., Song, G., Ye, H.-J., Liu, Y., and Li, H. Gen-l-video: Multi-text to long video generation via temporal co-denoising. arXiv preprint arXiv: 2305.18264, 2023.

Wang, H., Ma, C.-Y., Liu, Y.-C., Hou, J., Xu, T., Wang, J., Juefei-Xu, F., Luo, Y., Zhang, P., Hou, T., Vajda, P., Jha, N. K., and Dai, X. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. arXiv preprint arXiv: 2412.09856, 2024a.

Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024b.

Wang, Y., Xiong, T., Zhou, D., Lin, Z., Zhao, Y., Kang, B., Feng, J., and Liu, X. Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv: 2410.02757, 2024c.

Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., and Duan, N. Godiva: Generating opendomain videos from natural descriptions. arXiv preprint arXiv:2104.14806, 2021.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. Nüwa: Visual synthesis pre-training for neural visual world creation. In European conference on computer vision, pp. 720–736. Springer, 2022.

Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al. Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429, 2024.

Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.-T., and Shan, Y. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023.

Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157, 2021.

Yan, X., Cai, Y., Wang, Q., Zhou, Y., Huang, W., and Yang, H. Long video diffusion generation with segmented cross-attention and content-rich video data curation. arXiv preprint arXiv:2412.01316, 2024.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.

Yin, T., Zhang, Q., Zhang, R., Freeman, W. T., Durand, F., Shechtman, E., and Huang, X. From slow bidirectional to fast causal video generators. arXiv preprint arXiv:2412.07772, 2024.

Zhang, J., Huang, H., Zhang, P., Wei, J., Zhu, J., and Chen, J. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. arXiv preprint arXiv:2411.10958, 2024a.

Zhang, J., Wei, J., Huang, H., Zhang, P., Zhu, J., and Chen, J. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. arXiv preprint arXiv:2410.02367, 2024b.

Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., and Chen, J. Spargeattn: Accurate sparse attention accelerating any model inference. arXiv preprint arXiv:2502.18137, 2025.

Zhao, M., Bao, F., Li, C., and Zhu, J. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. Advances in Neural Information Processing Systems, 35:3609–3623, 2022.

Zhao, M., Wang, R., Bao, F., Li, C., and Zhu, J. Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098, 2(3), 2023.

Zhao, M., Zhu, H., Xiang, C., Zheng, K., Li, C., and Zhu, J. Identifying and solving conditional image leakage in image-to-video diffusion model. Advances in Neural Information Processing Systems, 37:30300–30326, 2024.

Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv: 2412.20404, 2024.

Zhou, Y., Wang, Q., Cai, Y., and Yang, H. Allegro: Open the black box of commercial-level video generation model. arXiv preprint arXiv: 2410.15458, 2024.

Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhao, L., Wang, F.-Y., Ma, Z., et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. Advances in Neural Information Processing Systems., 2024.

A. Related Work

Length extrapolation with RoPE. Position encoding, exemplified by the widely used RoPE, plays a crucial role in enabling length extrapolation in transformers. Prior research in both language and image domains has primarily focused on training-free methods and fine-tuning under target sequence length settings. For instance, position interpolation generally outperforms direct position extrapolation in fine-tuning efficiency, requiring fewer steps to adapt to the target length (Chen et al., 2023b), though it performs poorly in training-free settings (Zhuo et al., 2024). Advanced strategies such as NTK (bloc97, 2023) and YaRN (Peng et al., 2023) have demonstrated decent training-free performance while being more efficient in fine-tuning scenarios. Further refinements, such as optimizing RoPE frequencies (Ding et al., 2024) or modifying RoPE's extrapolation behavior (Hu et al., 2024), have shown additional improvements in language modeling. Our work provides new insights into the impact of RoPE in video diffusion transformers, introducing a length extrapolation strategy tailored for video generation. Unlike previous approaches, our proposed RIFLEx requires training only on the original pre-trained sequence length while also demonstrating strong potential in training-free settings.

Text-to-video diffusion models. Drawing upon the progress made in image generation, a burgeoning body of research has emerged, focusing on the utilization of diffusion models for video generation (Kong et al., 2024; Yang et al., 2024; Ho et al., 2022; Polyak et al., 2024; Brooks et al., 2024; Zhou et al., 2024; Team, 2024b; Zheng et al., 2024; Blattmann et al., 2023; Lin et al., 2024; Xing et al., 2023; Chen et al., 2023a; 2024; He et al., 2022; Zhao et al., 2023; 2022). By combining spatial and temporal attention, VDM (He et al., 2022) introduces a space-time factorized UNet for video synthesis, marking an early contribution to the field. Later, Make-A-Video extends the 2D-UNet with temporal modules (Singer et al., 2022), exploring the integration of prior knowledge from text-to-image diffusion models into video diffusion techniques. More recently, a surge of video diffusion models leveraging the expressive power of transformers has emerged (Lin et al., 2024; Zheng et al., 2024; Kong et al., 2024; Yang et al., 2024; Bao et al., 2024; Brooks et al., 2024; Team, 2024b). These diffusion transformer-based models have demonstrated remarkable performance. Our approach builds on these advancements by applying them to pre-trained video diffusion transformers, further enhancing their capabilities. Moreover, recent developments have also seen the emergence of video diffusion models that leverage efficient attention mechanisms to accelerate their performance (Zhang et al., 2024b;a; 2025). Our approach is also applicable to these models, further extending their capabilities.

Autoregressive video generation models. Unlike diffusion models, autoregressive video generation models typically quantize videos into discrete tokens and generate video content through next-token prediction in an autoregressive manner. Previous works have demonstrated great performance in such models (Wu et al., 2021; Yan et al., 2021; Hong et al., 2022; Wu et al., 2022; Kondratyuk et al., 2023; Wu et al., 2024; Sun et al., 2024; Wang et al., 2024b). For example, NÜWA (Wu et al., 2022) employs VQ-GAN for tokenization and generates videos using a 3D transformer encoder-decoder framework. More recently, VideoPoet (Kondratyuk et al., 2023) tokenizes images and videos with a MAGVIT-v2 encoder and autoregressively generates videos using a decoder-only transformer based on a pretrained large language model. While autoregressive video models can theoretically extend sequences indefinitely through next-token prediction (Wang et al., 2024c; Liang et al., 2022; Ge et al., 2022), recent studies reveal their tendency to degenerate into repetitive content generation (Kondratyuk et al., 2023; Ge et al., 2022). In this work, we present a principled approach to video length extrapolation that effectively generates novel temporal content in diffusion-based frameworks. Although our method is developed for video diffusion transformers, the underlying mechanism governing position embedding periodicity may also offer insights for addressing repetition challenges in autoregressive video generation.

Long video with diffusion models. Recent studies have explored long video generation with diffusion models from various angles (Lu et al., 2024a; Wang et al., 2023; 2024a;c; Lin et al., 2023; Li et al., 2024; Qiu et al., 2023; NVIDIA et al., 2025). For instance, Kim et al. (2024); Chen et al. (2025) propose diffusion sampling schemes that employ a queue of video frames with varying noise levels, progressively decoding new frames. Yan et al. (2024) introduce a cross-attention module to enhance the semantic fidelity and richness of long videos. Yin et al. (2024) distill a chunk-wise, few-step auto-regressive video diffusion transformer from a bidirectional teacher model, enabling efficient long video generation. In this work, we address long video generation with diffusion transformers through the lens of position encoding—a fundamental component for capturing sequential structure in video data. We propose a minimal yet general and effective strategy that requires no training on long video data.

B. Additional Results of RIFLEx

In this section, we present additional demos for temporal extrapolation in Fig. 6, spatial extrapolation in Fig. 7, and both extrapolations in Fig. 8.

Figure 6. **More results of** $2\times$ **temporal extrapolation from** 129 **to** 261 **frames.**

(a) 2× length extrapolation on both width and height

(b) 2× length extrapolation on width
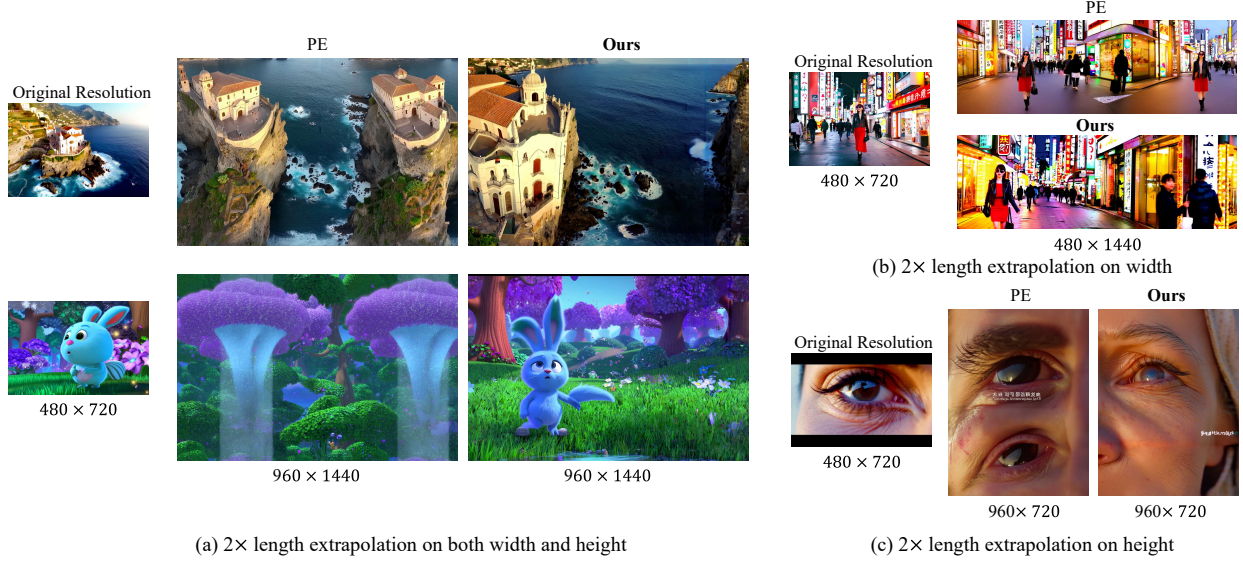
(c) 2× length extrapolation on height

Figure 7. **Visualization results of spatial resolution extrapolation method in image generation.** Our method outperforms the extrapolation by generating new content with better visual quality.

Figure 8. **More results of** 2× **temporal and spatial extrapolation**, extending video dimensions from $480 \times 720 \times 49$ to $960 \times 1440 \times 97$.

Figure 9. **Results of** 4× **temporal extrapolation from** 49 **to** 193 **frames.**

15

Table 2. **Code Links and Licenses.**

Method	Link	License
HunyuanVideo (Kong et al., 2024)	https://github.com/Tencent/HunyuanVideo	Tencent Hunyuan Community License
FastVideo (Team, 2024a)	https://github.com/hao-ai-lab/FastVideo	Apache License
CogVideoX (Yang et al., 2024)	https://github.com/THUDM/CogVideo	Apache License
Lumina-T2X (Zhuo et al., 2024)	https://github.com/Alpha-VLLM/Lumina-T2X	MIT License
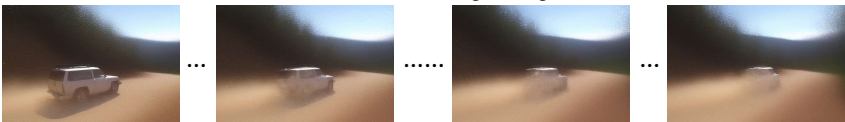
Figure 10. **Visualization of other existing methods for 2× extrapolation in video and image generation.** YaRN leads to slower motion. While TASR can successfully perform resolution extrapolation, it simultaneously causes slower motion and temporal repetition in video generation.

C. More Results of Failure Patterns of Existing Methods

As shown in Fig. 10, we present the results of other existing methods for 2× extrapolation in video and image generation. Specifically, YaRN results in slower motion, using parameters $\alpha = 1$ and $\beta = 32$ as set in previous studies (Lu et al., 2024b; Peng et al., 2023). TASR utilizes PI at larger timesteps and employing NTK at smaller timesteps. Consequently, it combines the characteristics of both PI and NTK, which leads to slower motion and temporal repetition in video generation.

D. Experimental Setup.

Used code and license. All used codes in this paper and its license are listed in Tab. 2.

Implementation details. For spatial extrapolation, following Algorithm 1, we identify the intrinsic frequency components whose periods closely match the repeating patterns observed in the height and width pixels, then adjust them to ensure unique encoding. For both spatial and temporal extrapolation, we simultaneously adjust the intrinsic frequency components for the time, width, and height dimensions. The training-free setting shares the same intrinsic frequency values as those in Tab. 3.

Evaluation metrics. For the NoRepeat Score, we identify the frame around N_k with the minimum L_2 distance to the first frame, marking it as the start of the possible repeated sequence. We then calculate the L_2 distance between each frame in the possible repeated sequence and the corresponding frame at the beginning of the video. If the average distance across frames exceeds a threshold, the video has a higher probability of being non-repetitive. We then calculate the proportion of videos with a higher probability of being non-repetitive. Empirically, we find that a threshold of 100 aligns better with human perception, so we set it to 100. For the human evaluation of the training-free setting, considering that several methods may share similar quality (e.g., slow motion with poor visual quality), we allow for ties. However, for the fine-tuning setting, ties are not permitted.

Table 3. **Fine-tuning settings for all experiments**. Both. denotes spatial and temporal extrapolation simultaneously. $b_k^{t'}$, $b_k^{h'}$, and $b_k^{w'}$ represent the base frequency for the intrinsic frequency in the time, height, and width dimensions, respectively. By adjusting these variables, we can modify the corresponding $\theta_k^{t'}$, $\theta_k^{h'}$, and $\theta_k^{w'}$ values accordingly (refer to Section 2.2 for details).

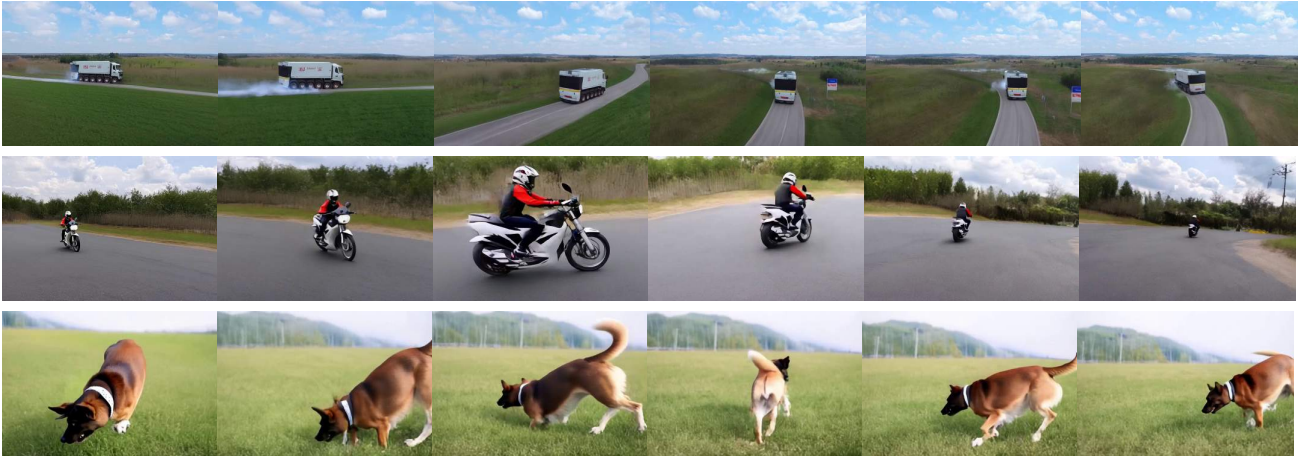| Config	$2\times$ Temporal	$2\times$ Temporal	$3\times$ Temporal	$2\times$ Spatial	$2\times$ Both.
Base model	CogVideoX-5B	HunyuanVideo	CogVideoX-5B	CogVideoX-5B	CogVideox-5B
Training iterations	2500	1000	5000	2000	10000
$b_k^{t'}$	$1e5$	560	$1e6$	-	$1e5$
$b_k^{h'}$	-	-	-	$1e6$	$1e6$
$b_k^{w'}$	-	-	-	$5e4$	$5e4$
Data size	$480 \times 720 \times 49$	$544 \times 960 \times 129$	$480 \times 720 \times 49$	$480 \times 720 \times 1$	$480 \times 720 \times 49$
Batch size	8	8	8	64	8
GPU	8 A100-80G	24 A100-80G	8 A100-80G	8 A100-80G	8 A100-80G

Figure 11. **The results of adjusting all frequency components lower than the intrinsic frequency.** See detailed analysis in Appendix E.

E. Details about RIFLEx

Robustness of the intrinsic frequency k. Empirically, we collected 20 videos and found that, although the first observed repetition frame may vary across videos within a certain range, the identified intrinsic frequencies remain consistent. For example, in HunyuanVideo, even though the first observed repetition frame range from 178 to 200, the closest intrinsic frequency is always $k = 4$, where $N_k = 200$.

Adjust all frequency components lower than the intrinsic frequency. In our preliminary experiments, we slow down all frequency components lower than the intrinsic frequency by increasing the base frequency b for $j \geq k$, where b is chosen to satisfy the non-repetition condition Eqn. (8) for intrinsic frequency k. As shown in Fig. 11, this approach effectively addresses the repetition issue while maintaining visual quality. It is important to note that, despite this choice, our RIFLEx, which reduces the intrinsic frequency, provides the minimal solution.

Ablations for other frequencies. As shown in Fig. 12, reducing the higher-frequency components slows down the video. Based on the analysis in Section 3.2, this may be because these components are crucial for capturing fast motion. Reducing their frequencies leads to a slower rate of change in the positional encoding, which weakens the model's ability to generate rapid movements.

On the other hand, reducing the lower frequencies has a negligible impact. This is likely because, for these frequencies, the encoding functions change very little across the training length, from $p = 1$ to $p = L$. Therefore, these frequencies may be less sensitive to positional encoding, and altering them results in minimal effect.
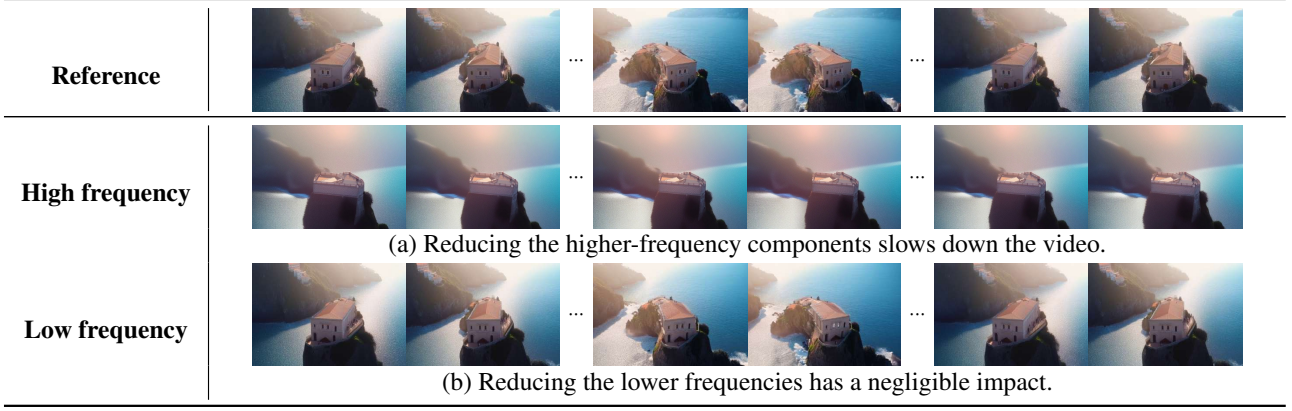
Reference	
High frequency	

(a) Reducing the higher-frequency components slows down the video.

Low frequency	

(b) Reducing the lower frequencies has a negligible impact.

Figure 12. **Ablations for reducing other frequencies.** Reference refers to the results of PE, where no frequencies are reduced, serving as the baseline.

F. More Results about Comparisons

In this section, we show the visualization comparisons of CogVideoX-5B. As shown in Fig. 13, PI and YaRN suffer from slow motion, while PE and NTK experience repetition issues. TASR suffers from both slow motion and repetition. By effectively addressing both challenges, our method significantly enhances motion quality.
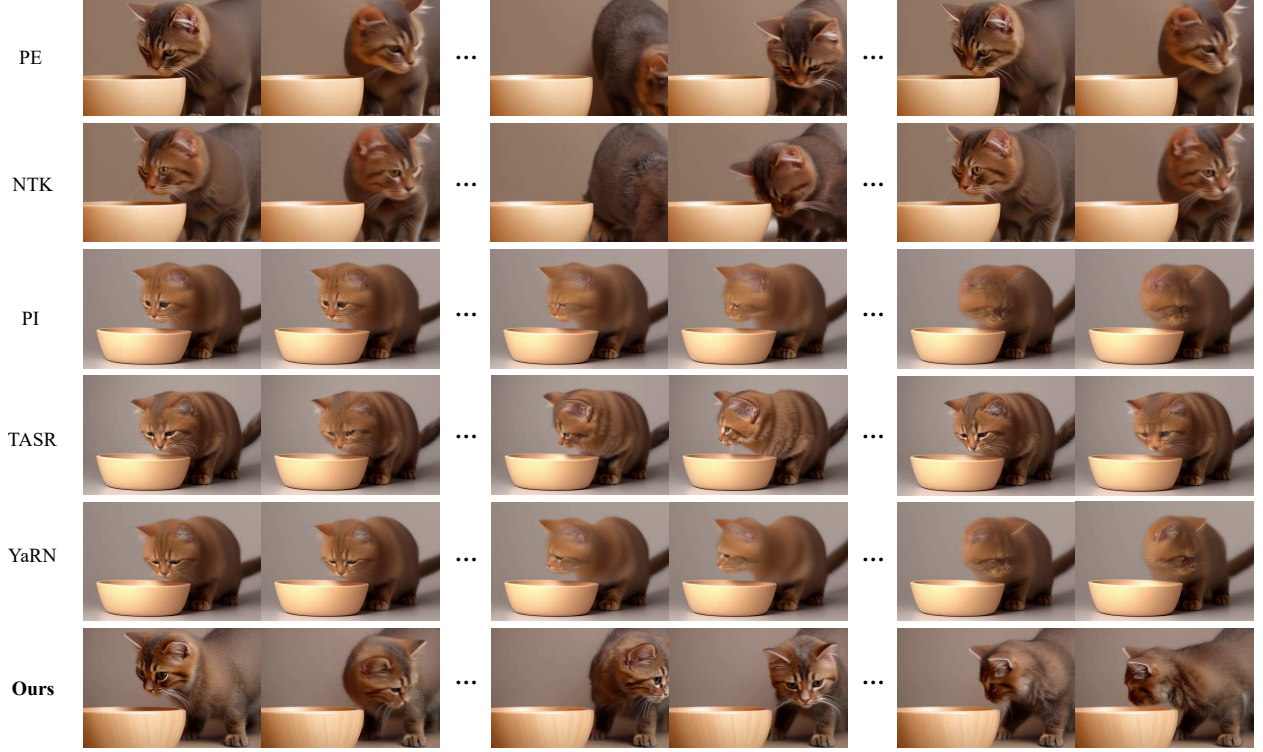
(a) Comparison of training-free methods for 2× extrapolation.

(b) Comparison of fine-tuning methods for 2× extrapolation.

Figure 13. **Visualization results of length extrapolation based on CogVideoX-5B.** We achieve better video quality by effectively addressing issues of slow motion and repetition.