

# Long-Context Autoregressive Video Modeling with Next-Frame Prediction

Yuchao Gu, Weijia Mao, Mike Zheng Shou\*

Show Lab, National University of Singapore  
<https://farlongctx.github.io>

## Abstract

Long-context autoregressive modeling has significantly advanced language generation, but video generation still struggles to fully utilize extended temporal contexts. To investigate long-context video modeling, we introduce **Frame AutoRegressive (FAR)**, a strong baseline for video autoregressive modeling. Just as language models learn causal dependencies between tokens (i.e., Token AR), FAR models temporal causal dependencies between continuous frames, achieving better convergence than Token AR and video diffusion transformers. Building on FAR, we observe that long-context vision modeling faces challenges due to visual redundancy. Existing RoPE lacks effective temporal decay for remote context and fails to extrapolate well to long video sequences. Additionally, training on long videos is computationally expensive, as vision tokens grow much faster than language tokens. To tackle these issues, we propose balancing locality and long-range dependency. We introduce *FlexRoPE*, an test-time technique that adds flexible temporal decay to RoPE, enabling extrapolation to  $16\times$  longer vision contexts. Furthermore, we propose long short-term context modeling, where a high-resolution short-term context window ensures fine-grained temporal consistency, while an unlimited long-term context window encodes long-range information using fewer tokens. With this approach, we can train on long video sequences with a manageable token context length. We demonstrate that FAR achieves state-of-the-art performance in both short- and long-video generation, providing a simple yet effective baseline for video autoregressive modeling.

## 1. Introduction

Advanced long-context autoregressive language models have demonstrated remarkable capabilities, enabling various applications that require test-time scaling, such as extended conversations [1], chain-of-thought reasoning [22,

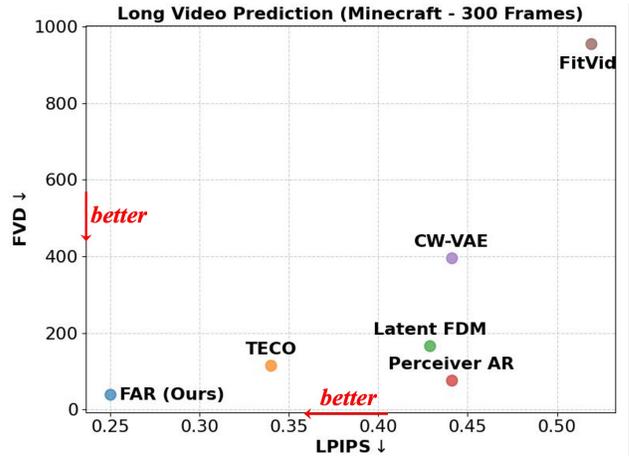


Figure 1. **Evaluation on Long Video Prediction.** FAR effectively exploits long video contexts and achieves accurate prediction.

42, 44], in-context learning [14], and retrieval-augmented generation [19]. However, video modeling has not achieved comparable progress. Recent video autoregressive modeling [32] directly adapts the paradigm of language models [8], where frames are factorized into discrete [16, 62] codes for next-token prediction (denoted as Token-AR). However, Token-AR still fails to achieve comparable quality to video diffusion transformers [7] due to the unidirectional modeling of visual tokens [18] and the irreparable information loss caused by vector quantization. On the other hand, video diffusion models [26, 27, 37, 55], which generate long videos using a progressive sliding window, struggle to effectively use earlier context.

In this paper, we introduce **Frame AutoRegressive (FAR)** model, specifically designed for video autoregressive modeling. FAR is trained using a frame-wise flow matching objective with autoregressive contexts. Unlike Token-AR, which learn causal dependencies between discrete tokens, FAR captures causal dependencies between continuous frames while still allowing full attention modeling within each frame. However, as a hybrid AR-Diffusion model [10, 31, 58], FAR also encounters a common issue observed in such models, namely, the discrepancy in ob-

\*Corresponding Author.

served contexts between training and inference. During training, later frames are exposed only to noised context frames due to diffusion objective, whereas inference relies on clean context frames. Recent methods [30, 65] mitigate this issue by appending a clean copy of the noised sequence during training, but this approach doubles the training cost.

To address the discrepancy of observed context, we propose training FAR with *stochastic clean context*. During training, we randomly replace a portion of noisy frames with clean frames and assign them a unique timestep embedding beyond the diffusion schedule to indicate representation extraction from clean context. During inference, this special embedding guides the model to effectively utilize clean context frames. We demonstrate FAR with stochastic clean context achieve same training efficiency to video diffusion transformers while achieving better convergence, served as a strong autoregressive video generation baseline.

Building on FAR, we investigate long-context video modeling and explore two common settings, similar to those in language modeling: (1) test-time lengthy extrapolation and (2) long-sequence training. Unlike language modeling, long-context video modeling suffers from visual redundancy. On one hand, RoPE [51] exhibits weak temporal decay, leading to the accumulation of redundant visual context and degrading test-time extrapolation performance. On the other hand, training on long videos is computationally expensive, as vision tokens grow significantly faster than language tokens.

To address these challenges, we propose to balance the locality and long-range dependency. For test-time temporal extrapolation, we introduce *FlexRoPE*, which incorporates a controllable temporal decay into RoPE using a linear bias. FlexRoPE is applied only at test time and remains compatible with models trained using RoPE. It enhances temporal locality and effectively reduces redundancy from distant contexts while still allowing the learned RoPE to model long-range dependencies. For long-video training, we introduce *long short-term context modeling*. Specifically, we maintain a high-resolution short-term context window to ensure fine-grained temporal consistency, while using a unlimited long-term context window with aggressive patchification to reduce redundant context tokens. This strategy enables efficient training on long video sequences with a manageable token context length.

Our contributions are summarized as follows:

1. We introduce FAR, an strong autoregressive video generation baseline, combined with stochastic clean context to bridge the training-inference gap in observed context.
2. Building on FAR, we introduce FlexRoPE to enable  $16\times$  test-time temporal extrapolation, and long short-term context modeling for efficient long-video training.
3. FAR achieves state-of-the-art performance in both short- and long-video generation.

## 2. Related Work

### 2.1. Video Generation

**Video Diffusion Models.** Recent advances in video generation have led to the scaling of video diffusion transformers [7, 33, 61] for text-to-video generation, resulting in superior visual quality. Pretrained text-to-video models are subsequently fine-tuned to incorporate images as conditions for image-to-video generation [23, 59, 61]. The trained image-to-video models can be utilized for autoregressive long-video generation using a sliding window [37, 55], but their ability to leverage visual context is limited by the sliding window’s size. In this work, we show that FAR achieves better convergence than video diffusion transformers for short-video generation while naturally supporting variable-length visual context.

**Token Autoregressive Models.** Video generation based on token autoregressive models (i.e., Token AR) aims to follow the successful paradigm of large language models. These models typically quantize continuous frames into discrete tokens [21, 62] and learn the causal dependencies between tokens using language models [28, 32]. While they achieve plausible performance, their generation quality remains inferior to that of video diffusion transformers due to information loss from vector quantization. Additionally, unidirectional visual token modeling may be suboptimal [18]. Subsequent studies have explored continuous tokens [34] without vector quantization but have not demonstrated their effectiveness in video generation. In this work, we show that FAR can learn causal dependencies from continuous frames and achieve better performance than Token AR in both short- and long-video modeling.

**Hybrid AR-Diffusion Models.** To leverage the strengths of both continuous latent spaces and autoregressive modeling, recent studies [40, 58, 64] have explored hybrid AR-Diffusion models. These models typically employ a diffusion objective for image-level modeling with autoregressive contexts. Hybrid AR-Diffusion models are widely applicable to both visual [10, 31, 58] and language generation [5, 57]. Recent research has also applied it in frame-level autoregressive modeling [10, 31] for video generation. However, they suffer from a training-inference discrepancy in the observed context. Some studies [30, 65] have attempted to mitigate this issue by maintaining a clean copy of the noised sequence during training, but this approach doubles the training cost. Among these methods, FAR efficiently addresses the training-inference gap through the proposed stochastic clean context, demonstrating its superior performance in long-context video modeling.

### 2.2. Long-Context Language Modeling

**Test-time Lengthy Extrapolation.** Lengthy extrapolation is an appealing characteristic of autoregressive models, al-

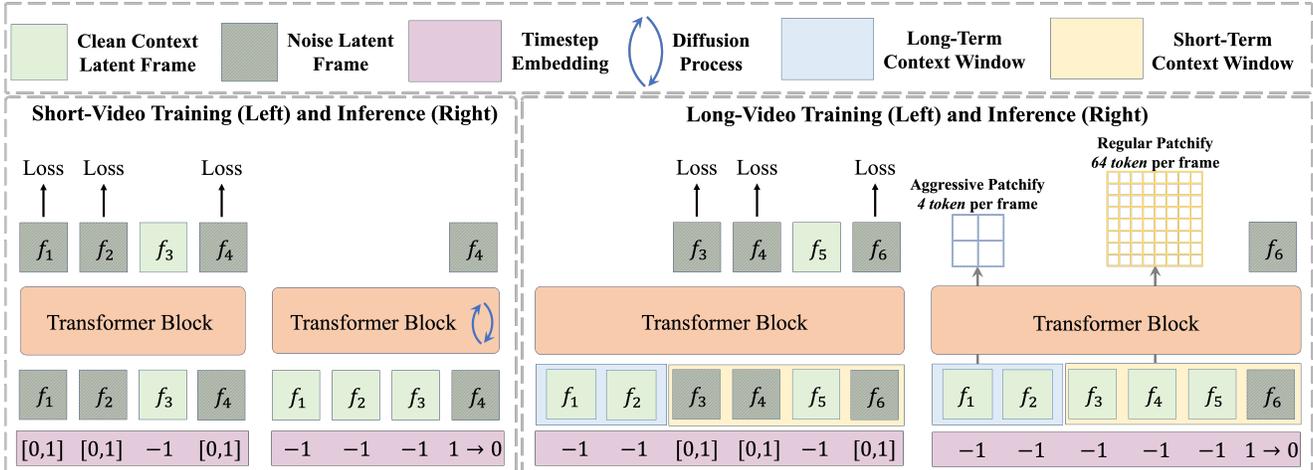


Figure 2. **Illustration of FAR’s Training and Inference Pipeline.** In short-video training, a portion of frames is randomly replaced with clean context frames, marked with a unique timestep embedding (e.g., -1) beyond the flow-matching scheduler. In long-video training, we adopt long short-term context modeling. A long-term context window with aggressive patchification is adopted to reduce redundant vision tokens, while a short-term context window is used to model fine-grained temporal consistency.

Table 1. **Model Variants of FAR.** We follow the model size configurations of DiT [46] and SiT [38].

Models	#Layers	Hidden Size	MLP	#Heads	Params
FAR-B	12	768	3072	12	130M
FAR-M	12	1024	4096	16	230M
FAR-L	24	1024	4096	16	457M
FAR-XL	28	1152	4608	18	674M
FAR-B-Long	12	768	3072	12	150M
FAR-M-Long	12	1024	4096	16	280M

lowing them to be trained on short sequences while performing inference on longer ones. However, extrapolation performance primarily depends on the characteristics of the position embedding. Two common relative position embeddings that support extrapolation are RoPE [51] and ALiBi [48]. RoPE encodes relative distance through dot-product operations, while ALiBi achieves this using attention bias. Subsequent studies [6, 47] have further advanced RoPE to enhance extrapolation performance. In this work, we introduce FlexRoPE and demonstrate its superior performance compared to RoPE and ALiBi in temporal extrapolation for video modeling.

**Long Sequence Training.** A straightforward approach to improving long-context modeling performance is to directly train the model on longer sequences. Recent work in language modeling has explored efficient long-sequence fine-tuning with position interpolation [12, 13]. However, vision tokens scale much faster than language tokens as context increases. To address this, we introduce long short-term context modeling to eliminate visual redundancy in long-video training.

### 2.3. Long-Context Video Modeling

Recent advancements in video generation models have enabled their use as interactive world simulators [9, 45, 53],

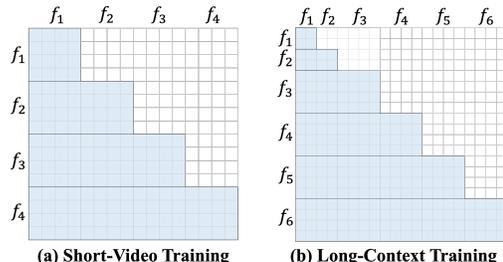


Figure 3. **Visualization of Attention Mask.** FAR enables full attention within a frame while maintaining causality at the frame level. In long-context training, we adopt aggressive patchification for long-term context frames to reduce tokens.

which require the ability to exploit long-range context and memorize the observed environment. However, existing video diffusion transformers lack effective mechanism to utilize long-range context. Although early work [60] has explored long-context video prediction, it has been limited in visual quality and long-range consistency. In this work, we introduce FAR, a scalable framework for both short- and long-context autoregressive video modeling.

## 3. Preliminary

### 3.1. Flow Matching

Flow Matching [2, 35, 36] is a simple alternative objective for training diffusion models. Rather than modeling the reverse process with stochastic differential equations, Flow Matching learns a continuous vector field that deterministically connect two distribution.

Specifically, given a data sample  $x_0 \sim p_{\text{data}}(x)$  and a noise sample  $x_1 \sim \mathcal{N}(0, I)$ , we construct a continuous tra-

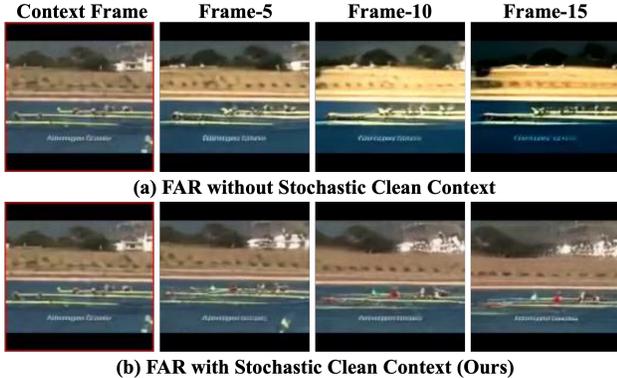


Figure 4. **Effect of Stochastic Clean Context.** This technique eliminate training-inference gap in observed context.

jectory connecting them via linear interpolation:

$$x(t) = (1-t)x_0 + tx_1, \quad t \in [0, 1]. \quad (1)$$

This formulation implies a constant velocity:

$$\frac{dx(t)}{dt} = v^* = x_1 - x_0. \quad (2)$$

To enable the model to learn the optimal transport between the data and noise distributions, we introduce a learnable time-dependent velocity field  $v_\theta(x, t)$ . During training, a random time  $t \sim U(0, 1)$  is sampled, and the model is optimized by minimizing the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, x_1, t} \left[ \|v_\theta(x(t), t) - v^*\|^2 \right]. \quad (3)$$

### 3.2. Autoregressive Models

Autoregressive models are a class of probabilistic models where each element in a sequence is conditioned on its preceding elements, denote as context. Formally, given a sequence of tokens  $(x_1, x_2, \dots, x_n)$ , an autoregressive model assumes that each token  $x_i$  is generated based on its previous tokens  $(x_1, x_2, \dots, x_{i-1})$ . The generative process can be expressed as a factorization of the joint probability:

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i | x_1, x_2, \dots, x_{i-1}). \quad (4)$$

By modeling each token conditioned on its preceding tokens, autoregressive models naturally capture the sequential dependencies inherent in data.

## 4. FAR

In this section, we first present the framework of FAR in Sec. 4.1. Then, we discuss the difficulties and solutions in training FAR in Sec. 4.2. In Sec. 4.3, we analyze the key design that enables FAR for long-context video modeling.

### 4.1. Framework Overview

**Architecture.** As shown in Fig. 2 (a), FAR is built upon the diffusion transformer [38, 46]. We adopt the model configuration of DiT [46] and Latte [39], as listed in Tab. 1. The key architectural difference between FAR and video

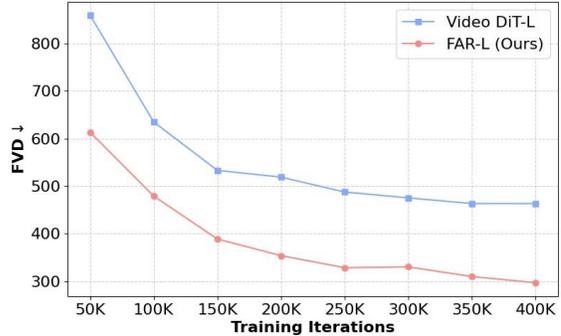


Figure 5. **Comparison of FAR and video diffusion transformer.** FAR achieves better convergence than video diffusion transformer in unconditional video generation on UCF-101.

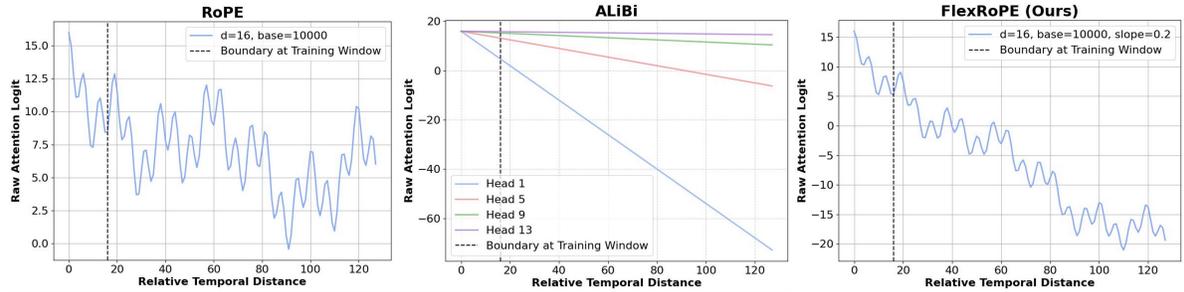
diffusion transformers (*e.g.*, Latte [39]) lies in the attention mechanism. As shown in Fig. 3(a), for each frame, we apply causal attention at the frame level while maintaining full attention within each frame. We adopt this causal spatiotemporal attention for all layers, instead of the interleaved spatial and temporal attention used in Latte. In FAR, image generation and image-conditioned video generation are jointly learned thanks to the causal mask, whereas video diffusion transformer [39] requires additional image-video co-training.

**Basic Training Pipeline.** The training pipeline of FAR is illustrated in Fig. 2 (a). Given a video sequence  $\mathbf{X}$ , we first employ a pretrained VAE to compress it into the latent space  $\mathbf{Z} \in \mathbb{R}^{T \times H \times W}$ , where  $T$ ,  $H$ , and  $W$  denote the number of frames, height, and width of the latent features, respectively. Note that although we primarily adopt an image VAE in this work, FAR can also be trained with a video VAE since our autoregressive unit is the latent frame. Following diffusion forcing [10], we independently sample a timestep for each frame. We then interpolate between the clean latent and the sampled noise using Eq. (1) and apply the frame-wise flow matching objective in Eq. (3) for learning. The key difference between FAR and image flow matching lies in that we adopt causal spatiotemporal attention, allowing each frame to access previous context frames during denoising.

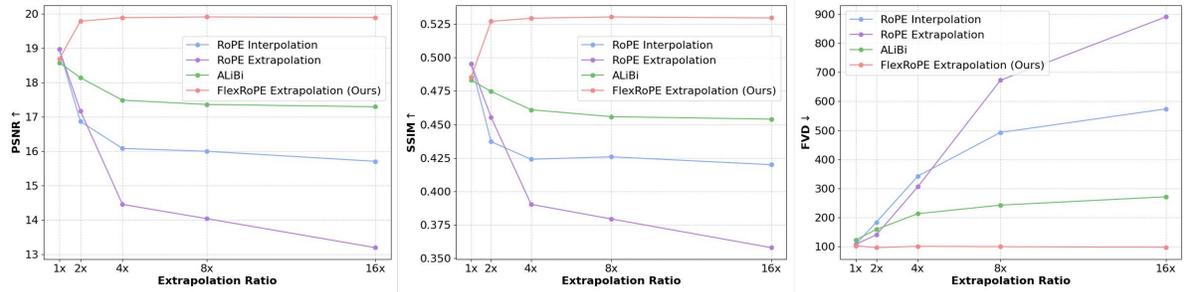
### 4.2. Short-Video Modeling

**Training-Inference Gap in Observed Context.** As a hybrid AR-diffusion model, FAR also encounters a training-inference gap in the observed context. As illustrated in Fig. 2(a), each clean latent is fused with sampled noise for the flow matching objective, as defined in Eq. (1). Consequently, later frames can only access the noised version of previous frames during training. However, during inference, this leads to a distribution shift when clean context frames is used.

As shown in the example in Fig. 4(a), the training-inference gap in the observed context leads to a distribution shift when inferring with a clean context. Although adding



(a) Visualization of Temporal Decay Pattern



(b) Lengthy Extrapolation Performance Comparison

Figure 6. **Visualization and Comparison of Various Temporal Position Embeddings.** The proposed FlexRoPE incorporates a linear bias to induce controllable temporal decay at test time, enhancing extrapolation performance as context increases. In contrast, other methods degrade when the inference context exceeds the training window.

mild noise to the context during inference can help mitigate this effect, it still causes low-level flickering, degrading the quality of the generated video. Recent works [30, 65] attempt to address this issue by maintaining a clean copy of the noised sequence during training. However, this approach doubles the training costs.

**Our Solution: Stochastic Clean Context.** To bridge the gap in observed context, we introduce stochastic clean context for training FAR. As illustrated in Fig. 2(a), we randomly replace a portion of the noised frames with their corresponding clean context and assign them a unique timestep embedding (e.g., -1) beyond the flow-matching timestep scheduler. These clean context frames are excluded from loss computation and are implicitly learned through later frames that use them as context. During inference, this unique timestep embedding guides the model to use clean context effectively.

Training FAR with stochastic clean context does not add extra computation and does not conflict with different timestep sampling strategies during training (e.g., logit-normal sampling [17]). It effectively resolves the training-inference discrepancy, as exemplified in Fig. 4(b).

**FAR vs. Video Diffusion Transformer.** FAR and video diffusion transformer differ only in their training schemes. FAR is trained with independent noise and causal attention, while the video diffusion transformer is trained with uniform noise and full attention. This raises an interesting question: *Can FAR surpass video diffusion transformers?*

To explore this, we convert FAR to video diffusion transformer as a baseline, denoted as Video DiT. We align the training settings to compare the two paradigms. As shown in Fig. 5, FAR achieves better convergence than the Video DiT, demonstrating its potential to become a strong baseline for autoregressive video modeling.

### 4.3. Long-Context Video Modeling

In this section, we discuss the challenge of long-context video modeling. We focus on two practical long-context settings, similar to those in language models: test-time temporal extrapolation (Sec. 4.3.1) and long-video training (Sec. 4.3.2).

#### 4.3.1. Test-Time Temporal Extrapolation

**Weak Temporal Decay.** An appealing characteristic of autoregressive models is their potential to be trained on short sequences while being tested on long sequences, enabling lengthy extrapolation at test time. This capability relies on effective positional embedding. Following 3D-RoPE [61] for video data, the spatial and temporal dimensions (i.e., height, width, and frame) are treated as independent 1D-RoPE embeddings. Consequently, we keep RoPE positional embedding for height and width unchanged while focusing only on the temporal position embedding. In this study, we examine RoPE [51] and ALiBi [48], two common positional embedding methods for capturing temporal relative distances. From the visualization in Fig. 6(a), RoPE does

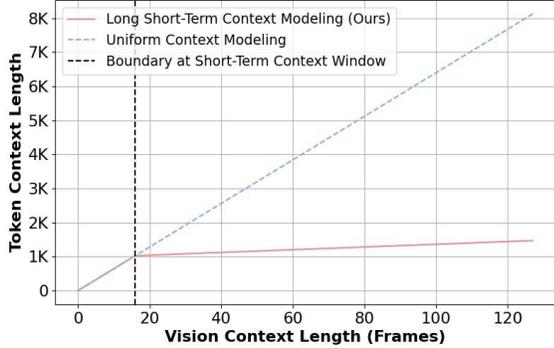


Figure 7. **Relation between Token Context Length and Vision Context Length.** With the proposed long short-term context modeling, the token context length scales more slowly as the vision context length increases compared to uniform context modeling.

not impose sufficient temporal decay, leading to accumulated redundant visual context. Similarly, ALiBi exhibits this issue in parts of attention heads with small slopes.

To evaluate extrapolation performance, we gradually increase the number of context frames at test time and measure the resulting improvement in predictions. From Fig. 6(b), position interpolation of RoPE performs slightly better than position extrapolation. Additionally, we retrain the model using ALiBi as temporal position embedding. ALiBi applies a linear decay based on temporal relative distance, with different decay rates assigned to each attention head. Our results suggest that ALiBi achieves slightly better extrapolation than RoPE due to its explicit temporal decay mechanism. However, all solutions exhibit performance degradation as vision context increases. Therefore, we aim to develop a more effective temporal position embedding method to improve temporal extrapolation.

**Our Solution: FlexRoPE.** To address this problem, we propose FlexRoPE, which explicitly controls temporal decay to suppress redundant visual context while still allowing the model to capture long-range dependencies using RoPE. The FlexRoPE is defined as:

$$\text{Attention}(q_i, k_j) = \text{Softmax} \left( \underbrace{\overbrace{\text{RoPE}(q_i, k_j)}^{\text{used in training}} - \lambda \cdot |i - j|}_{\text{FlexRoPE, used in inference}} \right), \quad (5)$$

where the temporal decay is flexibly controlled by the slope  $\lambda$ , and  $i$  and  $j$  represent the temporal indices of the frame. We visualize FlexRoPE with  $\lambda = 0.2$  in Fig. 6(a). FlexRoPE is inference-compatible with models trained using RoPE since it does not modify the dot-product computation but instead compensates for RoPE’s temporal decay.

We compare FlexRoPE with RoPE in Fig. 6(b). FlexRoPE effectively balances locality and long-range correspondence, leading to improved performance as the context frame increases, whereas RoPE interpolation and extrapolation exhibit poorer extrapolation performance.

Table 2. **Quantitative Comparison of Conditional and Unconditional Video Generation on UCF-101.** We follow the evaluation setup of Latte [39]. † denotes FVD reported on 10,000 videos.

Methods	Type	Params	Double Train Cost	Cond. Gen FVD <sub>2048</sub> ↓	Uncond. Gen FVD <sub>2048</sub> ↓
<b>Resolution-128×128</b>					
MAGViT <sub>v2</sub> -MLM [62]	Non-AR	307 M	✗	58†	-
MAGViT <sub>v2</sub> -AR [62]	Token-AR	840 M	✗	109†	-
TATS [20]	Token-AR	331 M	✗	332	420
FAR-L (Ours)	Frame-AR	457 M	✗	<b>99 (57†)</b>	<b>280</b>
<b>Resolution-256×256</b>					
LVDM [26]	Video-DiT	437 M	✗	-	372
Latte [39]	Video-DiT	674 M	✗	-	478
CogVideo [28]	Token-AR	9.4 B	✗	626	-
OmniTokenizer [56]	Token-AR	650 M	✗	191	-
ACDIT [30]	Frame-AR	677 M	✓	111	-
MAGI [65]	Frame-AR	850 M	✓	-	421
FAR-L (Ours)	Frame-AR	457 M	✗	113	303
FAR-XL (Ours)	Frame-AR	674 M	✗	<b>108</b>	<b>279</b>

### 4.3.2. Long-Video Training

**Token Redundancy in Long Video.** Visual data contains spatial redundancy, causing vision tokens to expand much faster than language tokens as context increases. For example, a video sequence of 128 frames requires more than 8K tokens, with 64 tokens per frame, as illustrated in Fig. 7. As a result, training and inference on long videos become computationally inefficient.

**Our Solution: Long Short-Term Context Modeling.**

To address the token redundancy in video, we introduce long short-term context modeling, which exploits the spatial and temporal locality in video data. As illustrated in Fig. 2(b), we maintain a high-resolution short-term context window to learn fine-grained temporal consistency and a low-resolution long-term context window, where we adopt aggressive patchification to reduce the number of context tokens. During training, given that the data has a maximum sequence length of  $m$  frames, we fix the short-term context window to  $n$  frames and randomly sample the long-term context frames from the range  $[0, m - n]$ . The attention mask with long short-term context modeling is shown in Fig. 3(b), where the long-term context uses fewer tokens. As demonstrated in Fig. 7, this strategy ensures that increasing the vision context length maintains a manageable token context length. To prevent interference between long-term and short-term contexts, we adopt separate projection layers for each context, inspired by MM-DiT [17]. This approach results in a slightly larger parameter size during long-video training, as shown in Tab. 1.

## 5. Experiment

### 5.1. Implementation Details

We follow the DiT’s structure [46] to implement FAR. To compress video latents, we train a series of image DC-AE [11] on the corresponding dataset, resulting in 64 tokens per frame. All models are trained from scratch without image pretraining. We provide detailed settings in Tab. 7.

Table 3. **Quantitative Comparison on Short Video Prediction.** We follow the evaluation setup of MCVD [54] and ExtDM [63], where  $c$  denotes the number of context frames and  $p$  denotes the number of predicted frames.

Methods	Params	$c = 4, p = 12$				Methods	Params	$c = 2, p = 14$				$c = 2, p = 28$			
		SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$			SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
RaMViD [29]	235 M	0.639	21.37	0.090	396.7	RaMViD [29]	235 M	0.758	17.55	0.085	166.5	0.691	16.51	0.109	238.7
LFDM [43]	108 M	0.627	20.92	0.098	698.2	LFDM [43]	108 M	0.770	17.45	0.084	167.6	0.730	16.68	0.106	276.8
MCVD-cp [54]	565 M	0.658	21.82	0.088	468.1	VIDM [41]	194 M	0.763	16.97	0.080	131.7	0.728	16.20	0.096	194.6
ExtDM-K2 [63]	119 M	0.754	23.89	0.056	394.1	MCVD-cp [54]	565 M	0.838	19.10	0.075	87.8	0.797	17.70	0.078	119.0
FAR-B (Ours)	130 M	<b>0.818</b>	<b>25.64</b>	<b>0.037</b>	<b>194.1</b>	ExtDM-K4 [63]	121 M	0.845	20.04	0.053	<b>81.6</b>	0.814	18.74	0.069	<b>102.8</b>
						FAR-B (Ours)	130 M	<b>0.849</b>	<b>20.87</b>	<b>0.038</b>	99.3	<b>0.819</b>	<b>19.40</b>	<b>0.049</b>	144.3

(a) Evaluation on UCF-101 ( $64 \times 64$ )

(b) Evaluation on BAIR ( $64 \times 64$ )

Table 4. **Quantitative Comparison on Long-Context Video Prediction.** We follow the evaluation setup of TECO [60], where  $c$  denotes the number of context frames and  $p$  denotes the number of predicted frames.

Methods	Params	$c = 144, p = 156$				Methods	Params	$c = 144, p = 156$				$c = 36, p = 264$			
		SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$			SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
FitVid [3]	165 M	0.356	12.0	0.491	176	FitVid [3]	176 M	0.343	13.0	0.519	956				
CW-VAE [49]	111 M	0.372	12.6	0.465	125	CW-VAE [49]	140 M	0.338	13.4	0.441	397				
Perceiver AR [25]	30 M	0.304	11.2	0.487	96	Perceiver AR [25]	166 M	0.323	13.2	0.441	76				
Latent FDM [24]	31 M	0.588	17.8	0.222	181	Latent FDM [24]	33 M	0.349	13.4	0.429	167				
TECO [60]	169 M	<b>0.703</b>	21.9	0.157	<b>48</b>	TECO [60]	274 M	0.381	15.4	0.340	116				
FAR-B-Long (Ours)	150 M	0.687	<b>22.3</b>	<b>0.104</b>	64	FAR-M-Long (Ours)	280 M	<b>0.448</b>	<b>16.9</b>	<b>0.251</b>	<b>39</b>				

(a) Evaluation on DMLab ( $64 \times 64$ )

(b) Evaluation on Minecraft ( $128 \times 128$ )

## 5.2. Quantitative Comparison

### 5.2.1. Video Generation

**Dataset and Evaluation Setting.** We benchmark both unconditional and conditional video generation on the UCF-101 dataset [50], which consists of approximately 13,000 videos. Following Latte [39], we use the entire dataset for training. For evaluation, we randomly sample 2,048 videos to compute FVD [52] against the ground-truth videos. For conditional video generation, we set the guidance scale to 2.0 during inference.

**Main Results.** From the results listed in Tab. 2, we achieve state-of-the-art performance in both unconditional and conditional video generation. Specifically, Latte [39] is based on video diffusion transformer, while OmniTokenizer [56] is based on Token AR. Our method significantly outperforms both. Furthermore, compared to recent frame-autoregressive models [30, 65], which require twice the training cost, FAR achieves superior performance without any additional training cost.

### 5.2.2. Short-Video Prediction

**Dataset and Evaluation Settings.** We evaluate FAR on the UCF-101 [50] and BAIR [15] datasets, following the evaluation settings in MCVD [54] and ExtDM [63]. We randomly sample 256 videos based on provided context frames, each with 100 different trajectories, and select the best trajectory to compute pixel-wise metrics. For FVD, we report the average over all trajectories.

**Main Results.** We summarize the results in Tab. 3. Unlike previous works such as MCVD [54] and ExtDM [63], which introduce complex multi-scale fusion strategies and optical flow, FAR achieves superior results on both datasets without requiring additional design.

### 5.2.3. Long-Video Prediction

**Dataset and Evaluation Settings.** We benchmark long-context video modeling results on action-conditioned video prediction using the Minecraft and DMLab datasets [60]. The Minecraft dataset contains approximately 200K videos, while the DMLab dataset contains about 40K videos. Each video consists of 300 frames with action annotations. We follow the evaluation setup in TECO [60], which uses 144 observed context frames to predict 156 future frames and compute pixel metrics. Additionally, we compute FVD on 264 generated frames based on 36 context frames.

**Main Results.** We summarize the results in Tab. 4. The previous work, TECO [60], adopts aggressive downscaling for all frames to reduce tokens for temporal modeling, creating a trade-off between training efficiency and prediction accuracy. In contrast, FAR employs long short-term context modeling, effectively achieving the lowest prediction error (*i.e.*, LPIPS) without prohibitive computation cost.

## 5.3. Qualitative Comparison

We present a qualitative comparison of long-video prediction in Fig. 8. Compared to previous methods, FAR effectively utilizes the observed context and generates predictions that most closely resemble the ground truth, demonstrating its ability to leverage long-range context.

## 5.4. Ablation Study

**Stochastic Clean Context.** We have visualized the effectiveness of stochastic clean context in Fig. 4. Based on the quantitative evaluation of video prediction in Tab. 5, FAR with stochastic clean context achieves significantly improved performance.

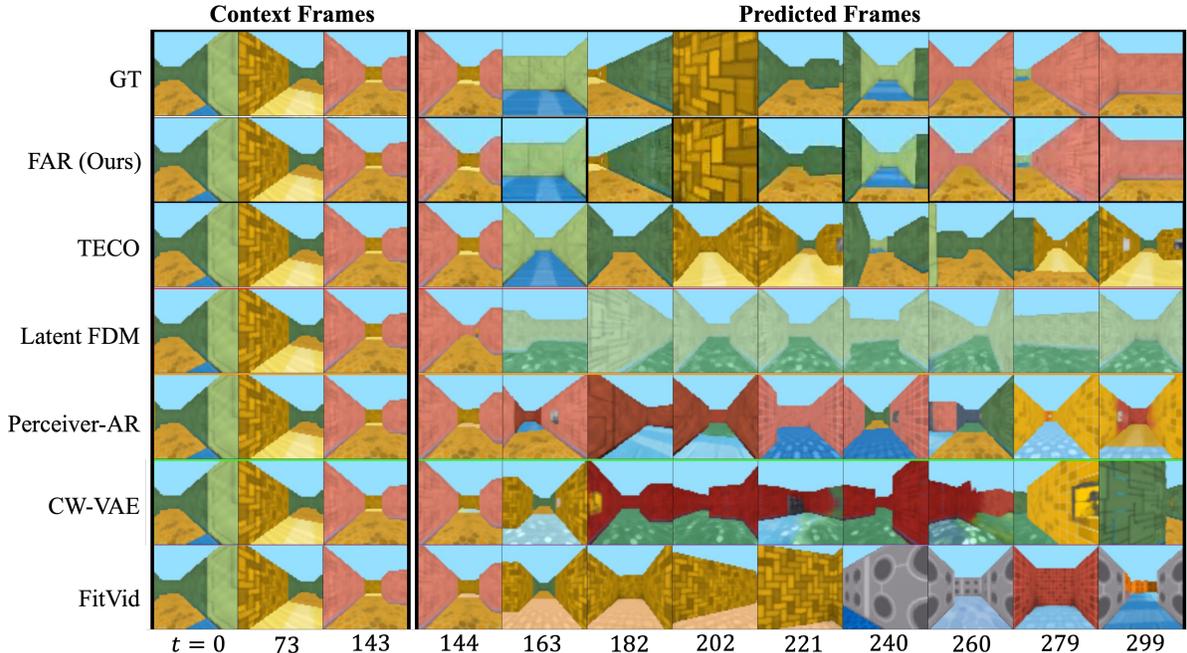


Figure 8. **Qualitative Comparison of Long-Context Video Prediction on DMLab.** FAR fully utilizes the long-range context (144 frames), resulting in more consistent prediction (156 frames) compared to previous methods.

Table 5. **Ablation Study of Stochastic Clean Context on UCF-101.** Stochastic clean context mitigates the training-inference discrepancy in observed context, leading to improved performance.

Methods	$c = 1, p = 15$			
	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
FAR w/o. Stochastic Clean Context	0.540	16.42	0.211	399
FAR w/. Stochastic Clean Context	<b>0.596</b>	<b>18.46</b>	<b>0.187</b>	<b>347</b>

Table 6. **Ablation Study on the Resolution of Long-Term Context.** The speed is averaged over the generation of 300 frames.

Context Resolution	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	Speed (fps)
$1 \times 1$	0.411	15.25	0.312	40	2.94
$2 \times 2$	0.423	15.84	0.291	40	2.88
$4 \times 4$	0.433	16.26	0.276	37	1.42

**Long-Term Context Resolution.** We investigate the impact of long-term context resolution on prediction accuracy and inference speed. As the context resolution increases, pixel-level metrics improve; however, the overall video quality remains similar. Nonetheless, inference speed significantly degrades at higher context resolutions due to the increased number of tokens involved in computation. Therefore, we select a  $2 \times 2$  resolution for the long-term context as a balance between computational efficiency and long-term context performance.

**Short-Term Context Window Size.** We evaluate the impact of the short-term context window size on performance. As shown in Fig. 9, video quality (FVD) quickly saturates as the short-term context window size increases, while pixel-level metrics continue to improve but also approach saturation at a window size of 16. Therefore, we set the short-term context window size to 16 by default.

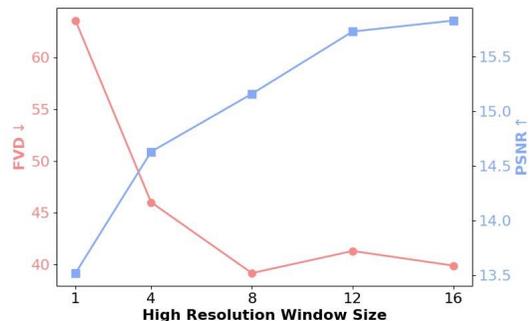


Figure 9. **Ablation Study of the Short-Term Context Window Size.** Performance saturates as the window size increases.

## 6. Conclusion

In this paper, we introduce FAR for autoregressive video modeling. FAR enables learning causal dependency of continuous frames and demonstrates better convergence than video diffusion transformers. Building upon FAR, we explore its application in long-context video modeling. We identify visual redundancy as a key challenge, leading to poor temporal extrapolation performance at test time and inefficient long-video training. To address these issues, we propose balancing locality and long-range dependency. Specifically, we introduce FlexRoPE to enable  $16 \times$  temporal extrapolation at test time and incorporate long short-term context modeling for efficient long-video training. FAR achieves state-of-the-art performance on both short- and long-video modeling, highlighting its potential as a new foundation model for video generation.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 3
- [3] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021. 7
- [4] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 12
- [5] Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv e-prints*, pages arXiv–2412, 2024. 2
- [6] bloc97. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., 2023. 3
- [7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [9] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025. 1, 2, 4
- [11] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 6, 13
- [12] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 3
- [13] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhi-jian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023. 3
- [14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 1
- [15] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *CoRL*, 12(16):23, 2017. 7
- [16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1
- [17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 5, 6
- [18] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. 1, 2
- [19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023. 1
- [20] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 6
- [21] Yuchao Gu, Xintao Wang, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Rethinking the objectives of vector-quantized tokenizers for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7631–7640, 2024. 2
- [22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1
- [23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2
- [24] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 7
- [25] Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander

- Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. General-purpose, long-context autoregressive modeling with perceiver ar. In *International Conference on Machine Learning*, pages 8535–8558. PMLR, 2022. 7
- [26] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1, 6
- [27] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1
- [28] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 6
- [29] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 7
- [30] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acddit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024. 2, 5, 6, 7
- [31] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 1, 2
- [32] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vignesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 1, 2
- [33] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2
- [34] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2025. 2
- [35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3
- [36] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [37] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 1, 2
- [38] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 3, 4
- [39] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 4, 6, 7, 13
- [40] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024. 2
- [41] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9117–9125, 2023. 7
- [42] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. 1
- [43] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023. 7
- [44] OpenAI. Learning to reason with llms, 2024. 1
- [45] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 3
- [46] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3, 4, 6
- [47] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023. 3
- [48] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 3, 5, 12
- [49] Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *Advances in Neural Information Processing Systems*, 34:29246–29257, 2021. 7
- [50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7
- [51] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 3, 5, 12
- [52] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 7

- [53] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 3
- [54] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 7
- [55] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 1, 2
- [56] Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024. 6, 7
- [57] Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974, 2023. 2
- [58] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 2
- [59] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 2
- [60] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *International Conference on Machine Learning*, pages 39062–39098. PMLR, 2023. 3, 7, 13
- [61] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 5
- [62] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion—tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 1, 2, 6
- [63] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19310–19320, 2024. 7, 13
- [64] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2
- [65] Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, Xiangyu Zhang, Lionel M Ni, et al. Taming teacher forcing for masked autoregressive video generation. *arXiv preprint arXiv:2501.12389*, 2025. 2, 5, 6, 7

## 7. Appendix

### 7.1. Experimental Settings

As shown in Tab. 7, we list the detailed training and evaluation configurations of FAR. For the ablation study in this paper, we halve the training iterations while keeping other settings the same.

### 7.2. Qualitative Comparison

We provide additional visualization of long-video prediction results on DMLab and Minecraft in Fig. 10 and Fig. 11. From the results, FAR better exploits the provided context and provides more consistent results in later predictions compared to previous works.

### 7.3. Ablation Study of FlexRoPE

We compare different position embeddings on  $16\times$  temporal extrapolation. We focus on two settings: The first is unique in video generation, where we directly unroll a  $16\times$  longer sequence from 1 context frame. Second, we follow the long-context language model [48], gradually adding more context frames and comparing the last 16-frame predictions.

**$16\times$  Extrapolation:  $c = 1$ ,  $p = 255$ .** In this evaluation, we directly generate videos that are  $16\times$  longer than the training sequence length (*i.e.*, 16 frames), using only the first frame as a condition. As shown in Fig. 12, while RoPE extrapolation can adapt to periodic motion extrapolation, our proposed FlexRoPE achieves superior results in both periodic and non-periodic extrapolation.

**$16\times$  Extrapolation:  $c = 240$ ,  $p = 16$ .** Following the common practice to evaluate long-context ability in language models, we collect 256 frames, use different context frames [0, 240], and allow the model to infer the last 16 frames to test performance. The quantitative results are demonstrated in Fig. 6 in the main paper. In Fig. 13, we visualize the  $16\times$  temporal extrapolation inference results. We can see that RoPE [51] (PE, position extrapolation) at test time results in the worst performance, accumulating redundant context and failing to extrapolate. Meanwhile, RoPE (PI, positional interpolation) breaks the learned video speed, resulting in poor motion. Although ALiBi [48] performs better than RoPE (PI and PE), it still influences the learned motion distribution and falls far from the GT. Compared to these methods, FlexRoPE achieves the best temporal extrapolation results.

### 7.4. Limitations and Future Work

#### 7.4.1. Limitations

The primary limitation lies in the lack of scaled-up experiments. Although FAR demonstrates great potential, we still lack large-scale training on million-level text-to-video

generation datasets. Additionally, restricted by the available datasets, we only experiment with FAR on up to 300 frames (about 20 seconds), not fully investigating its ability on minute-level videos.

#### 7.4.2. Future Work

One future direction is to scale up FAR to benchmark it against video diffusion transformers on large-scale text-to-video generation tasks. Additionally, we plan to simulate a longer video dataset (minute-level) for better evaluating the model’s long-context ability. Finally, it would be interesting to explore whether FAR can adapt to sequential vision modeling beyond videos, for example, the image sequences as demonstrated in LVM [4].

Table 7. **Experimental Configurations of FAR.** We follow the evaluation settings from Latte [39], MCVD [63], and TECO [60].

Hyperparameters	Short-Video Generation		Short-Video Prediction		Long-Video Prediction	
	Cond. UCF-101	Uncond. UCF-101	BAIR	UCF-101	Minecraft	DMLab
<b>Dataset Configuration</b>						
Resolution	256/128	256/128	64	64	128	64
Total Training Samples	13,320	13,320	43,264	9,624	194,051	39,375
<b>Training Configuration</b>						
Batch Size	32	32	32	32	32	32
Latent Size	8×8 (DC-AE [11])	8×8 (DC-AE [11])	8×8 (DC-AE [11])	8×8 (DC-AE [11])	8×8 (DC-AE [11])	8×8 (DC-AE [11])
Training Sequence Length	16	16	32	16	300	300
LR	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$
LR Schedule	constant	constant	constant	constant	constant	constant
Warmup Steps	-	-	-	-	10K	10K
Total Training Steps	400K	400K	200K	200K	1M	1M
Stochastic Clean Context	0.1	0.1	0.1	0.1	0.1	0.1
Short-Term Context Window	16	16	32	16	16	16
Long-Term Context Resolution	-	-	-	-	2×2	2×2
<b>Evaluation Configuration</b>						
Samples	4×2048	4×2048	100×256	100×256	4×256	4×256
Guidance Scale	2.0	-	-	-	1.5	1.5
Reference Work	Latte [39]	Latte [39]	MCVD [63]	MCVD [63]	TECO [60]	TECO [60]

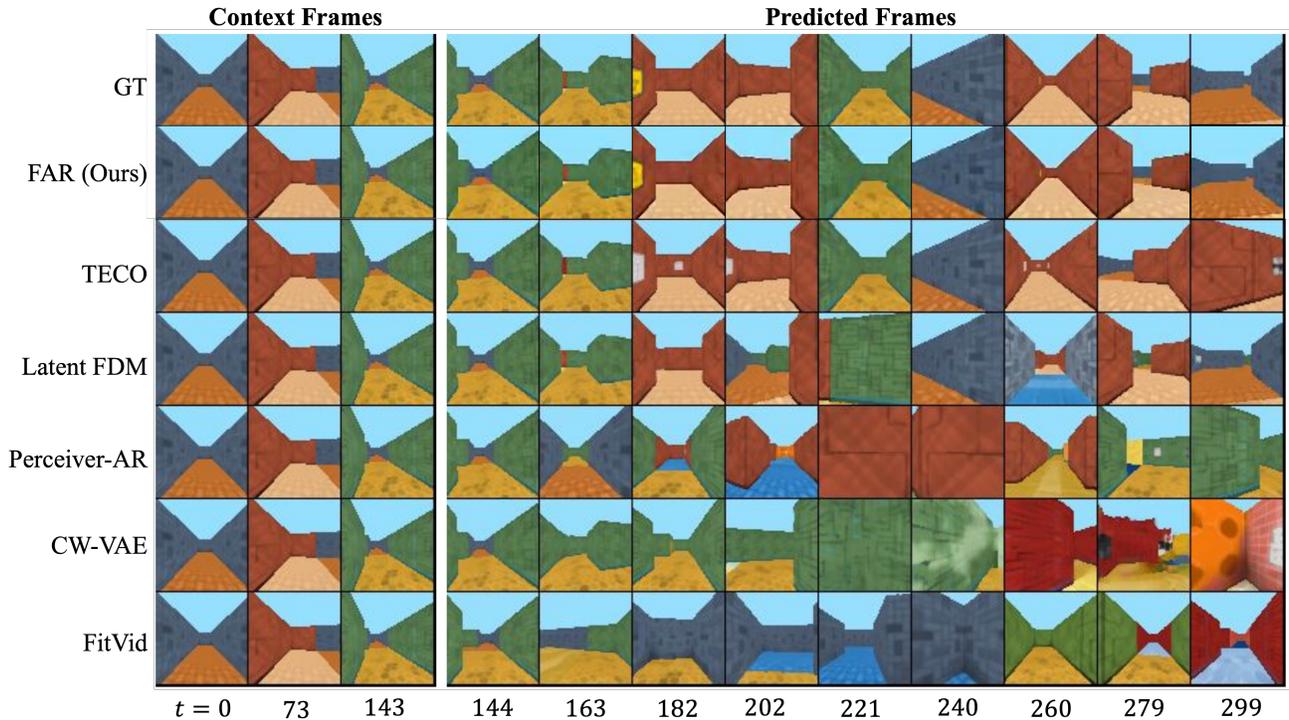


Figure 10. **Qualitative Comparison of Long-Context Video Prediction on DMLab.** FAR fully utilizes the long-range context (144 frames), resulting in more consistent prediction (156 frames) compared to previous methods.

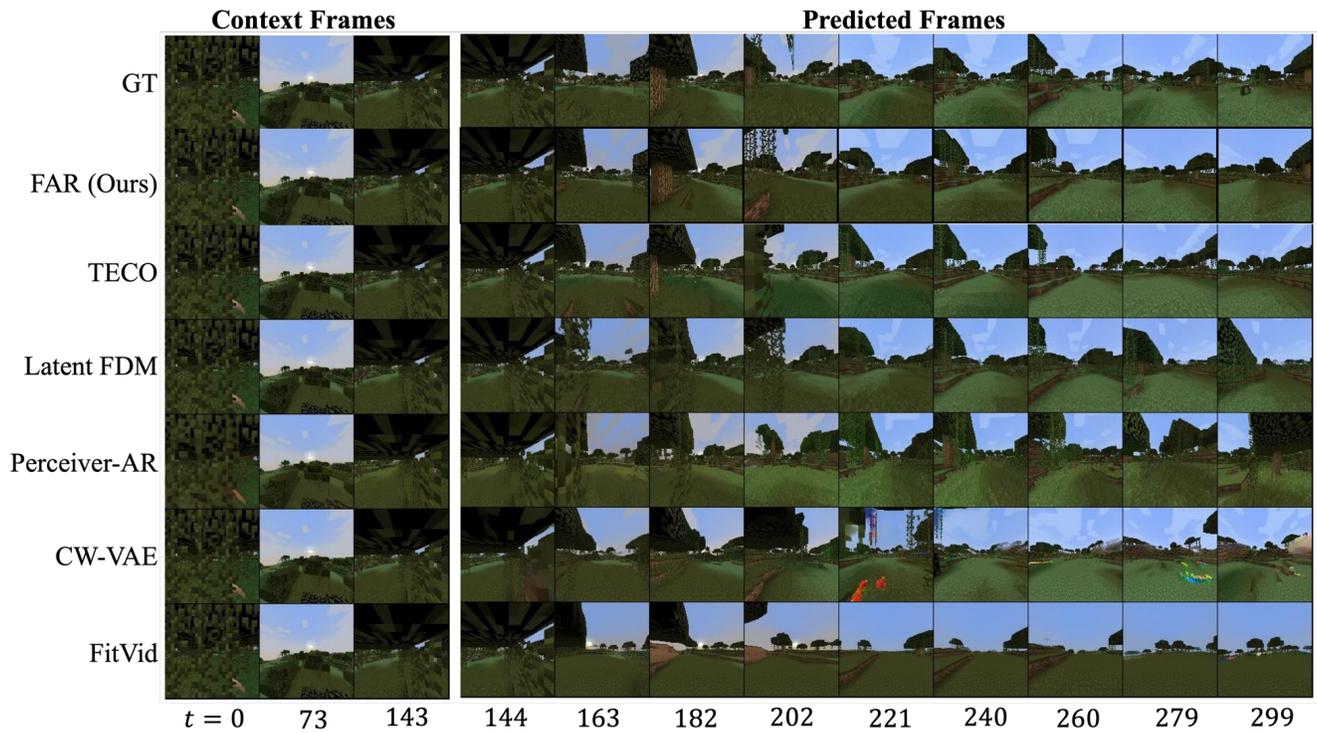


Figure 11. **Qualitative Comparison of Long-Context Video Prediction on Minecraft.** FAR fully utilizes the long-range context (144 frames), resulting in more consistent prediction (156 frames) compared to previous methods.

Figure 12. **Comparison of Position Embeddings for  $16\times$  Temporal Extrapolation.** We leverage the model (trained on 16 frames) to infer 255 future frames based on the provided 1 context frames. PE denotes position extrapolation. We encourage readers to [click and play](#) the video clips in this figure using Adobe Acrobat.

Figure 13. **Comparison of Position Embeddings for  $16\times$  Temporal Extrapolation.** We leverage the model (trained on 16 frames) to infer 16 future frames based on the provided 240 context frames. PI denotes position interpolation, and PE denotes position extrapolation. We encourage readers to [click and play](#) the video clips in this figure using Adobe Acrobat.