End-to-End Training for Autoregressive Video Diffusion via Self-Resampling

Yuwei Guo[1], Ceyuan Yang[2†], Hao He[1], Yang Zhao[2], Meng Wei[3], Zhenheng Yang[3], Weilin Huang[2], and Dahua Lin[1]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] ByteDance Seed, China
[3] ByteDance, China

Abstract. Autoregressive video diffusion models hold promise for world simulation but are vulnerable to exposure bias arising from the train–test mismatch. While recent works address this via post-training, they typically rely on a bidirectional teacher model or discriminator. To achieve an end-to-end solution, we introduce *Resampling Forcing*, a teacher-free framework that enables training autoregressive video models from scratch and at scale. Central to our approach is a self-resampling scheme that simulates inference-time model errors on history frames during training. Conditioned on these degraded histories, a sparse causal mask enforces temporal causality while enabling parallel training with frame-level diffusion loss. To facilitate efficient long-horizon generation, we further introduce history routing, a parameter-free mechanism that dynamically retrieves the top-k most relevant history frames for each query. Experiments demonstrate that our approach achieves performance comparable to distillation-based baselines while exhibiting superior temporal consistency on longer videos owing to native-length training. See our Project Page for more details.

1 Introduction

Recent advances in generative video models have demonstrated strong potential for world modeling by approximating physical dynamics and predicting future states conditioned on current observations [2, 7, 28, 36]. Realizing this vision necessitates an autoregressive video generation paradigm that predicts the next frame conditioned on past context, thereby mirroring the *strict causal* nature of the physical world. Beyond world simulation, such a paradigm enables a diverse array of applications, spanning game simulation [1, 62, 78, 85], interactive content creation [42, 54], and temporal reasoning [68].

Despite its conceptual elegance, autoregressive video generation poses significant challenges. The primary hurdle is exposure bias [48, 53]: under teacher forcing, the model conditions on ground-truth histories during training, yet must rely on its own generated outputs during inference. This train–test mismatch can

† Corresponding author.

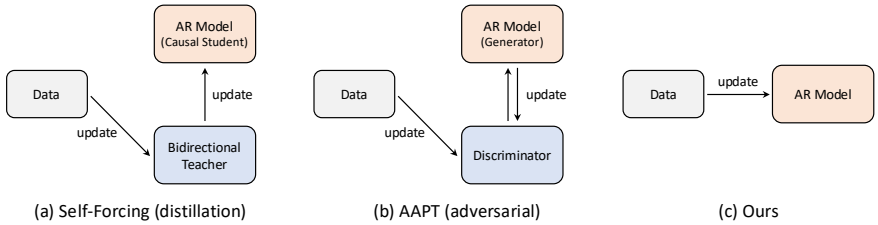

Fig. 1: To obtain an autoregressive video diffusion model (AR model), (a) Self Forcing [32] first trains a bidirectional teacher and then distills it into a causal student, while (b) AAPT [42] adopts adversarial training by jointly optimizing the generator and discriminator. (c) We propose an end-to-end approach that directly trains the target model, without relying on any auxiliary model.

induce error accumulation, where small artifacts in model predictions are amplified across the autoregressive rollout, potentially leading to catastrophic video collapse (see Fig. 2 top). Furthermore, the ever-expanding historical context in autoregressive generation exacerbates attention complexity, posing practical obstacles for both training and inference over long horizons.

To mitigate the train–test mismatch, recent works [32, 42] employ post-training strategies aimed at aligning the generated video distribution with real data (Fig. 1). For instance, Self Forcing [32] first autoregressively rolls out full videos, subsequently applying distillation or adversarial objectives to enforce distribution matching. By simulating inference during training, they reduce the discrepancy between training and test conditions. However, the reliance on a bidirectional teacher or an online discriminator impedes scalable training of autoregressive video models from scratch. A bidirectional teacher can also leak future information, compromising the strict temporal causality of the student model. Additionally, extensions to longer sequences typically use simple sliding-window attentions that disregard the varying importance of historical context, which may undermine long-term consistency.

In this work, we present *Resampling Forcing*, an end-to-end training framework for autoregressive video diffusion models. Drawing inspiration from the next-token prediction objective in LLMs, we condition each frame on its clean history and train in parallel via a per-frame diffusion loss under causal masking. We posit that, to mitigate error propagation and amplification, the model must be trained for robustness against input perturbations while retaining a clean prediction objective. To this end, the core of our method is a self-resampling mechanism: the model first induces errors in the history frames, then uses this degraded history to condition next-frame prediction. To simulate inference-time model errors, we autoregressively resample the latter segment of each frame's denoising trajectory with the online model weights. This process is detached from gradient backpropagation to avoid shortcut learning. In addition, we introduce a history routing mechanism that dynamically retrieves the top-k most

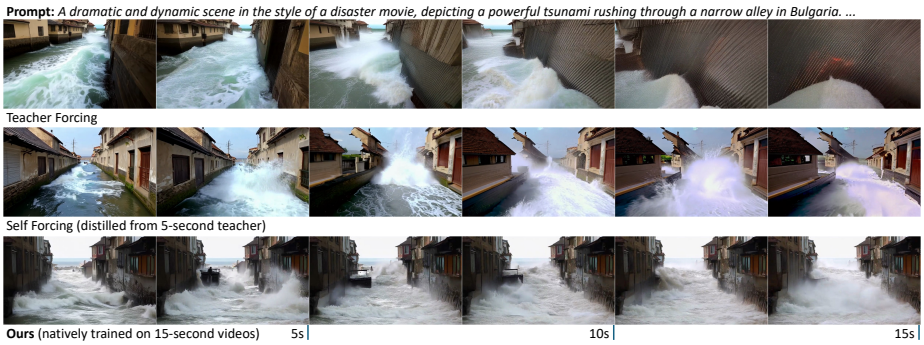

Fig. 2: We introduce *Resampling Forcing*, an end-to-end, teacher-free training framework for autoregressive video diffusion models. *Top*: The teacher forcing accumulates errors and leads to video collapse. *Middle*: Distilled from a short bidirectional teacher, Self Forcing suffers from the degraded quality on longer videos. *Bottom*: Our method offers stable quality by native training on long videos.

relevant history frames via a parameter-free router, maintaining a near-constant attention complexity in long-horizon rollout.

Empirical results demonstrate that *Resampling Forcing* effectively mitigates error accumulation in autoregressive video diffusion models, achieving generation quality comparable to state-of-the-art distilled models. Leveraging native long-video training, our approach outperforms extrapolated baselines in longer video generation. Furthermore, our model exhibits stricter adherence to causal dependencies compared to distillation baselines. We also demonstrate that our history routing mechanism attains a sparse context with negligible quality loss, offering a viable memory design for long-horizon generation. We anticipate that this work will advance scalable training and long-term memory for future video world models.

2 Related Works

Bidirectional Video Generation. Bidirectional video generation refers to non-causal models that synthesize all frames jointly, allowing each frame to attend to both past and future context. Early efforts leveraged GANs [55, 61] or adapted UNet-based text-to-image diffusion models [3, 5, 6, 25]. Inspired by Sora [7], the field shifted toward 3D autoencoders and scalable Diffusion Transformers (DiT) [49], where all video tokens interact via self-attention, with text conditions injected through MMDiT-style fusion [18, 26, 37] or separate cross-attention layers. State-of-the-art systems include commercial models such as Veo [23], Seedance [21], Kling [38], as well as open-source ones like CogVideoX [74], LTX-Video [27], HunyuanVideo [37], and Wan [64]. Our approach is built with a DiT-based video diffusion backbone.

Autoregressive Video Generation. Recently, autoregressive video generation [24, 29, 41, 44, 45, 51, 67, 79, 86] has attracted increasing attention due to its potential in world and game simulation. It generates a video sequentially under a causal factorization, conditioning each frame on its historical context.

A pivotal challenge for autoregressive video diffusion is the train–test mismatch that leads to error accumulation [66]. Earlier attempts that directly use teacher forcing suffer from degraded quality as video length increases [20,31,84]. To counteract this, prior work injects small noise into history frames to approximate inference-time degradation [62, 67]. Another avenue adopts Diffusion Forcing [10,11,57], assigning each frame an independent noise level to enable conditioning at arbitrary noise during autoregressive rollout. Other works explore relaxing strict causality via a rolling denoising framework [52, 58, 60, 72]. In this setup, video frames within a sliding window maintain non-decreasing noise levels and initiate generation when the preceding frame reaches a target timestep. Some works explore a plan-and-interpolate strategy that first generates a future keyframe and then interpolates the intermediate frames [82].

Recently, Self Forcing-style post-training [32, 42] has emerged as a promising solution for train–test alignment. It first autoregressively rolls out the entire video, then computes a holistic distribution matching loss. However, its reliance on online discriminators for adversarial loss [22] or pretrained bidirectional teachers for distillation [75,76,87,88] limits scalability and hinders training-from-scratch viability. While sharing the insight of inference simulation, our method uniquely supports end-to-end training without recourse to auxiliary models.

Conditioning on Model Predictions. Conditioning the model on its own output is a strategy widely adopted to mitigate exposure bias in autoregressive systems. Scheduled Sampling [4, 9, 47] in language models exemplifies this approach, replacing portions of the ground-truth sequence with model-predicted tokens. In diffusion models, Self-Conditioning [13] improves sample quality by conditioning the current denoising step on previous estimation. For video generation, Stable Video Infinity [39] adopts an error-recycling strategy to reduce drift in autoregressive long video inference.

Efficient Attention for Video Generation. As high-dimensional spatiotemporal signals, videos require a large number of tokens to represent, making quadratic-cost attention a computational bottleneck. To alleviate complexity, numerous works explore efficient attention design for video generation. One line of work employs linear complexity attention [12, 33, 50, 65]. Others exploit the sparsity in the attention score, pruning less activated tokens via defined attention masks [59, 70, 71, 81, 83]. For instance, Radial Attention [40] observes a spatiotemporal decay and proposes a sparse mask that shrinks importance as temporal distance grows. Recent works also adapt advanced sparse-attention designs from LLMs [46, 80]. For example, MoC [8] and VMoBA [69] integrate Mixture of Block Attention [46] into DiT blocks, where the key and value in attention are dynamically selected via a top-k router. Our work integrates a similar routing mechanism, specifically tailored for handling long histories in autoregressive video generation.

3 Method

We begin by reviewing the background of autoregressive video diffusion models and analyzing the exposure bias issue in Sec. 3.1. We then present our *Resampling Forcing* algorithm for end-to-end training in Sec. 3.2, and the dynamic history routing for efficient long-horizon attention in Sec. 3.3.

3.1 Background

Definition. Autoregressive (AR) video diffusion models factorize video generation into inter-frame autoregression and intra-frame diffusion [30,56]. Specifically, given condition c, the joint distribution of an N-frame video sequence $\boldsymbol{x}^{1:N}$ is expressed as

$$p(\boldsymbol{x}^{1:N}|c) = \prod_{i=1}^{N} p(\boldsymbol{x}^i|\boldsymbol{x}^{<i}, c). \tag{1}$$

To sample from each conditional distribution, the i-th frame $\boldsymbol{x}^i$ (denoted as $\boldsymbol{x}_0^i$ at timestep $t = 0$) is synthesized by solving the reverse-time ODE starting from Gaussian noise $\boldsymbol{x}_1^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ at timestep $t = 1$ through

$$\boldsymbol{x}^i = \boldsymbol{x}_1^i + \int_1^0 \boldsymbol{v}_\theta(\boldsymbol{x}_t^i, \boldsymbol{x}^{<i}, t, c)\,\mathrm{d}t, \tag{2}$$

which can be calculated via numerical solvers such as Euler. Here, the neural network $\boldsymbol{v}_\theta(\cdot)$ with parameter θ parameterizes the velocity field $\mathrm{d}\boldsymbol{x}_t^i/\mathrm{d}t$ and is conditioned on history frames $\boldsymbol{x}^{<i}$, *i.e.*, $\mathrm{d}\boldsymbol{x}_t^i/\mathrm{d}t = \boldsymbol{v}_\theta(\boldsymbol{x}_t^i, \boldsymbol{x}^{<i}, t, c)$. Modern diffusion models typically employ the Diffusion Transformer (DiT) [49] architecture, where videos are patchified and processed via attention mechanisms [63]. In practice, it is also common to generate a chunk of frames per autoregressive step. For simplicity, we refer to each chunk as a frame throughout this paper.

Teacher Forcing. To train such sequence models, a common approach is teacher forcing, where the model is trained to predict the current frame $\boldsymbol{x}^i$ given its ground-truth history $\boldsymbol{x}^{<i}$. In Flow Matching [43], the sample $\boldsymbol{x}_t^i$ at timestep t is an interpolation between Gaussian noise $\boldsymbol{\epsilon}^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and the clean frame $\boldsymbol{x}^i$ via

$$\boldsymbol{x}_t^i = (1-t)\cdot\boldsymbol{x}^i + t\cdot\boldsymbol{\epsilon}^i. \tag{3}$$

The network $\boldsymbol{v}_\theta(\cdot)$ is then trained to regress the velocity $\mathrm{d}\boldsymbol{x}_t^i/\mathrm{d}t = \boldsymbol{\epsilon}^i - \boldsymbol{x}^i$ by minimizing

$$\mathcal{L} = \mathbb{E}_{i,t,\boldsymbol{x},\boldsymbol{\epsilon}}\left[\|(\boldsymbol{\epsilon}^i - \boldsymbol{x}^i) - \boldsymbol{v}_\theta(\boldsymbol{x}_t^i, \boldsymbol{x}^{<i}, t, c)\|_2^2\right]. \tag{4}$$

Here, $\boldsymbol{v}_\theta(\cdot)$ takes two separate sequences as inputs: the noisy frames $\boldsymbol{x}_t^i$ as diffusion samples and the noise-free ground-truth frames $\boldsymbol{x}^{<i}$. A causal mask restricts each frame to attend only to its clean history, enabling parallel training for all frames (see Fig. 4 (b,c)). During inference, once a frame is generated, its clean features can be cached and reused for subsequent frame generation (KV cache).

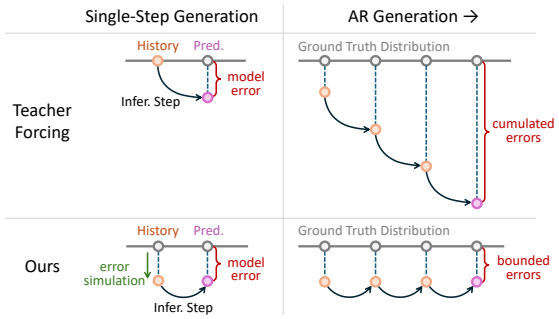

Fig. 3: Error Accumulation. *Top*: Models trained with ground-truth input add and compound errors autoregressively. *Bottom*: We train the model on imperfect input with simulated model errors, stabilizing the long-horizon autoregressive generation. The gray circle represents the closest match in the ground-truth distribution.

Therefore, the number of attention queries remains constant, while the keys and values grow as the video becomes longer.

Error Accumulation. Under teacher forcing, each frame's generation is conditioned on its ground-truth history. However, during inference, model predictions are inevitably *imperfect*. In other words, model generations always have a non-zero discrepancy from the ground-truth distribution, which we refer to as *model error*. The model trained on *perfect* inputs will propagate and accumulate such errors through the autoregressive loop, leading to quality degradation and eventual failure in long-horizon rollouts (see Fig. 3 top).

3.2 Enhancing Error Robustness

As analyzed above, the failure of teacher forcing stems from the distributional mismatch between training and inference inputs, driven by irreducible model errors. With finite model capacity and training samples, eliminating such error is intractable. Instead, we propose training the model to condition on degraded histories while maintaining error-free targets for prediction. This relaxes the model from strict adherence to input conditions and enables correction of input errors. As illustrated at the bottom of Fig. 3, although the model predictions remain imperfect, errors no longer compound. Instead, the errors are stabilized to a near-constant level over the autoregressive process.

Simulating the Model Error. To train error-robust autoregressive models, we must simulate inference-time model errors on the input conditions. We pursue an end-to-end, teacher-free approach to achieve this, prioritizing simplicity and scalability. There are two major factors that drive the distribution shift in autoregressive diffusion: (1) intra-frame generation errors that arise from imperfect score estimation and discretization, which mainly affect high-frequency details [19, 66]; and (2) inter-frame accumulated errors that propagate through the autoregressive loop [66, 77].

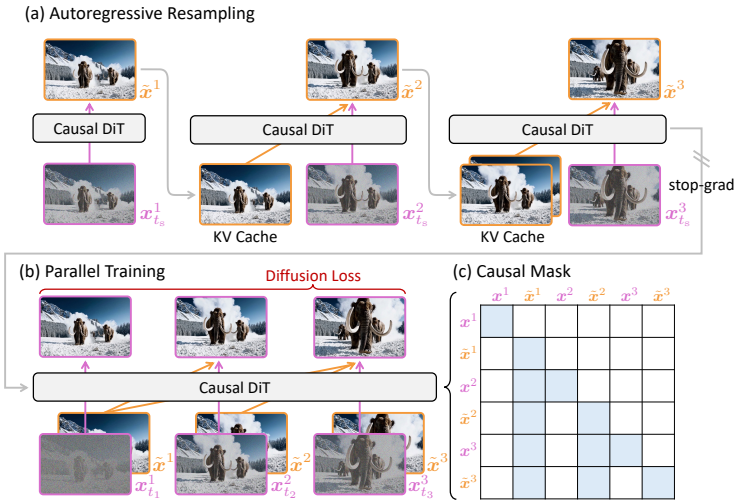

Fig. 4: ***Resampling Forcing.*** (a) To simulate inference-time model error, we add noise to clean videos up to a sampled timestep t_s, then use the online model weights to autoregressively complete the remaining denoising steps. (b) The model is parallel trained with frame-level diffusion loss. (c) A sparse causal mask restricts each frame to attend only to its clean history frames.

To simulate errors from both aspects, we introduce *autoregressive self-resampling* on the history condition. To mimic intra-frame errors, we resample the latter denoising trajectory, where high-frequency details are typically synthesized. Specifically, we corrupt ground-truth video frame $\boldsymbol{x}^i$ to a sampled timestep $t_s \in (0, 1)$ via Eq. (3) to obtain $\boldsymbol{x}^i_{t_s}$. Subsequently, we employ the online model $\boldsymbol{v}_\theta(\cdot)$ to complete the remaining denoising steps and produce a degraded noise-free frame $\tilde{\boldsymbol{x}}^i$ that contains model errors. The timestep t_s controls how close $\tilde{\boldsymbol{x}}^i$ is to its ground-truth version $\boldsymbol{x}^i$. To mimic inter-frame error accumulation, we resample each frame autoregressively, conditioning on degraded history frames $\tilde{\boldsymbol{x}}^{<i}$ (see Fig. 4 (a)), *i.e.*,

$$\tilde{\boldsymbol{x}}^i = \boldsymbol{x}^i_{t_s} + \int_{t_s}^{0} \boldsymbol{v}_\theta(\boldsymbol{x}^i_t, \tilde{\boldsymbol{x}}^{<i}, t, c)\, \mathrm{d}t. \tag{5}$$

Using online model weights ensures that the error distribution evolves alongside training, thereby compelling the model to continuously learn to correct its current imperfections. Gradients are detached from this process to prevent shortcut learning. In practice, this procedure can be efficiently implemented with KV cache.

Sampling Simulation Timestep. The timestep t_s in Eq. (5) governs the trade-off between history faithfulness and error correction flexibility. A small t_s yields low resampling strength, resulting in a degraded sample $\tilde{\boldsymbol{x}}^i$ that closely resembles its ground-truth $\boldsymbol{x}^i$. This encourages the model to stay faithful to the

history frames and risks error accumulation (teacher forcing is a limiting case where $t_s = 0$). On the other hand, a large t_s grants greater flexibility for error correction but raises the chance of content drift, as the model is permitted to deviate significantly from the historical context. Consequently, the distribution of t_s should concentrate density on intermediate values while suppressing extremes. To model this, we choose to sample t_s from a logit-normal distribution LogitNormal$(0, 1)$ that satisfies the above properties:

$$\text{logit}(t_s) \sim \mathcal{N}(0, 1). \tag{6}$$

Generally, stronger models induce fewer errors, allowing for a greater emphasis on low resampling strength, and vice versa. To bias the distribution of t_s, inspired by [18], we apply a timestep shifting with parameter s after sampling t_s from the standard logit normal distribution via

$$t_s \leftarrow \frac{s \cdot t_s}{1 + (s-1) \cdot t_s}. \tag{7}$$

In implementation, we set $s < 1$ to put more weight on the low-noise region. After resampling, we use the degraded video $\tilde{\boldsymbol{x}}^{1:N}$ as history condition, and use the ground-truth video $\boldsymbol{x}^{1:N}$ as the training objective. The complete pseudo code for *Resampling Forcing* is shown in Algorithm 1.

Teacher Forcing Warmup. In the initial training phase, the model has not yet converged to the causal architecture and is incapable of generating meaningful content autoregressively. The model errors at this stage are dominated by random initialization rather than specific intra-frame imperfections or inter-frame accumulation. Therefore, performing history self-resampling can lead to uninformative learning signals and will hinder convergence. We therefore first warm up the model using teacher forcing. Once the model acquires basic autoregressive capabilities (though imperfect), we transition to *Resampling Forcing* and continue training.

3.3 Routing History Context

In autoregressive generation, the number of history frames grows as the video becomes longer. This decelerates generation for subsequent frames, as dense causal attention necessitates attending to the entire historical context. To resolve this, a common solution is to restrict the attention receptive field to a local sliding window [32, 42, 73]. However, this approach compromises long-term dependency, sacrificing global consistency and exacerbating the drifting problem.

To maintain stable attention complexity, we opt to *optionally* replace the dense causal attention with a dynamic routing mechanism, inspired by the advanced sparse attention in LLMs [46, 80]. Specifically, for the query token $\boldsymbol{q}_i$ of the i-th frame, rather than attending to the full history, we dynamically retrieve and attend to the top-k most relevant history frames (see Fig. 5), *i.e.*,

$$\text{Attention}(\boldsymbol{q}_i, \boldsymbol{K}_{<i}, \boldsymbol{V}_{<i}) = \text{Softmax}\left(\frac{\boldsymbol{q}_i \boldsymbol{K}_{\Omega(\boldsymbol{q}_i)}^{\top}}{\sqrt{d}}\right) \cdot \boldsymbol{V}_{\Omega(\boldsymbol{q}_i)}, \tag{8}$$

Algorithm 1 *Resampling Forcing*

```
# v(x_t, hist, t, c): causal diffusion model
# x: clean video frames; c: condition
# x_hat: resampled history
# s: shift parameter

def denoise(v, z, hist, t_s, c):
    for t, t_next in schedule(t_s, 0):
        z = ode_step(v, z, hist, t, t_next, c)
    return z

# (a) autoregressive resampling
t_s = logit_normal()
t_s = s * t_s / (1 + (s - 1) * t_s)    # timestep shift
with no_grad():
    x_hat = []
    for i in range(N):
        z = (1 - t_s) * x[i] + t_s * randn_like(x[i])
        x_hat.append(denoise(v, z, x_hat, t_s, c))

# (b) diffusion loss
t = sample_t(N)
e = randn_like(x)
z = (1 - t) * x + t * e
target = e - x

pred = v(z, x_hat, t, c)
loss = mse_loss(pred, target)
```

where $\Omega(\boldsymbol{q}_i)$ is set of selected indices of k history frames for query $\boldsymbol{q}_i$. For selection metric, we use the dot product of $\boldsymbol{q}_i$ and a frame descriptor $\phi(\boldsymbol{K}_j)$ (for j-th frame), *i.e.*,

$$\Omega(\boldsymbol{q}_i) = \arg\max_{|\Omega^*|=k} \sum_{j\in\Omega^*} \left(\boldsymbol{q}_i^\top \phi(\boldsymbol{K}_j)\right). \tag{9}$$

Following [8, 46, 69], we use mean pool as the descriptor transformation $\phi(\cdot)$ since it adheres to the attention score computation and is parameter-free. This reduces the per-token attention complexity from linear $\mathcal{O}(L)$ to constant $\mathcal{O}(k)$ as the number of history frames L grows, achieving an attention sparsity of $1-k/L$.

Notably, while k may be set small for high sparsity, the routing mechanism operates in a *head-wise* and *token-wise* manner, implying that tokens across different attention heads and spatial locations can route to distinct history mixtures, and collectively yield a receptive field significantly larger than k frames.

In implementation, following MoBA [46], we adopt an efficient two-branch attention that fuses an intra-frame pathway and a sparse history pathway via a global log-sum-exp. In the intra-frame branch, each query attends only to tokens

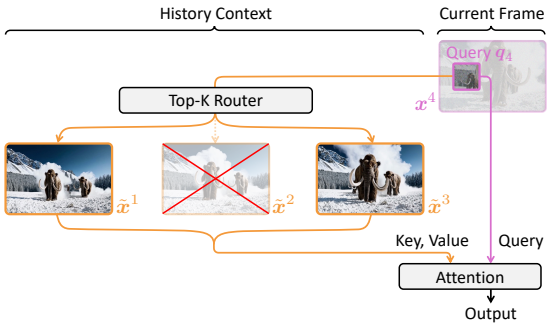

Fig. 5: History Routing Mechanism. Our routing mechanism dynamically selects the top-k important frames to attend. In this illustration, we show a $k = 2$ example, where only the 1st and 3rd frames are selected for the 4th frame's query token $\boldsymbol{q}_4$.

within its own frame. In the history branch, we select up to top-k relevant past frames for each query token. Both branches are implemented efficiently with the `flash_attn_varlen_func()` interface from FlashAttention [15,16]. Outputs are combined by aligning their log-sum-exp terms, yielding a result equivalent to a single softmax over the union of keys.

4 Experiments

Model. We build our method upon `Wan2.1-1.3B` [64] architecture. We modify timestep conditioning to support per-frame noise levels, and implement the sparse causal attention in Fig. 4 (c) with `flex_attention()` in PyTorch, incurring no additional parameters. Following [14, 32, 73], we use a chunk size of 3 latent frames as the autoregressive unit.

Training. After switching to causal attention, the model was trained on 5 s videos with the teacher forcing objective for 10K steps to warm up, then transitioned to *Resampling Forcing* and trained sequentially on 5 s and 15 s (249 frames) videos, for 15K and 5K steps respectively. We then fine-tuned with sparse history routing enabled for 1.5K iterations on 15 s videos. The training batch size is 64 and the learning rate for the AdamW optimizer is 5×10^{-5}. We set the timestep shifting factor $s = 0.6$ (Sec. 3.2), and $k = 5$ in top-k history routing (Sec. 3.3). For efficiency, we use a 1-step Euler solver for history resampling (Eq. (5)).

Inference. We use the same inference settings for all generated frames. We use an Euler sampler with 32 steps and a timestep shifting factor of 5.0. The classifier-free guidance scale is 5.0 for all frames.

Fig. 6: Qualitative Comparisons. *Top*: We compare with representative autoregressive video generation models, showing our method's stable quality on long video generation. *Bottom*: Compared with LongLive [73] that distilled from a short bidirectional teacher, our method exhibits better causality. We use dashed lines to denote the highest liquid level, and red arrows to highlight the liquid level in each frame.

4.1 Comparisons

Baselines. We compare our method with recent autoregressive video generation baselines, including SkyReels-V2 [11], MAGI-1 [60], NOVA [17], Pyramid Flow [35], CausVid [77], Self Forcing [32], and a concurrent work, LongLive [73]. Notably, SkyReel-V2 operates as a clip-level autoregressive model, generating 5 s video segments sequentially. MAGI-1 relaxes the strict causal constraint, initiating next-chunk denoising prior to the completion of the current chunk's generation. LongLive rolls out longer videos and takes 5 s sub-clips and computes distillation loss with teacher models.

Qualitative Comparison. We provide a visual qualitative comparison across different methods in Fig. 6 on generated 15-second videos.

In the upper panel, we compare the visual quality, observing that most strict autoregressive models (*e.g.*, Pyramid Flow [35], CausVid [77], and Self Forcing [32]) exhibit error accumulation, manifested as progressive degradation in color, texture, and overall sharpness. By contrast, approaches that relax strict causality (MAGI-1 [60]) or use large autoregressive chunks (SkyReels-V2 [11]) alleviate long-horizon degradation. However, these relaxed settings compromise intrinsic advantages of strict autoregression, such as per-frame interactivity and faithful causal dependency for future-state prediction. Within the strict autore-

Table 1: Quantitative Comparisons. We split the generated 15-second videos into three parts, *i.e.*, 0–5 s, 5–10 s, and 10–15 s, and separately evaluate the videos with VBench [34].

Method	#Param	Teacher	Length = 0–5 s			Length = 5–10 s			Length = 10–15 s		
			Temp.	*Visual*	*Text*	*Temp.*	*Visual*	*Text*	*Temp.*	*Visual*	*Text*
SkyReels-V2 [11]	1.3B	-	81.93	60.25	21.92	84.63	59.71	21.55	87.50	58.52	21.30
MAGI-1 [60]	4.5B	-	87.09	59.79	26.18	89.10	59.33	25.40	86.66	59.03	25.11
NOVA [17]	0.6B	-	87.58	44.42	25.47	88.40	35.65	20.15	84.94	30.23	18.22
Pyramid Flow [35]	2.0B	-	81.90	62.99	27.16	84.45	61.27	25.65	84.27	57.87	25.53
CausVid [77]	1.3B	`Wan-14B(5s)`	89.35	65.80	23.95	89.59	65.29	22.90	87.14	64.90	22.81
Self Forcing [32]	1.3B	`Wan-14B(5s)`	90.03	67.12	25.02	84.27	66.18	24.83	84.26	63.04	24.29
LongLive [73]	1.3B	`Wan-14B(5s)`	81.84	66.56	24.41	81.72	67.05	23.99	84.57	67.17	24.44
Ours (75% sparsity)	1.3B	-	90.18	63.95	24.12	89.80	61.95	24.19	87.03	61.01	23.35
Ours	1.3B	-	91.20	64.72	25.79	90.44	64.03	25.61	89.74	63.99	24.39

gressive paradigm, our method demonstrates superior robustness in long-term visual quality compared to baselines.

In the lower panel, we further compare with LongLive [73], which first generates long videos and then performs sub-clip distillation using a short-horizon teacher. Although LongLive attains strong long-range visual quality, we observe that distillation from a short bidirectional teacher fails to ensure strict causality, even with a temporally causal student architecture. In the "milk pouring" example in Fig. 6, LongLive produces a liquid level that rises and then falls despite continuous pouring, which violates physical laws. By contrast, our model maintains strict temporal causality: the liquid level monotonically increases while the source container empties. We attribute this non-causal behavior to two factors. First, the bidirectional teacher is inherently non-causal, allowing future information to influence earlier frames via attention, thereby leaking future context to the student during distillation. Second, sub-clip distillation emphasizes local appearance quality and neglects global causality. Conversely, our training strictly precludes information leakage from the future.

Quantitative Comparison. We evaluate methods using the automatic metrics provided by VBench [34]. All models are prompted to generate 15-second videos, which we partition into three segments and evaluate them separately to better assess long-term quality. Results are summarized in Tab. 1. As evidenced in the table, our method maintains comparable visual quality and superior temporal quality on all video lengths to baselines. On longer video lengths, our method's performance also matches the long video distillation baseline LongLive. Given that the distill-based methods (*i.e.*, CausVid [77], Self Forcing [32], and LongLive [73]) necessitate a pretrained 14B-parameter bidirectional teacher, our method offers substantial efficiency and practicality for training autoregressive video models. Moreover, the history routing mechanism achieves an attention sparsity of 75% while incurring only a negligible drop relative to the dense-attention baseline, demonstrating its strong potential for long-horizon generation under constrained compute and memory budgets.

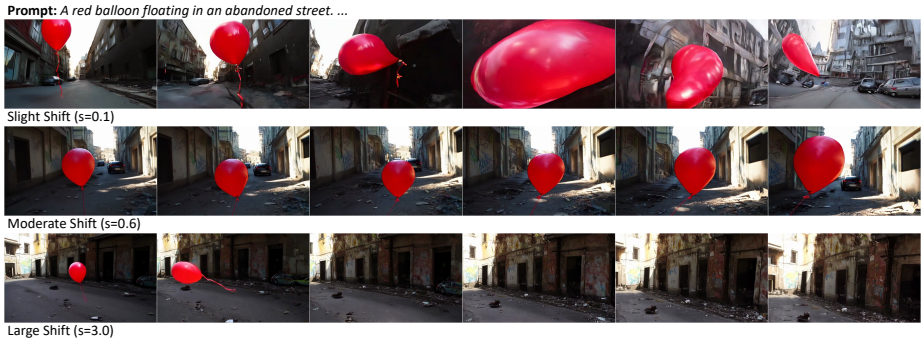

Fig. 7: Comparing Timestep Shifting. A moderate shifting scale for resampling timestep t_s is necessary to balance between error accumulation and content drift.

Table 2: Error Simulation Strategies. We compare different variants and find that autoregressive resampling yields the best quality.

Simulation Strategies	Length = 0–15 s		
	Temp.	*Visual*	*Text*
Noise Augmentation	87.15	61.90	21.44
Parallel Resampling	88.01	62.51	24.51
Autoregressive Resampling	90.46	64.25	25.26

4.2 Analytical Studies

Error Simulation Strategies. In Sec. 3.2, we hypothesized that exposing the model to imperfect historical contexts during training mitigates error accumulation, and propose autoregressive self-resampling to simulate model errors. We compare against two alternatives: noise augmentation [62] and parallel resampling. In the first, small Gaussian noise is added to the history frames to improve robustness to inference errors. In the second, all historical frames are resampled in parallel rather than autoregressively. As shown in Tab. 2, the autoregressive resampling strategy achieves the highest quality, followed by parallel resampling and noise augmentation. We attribute this to a mismatch between additive noise and the model's inference-time error mode, as well as the fact that parallel resampling captures only per-frame degradation while neglecting autoregressive accumulation across time.

Simulation Timestep Shifting. We ablate the shifting factor s that biases the t_s distribution. As defined in Eq. (7), a small s concentrates t_s in the low-noise region, and a large s shifts t_s toward higher noise. Equivalently, a small s corresponds to weaker history resampling, encouraging faithfulness to past content, whereas a large s enforces stronger resampling, promoting content modifications that may induce drift. We observe that model performance is robust to the choice of s; therefore, we adopt extreme values in this ablation to better visualize the impact of the shifting factor. In Fig. 7, the model trained with small s exhibits

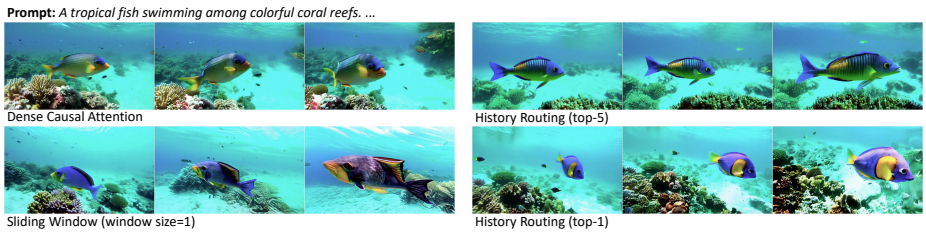

Fig. 8: Sparse History Strategies. We compare dense causal attention, dynamic history routing, and sliding window attention in terms of appearance consistency.

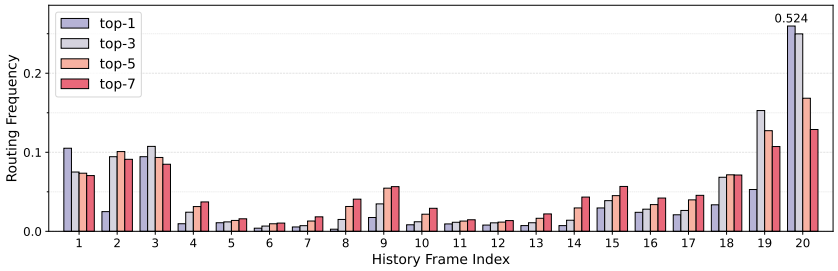

Fig. 9: History Routing Frequency. We visualize the beginning 20 frames' frequency of being selected when generating the 21st frame. For readability, the maximum bar is truncated and labeled with its exact value.

error accumulation and quality degradation, while a very large s reduces semantic consistency with history, increasing the risk of initial content drift. Thus, a moderate value for s is essential to strike a balance between mitigating error accumulation and preventing drift.

Sparse History Strategies. We compare three mechanisms for leveraging historical context in autoregressive generation: dense causal attention, dynamic history routing, and sliding-window attention [32, 42]. We show qualitative results for dense attention, top-5 and top-1 routing, and sliding-window attention in Fig. 8. As shown in the figure, routing to the top-5 of 20 history frames yields 75% sparsity with quality comparable to dense attention. Reducing from top-5 to top-1 (95% sparsity) causes only minor quality degradation, demonstrating the robustness of the routing mechanism. We further contrast top-1 routing with a sliding window of size 1. Despite equal sparsity, the routing mechanism maintains superior consistency in the fish's appearance. We hypothesize that sliding-window attention's fixed and localized receptive field exacerbates the risk of drift. By contrast, our dynamic routing enables each query token to select diverse historical context combinations, collectively yielding a larger effective receptive field that better preserves global consistency.

History Routing Frequency. To provide deeper insights into history routing, we experiment with $k = 1, 3, 5, 7$ and visualize the selection frequency of each

history frame during the generation of the current frame. As shown in Fig. 9, the selection frequencies exhibit a hybrid "sliding-window" and "attention-sink" pattern: the router prioritizes initial frames alongside the most recent frames preceding the target. This effect is most pronounced under extreme sparsity ($k = 1$) and becomes more distributed as sparsity decreases ($k = 1 \rightarrow 7$), encompassing a broader range of intermediate frames. The observation offers empirical support for recent attention designs combining "frame sinks" with sliding windows for long video rollout [73], which can be viewed as a special case of our approach. Our results suggest an alternative path to attention sparsity: replacing fixed heuristic masks with a dynamic, content-aware routing mechanism capable of exploring a vastly larger space of context combinations.

5 Discussions

We presented *Resampling Forcing*, an end-to-end, teacher-free framework for training autoregressive video diffusion models. Identifying the root cause of error accumulation, we proposed a history self-resampling strategy that effectively mitigates this issue, ensuring stable long-horizon generation. Furthermore, we introduced a history routing mechanism designed to maintain near-constant attention complexity despite the ever-growing historical context. Experiments demonstrate our method's superior visual quality and robustness under high attention sparsity.

References

1. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024)
2. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., Baetu, C., Berbel, J., Bridson, D., Bruce, J., Buttimore, G., Chakera, S., Chandra, B., Collins, P., Cullum, A., Damoc, B., Dasagi, V., Gazeau, M., Gbadamosi, C., Han, W., Hirst, E., Kachra, A., Kerley, L., Kjems, K., Knoepfel, E., Koriakin, V., Lo, J., Lu, C., Mehring, Z., Moufarek, A., Nandwani, H., Oliveira, V., Pardo, F., Park, J., Pierson, A., Poole, B., Ran, H., Salimans, T., Sanchez, M., Saprykin, I., Shen, A., Sidhwani, S., Smith, D., Stanton, J., Tomlinson, H., Vijaykumar, D., Wang, L., Wingfield, P., Wong, N., Xu, K., Yew, C., Young, N., Zubov, V., Eck, D., Erhan, D., Kavukcuoglu, K., Hassabis, D., Gharamani, Z., Hadsell, R., van den Oord, A., Mosseri, I., Bolton, A., Singh, S., Rocktäschel, T.: Genie 3: A new frontier for world models (2025)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
4. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. Advances in neural information processing systems **28** (2015)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
7. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`, last accessed 2026-06-30
8. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., et al.: Mixture of contexts for long video generation. arXiv preprint arXiv:2508.21058 (2025)
9. Cen, Z., Liu, Y., Zeng, S., Chaudhari, P., Rangwala, H., Karypis, G., Fakoor, R.: Bridging the training-inference gap in llms by leveraging self-generated tokens. arXiv preprint arXiv:2410.14655 (2024)
10. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2025)
11. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
12. Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., et al.: Sana-video: Efficient video generation with block linear diffusion transformer. arXiv preprint arXiv:2509.24695 (2025)

13. Chen, T., Zhang, R., Hinton, G.: Analog bits: Generating discrete data using diffusion models with self-conditioning. arXiv preprint arXiv:2208.04202 (2022)
14. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
15. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (ICLR) (2024)
16. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
17. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024)
18. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
19. Falck, F., Pandeva, T., Zahirnia, K., Lawrence, R., Turner, R., Meeds, E., Zazo, J., Karmalkar, S.: A fourier space perspective on diffusion models. arXiv preprint arXiv:2505.11278 (2025)
20. Gao, K., Shi, J., Zhang, H., Wang, C., Xiao, J., Chen, L.: Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. arXiv preprint arXiv:2411.16375 (2024)
21. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
23. Google DeepMind: Veo (2025), `https://deepmind.google/models/veo`, last accessed 2026-06-30
24. Gu, J., Shen, Y., Chen, T., Dinh, L., Wang, Y., Bautista, M.A., Berthelot, D., Susskind, J., Zhai, S.: Starflow-v: End-to-end video generative modeling with normalizing flow. arXiv preprint arXiv:2511.20462 (2025)
25. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
26. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long context tuning for video generation. arXiv preprint arXiv:2503.10589 (2025)
27. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
28. He, H., Zhang, Y., Lin, L., Xu, Z., Pan, L.: Pre-trained video generative models as world simulators. arXiv preprint arXiv:2502.07825 (2025)
29. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

31. Hu, J., Hu, S., Song, Y., Huang, Y., Wang, M., Zhou, H., Liu, Z., Ma, W.Y., Sun, M.: Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. arXiv preprint arXiv:2412.07720 (2024)
32. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
33. Huang, Y., Ge, X., Gong, R., Lv, C., Zhang, J.: Linvideo: A post-training framework towards o (n) attention in efficient video generation. arXiv preprint arXiv:2510.08318 (2025)
34. 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)
35. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
36. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
37. 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)
38. Kuaishou: Kling video model. `https://kling.kuaishou.com/en` (2024), last accessed 2026-06-30
39. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. arXiv preprint arXiv:2510.09212 (2025)
40. Li, X., Li, M., Cai, T., Xi, H., Yang, S., Lin, Y., Zhang, L., Yang, S., Hu, J., Peng, K., et al.: Radial attention: $\mathcal{O}(n \log n)$ sparse attention with energy decay for long video generation. arXiv preprint arXiv:2506.19852 (2025)
41. Li, Z., Hu, S., Liu, S., Zhou, L., Choi, J., Meng, L., Guo, X., Li, J., Ling, H., Wei, F.: Arlon: Boosting diffusion transformers with autoregressive models for long video generation. arXiv preprint arXiv:2410.20502 (2024)
42. Lin, S., Yang, C., He, H., Jiang, J., Ren, Y., Xia, X., Zhao, Y., Xiao, X., Jiang, L.: Autoregressive adversarial post-training for real-time interactive video generation. arXiv preprint arXiv:2506.09350 (2025)
43. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
44. Liu, H., Liu, S., Zhou, Z., Xu, M., Xie, Y., Han, X., Pérez, J.C., Liu, D., Kahatapitiya, K., Jia, M., et al.: Mardini: Masked autoregressive diffusion for video generation at scale. arXiv preprint arXiv:2410.20280 (2024)
45. Liu, J., Han, J., Yan, B., Wu, H., Zhu, F., Wang, X., Jiang, Y., Peng, B., Yuan, Z.: Infinitystar: Unified spacetime autoregressive modeling for visual generation. arXiv preprint arXiv:2511.04675 (2025)
46. Lu, E., Jiang, Z., Liu, J., Du, Y., Jiang, T., Hong, C., Liu, S., He, W., Yuan, E., Wang, Y., et al.: Moba: Mixture of block attention for long-context llms. arXiv preprint arXiv:2502.13189 (2025)
47. Mihaylova, T., Martins, A.F.: Scheduled sampling for transformers. arXiv preprint arXiv:1906.07651 (2019)
48. Ning, M., Li, M., Su, J., Salah, A.A., Ertugrul, I.O.: Elucidating the exposure bias in diffusion models. arXiv preprint arXiv:2308.15321 (2023)

49. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
50. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. arXiv preprint arXiv:2505.20171 (2025)
51. Ren, S., Chen, C., Wang, Z., Song, L., Zhu, X., Yuille, A., Yang, Y., Lu, J.: Autoregressive video generation beyond next frames prediction. arXiv preprint arXiv:2509.24081 (2025)
52. Ruhe, D., Heek, J., Salimans, T., Hoogeboom, E.: Rolling diffusion models (2024), `https://arxiv.org/abs/2402.09470`, last accessed 2026-06-30
53. Schmidt, F.: Generalization in generation: A closer look at exposure bias. arXiv preprint arXiv:1910.00292 (2019)
54. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Schechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
55. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3626–3636 (2022)
56. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
57. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
58. Sun, M., Wang, W., Li, G., Liu, J., Sun, J., Feng, W., Lao, S., Zhou, S., He, Q., Liu, J.: Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7364–7373 (2025)
59. Sun, W., Tu, R.C., Ding, Y., Jin, Z., Liao, J., Liu, S., Tao, D.: Vorta: Efficient video diffusion via routing sparse attention. arXiv preprint arXiv:2505.18809 (2025)
60. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
61. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
62. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
64. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
65. Wang, H., Ma, C.Y., Liu, Y.C., Hou, J., Xu, T., Wang, J., Juefei-Xu, F., Luo, Y., Zhang, P., Hou, T., et al.: Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2578–2588 (2025)

66. Wang, J., Zhang, F., Li, X., Tan, V.Y., Pang, T., Du, C., Sun, A., Yang, Z.: Error analyses of auto-regressive video diffusion models: A unified framework. arXiv preprint arXiv:2503.10704 (2025)
67. Weng, W., Feng, R., Wang, Y., Dai, Q., Wang, C., Yin, D., Zhao, Z., Qiu, K., Bao, J., Yuan, Y., et al.: Art-v: Auto-regressive text-to-video generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7395–7405 (2024)
68. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
69. Wu, J., Hou, L., Yang, H., Tao, X., Tian, Y., Wan, P., Zhang, D., Tong, Y.: Vmoba: Mixture-of-block attention for video diffusion models. arXiv preprint arXiv:2506.23858 (2025)
70. Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., et al.: Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. arXiv preprint arXiv:2502.01776 (2025)
71. Xia, Y., Ling, S., Fu, F., Wang, Y., Li, H., Xiao, X., Cui, B.: Training-free and adaptive sparse attention for efficient long video generation. arXiv preprint arXiv:2502.21079 (2025)
72. Xie, D., Xu, Z., Hong, Y., Tan, H., Liu, D., Liu, F., Kaufman, A., Zhou, Y.: Progressive autoregressive video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6322–6332 (2025)
73. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
74. 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)
75. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
76. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
77. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast causal video generators. arXiv e-prints pp. arXiv–2412 (2024)
78. Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., Liu, X.: Gamefactory: Creating new games with generative interactive videos. arXiv preprint arXiv:2501.08325 (2025)
79. Yuan, H., Chen, W., Cen, J., Yu, H., Liang, J., Chang, S., Lin, Z., Feng, T., Liu, P., Xing, J., et al.: Lumos-1: On autoregressive video generation from a unified model perspective. arXiv preprint arXiv:2507.08801 (2025)
80. Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., Xie, Z., Wei, Y., Wang, L., Xiao, Z., et al.: Native sparse attention: Hardware-aligned and natively trainable sparse attention. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 23078–23097 (2025)
81. Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., Chen, J.: Spargeattn: Accurate sparse attention accelerating any model inference. arXiv preprint arXiv:2502.18137 (2025)

82. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:2504.12626 **2**(3), 5 (2025)
83. Zhang, P., Chen, Y., Huang, H., Lin, W., Liu, Z., Stoica, I., Xing, E., Zhang, H.: Vsa: Faster video diffusion with trainable sparse attention. arXiv preprint arXiv:2505.13389 (2025)
84. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)
85. Zhang, Y., Peng, C., Wang, B., Wang, P., Zhu, Q., Kang, F., Jiang, B., Gao, Z., Li, E., Liu, Y., et al.: Matrix-game: Interactive world foundation model. arXiv preprint arXiv:2506.18701 (2025)
86. Zhang, Y., Jiang, J., Ma, G., Lu, Z., Huang, H., Yuan, J., Duan, N.: Generative pre-trained autoregressive diffusion transformer. arXiv preprint arXiv:2505.07344 (2025)
87. Zhou, M., Zheng, H., Gu, Y., Wang, Z., Huang, H.: Adversarial score identity distillation: Rapidly surpassing the teacher in one step. arXiv preprint arXiv:2410.14919 (2024)
88. Zhou, M., Zheng, H., Wang, Z., Yin, M., Huang, H.: Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In: Forty-first International Conference on Machine Learning (2024)