HiAR: Efficient Autoregressive Long Video Generation via Hierarchical Denoising

Kai Zou[1,5], Dian Zheng[2], Hongbo Liu[3], Tiankai Hang[4], Bin Liu[1,5⋆], and Nenghai Yu[1,5]

[1] University of Science and Technology of China, School of Cyberspace and Technology, Hefei, China
kzou@mail.ustc.edu.cn
[2] The Chinese University of Hong Kong, Hong Kong, China
[3] Tongji University, School of Computer Science and Technology, Shanghai, China
[4] Tencent Hunyuan, Shenzhen, China
[5] Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China, Hefei, China
Project page: https://jacky-hate.github.io/HiAR/

Abstract. Autoregressive (AR) diffusion offers a promising framework for generating videos of theoretically infinite length. However, a major challenge is maintaining temporal continuity while preventing the progressive quality degradation caused by error accumulation. To ensure continuity, existing methods typically condition on highly denoised contexts; yet, this practice propagates prediction errors with high certainty, thereby exacerbating degradation. In this paper, we argue that a highly clean context is unnecessary. Drawing inspiration from bidirectional diffusion models, which denoise frames at a shared noise level while maintaining coherence, we propose that conditioning on context at the same noise level as the current block provides sufficient signal for temporal consistency while effectively mitigating error propagation. Building on this insight, we propose **HiAR**, a hierarchical denoising framework that reverses the conventional generation order: instead of completing each block sequentially, it performs causal generation across all blocks at every denoising step, so that each block is always conditioned on context at the same noise level. This hierarchy naturally admits pipelined parallel inference, yielding a $\sim$1.8$\times$ wall-clock speedup in our 4-step setting. We further observe that self-rollout distillation under this paradigm amplifies a low-motion shortcut inherent to the mode-seeking reverse-KL objective. To counteract this, we introduce a forward-KL regulariser in bidirectional-attention mode, which preserves motion diversity for causal inference without interfering with the distillation loss. On VBench (20 s generation), HiAR achieves the best overall score and the lowest temporal drift among all compared methods.

Keywords: autoregressive diffusion · long video generation · distillation

⋆ Corresponding author.

1 Introduction

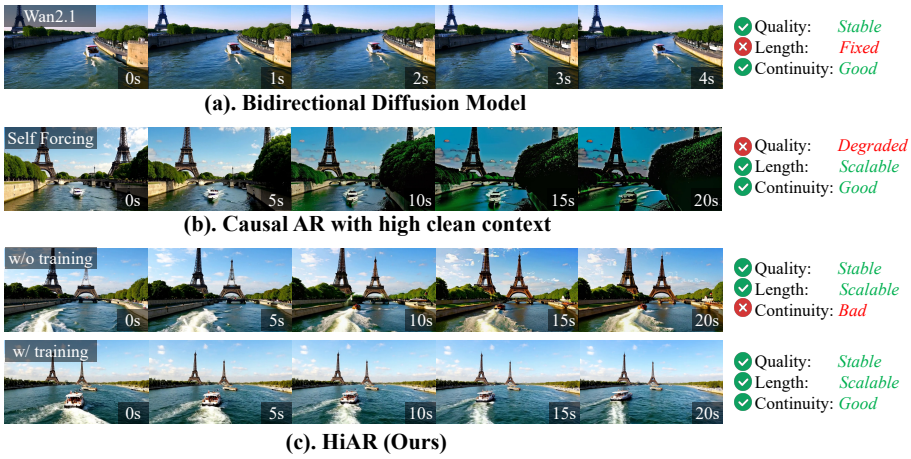

Fig. 1: Motivation. (a) Bidirectional diffusion (Wan2.1) shows that a shared noise level provides sufficient context for temporal coherence within a fixed generation window, while longer videos require conditioning schemes. (b) Standard AR (Self-Forcing) is naturally structured for streaming extension but suffers quality drift, as conditioning on fully clean context amplifies error propagation. (c) Applying our hierarchical denoising (matched-noise context) only at inference (*w/o training*) mitigates drift but breaks continuity due to train–test mismatch; **HiAR** retrains under the hierarchical pipeline (*w/ training*), achieving scalable long-video generation with stable quality and seamless continuity.

Recent years have witnessed rapid progress in video generation, with Diffusion Transformer (DiT) [32] backbones powering strong foundation models [3, 4, 14, 34, 42, 48, 57] and conditional paradigms—including image-to-video and video-to-video—further broadening controllable generation. A remaining frontier is long-horizon, and ultimately open-ended, video generation, central to interactive agents and world models [12, 15, 29, 39, 40, 49]. To scale video duration, causal autoregressive (AR) generation [6, 19, 41, 45] is increasingly attractive: it supports streaming output, indefinite extension, and real-time interaction.

However, a critical challenge in this pipeline is maintaining strict temporal continuity between consecutive video blocks while simultaneously preventing distribution drift (*e.g.*, oversaturation, over-sharpening, motion repetition, and semantic drift) caused by error accumulation. In the Self-Forcing/DMD-style distilled AR setting that we focus on, the previous frames are typically denoised into a *highly clean* context before generating the next. Consequently, every denoising step of the current block is conditioned on a context with noise level $t_c = 0$ (maximal SNR). While this highly clean context anchors the temporal consistency, it inadvertently causes the model to propagate accumulated predic-

tion errors forward with high confidence, thereby exacerbating the degradation, as illustrated in Fig. 1(b).

In this work, we recognize that **a highly clean context is not a prerequisite**. Drawing inspiration from bidirectional diffusion models, which denoise all frames concurrently from a shared noise level yet still yield temporally coherent videos, we observe that noisy context already provides sufficient signal for continuity while reducing error propagation, as shown in Fig. 1(a). Based on this principle, we introduce **HiAR**, a Hierarchical Denoising paradigm that swaps the denoising order: instead of fully denoising previous blocks first, we perform causal generation across all blocks within each denoising step, then move to the next step. This simple yet fundamental change substantially reduces inter-block error transmission and improves long-horizon stability as shown in Fig. 1(c, *w/o training*). Moreover, the hierarchical structure enables pipelined parallelism across denoising steps at inference time, improving wall-clock efficiency ($\times 1.8$).

To maintain train–test consistency, we retrain under the hierarchical denoising pipeline. However, we find that self-rollout distillation [17] exhibits a low-motion shortcut that worsens over training—consistent with the mode-seeking tendency of Distribution Matching Distillation (DMD)-style reverse-KL objectives [27,51,52]: the model gradually collapses into near-static outputs that minimise distillation loss but sacrifice dynamics. Hierarchical denoising amplifies this effect, as the increased learning difficulty of conditioning on multi-level noisy contexts requires more training steps. Empirically, we find that motion diversity under bidirectional-attention denoising is strongly correlated with that under causal AR inference. Motivated by this observation, we introduce a distillation-based forward-KL regulariser computed in bidirectional-attention denoising mode, effectively preventing dynamics collapse for the *causal* inference path and enabling stable long-step training.

We conduct extensive evaluation on VBench [18, 56] and a dedicated drift metric tailored to long-horizon rollouts, together with thorough ablations, demonstrating that HiAR yields more stable long video generation and validating the contribution of each component. The visual result is shown in Fig. 1(c, *w/ training*).

We highlight the main contributions of this paper below:

- We propose **HiAR**, a hierarchical denoising pipeline that performs causal generation across blocks within each denoising step, substantially reducing inter-block error transmission and enabling pipelined inference across hierarchy levels for $\sim 1.8\times$ wall-clock speedup in our implementation.
- We introduce a simple forward-KL regulariser via bidirectional-attention distillation to prevent low-motion shortcuts in self-rollout training, enabling stable scaling to long training schedules while preserving dynamics.
- Extensive experiments on VBench and a dedicated drift metric, together with thorough ablations, demonstrate the long-horizon stability and the effectiveness of each component.

2 Background

2.1 Diffusion Models and Flow Matching

Diffusion-based generative models [13,37] learn to reverse a forward noising process that gradually corrupts data into Gaussian noise. In this work we adopt the flow matching formulation [1,24,26]. Let $x_0 \sim p_{\text{data}}$ denote a clean data sample and $\epsilon \sim \mathcal{N}(0, I)$ standard Gaussian noise. The forward interpolation (corruption) at continuous time $t \in [0, 1]$ is defined as

$$x_t \;=\; (1 - \sigma_t)\, x_0 \;+\; \sigma_t\, \epsilon, \qquad \sigma_t \;=\; \frac{s \cdot t}{1 + (s - 1) \cdot t}, \tag{1}$$

where $s > 0$ is a shift parameter that controls the noise schedule curvature. At $t = 0$ we recover x_0; at $t = 1$ we obtain (approximately) pure noise. A neural network $v_\theta(x_t, t)$ is trained to predict the velocity field

$$v^*(x_t, t) \;=\; \epsilon - x_0, \tag{2}$$

so that clean data can be recovered by integrating the probability-flow ODE backward from $t = 1$ to $t = 0$. In practice, one discretises the trajectory into S steps $1 = t_1 > t_2 > \cdots > t_S > 0$ and applies the Euler update

$$x_{t_{j+1}} \;=\; x_{t_j} \;+\; v_\theta(x_{t_j}, t_j)\, \big(\sigma_{t_{j+1}} - \sigma_{t_j}\big). \tag{3}$$

2.2 Autoregressive Video Diffusion

Bidirectional-attention diffusion models [9, 22, 30, 35, 42] operate on fixed windows and require conditioning schemes for longer videos, whereas causal AR generation [23, 25, 28, 33, 47, 55] is naturally structured for streaming extension: it allows real-time intervention and provides a principled interface for interactive control—making it a key building block toward world models [12, 29, 49]. To generate videos beyond a fixed temporal window, recent work partitions the video latent sequence into N successive blocks $\{B_1, \ldots, B_N\}$, each containing k frames, and generates them autoregressively: for $n = 2, \ldots, N$, block B_n is denoised conditioned on the previously generated blocks $B_{<n}$.

Concretely, let $x_t^{(n)}$ denote the noisy latent of block n at timestep t. The denoiser is queried as

$$v_\theta\big(x_t^{(n)},\, t \;\big|\; c_{<n}\big), \tag{4}$$

where $c_{<n}$ is the context representation of blocks $B_1, \ldots, B_{n-1}$ injected through causal attention: the query tokens come from $x_t^{(n)}$ while the key/value tokens include $c_{<n}$.

Under teacher forcing [8, 16, 19, 44, 55], training conditions on ground-truth context ($c_{<n} = x_0^{(<n)}$), whereas at inference $c_{<n}$ consists of model predictions $\hat{x}_0^{(<n)}$; this train–test mismatch causes per-step errors to accumulate along

the autoregressive chain—exposure bias [2]—manifesting as progressive oversaturation, motion repetition, and semantic drift, collectively termed distribution drift. Diffusion Forcing [5,6,10,33,36,41,53] mitigates this by training with independent per-token noise levels, so the model learns to denoise under heterogeneous noise conditions and gains robustness to partially noisy contexts at inference.

Several recent long-video methods also exploit noisy or progressively denoised contexts. AR-Diffusion [38] imposes non-decreasing timesteps for schedule consistency, PA-VDM [46] uses progressively increasing noise levels for smoother autoregressive transitions, and FIFO-Diffusion [21] converts a pretrained bidirectional model into a training-free streaming generator through diagonal denoising. HiAR instead targets error accumulation in the Self-Forcing/DMD-style distilled AR setting: it derives the matched context noise level from a bias–information trade-off and retrains the model under this schedule.

Self-Forcing [17, 50–52] further closes the train–test gap through self-rollout training: during each training iteration, a block is first rolled out with the student model v_θ to obtain $\hat{x}_0^{(n-1)}$, which is then used as context for the next block's denoising. The training objective is an asymmetric Distribution Matching Distillation (DMD) loss [51, 52], formulated as a reverse KL divergence between the student's one-step output distribution and the teacher's multi-step output distribution:

$$\mathcal{L}_{\mathrm{DMD}} \;=\; \mathbb{E}_{t,\,x_t}\Big[\, D_{\mathrm{KL}}\big(\, p_\theta(x_0 \mid x_t) \,\|\, p_{\mathrm{teacher}}(x_0 \mid x_t)\,\big)\,\Big], \tag{5}$$

where $p_\theta(x_0 \mid x_t)$ denotes the distribution induced by the student's single Euler step from x_t, and p_{teacher} is the distribution obtained by multi-step ODE integration with the teacher model. This reverse KL encourages the student to mode-seek toward the teacher's high-density regions. In practice the gradient is computed via a learned score difference between student and teacher distributions. While Self-Forcing achieves notable improvements at moderate horizons, it still employs $t_c = 0$ for context (i.e., the predicted clean frame), propagating errors with maximal confidence.

3 Method

We now formalise the intuition developed in Sec. 1: the context noise level t_c governs a bias–information trade-off, and the optimal choice is $t_c^* = t_{j+1}$—the output noise level of the current denoising step. We first derive this result analytically and then build upon it to design Hierarchical Denoising.

3.1 Context Noise Level and Error Propagation

Error decomposition. Consider block B_n being denoised at step j (from noise level t_j to t_{j+1}). Let $x_0^{(n-1)}$ denote the ground-truth clean latent of the preceding block and $\hat{x}_0^{(n-1)} = x_0^{(n-1)} + \delta^{(n-1)}$ the model's prediction, where $\delta^{(n-1)}$ is the

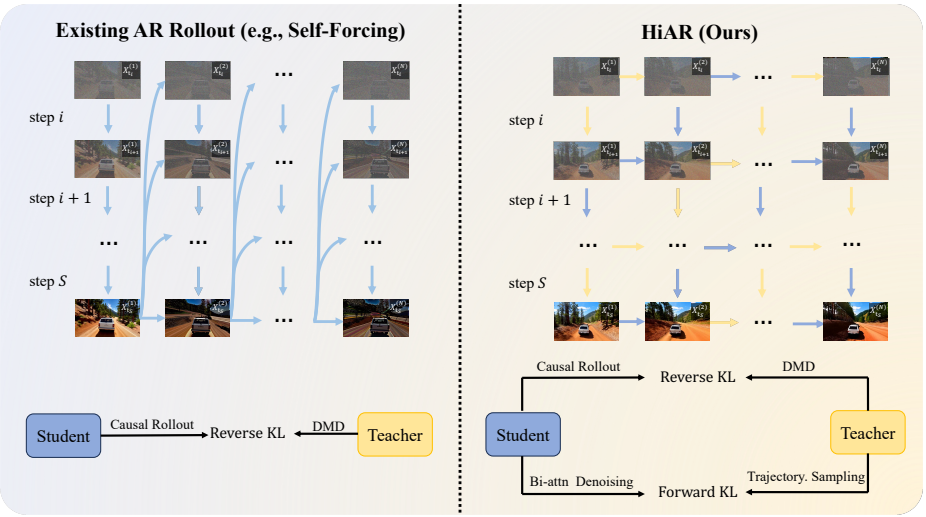

Fig. 2: Overview of HiAR. Left: Existing block-first AR (*e.g.*, Self-Forcing) fully denoises each block before generating the next, conditioning every step on predicted clean context and thus amplifying inter-block error propagation. **Right:** Our hierarchical denoising performs causal generation across all blocks within each denoising step, conditioning on context at the matched noise level to suppress error accumulation. **Bottom:** Training combines causal self-rollout with a reverse-KL (DMD) loss for distillation, and a forward-KL regulariser computed in bidirectional-attention mode via teacher trajectory sampling to preserve motion diversity.

accumulated prediction error. In AR diffusion, the context for B_n is derived from $\hat{x}_0^{(n-1)}$ and presented at some noise level $t_c \in [0, 1]$:

$$c_{n-1}^{(t_c)} \;=\; (1-\sigma_{t_c})\,\hat{x}_0^{(n-1)} + \sigma_{t_c}\,\eta, \quad \eta \sim \mathcal{N}(0, I). \tag{6}$$

Expanding $\hat{x}_0^{(n-1)}$ decomposes the context into three terms:

$$c_{n-1}^{(t_c)} \;=\; \underbrace{(1-\sigma_{t_c})\,x_0^{(n-1)}}_{\text{true signal}} + \underbrace{(1-\sigma_{t_c})\,\delta^{(n-1)}}_{\text{propagated bias}} + \underbrace{\sigma_{t_c}\,\eta}_{\text{stochastic perturbation}}. \tag{7}$$

The true-signal and propagated-bias terms share the same coefficient $(1-\sigma_{t_c})$, while the stochastic term carries the complementary coefficient σ_{t_c}. The noise level t_c thus controls a *bias–information trade-off*: raising t_c attenuates the bias but simultaneously reduces the useful conditioning signal by the same factor. In particular, prior AR methods [17] use $t_c = 0$, which reduces Eq. 7 to $c_{n-1}^{(0)} = x_0^{(n-1)} + \delta^{(n-1)}$ and propagates the full prediction error with no attenuation.

Temporal causality. To produce temporally coherent continuations, the context must carry at least as much information as the current block possesses after step j. Under Eq. 1, the signal-to-noise ratio $\mathrm{SNR}(t) = (1-\sigma_t)^2/\sigma_t^2$ increases monotonically as t decreases, so after step j the current block at t_{j+1} contains

Algorithm 1 Hierarchical Denoising

Require: Schedule $t_1 > t_2 > \cdots > t_S \approx 0$; initial noise $\{x_{t_1}^{(n)}\}_{n=1}^{N}$
Ensure: Generated blocks $\{\hat{x}_0^{(n)}\}_{n=1}^{N}$
for $j = 1, \ldots, S$ **do** ▷ denoising steps
  for $n = 1, \ldots, N$ **do** ▷ causal block sweep with KV cache
    $x_{t_{j+1}}^{(n)} \leftarrow x_{t_j}^{(n)} + v_\theta\big(x_{t_j}^{(n)}, t_j \mid x_{t_{j+1}}^{(<n)}\big)\,(\sigma_{t_{j+1}} - \sigma_{t_j})$
  end for
  Update KV cache with $\{x_{t_{j+1}}^{(n)}\}_{n=1}^{N}$
end for
return $\hat{x}_0^{(n)} \leftarrow x_{t_S}^{(n)}$ for $n = 1, \ldots, N$

strictly more information than at t_j. Temporal causality therefore requires

$$\mathrm{SNR}(t_c) \;\geq\; \mathrm{SNR}(t_{j+1}) \quad\Longleftrightarrow\quad t_c \;\leq\; t_{j+1}. \tag{8}$$

Any t_c satisfying this bound provides sufficient information for step j. Since the bias coefficient $(1-\sigma_{t_c})$ decreases monotonically in t_c, choosing $t_c < t_{j+1}$ only transmits more prediction error without additional benefit. The optimum is therefore the boundary of the constraint:

$$\boxed{t_c^* \;=\; t_{j+1},} \tag{9}$$

the noisiest context level that still fulfills temporal causality— attenuating inter-block bias while retaining all information the denoiser needs at step j.

3.2 Hierarchical Denoising

The analysis above motivates a simple but fundamental change to the autoregressive denoising pipeline: instead of fully denoising each block before moving to the next, we perform causal generation across all blocks at each denoising step. We call this *Hierarchical Denoising* (Fig. 2).

Inference procedure. The complete procedure is summarised in Alg. 1. At each step j, block B_n is denoised with blocks $B_{<n}$ at noise level t_{j+1} as context—the noisiest level that still preserves temporal causality (Sec. 3.1).

Pipelined parallelism. In Alg. 1, block B_n at step j depends only on $B_{<n}$ at step j and on B_n at step $j-1$, so blocks at different (n, j) positions that lie on the same anti-diagonal of the $N \times S$ grid are mutually independent. We exploit this by assigning each denoising step to a dedicated process and traversing the grid along its $N+S-1$ anti-diagonals, with inter-stage latents exchanged via asynchronous point-to-point communication. Within each stage, naïvely updating the KV cache for block B_n and denoising block B_{n+1} are two separate forward passes, totalling $2N$ per stage. We observe that under causal attention the two operations can be fused into one forward call by concatenating $[c^{(n)},\, x_{t_j}^{(n+1)}]$ along the frame dimension with per-frame timesteps $[t_{j+1}, \ldots, t_{j+1},\, t_j, \ldots, t_j]$:

the first segment writes B_n's context into the KV cache while the second segment denoises B_{n+1} attending to the freshly written keys and values. This fusion reduces the cost to $N+2$ passes per stage (one standalone denoise for the first block, $N-1$ fused passes, and one trailing cache write), yielding an overall $\sim 1.8\times$ wall-clock speedup in our 4-step setting.

3.3 Training with Forward-KL Regulation

Although hierarchical denoising already mitigates degradation at test time, a train–test gap remains when the model has been trained under the conventional block-first rollout. We therefore retrain with self-rollout under the hierarchical schedule, optimising the DMD reverse-KL objective (Eq. 5) following Self-Forcing [17]. The overall training pipeline is illustrated in Fig. 2 (bottom).

The low-motion shortcut. As training progresses, temporal coherence improves yet motion diversity collapses: the model increasingly produces near-static videos. The root cause is the mode-seeking nature of the reverse-KL objective: $D_{\mathrm{KL}}(p_\theta \| p_{\mathrm{teacher}})$ is minimised when the student concentrates its mass on a single high-density mode, so it can reduce loss by generating low-motion outputs that are inherently easier to denoise and less prone to rollout errors. Hierarchical denoising amplifies this shortcut, because conditioning on contexts at varying noise levels—rather than only clean ones—increases learning difficulty and demands more training steps, giving the mode-seeking objective more iterations to collapse onto the low-motion mode.

Forward-KL regularisation via distillation. To counteract this shortcut, we introduce a complementary loss that penalises mode dropping. We first run the teacher for a large number of ODE steps to obtain a dense denoising trajectory, from which we extract checkpoints $\{x^{\mathrm{ref}}_{t_1}, \ldots, x^{\mathrm{ref}}_{t_S}\}$ aligned with the student's S-step schedule. The student is then supervised to match each consecutive pair via a single Euler step:

$$\mathcal{L}_{\mathrm{FKL}} \;=\; \mathbb{E}_i\Big[\, \big\| \, v_\theta(x^{\mathrm{ref}}_{t_i},\, t_i) \;-\; \tfrac{x^{\mathrm{ref}}_{t_{i+1}} - x^{\mathrm{ref}}_{t_i}}{\sigma_{t_{i+1}} - \sigma_{t_i}} \big\|^2 \,\Big]. \tag{10}$$

Because the targets x^{ref}_t are drawn from the teacher's distribution, optimising Eq. 10 amounts to minimising a forward-KL-direction objective that encourages the student to cover the teacher's output modes rather than mode-seek, thereby preserving motion diversity.

Decoupling from DMD. To prevent interference between $\mathcal{L}_{\mathrm{FKL}}$ and the DMD objective, we adopt two design choices:

1. **Bidirectional-attention mode only.** Motion dynamics under bidirectional and causal attention are strongly positively correlated (Sec. 4.4), so regularising the former effectively constrains the latter. We therefore compute $\mathcal{L}_{\mathrm{FKL}}$ exclusively in bidirectional-attention mode, leaving the causal self-rollout DMD loss unmodified and minimising gradient interference.
2. **Early-step restriction.** Motion dynamics are governed by low-frequency structures established during the earliest denoising steps. We thus apply

$\mathcal{L}_{\text{FKL}}$ only to the first K of S steps, leaving subsequent high-frequency refinement steps unconstrained.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{DMD}} + \lambda \mathcal{L}_{\text{FKL}}, \tag{11}$$

where $\lambda > 0$ balances the two terms. We ablate the choice of K and the attention-mode decoupling strategy in Sec. 4.4.

4 Experiments

4.1 Setups

Table 1: Quantitative comparison on 20 s generation. Throughput is in frames/s; Latency is in seconds; VBench scores (Total/Quality/Semantic/Dynamic) are on a 0–1 scale; Drift is our proposed drift metric. "–" indicates the model is non-autoregressive and drift is not applicable. Best distilled AR results are **bolded**.

Model	Thru.	Lat.	Total↑	Quality↑	Semantic↑	Dynamic↑	Drift↓
Bidirectional video diffusion models							
LTX-Video [11]	8.98	13.5	0.766	0.789	0.685	0.458	–
Wan2.1-1.3B [42]	0.78	103	0.802	0.813	0.766	**0.690**	–
Autoregressive video diffusion models							
NOVA [7]	0.88	4.1	0.773	0.777	0.757	0.444	–
Pyramid Flow [19]	6.70	2.5	0.775	0.804	0.670	0.161	–
SkyReels-V2-1.3B [6]	0.49	112	0.788	0.808	0.707	0.333	–
MAGI-1-4.5B [41]	0.19	282	0.757	0.785	0.647	0.486	–
Distilled autoregressive video models							
CausVid [53]	17	0.69	0.764	0.771	0.740	0.621	0.842
Self-Forcing [17]	17	0.69	0.805	0.829	0.708	0.542	0.355
Causal Forcing [58]	17	0.69	0.810	0.837	0.701	0.672	0.615
HiAR (Ours)	**30**	**0.30**	**0.821**	**0.846**	**0.723**	0.686	**0.257**

Implementation details. We use the Wan2.1-1.3B backbone [42] as our base model. Following Self-Forcing [17], we fine-tune the model with causal attention masking on 16k ODE solution pairs sampled from the base model. We adopt a 4-step denoising schedule ($S = 4$) and use Wan2.1-14B as the teacher model for the DMD critic. All methods are implemented in a *chunk-wise* manner, where each chunk contains 3 latent frames. For the forward-KL regulariser, we sample 20 k denoising trajectories (50 ODE steps each) from the Wan2.1-1.3B base model, and restrict $\mathcal{L}_{\text{FKL}}$ to the first denoising step only ($K = 1$), with a balancing weight $\lambda = 0.1$. The critic model and generator are updated at a 5:1 ratio. We

train with a learning rate of 2×10^{-6} and a total batch size of 64 for 20 k steps on 5-second clips. At inference time, we employ a sliding-window KV cache with a constant attention window of 5 s.

Evaluation metrics. We adopt the VBench [18] protocol, which measures 16 dimensions grouped into a Quality score and a Semantic score, providing a comprehensive assessment of average generation quality. All models are sampled to 20 s to evaluate long-video capability. To quantify temporal degradation beyond aggregate scores, we introduce a drift metric suite specifically designed for long-horizon evaluation. Each 20-second video is evenly divided into five temporal segments, and the following per-segment statistics are computed: perceptual quality via MUSIQ [20] and CLIP-IQA [43]; temporal coherence via DINOv2 [31] consecutive-frame cosine similarity and LPIPS [54] consecutive-frame distance; and low-level statistics including HSV saturation mean and Laplacian variance (sharpness). For each metric, we report the slope of a linear fit over the five segments as a measure of drift rate. All per-metric slopes are then normalised and aggregated via a weighted sum into a single *Drift Score* (lower is better) that summarises overall temporal stability.

Baselines. We compare against recent open-source video generation methods spanning three categories: (i) bidirectional diffusion models—**LTX-Video** [11] (real-time Video-VAE + spatiotemporal transformer) and **Wan2.1-1.3B** [42] (the foundation model shared by all distilled methods below); (ii) autoregressive diffusion models—**NOVA** [7] (non-quantised temporal AR with spatial diffusion), **Pyramid Flow** [19] (pyramidal flow matching with temporal pyramid), **SkyReels-V2** [6] (1.3B; diffusion forcing with non-decreasing noise schedules), and **MAGI-1** [41] (4.5B; block-causal attention); (iii) distilled AR models, all distilling Wan2.1 into a 4-step causal generator—**CausVid** [53] (bidirectional-to-AR DMD distillation), **Self-Forcing** [17] (self-rollout DMD training), and **Causal Forcing** [58] (use diffusion forcing as mid-training before self-rollout distillation). All baselines use official checkpoints and are evaluated under identical prompts and generation lengths (5s for bidirectional models).

4.2 Quantitative Results

Table 1 reports VBench scores, drift, and inference efficiency for all methods.

VBench results. HiAR achieves the highest Total score (0.821) among all methods, surpassing both bidirectional and autoregressive baselines. Notably, it attains the best Quality score (0.846) while maintaining a strong Semantic score (0.723), indicating that hierarchical denoising does not sacrifice semantic fidelity for visual quality. On the Dynamic dimension, HiAR scores 0.686, closely preserving the motion diversity of the bidirectional Wan2.1-1.3B teacher (0.690) and substantially outperforming all other AR methods—including Causal Forcing (0.672) and Self-Forcing (0.542)—demonstrating the effectiveness of our forward-KL regulariser in preventing motion collapse.

Temporal stability. On our proposed Drift metric, HiAR achieves 0.257, the lowest among all distilled AR models, indicating minimal quality degradation over the 20 s horizon. By contrast, CausVid exhibits the highest drift (0.842),

consistent with its visible colour oversaturation at later segments; Self-Forcing (0.355) and Causal Forcing (0.615) show intermediate degradation. HiAR reduces drift by 27.6% relative to Self-Forcing (0.257 vs. 0.355), confirming that hierarchical denoising with matched context noise levels substantially mitigates the compounding inter-block error that drives long-horizon degradation.

Inference efficiency. Owing to pipelined parallelism across hierarchy levels, HiAR achieves 30 fps throughput and 0.30 s per-chunk latency—a $\sim$1.8$\times$ wall-clock speedup over other distilled AR models (17 fps, 0.69 s) that share the same Wan2.1-1.3B backbone and 4-step denoising schedule. This speedup comes at no cost to generation quality; in fact, HiAR simultaneously achieves the best VBench scores and the lowest drift. Compared with full-step bidirectional Wan2.1-1.3B (0.78 fps, 103 s latency), HiAR is 38$\times$ faster in throughput while producing higher Total and Quality scores, demonstrating that the combination of few-step distillation and pipelined hierarchical denoising offers a compelling quality–efficiency trade-off for long video generation.

4.3 Qualitative Results

Fig. 3 presents visual comparisons among all distilled autoregressive models on 20 s generation across six diverse prompts spanning natural scenery (beach, mountain landscape), objects (umbrellas), and human subjects (rock climbing, woman reading, baby portrait).

CausVid exhibits the most severe degradation: frames progressively shift toward neon green and yellow tints, with scene content largely unrecognisable by 20 s. Self-Forcing and Causal Forcing alleviate this to some extent, yet still develop visible colour oversaturation and hue drift over time. The degradation is particularly pronounced on human-centric content—facial regions suffer from unnatural colour casts and loss of fine detail (e.g., skin texture, facial features), which are perceptually salient and difficult to mask. By contrast, HiAR maintains stable colour fidelity, sharpness, and structural coherence from the first frame to the last across all content types, with no perceptible drift in either scenery or portrait prompts. These qualitative observations align with the drift scores in Table 1 and corroborate the effectiveness of hierarchical denoising in suppressing long-horizon error accumulation.

4.4 Ablation Studies

We conduct ablations along two axes: the context noise level t_c (Table 2) and the design choices of the forward-KL regulariser (Table 3). All variants are retrained under the same rollout mode used at inference to ensure train–test consistency, unless stated otherwise.

Context noise level. Table 2 compares three context noise configurations. We evaluate overall video quality (Quality, Semantic), temporal smoothness approximated by the VBench motion smoothness score, and long-horizon stability (Drift).

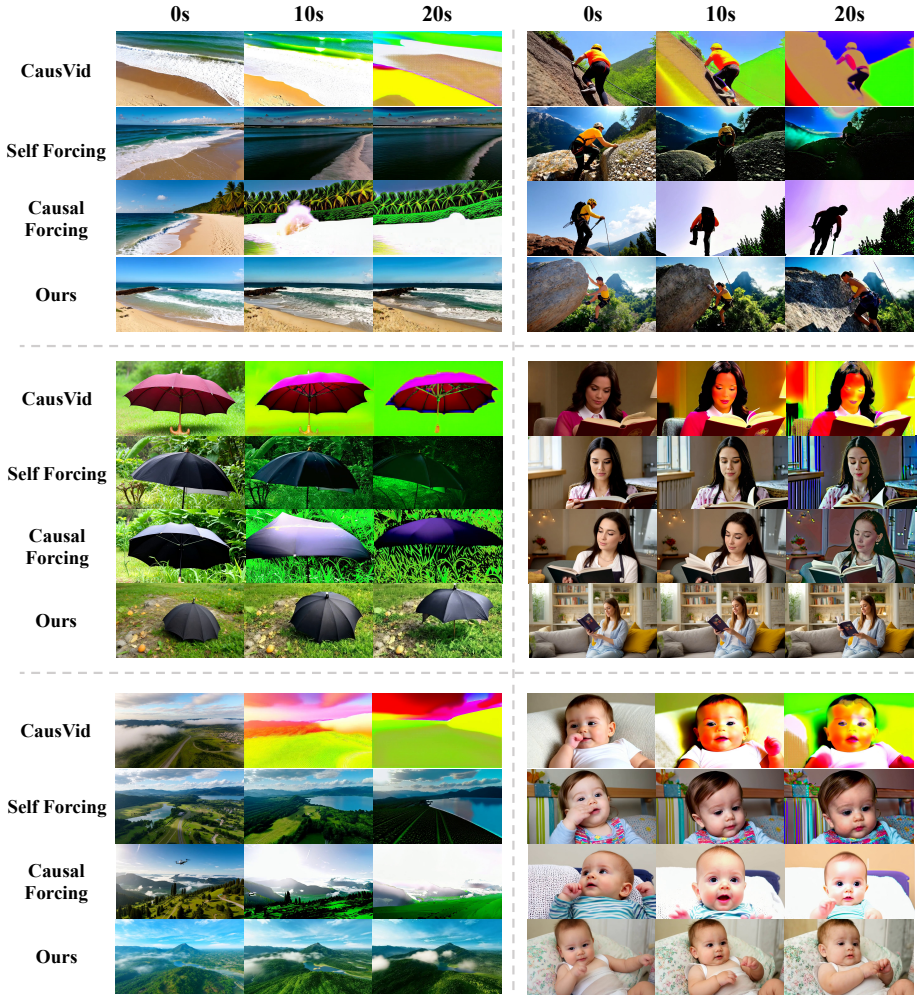

Fig. 3: Qualitative comparison of distilled AR models at 20 s. We show temporally sampled frames from six diverse prompts covering natural scenery, objects, and human subjects. HiAR maintains consistent colour and detail throughout, while baselines exhibit progressive degradation.

When $t_c = t_j$ (the input noise level of the current step), the context carries the same noise level as the current block's input, meaning that block B_n cannot observe the result of denoising step j on block B_{n-1}—effectively removing intra-step causality. While this yields the lowest drift (0.184), the lack of any one-step-ahead information substantially degrades generation quality (Quality 0.799 vs. 0.846) and produces noticeably unsmooth motion (Smooth. 0.978). At the other extreme, $t_c = 0$ (the standard Self-Forcing setting) fully denoises the context, exposing the model to maximum error propagation and the highest drift (0.355).

Table 2: Ablation study on context noise level t_c. Quality, Semantic, and Smooth are VBench sub-scores; Drift is our proposed drift metric.

Context noise	Quality↑	Semantic↑	Smooth.↑	Drift↓
$t_c = t_j$ (input level)	0.799	0.692	0.978	**0.184**
$t_c = t_{j+1}$ (output level; default)	**0.846**	**0.723**	0.988	0.257
$t_c = 0$ (Self-Forcing)	0.829	0.708	**0.991**	0.355

Table 3: Ablation on forward-KL regulariser design. "bi-attn", "causal" denotes the attention mode used for $\mathcal{L}_{\text{FKL}}$; "Kstep" is the number of denoising steps.

Configuration	Quality↑	Semantic↑	Dynamic↑	Drift↓
bi-attn + 1 step (default)	**0.846**	**0.723**	0.686	**0.257**
causal + 1 step	0.828	0.701	0.625	0.271
bi-attn + 2 steps	0.835	0.708	**0.693**	0.296
bi-attn + 4 steps	0.813	0.684	0.691	0.306
w/o $\mathcal{L}_{\text{FKL}}$	0.839	0.732	0.445	0.218
w/o re-training	0.767	0.559	0.512	0.309
w/o hier. denoising (Self-Forcing)	0.829	0.708	0.542	0.355

Our default $t_c = t_{j+1}$ (the output noise level)—where each block conditions on the context that has been denoised through the current step—strikes the optimal balance: it preserves nearly the same temporal smoothness as Self-Forcing (0.988 vs. 0.991) while substantially reducing drift and improving overall quality.

Forward-KL regulation design. Table 3 ablates the attention mode, number of constrained denoising steps K, and the necessity of each component. We focus on motion dynamics (Dynamic), overall quality (Quality, Semantic), and drift.

Attention mode. Applying $\mathcal{L}_{\text{FKL}}$ in causal mode ("causal + 1 step") leads to lower dynamics (0.625 vs. 0.686) and reduced quality compared with the bidirectional-attention default. To empirically justify our design of applying $\mathcal{L}_{\text{FKL}}$ in bidirectional-attention mode, we track the dynamic scores under both attention modes across training checkpoints (without $\mathcal{L}_{\text{FKL}}$). As shown in Fig. 4, both modes exhibit a consistent decline in dynamics over training, and the two scores are strongly positively correlated (Pearson $r = 0.968$). This confirms that regularising the bidirectional mode serves as an effective and non-intrusive proxy for preserving causal-mode motion diversity. Fig. 5 visualises single-step denoising outputs under both modes. Under bidirectional attention, all frames exhibit a uniform level of quality and blur, since the full-sequence attention treats every position symmetrically. In contrast, causal denoising produces frames that become progressively sharper along the temporal axis: as preceding frames fix the low-frequency structure, the conditional distribution of later frames concentrates, resulting in higher-frequency details. This asymmetry means that a distillation target derived from bidirectional denoising provides a spatiotemporally uniform supervision signal well suited to regularising global dynamics,

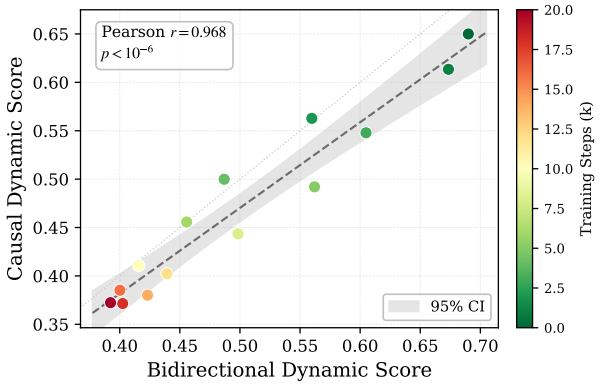

Fig. 4: Correlation between bidirectional and causal dynamics during training (w/o $\mathcal{L}_{\mathbf{FKL}}$). Each point represents one training checkpoint; colour encodes the training step. A strong positive correlation (Pearson $r = 0.968$) confirms that the low-motion shortcut affects both attention modes simultaneously and that regularising the bidirectional mode effectively constrains causal-mode dynamics.

whereas directly constraining causal outputs introduces mismatched targets that are tightly coupled with the model's autoregressive generation pathway, degrading overall quality. Bidirectional-mode regularisation is therefore the preferred configuration.

Number of constrained steps. Increasing K from 1 to 2 or 4 brings marginal gains in dynamics (0.693, 0.691 vs. 0.686) but monotonically degrades both quality and drift. This confirms that motion diversity is primarily governed by the low-frequency structure laid down in the first denoising step; constraining subsequent high-frequency refinement steps provides diminishing returns while interfering with the model's denoising capacity. A single constrained step ($K = 1$) is therefore sufficient and optimal.

Component necessity. Removing $\mathcal{L}_{\text{FKL}}$ entirely ("w/o $\mathcal{L}_{\text{FKL}}$") yields competitive quality and the lowest drift, but dynamics collapse drastically (0.445), confirming that the model falls into the low-motion shortcut without forward-KL regulation. "w/o re-training" applies hierarchical denoising only at inference without corresponding training, which significantly reduces drift compared with Self-Forcing (0.309 vs. 0.355) yet at a substantial cost to visual quality (Quality 0.767), highlighting the importance of train–test alignment. Finally, removing hierarchical denoising altogether recovers the standard Self-Forcing baseline, which exhibits the highest drift (0.355) and lower dynamics (0.542), validating the contribution of hierarchical denoising to long-horizon stability.

5 Conclusion

We presented **HiAR**, a hierarchical denoising framework that addresses the distribution drift problem in autoregressive long video generation. Our key insight

Fig. 5: Comparison of single-step denoising under bidirectional vs. causal attention. Bidirectional attention produces frames of uniform quality and blur across all positions, while causal attention yields progressively sharper frames as preceding context reduces uncertainty for later positions.

is that a fully clean context is unnecessary and, in fact, harmful: by conditioning each block on context at matched noise level rather than predicted clean frames, hierarchical denoising attenuates inter-block error propagation while preserving temporal causality. This simple reordering—from the conventional block-first pipeline to a step-first paradigm—also enables pipelined parallel inference, achieving $\sim$1.8$\times$ wall-clock speedup in our 4-step setting. To stabilise training, we introduced a forward-KL regulariser in bidirectional-attention mode that counteracts the low-motion shortcut inherent to reverse-KL distillation, preserving motion diversity without interfering with the DMD objective. Experiments on VBench and a dedicated drift metric confirm that HiAR achieves the best overall quality and the lowest temporal degradation among all compared methods on 20-second generation.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: ICLR (2023)
2. 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)
3. 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)
4. 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. OpenAI Technical Report (2024)
5. 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 (2024)

6. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., Xiong, W., Wang, W., Pang, N., Kang, K., Xu, Z., Jin, Y., Liang, Y., Song, Y., Zhao, P., Xu, B., Qiu, D., Li, D., Fei, Z., Li, Y., Zhou, Y.: SkyReels-V2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
7. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. In: ICLR (2025)
8. 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)
9. Google: Introducing veo 3, our video generation model with expanded creative controls – including native audio and extended videos. `https://deepmind.google/models/veo/` (2025), accessed: 2026-06-26
10. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025)
11. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: LTX-Video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2025)
12. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., Xu, B., Guo, H.X., Gong, K., Wu, S., Li, W., Song, X., Liu, Y., Li, Y., Zhou, Y.: Matrix-game 2.0: An open-source real-time and streaming interactive world model (2025), `https://arxiv.org/abs/2508.13009`
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
15. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., Sunkavalli, K., Liu, F., Li, Z., Tan, H.: Relic: Interactive video world model with long-horizon memory (2025), `https://arxiv.org/abs/2512.04040`
16. 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)
17. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion (2025), `https://arxiv.org/abs/2506.08009`
18. 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: CVPR (2024)
19. 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)
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: ICCV (2021)
21. Kim, J., Kang, J., Choi, J., Han, B.: FIFO-Diffusion: Generating infinite videos from text without training. In: Advances in Neural Information Processing Systems (2024)
22. Kling: Kling video 2.6 – kling's first "native audio" model official launched! `https://app.klingai.com/global/release-notes/c605hp1tzd` (2025), accessed: 2026-06-26

23. 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)
24. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M.: Flow matching for generative modeling. In: ICLR (2023)
25. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
27. Lu, Y., Ren, Y., Xia, X., Lin, S., Wang, X., Xiao, X., Ma, A.J., Xie, X., Lai, J.H.: Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16818–16829 (2025)
28. Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K.L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al.: Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. arXiv preprint arXiv:2512.04678 (2025)
29. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., Zhang, K.: Yume-1.5: A text-controlled interactive world generation model (2025), https://arxiv.org/abs/2512.22096
30. OpenAI: Sora 2 is here. https://openai.com/index/sora-2/ (2025), accessed: 2026-06-26
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
33. Po, R., Chan, E.R., Chen, C., Wetzstein, G.: Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. arXiv preprint arXiv:2512.12080 (2025)
34. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
35. Runway: Introducing runway gen-4.5: A new frontier for video generation. https://runwayml.com/research/introducing-runway-gen-4.5 (2025), accessed: 2026-06-26
36. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
38. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
39. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling (2025), https://arxiv.org/abs/2512.14614
40. Tang, J., Liu, J., Li, J., Wu, L., Yang, H., Zhao, P., Gong, S., Yuan, X., Shao, S., Zhang, L., Lu, Q.: Hunyuan-gamecraft-2: Instruction-following interactive game world model (2025), https://arxiv.org/abs/2511.23429

41. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W.Q., Luo, W., Kang, X., Sun, Y., Cao, Y., Huang, Y., Lin, Y., Fang, Y., Tao, Z., Zhang, Z., Wang, Z., Liu, Z., et al.: MAGI-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
42. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
43. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: AAAI (2023)
44. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Computation **1**(2), 270–280 (1989)
45. Wu, X., Zhang, G., Xu, Z., Zhou, Y., Lu, Q., He, X.: Pack and force your memory: Long-form and consistent video generation (2025), `https://arxiv.org/abs/2510.01784`
46. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2025)
47. 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)
48. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Gu, X., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
49. Ye, D., Zhou, F., Lv, J., Ma, J., Zhang, J., Lv, J., Li, J., Deng, M., Yang, M., Fu, Q., Yang, W., Lv, W., Yu, Y., Wang, Y., Guan, Y., Hu, Z., Fang, Z., Sun, Z.: Yan: Foundational interactive video generation (2025), `https://arxiv.org/abs/2508.08601`
50. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
51. 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)
52. 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)
53. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22963–22974 (2025)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

55. 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)
56. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness (2025), https://arxiv.org/abs/2503.21755
57. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
58. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)