Direct Autoregressive Diffusion Distillation via Error-aware Causal Pretraining

Jiaxing Li[*1,2], Kaichen Huang[*1,2], Baixin Xu[*1,2], Zexiang Liu[2], Xianglong He[2], Zile Wang[2], Junyao Gao[3], Yang Liu[2], Ying He[1], Bo An[1], and Yangguang Li[†2]

[1] Nanyang Technological University
[2] Skywork AI
[3] Tongji University

Abstract. Real-time autoregressive (AR) video diffusion has progressed rapidly. Existing approaches typically distill pretrained bidirectional video foundation models into few-step causal students; however, naive distillation often collapses due to architectural mismatch. To obtain stable initialization, prior methods rely on ordinary differential equation (ODE) trajectory matching, which incurs substantial computation and can introduce errors by forcing the student to imitate global trajectories. In this paper, we bypass trajectory matching stage by pretraining a robust causal AR model that equips the student for direct few-step distillation. Moreover, to bridge the gap between pretraining and distillation, we propose **Error Forcing**, an error-aware and parallelizable AR video diffusion framework. During training, Error Forcing injects controlled residual errors into the conditioning context, approximating the inference-time history distribution. This enables the causal model to learn from degraded yet clean contexts, improving robustness to accumulated errors without sacrificing parallel training throughput. With this initialization, we can perform direct distillation, and the student and teacher/critic naturally operate under consistent conditioning distributions. Extensive experiments show that our method outperforms existing baselines in generation quality during both the pre-training and distillation stages.

Keywords: autoregressive video diffusion, real-time video generation, diffusion distillation

1 Introduction

In recent years, the emerging vision of world models [32,42] and interactive content creation [4,17,33] has rapidly driven the demand for high-quality real-time autoregressive (AR) video diffusion models [10,60,66,68]. Unlike conventional full-sequence diffusion models [47,62], these AR models need robustly exploit historical context to predict the next frame under the autoregressive formulation, while also maintaining fast per-step inference to satisfy the low-latency requirements of real-time applications.

[*]Equal contribution. [†]Corresponding author.

To meet these demands, recent research has focused on distilling high performance pretrained bidirectional video foundation models into few-step autoregressive student models [10, 66]. However, due to inherent architectural discrepancies, directly initializing an autoregressive student with bidirectional weights often leads to training instability or even collapse, as illustrated in Fig. 4. To address this issue, existing methods [66] typically introduce an intermediate trajectory-matching phase, where the student regresses ODE trajectories simulated by a multi-step teacher [36] for stable initialization. Subsequently, Distribution Matching Distillation (DMD) is applied to align the generation distribution. Although effective in bridging the architectural gap, the substantial computational cost of generating ODE trajectories limits the scalability of such approaches to a certain extent. Moreover, conventional ODE distillation has been shown to violate frame-level injectivity [68], which introduces additional error accumulation throughout the pipeline.

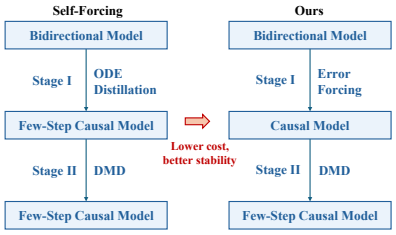

Fig. 1: Comparison of Self Forcing [10] and our pipeline.

In this work, we argue that distillation collapse under architectural mismatch stems from the entangled optimization of causal alignment and few-step adaptation. This motivates us to decouple these tasks by introducing a structurally aligned causal initialization, allowing the student to bypass forced ODE trajectory matching [10, 66, 68] and directly optimize the few-step generation path at the distribution level. However, naively distilling from standard causal initializations introduces a distribution gap, leading to sub-optimal convergence and biased gradients. Specifically, pretraining paradigms like Teacher Forcing [7, 12, 68] and Diffusion Forcing [3] condition on perfectly clean or Gaussian-corrupted contexts, whereas both teacher and student in the distillation stage operate on *clean-but-biased* histories with accumulated prediction errors. This mismatch degrades the overall pipeline. Although Self Forcing [10] mitigates this exposure bias via sequential rollout, its lack of parallelizability makes it fundamentally incompatible with the large-scale supervised pretraining required for stable convergence.

To address these limitations, we propose **Error Forcing**, an autoregressive video diffusion framework designed for efficient parallel training with strict train-test consistency. Specifically, we introduce error estimation and injection into the AR video diffusion pipeline. This enables an error-aware causal pretraining stage that yields a robust causal initialization, thereby enabling direct distillation of the few-step causal student. The key idea is to approximate the inference-time history distribution by injecting controlled, simulated residual errors into the ground-truth context during training. Consequently, the model learns to predict from a degraded yet clean context, significantly improving its robustness against accumulated errors while preserving the high throughput of parallel training. Fig. 2 illustrates the comparison between Error Forcing and existing autoregressive video diffusion paradigms. As depicted in our complete

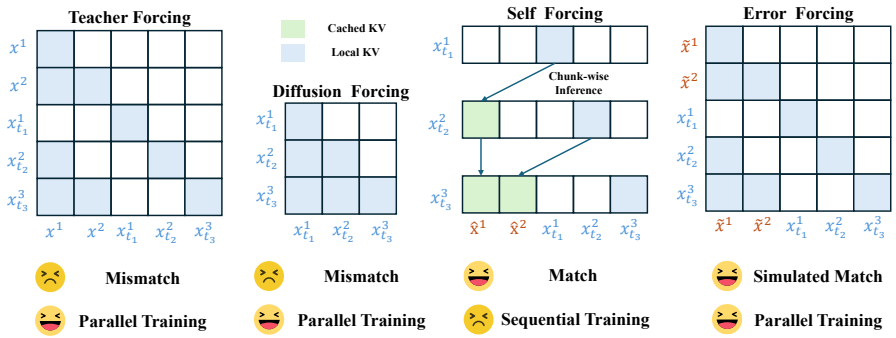

Fig. 2: Comparison of different autoregressive video diffusion methods. Error Forcing ensures train-test consistency while maintaining high parallel training efficiency and supporting supervised training. Here, x_i, $\hat{x}_i$, and $\tilde{x}_i$ denote the ground-truth frames, predicted frames, and error-simulated frames, respectively.

pipeline (Fig. 1), our method effectively replaces the ODE distillation with error-aware causal pre-training, delivering superior distillation stability at a reduced computational cost.

Extensive experimental results show that our full pipeline consistently outperforms the baselines across multiple benchmark metrics. In particular, compared with existing distillation methods, our approach achieves better performance with lower training cost. Furthermore, ablation studies demonstrate that Error Forcing effectively mitigates error accumulation in autoregressive video diffusion models, improves training–inference consistency, and provides a stronger initialization for the subsequent distillation stage. The contributions are summarized as follows:

- A simple yet effective pipeline that directly performs few-step distillation from an error-aware causal initialization, avoiding the cost of ODE trajectory generation and the errors introduced by forced trajectory matching.
- We propose Error Forcing, an autoregressive video diffusion framework that enables efficient parallel training with improved training–inference consistency, providing a strong initialization for subsequent distillation.
- We conduct detailed analyses to determine the level of initialization required for a causal student to start distillation.

2 Related Work

2.1 Autoregressive Video Generation

Autoregressive (AR) frameworks have become a core paradigm for video generation due to their ability to model inherent temporal dependencies for interactive and long-form synthesis. Early efforts evolved from regression and GAN-based prediction [27, 31, 44, 46] to "tokenize-then-predict" methods using AR Transformers [8,15,53,54,59]. While these discrete models are powerful, diffusion-based

AR models [1,2,12,62] have recently emerged as the dominant approach, factorizing the video distribution into a chain of conditional distributions to allow for sequential frame generation.

To address the "exposure bias" in AR rollouts, two primary training paradigms have been developed. Teacher Forcing (TF) methods [7, 12, 43] condition on clean historical frames, sometimes incorporating noise injection [45, 52] or relaxed causality [55, 67] to improve temporal consistency. Conversely, Diffusion Forcing (DF) [3,5,37] trains models to condition on frames at varying noise levels, significantly enhancing sampling robustness. These advancements have enabled "World Models" [6, 40, 41] to simulate interactive environments and generate infinite-length sequences through rolling mechanisms [24, 60, 63].

2.2 Diffusion Distillation

Despite their quality, diffusion models suffer from the high computational cost of iterative denoising [9, 36]. Diffusion distillation addresses this by compressing multi-step teacher models into few-step students [13, 28, 34, 65], categorized into trajectory-preserving and distribution-matching approaches. Trajectory-preserving methods, including Consistency Distillation [14, 38, 39] and Rectified Flow [25, 26], ensure the student mimics the teacher's sampling path. Alternatively, Score Distillation [29, 51] and Distribution Matching Distillation (DMD) [64, 65] supervise the student at the distribution level, often employing GAN-based losses [19, 35, 57] to refine perceptual quality.

Recently, these techniques have been adapted for the temporal domain to enable efficient video synthesis. Beyond distilling non-causal models for short clips [20, 30, 48, 50], new research focuses on distilling teachers into causal AR students [10, 61, 66] for real-time interaction. These distillation strategies are often complemented by system-level optimizations, such as temporal attention decomposition [22] and feature caching [69], which together bridge the gap between heavy diffusion architectures and the low-latency requirements of streaming video applications.

3 Method

Our pipeline is shown in Fig. 3. With robust error-aware causal pretraining, we bypass ODE initialization and directly perform DMD distillation, reducing both computational cost and error accumulation. Section 3.2 introduces Error Forcing, which preserves training–inference consistency in autoregressive video diffusion while maintaining parallel efficiency. Section 3.3 presents the analysis and implementation of direct distillation from the resulting causal initialization.

3.1 Preliminaries

Autoregressive Video Diffusion Models. Let $x^{1:N} = (x^1, x^2, \ldots, x^N)$ denote a video sequence consisting of N frames, AR video diffusion models factorize the

Fig. 3: Illustration of our overall pipeline. (a) We first employ the Error Forcing paradigm, proactively degrading the ground-truth (GT) context via error injection to construct an error-aware, robust causal pre-trained model. (b) Subsequently, we bypass the conventional ODE-based adaptation and directly apply distillation for few-step generation. Benefiting from Error Forcing, train-test consistency is inherently satisfied during distillation.

joint data distribution into a product of conditional distributions: $p(x^{1:N}) = p(x^1)\prod_{i=2}^{N} p(x^i|x^{<i})$. Given a noise schedule defined by α_t and σ_t for $t \in [0,1]$, the forward process for the i-th frame is $x_t^i = \alpha_t x^i + \sigma_t \epsilon^i$, where $\epsilon^i \sim \mathcal{N}(0, I)$. In the flow matching paradigm [21, 25] , the model learns a velocity $v = \epsilon^i - x^i$ by minimizing the mean squared error:

$$\mathcal{L}_{AR} = \mathbb{E}_{x,\epsilon,t}\left[w(t)||v_\theta(x_t^i, t, c) - v||^2\right], \tag{1}$$

where c represents the conditioning context from preceding frames and $w(t)$ is a temporal weighting function.

Distillation and the Bottleneck of Causal Initialization. To achieve real-time interactivity, multi-step AR models must be distilled into few-step models. Current state-of-the-art frameworks typically distill a powerful bidirectional teacher model into an AR student model [10]. However, this introduces a severe *architectural gap*.

For ODE distillation to be well-defined, it must satisfy *frame-level injectivity*: each noisy frame x_t^i must map to a unique clean frame x_0^i via the flow map $\phi : (x_t^i, t) \mapsto x_0^i$ [68]. Distilling an AR student G_θ directly from a bidirectional teacher violates this injectivity condition, causing the regressive student to fail to recover the teacher's flow map and instead collapse to a blurred conditional expectation:

$$G_\theta^*(x_t^i, x_t^{<i}, t) = \mathbb{E}[x_0|x_t^i, x_t^{<i}, t]. \tag{2}$$

While recent efforts like Causal Forcing [68] mitigate this by using causal ODE distillation, these methods strictly require simulating and storing massive ODE trajectories from a natively AR teacher model to supervise the student. This reliance on explicitly high-fidelity ODE paths imposes prohibitive computational overhead, fundamentally limiting training efficiency and scalability.

3.2 Error Forcing: Robust Causal Pre-training

Exposure Bias Analysis In the pretraining-to-distillation pipeline, the train–test discrepancy introduces two coupled biases. First, the generator exhibits autoregressive drift. Under teacher-forcing pretraining, the model conditions on the ground-truth history $x^{<i}$, whereas inference relies on self-generated histories $\hat{x}^{<i}$. This leads to a mismatch between the inference and pretraining factorization:

$$\underbrace{p(\hat{x}^{1:K}) = p(\hat{x}^1)\prod_{j=2}^{K} p(\hat{x}^j \mid \hat{x}^{<j})}_{\text{Inference}} \neq \underbrace{p(\hat{x}^1)\prod_{j=2}^{K} p(\hat{x}^j \mid x^{<j})}_{\text{Pretraining}}. \tag{3}$$

As a consequence, prediction errors accumulate along the autoregressive chain and progressively move the generator away from the data manifold.

Second, the teacher and critic are also exposed to imperfect histories. Let $\hat{x}^{<i} = x_{\mathcal{M}}^{<i} + \delta^{<i}$, where $x_{\mathcal{M}}^{<i}$ lies on the data manifold and $\delta^{<i}$ denotes the off-manifold deviation. The conditional score can then be approximated by

$$s(x_t^i \mid \hat{x}^{<i}) \approx s(x_t^i \mid x_{\mathcal{M}}^{<i}) + J_s \delta^{<i}, \tag{4}$$

where J_s is the Jacobian with respect to the conditioning history. Therefore, the ideal score is perturbed by the history drift, which in turn corrupts the distillation gradient.

To mitigate this issue, we propose Error Forcing, which deliberately perturbs the ground-truth context during pretraining so as to learn a robust conditional mapping satisfying $p(x_t^i \mid \hat{x}^{<i}) \approx p(x_t^i \mid x^{<i})$.

Error Collection To mimic realistic rollout errors, we collect empirical residuals from the model's early predictions [16] to construct a clean-but-biased history. Such histories stay close to the data manifold while preserving the structured deviations induced by autoregressive rollout.

Specifically, during pretraining, we first convert the model output to the corresponding clean estimate $\hat{x}^i$, i.e., the x_0 prediction implied by the predicted flow. We then define the residual as

$$\delta = \hat{x}^i - x^i. \tag{5}$$

All residuals are stored in a buffer $\mathcal{B}$, which defines an empirical error distribution $\mathcal{D}_{\text{err}}$:

$$\delta \sim \mathcal{D}_{\text{err}}, \qquad \mathcal{D}_{\text{err}} \approx \text{Uniform}(\mathcal{B}). \tag{6}$$

Since $\mathcal{D}_{\text{err}}$ is derived from actual predictions, it captures structured artifacts such as local blur and motion misalignment, making it more suitable than isotropic noise for approximating the inference-time distribution shift.

Algorithm 1: Error Forcing (EF) for Causal Pre-training

Require: Video dataset $\mathcal{D}$
Require: Autoregressive video flow model $v_\theta(\cdot)$
Require: Residual buffer $\mathcal{B}$
Require: Error scale γ
Require: Warm-up steps T_{warm}
Return : Pretrained causal weights θ
$k \leftarrow 0$
while *not converged* **do**
  $k \leftarrow k + 1$
  $\boldsymbol{t} \sim \text{LogitNormal}(0,1), \quad (x^{1:N}, c) \sim \mathcal{D}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
  if $k \leq T_{\text{warm}}$ **then**
    $\tilde{x}^{1:N} \leftarrow x^{1:N}$ // Warm-up Phase
  else
    Sample residual $\delta^{1:N} \sim \text{Uniform}(\mathcal{B})$
    $\tilde{x}^{1:N} \leftarrow x^{1:N} + \gamma\delta^{1:N}$ // Error Injection
  $x_t^{1:N} \leftarrow (1 - \boldsymbol{t}) \cdot \tilde{x}^{1:N} + \boldsymbol{t} \cdot \boldsymbol{\epsilon}$
  $\mathcal{L}_{\text{EF}} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \left\| (\epsilon^i - x^i) - v_\theta(x_{t_i}^i, \tilde{x}^{<i}, t_i, c) \right\|_2^2$
  Update θ with gradient descent
  with gradient disabled do
  $\hat{x}^{1:N} \leftarrow \left\{ x_{t_i}^i + \int_{t_i}^0 v_\theta(x_t^i, \tilde{x}^{<i}, t, c)\, dt \right\}_{i=1}^{N}$
  $\delta^{1:N} \leftarrow \hat{x}^{1:N} - x^{1:N}$
  for $i = 1$ **to** N **do**
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{\delta^i\}$ // Error Collection
  end with
return θ

Error Injection Based on $\mathcal{D}_{\text{err}}$, we inject residual perturbations into the historical context during pretraining. Specifically, we sample $\delta \sim \mathcal{D}_{\text{err}}$ and construct the perturbed history as

$$\tilde{x}^i = x^i + \gamma\delta, \tag{7}$$

where γ controls the perturbation strength. The model is then trained with $\tilde{x}^{<i}$ in place of the clean history $x^{<i}$:

$$\mathcal{L}_{\text{EF}} = \mathbb{E}_{x,t,\epsilon,\delta} \left[\left\| \left(\epsilon^i - x^i\right) - v_\theta\left(x_t^i,\, t \mid \tilde{x}^{<i}\right) \right\|_2^2 \right]. \tag{8}$$

By replacing clean histories with structured residual perturbations, EF exposes the model to rollout-style imperfect contexts during pretraining, thereby reducing the train–test gap in few-step distillation.

Implementation Details. The overall error forcing procedure is summarized in Algorithm 1. To avoid unstable training caused by large early-stage prediction errors, we adopt a warm-up phase that collects residuals without injecting them,

reducing error forcing to standard teacher forcing with ground-truth (GT) contexts. In addition, the error buffer $\mathcal{B}$ is implemented as a fixed-capacity FIFO queue of size V, so that the induced empirical distribution $\mathcal{D}_{\text{err}}$ always reflects the model's most recent prediction errors.

3.3 Direct Distillation from a Causal Initialization

Unlike bidirectional initialization, a causal initialization θ_{init} is already aligned with the student's autoregressive inference structure, eliminating the need for ODE-based initialization that forces the student to mimic the multi-step teacher trajectory, as shown in Fig. 4.

Fig. 4: Unlike bidirectional init, a robust causal init enables direct DMD without additional ODE-distillation.

Based on this, we directly perform distillation under self-rollout. Under perturbed contexts, the DMD gradient can be written in a unified form as

$$\nabla_\phi \mathcal{L}_{\text{DMD}} \approx -\mathbb{E}_{i,t,\hat{x}^i}\left[\left(s_T(x_t^i \mid c_T^i, t) - s_F(x_t^i \mid c_S^i, t)\right)\frac{\partial x_t^i}{\partial \phi}\right]. \tag{9}$$

Here, s_T and s_F denote the real score and fake score, respectively, while c_T^i and c_S^i denote the conditioning contexts used by the two score branches. These contexts are only used for score estimation in DMD; the student generator itself remains causal during rollout. We define them according to the teacher/critic structure as

$$(c_T^i, c_S^i) = \begin{cases} (x^{1:N},\, x^{1:N}), & \text{bidirectional teacher/critic,} \\ (\hat{x}^{<i},\, \hat{x}^{<i}), & \text{causal teacher/critic.} \end{cases} \tag{10}$$

With bidirectional teacher/critic, both branches are evaluated under global context and provide stronger global corrective guidance. With causal teacher/critic, both branches are evaluated under the same autoregressive context $\hat{x}^{<i}$, yielding an internally consistent distillation configuration, which we term *Internal DMD*.

4 Experiments

4.1 Experiment Setup

Basic Setup. We adopt Wan2.1-T2V-1.3B [47] as our base model for fine-tuning, which generates 81-frame videos at a resolution of 832×480. During the causal pre-training phase, the network is trained for 3,000 steps on 32 NVIDIA A100 GPUs, with the first 500 steps serving as the warm-up phase. The training timesteps are configured in a chunk-wise manner, where each chunk comprises 3 latent frames. We set the error buffer $\mathcal{B}$ to a capacity of $V = 3,000$, storing and retrieving residuals on a global per-frame basis, and randomly sample the

Fig. 5: Qualitative comparisons with existing methods. Our method exhibits significantly stronger temporal dynamics and higher visual realism than existing distilled autoregressive video models such as Causvid and Self Forcing, and even surpasses the representative bidirectional diffusion model Wan2.1 in both dynamics and realism.

error injection scaling factor γ from $\{0, 0.1, 1\}$. For the subsequent distillation stage, the student generator is directly initialized with the causal weights obtained from pre-training and optimized for 2,700 steps across 32 A100 GPUs, constructing training trajectories via a self-rollout mechanism. We utilize the larger Wan2.1-T2V-14B model to instantiate both the teacher and the critic; notably, in the Internal DMD configuration, both the teacher and critic share the identical causal initialization as the student.

Dataset. We construct our pre-training dataset $\mathcal{D}$ by aggregating curated existing datasets and synthetically generated videos. Specifically, we carefully select 50K high-quality video clips from the MixKit [18] and UltraVideo [58] datasets. To further augment the training data, we employ the Wan2.1-T2V-1.3B model to generate approximately 70K synthetic videos, conditioned on text prompts sampled from the VidProM [49] dataset. All collected videos are standardized to a sequence length of 81 frames at a spatial resolution of 832×480. After a rigorous filtering process to discard low-quality and corrupted data, we obtain a final pre-training dataset comprising 100K high-fidelity text-video pairs.

Evaluation. Following Self Forcing, we evaluate on the VBench prompt list [11] and additionally adopt the TA-Hard prompt set proposed in [23], which emphasizes more complex semantic alignment. We report metrics from both VBench and VisionReward [56]. Specifically, from VBench, we report the Total, Quality, and Semantic scores. From VisionReward, we report the overall score, the

Table 1: Quantitative Comparison with baselines on VBench. We report VBench and VisionReward metrics on the standard evaluation prompts. For readability, all evaluation metrics are scaled by 100.

Model	VBench metrics ↑			VisionReward metrics ↑			
	Total Score	Quality Score	Semantic Score	Overall Score	Detail Quality	Motion Realism	Text Alignment
Wan2.1 [47]	75.53	73.86	67.17	3.94	12.26	-40.49	39.18
SkyReels-V2 [5]	78.48	81.37	66.93	7.71	29.32	-31.87	47.57
MAGI-1 [43]	78.15	80.50	68.77	7.18	28.40	-40.43	49.47
CausVid [66]	79.12	81.91	68.00	10.33	40.10	-24.47	56.73
Self Forcing [10]	<u>80.79</u>	<u>83.38</u>	<u>70.46</u>	11.33	41.86	-10.20	<u>64.76</u>
Longlive [60]	80.25	82.94	69.48	<u>11.35</u>	<u>42.71</u>	-18.71	64.27
Causal Forcing [68]	80.52	83.11	70.15	10.74	41.01	**-6.50**	62.79
Ours	**81.35**	**83.92**	**71.03**	**12.00**	**45.74**	<u>-6.98</u>	**64.83**

text-alignment score, and two additional discriminative sub-metrics. Since VisionReward scores can be negative, we note that larger values still indicate better performance. To evaluate real-time capability, we further report throughput in FPS and latency in seconds on a single A100 GPU. More details are provided in supplementary material.

4.2 Comparisons

For comparison, we consider a diverse set of baselines with comparable parameter scales. The evaluated methods cover three categories: bidirectional video diffusion models (Wan2.1 [47]), autoregressive video diffusion models (SkyReels-V2 [5], and MAGI-1 [43]), and distilled autoregressive video models (CausVid [66], Self Forcing [10], Longlive [60] and Causal Forcing [68]). Notably, while MAGI-1 utilizes a larger 4.5B parameter architecture, our method and all other baselines maintain a consistent and more efficient scale of 1.3B parameters.

As shown in Tabs. 1 and 2, our method consistently achieves the strongest overall performance on both the standard prompt set and the more challenging TA-Hard prompt set, highlighting its robustness under varying prompt difficulties. On the standard prompt set (Tab. 1), our approach achieves the best results across all VBench metrics, the highest VisionReward overall score, and the top Detail Quality score, surpassing strong distilled autoregressive baselines such as Self Forcing and Causal Forcing at a comparable parameter scale. On the TA-Hard prompt set (Tab. 2), where prompt understanding and generation are substantially more demanding, our advantage becomes even more pronounced: our model attains the best VBench total score, the highest semantic score by a clear margin, the best VisionReward overall score, the best Detail Quality score, and the best Motion Realism score. Notably, across both prompt sets, our method ranks either first or second on all reported metrics, demonstrating a consistently strong and well-balanced performance profile rather than gains on only

Table 2: Quantitative Comparison with baselines on TA-Hard. We report VBench and VisionReward metrics on the more challenging TA-Hard evaluation prompts. For readability, all evaluation metrics are scaled by 100.

Model	VBench metrics ↑			VisionReward metrics ↑			
	Total Score	Quality Score	Semantic Score	Overall Score	Detail Quality	Motion Realism	Text Alignment
Wan2.1	64.45	75.62	19.78	-6.96	-1.85	-86.81	-31.48
SkyReels-V2	61.88	73.08	17.07	-2.49	16.67	-81.25	-15.74
MAGI-1	62.60	74.58	14.70	-2.38	14.81	-93.75	-1.85
CausVid	63.09	74.84	16.09	1.20	<u>33.33</u>	-79.17	12.96
Self Forcing	67.69	78.98	22.56	1.43	31.48	-79.86	<u>21.30</u>
Longlive	63.72	76.31	13.38	<u>1.97</u>	32.41	<u>-78.47</u>	17.59
Causal Forcing	<u>69.42</u>	**81.03**	<u>22.99</u>	0.77	27.78	-81.94	**24.07**
Ours	**71.22**	<u>80.96</u>	**32.24**	**2.70**	**34.26**	**-69.44**	<u>21.30</u>

a small subset of metrics. Together with the qualitative results in Fig. 5, these results suggest that the improvements mainly arise from our architectural design rather than simple scaling, leading to better video quality, stronger semantic consistency, and improved robustness under harder prompts.

4.3 Analysis and Ablation

Effectiveness of Error Forcing. To validate the effectiveness of Error Forcing, we conduct comprehensive ablation studies across both the pre-training and distillation phases. Fig. 6 and Tab. 3 present our qualitative and quantitative results, respectively. As illustrated in Fig. 6(a), during the pre-training phase, Error Forcing yields stable generation, effectively mitigates the impact of accumulated errors, and demonstrates robust anti-drift capabilities. Furthermore, we observe that Teacher Forcing and Diffusion Forcing tend to produce over-smoothed, artificial textures that deviate from real-world aesthetics; Error Forcing successfully overcomes this limitation to preserve realistic details. For the distillation phase, we perform direct DMD initialized with the causal weights pre-trained via Error Forcing, Teacher Forcing, and Diffusion Forcing. As depicted in Fig. 6(b), our approach generates videos with superior visual quality and richer content. The quantitative results in Tab. 3 further corroborate the efficacy of our method, demonstrating that Error Forcing consistently outperforms the baselines in both the pre-training and distillation stages.

Error Setting. We conduct extensive experiments to investigate the optimal configuration for Error Forcing. Tab. 4, 6, and 5 present our ablation results across three distinct dimensions: (1) error storage granularity, (2) error injection region, and (3) error sampling strategy. Specifically, for error storage granularity (1), we evaluate the scale at which errors are stored and retrieved,

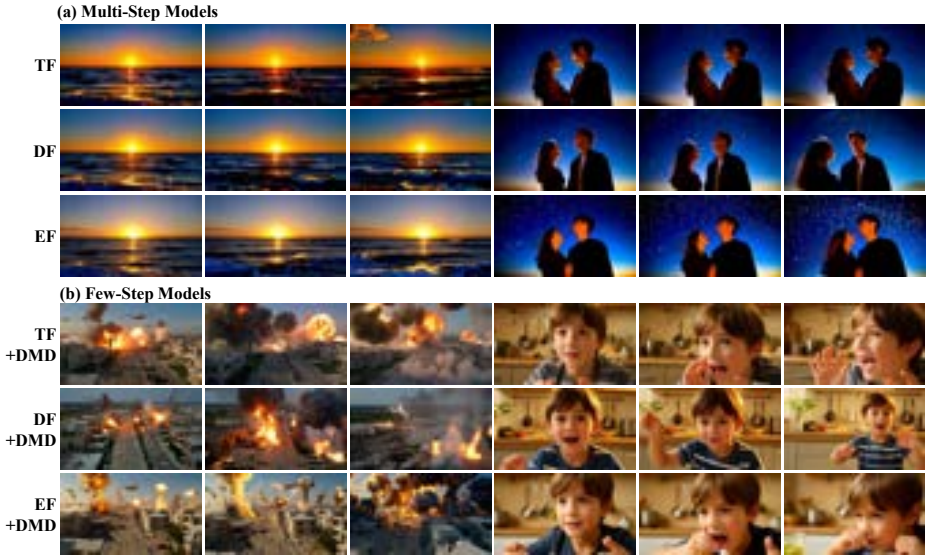

Fig. 6: Qualitative comparison between Error Forcing and other baselines. TF, DF, and EF denote Teacher Forcing, Diffusion Forcing, and Error Forcing, respectively. In few-step settings, "TF/DF/EF + DMD" represents initializing the student with the corresponding pre-trained weights and distilling it via self-forcing rollout.

where the "Window-wise" setting denotes processing the errors of an entire 21-frame latent simultaneously. For the error injection region (2), we investigate whether the errors should be exclusively applied to the conditioning context, or simultaneously injected into the target latents. When both regions are perturbed, we further compare the efficacy of applying independent errors versus a shared error, given that both the context and the target originate from the same continuous ground-truth sequence. For the error sampling strategy (3), we examine whether timestep indices should be recorded during collection to enable "Timestep-aware" retrieval corresponding to the sampled t, or if errors can be sampled randomly. As the results demonstrate, even though our training timesteps are chunk-wise, operating at a "Frame-wise" storage granularity yields the best overall performance. Furthermore, the optimal configuration during sampling is to randomly sample a shared error and inject it simultaneously into both the history context and the target latent.

When is a Causal Student Ready for Distillation? In this section, we investigate the optimal initialization strategy for the DMD process of a causal student. Existing methods typically incorporate ODE distillation to learn the generation trajectories of a multi-step teacher. Under a bidirectional initialization, ODE distillation is forced to simultaneously adapt to causal generation and few-step sampling capabilities. Conversely, under a causal initialization, the ODE distillation solely focuses on acquiring the few-step capability. This distinction motivates us to introduce *Internal DMD* (refer to Section 3.3), where we employ

Table 3: Quantitative comparison between Error Forcing and other baselines. In few-step settings, "TF/DF/EF + DMD" represents initializing the student with the corresponding pre-trained weights and distilling it via self-forcing rollout.

Models	VBench metrics ↑		
	Quality Score	Semantic Score	Total Score
Multi (50)-step models			
Diffusion Forcing (DF)	80.98	67.16	78.22
Teacher Forcing (TF)	81.34	68.19	79.15
Error Forcing (EF)	**81.89**	**68.95**	**79.64**

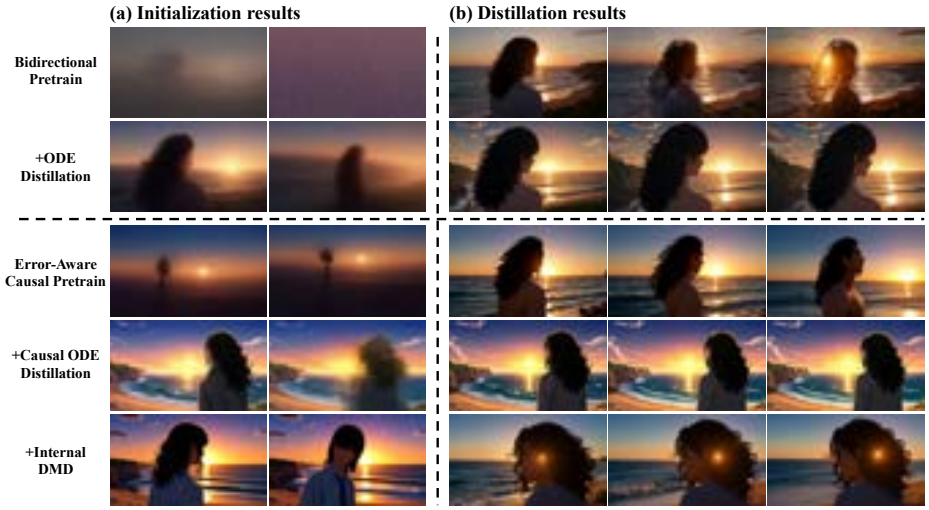

Fig. 7: Qualitative comparison of different initializations. (a) Few-step inference results of various bidirectional and causal initializations before distillation. Note that *Internal DMD* employs an architecturally identical teacher and critic to solely distill the few-step capability. (b) Corresponding results after external model (Wan2.1-T2V-14B) distillation.

a teacher and a critic that are architecturally identical to the student to exclusively distill the few-step generation ability. The results are presented in Tab. 7

Table 6: Ablation on error injection region.

Injection Region	VBench metrics ↑		
	Quality Score	Semantic Score	Total Score
Context only	79.91	67.15	77.36
Context + Target (independent errors)	80.49	61.45	77.40
Context + Target (shared error)	**81.89**	**68.95**	**79.64**

Table 7: Quantitative comparison of different initializations. (a) Few-step inference results of various causal initializations before distillation. (b) Corresponding results after external model distillation.

Models	VBench metrics ↑		
	Quality Score	Semantic Score	Total Score
(a) Initialization Results			
Bidirectional Pretrain	64.40	19.25	55.37
+ ODE	63.59	52.44	61.36
Error-Aware Causal Pretrain	56.89	43.10	54.14
+ Causal ODE	<u>76.59</u>	**70.52**	<u>75.38</u>
+ Internal DMD	**80.96**	<u>69.67</u>	**78.70**
(b) Distillation Results			
Bidirectional Pretrain	77.86	69.82	76.25
+ ODE	83.37	70.46	80.79
Error-Aware Causal Pretrain	<u>83.92</u>	**71.03**	<u>81.35</u>
+ Causal ODE	83.53	70.89	80.99
+ Internal DMD	**84.12**	<u>70.92</u>	**81.49**

and Fig. 6(a). We observe that causal ODE distillation frequently yields visual artifacts and abrupt temporal mutations. In contrast, Internal DMD effectively mitigates these issues, providing a superior few-step initialization.

Subsequently, we proceed to the distillation phase using different student initializations, adopting the Wan2.1-T2V-14B model as our teacher. As shown in Tab. 7, we find that directly applying DMD on a bidirectional initialization often leads to training collapse, thus necessitating an intermediate ODE distillation phase for optimization. For causal initialization, however, the ODE process can be safely bypassed. Our standard pipeline achieves comparable or even superior performance at a significantly reduced computational cost. Fig. 6(b) further corroborates this conclusion. Furthermore, while we note a marginal performance improvement when employing Internal DMD, we provide this configuration as an optional setting due to its doubled training cost.

5 Conclusion

In this work, we propose Error Forcing, which injects structured prediction errors into training history to better match inference conditions and stabilize distillation into real-time causal students. This yields a strong initialization that

enables direct DMD without expensive ODE trajectory matching, reducing training cost while outperforming prior distillation baselines on various benchmarks. Additional experimental results, discussion of limitations and future work will be provided in the supplementary material. All models, code, and data will be released.

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. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
3. 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)
4. Chen, F., Zhuang, B., Wu, Q.: Streaming video diffusion: Online video editing with diffusion models. In: 2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–9. IEEE (2025)
5. 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)
6. Feng, Y., Xiang, C., Mao, X., Tan, H., Zhang, Z., Huang, S., Zheng, K., Liu, H., Su, H., Zhu, J.: Vidarc: Embodied video diffusion model for closed-loop control. arXiv preprint arXiv:2512.17661 (2025)
7. 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)
8. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: European Conference on Computer Vision. pp. 102–118. Springer (2022)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. 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)
11. 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)
12. 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)
13. Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., Shechtman, E., Zhu, J.Y., Park, T.: Distilling diffusion models into conditional gans. In: European Conference on Computer Vision. pp. 428–447. Springer (2024)
14. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. arXiv preprint arXiv:2310.02279 (2023)

15. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)
16. 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)
17. Liang, F., Kodaira, A., Xu, C., Tomizuka, M., Keutzer, K., Marculescu, D.: Looking backward: Streaming video-to-video translation with feature banks. arXiv preprint arXiv:2405.15757 (2024)
18. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
19. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
20. Lin, S., Yang, X.: Animatediff-lightning: Cross-model diffusion distillation. arXiv preprint arXiv:2403.12706 (2024)
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
22. Liu, H., Zhang, W., Xie, J., Faccio, F., Xu, M., Xiang, T., Shou, M.Z., Perez-Rua, J.M., Schmidhuber, J.: Faster diffusion via temporal attention decomposition. arXiv preprint arXiv:2404.02747 (2024)
23. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
24. 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)
25. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
26. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: The Twelfth International Conference on Learning Representations (2023)
27. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE international conference on computer vision. pp. 4463–4471 (2017)
28. Luhman, E., Luhman, T.: Knowledge distillation in iterative generative models for improved sampling speed. arXiv preprint arXiv:2101.02388 (2021)
29. Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., Zhang, Z.: Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. Advances in Neural Information Processing Systems **36**, 76525–76546 (2023)
30. Mao, X., Jiang, Z., Wang, F.Y., Zhang, J., Chen, H., Chi, M., Wang, Y., Luo, W.: Osv: One step is enough for high-quality image to video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12585–12594 (2025)
31. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
32. Parker-Holder, J., Fruchter, S.: Genie 3: A new frontier for world models (Aug 2025), https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
33. PixVerse Research: Pixverse-r1: Next-generation real-time world model (Jan 2026), https://pixverse.ai/en/blog/pixverse-r1-next-generation-real-time-world-model

34. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
35. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
38. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. arXiv preprint arXiv:2310.14189 (2023)
39. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
40. 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. arXiv preprint arXiv:2512.14614 (2025)
41. Tang, J., Liu, J., Li, J., Wu, L., Yang, H., Zhao, P., Gong, S., Yuan, X., Shao, S., Lu, Q.: Hunyuan-gamecraft-2: Instruction-following interactive game world model. arXiv preprint arXiv:2511.23429 (2025)
42. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
43. 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)
44. 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)
45. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
46. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1020–1028 (2017)
47. 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)
48. Wang, F., et al.: Animatelcm: accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning. arxiv, 1 feb 2024 (2024)
49. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems **37**, 65618–65642 (2024)
50. Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., Sang, N.: Videolcm: Video latent consistency model. arXiv preprint arXiv:2312.09109 (2023)
51. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023)
52. 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)
53. Wu, C., Liang, J., Hu, X., Gan, Z., Wang, J., Wang, L., Liu, Z., Fang, Y., Duan, N.: Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. arXiv preprint arXiv:2207.09814 (2022)

54. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024)
55. 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)
56. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
57. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8196–8206 (2024)
58. Xue, Z., Zhang, J., Hu, T., He, H., Chen, Y., Cai, Y., Wang, Y., Wang, C., Liu, Y., Li, X., et al.: Ultravideo: High-quality uhd video dataset with comprehensive captions. arXiv preprint arXiv:2506.13691 (2025)
59. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
60. 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)
61. Yang, Y., Huang, H., Peng, X., Hu, X., Luo, D., Zhang, J., Wang, C., Wu, Y.: Towards one-step causal video generation via adversarial self-distillation. arXiv preprint arXiv:2511.01419 (2025)
62. 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)
63. 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)
64. 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)
65. 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)
66. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
67. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models for video generation. arXiv e-prints pp. arXiv–2504 (2025)
68. 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)
69. Zou, C., Liu, X., Liu, T., Huang, S., Zhang, L.: Accelerating diffusion transformers with token-wise feature caching. arXiv preprint arXiv:2410.05317 (2024)