CONDITIONAL RESIDUAL PREDICTION: IMPROVING AUTOREGRESSIVE VIDEO DIFFUSION WITHOUT A BIDIRECTIONAL TEACHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal video diffusion models generate video autoregressively, which suits streaming, interactive, and long-video generation. Under standard training, however, they often yield lower generation quality than bidirectional models of the same size. Many existing approaches address this gap by initializing from or distilling a pretrained bidirectional teacher. We instead train a causal model from an image-model initialization, with no bidirectional video model at any stage. Because this path requires neither a large bidirectional teacher nor a complex distillation pipeline, it is simpler and more scalable. On this path, we find that a causal model trained on ground-truth history becomes strongly dependent on it, so that at inference errors in its own generated history propagate forward. We hypothesize that much of this dependence is unnecessary, because the current input already determines much of what the history provides. We propose Conditional Residual Prediction (CRP), a simple recipe for reducing a model's reliance on a condition: the model first predicts the target without the condition, and the condition may only add a residual on top of this prediction. Applied to history, CRP makes the model predict each chunk from the present as far as it can and use the past only for what the present cannot supply. In controlled experiments, CRP nearly closes the 6.14-point gap to a bidirectional model trained under the same setup. Scaling this recipe, we train Optica, a 2B-parameter causal video model that autoregressively generates 5-second 480p videos and reaches 82.78 on VBench with only about 15M training videos.

1 INTRODUCTION

Many high-quality video diffusion systems are fixed-window, non-causal generators: they jointly denoise all latent frames in a clip, and their temporal operators can exchange information in both directions within that window (Ho et al., 2022b; Blattmann et al., 2023b;a; Yang et al., 2025; Kong et al., 2024; Team Wan et al., 2025; Chen et al., 2026a). Throughout this paper, "bidirectional" refers specifically to that connectivity, not to the forward and reverse processes of diffusion. Fixed-window generation is effective for short clips, but it must complete a window before emitting output and does not expose the temporal KV-cache structure used by streaming causal models (Yin et al., 2025; Huang et al., 2025; Yang et al., 2026b; Liu et al., 2026). By contrast, the autoregressive generation process of causal video models is naturally compatible with long-video and streaming settings (Yin et al., 2025; Chen et al., 2025a; Sand.ai et al., 2025; Huang et al., 2025; Yang et al., 2026b; Liu et al., 2026), with direct applications to practical scenarios such as world models and neural renderers (Ha & Schmidhuber, 2018; Bruce et al., 2024; Valevski et al., 2025; Liang et al., 2025).

Despite these theoretical advantages, causal video models often remain unsatisfactory in practice. More specifically, even for short durations such as 5 seconds, a causal video model produces videos with substantially lower visual quality and semantic fidelity than a bidirectional model with a comparable parameter count.

Recent works have recognized this phenomenon and proposed several solutions. Early methods initialize a causal student from a pretrained bidirectional video model and retain a bidirectional teacher

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

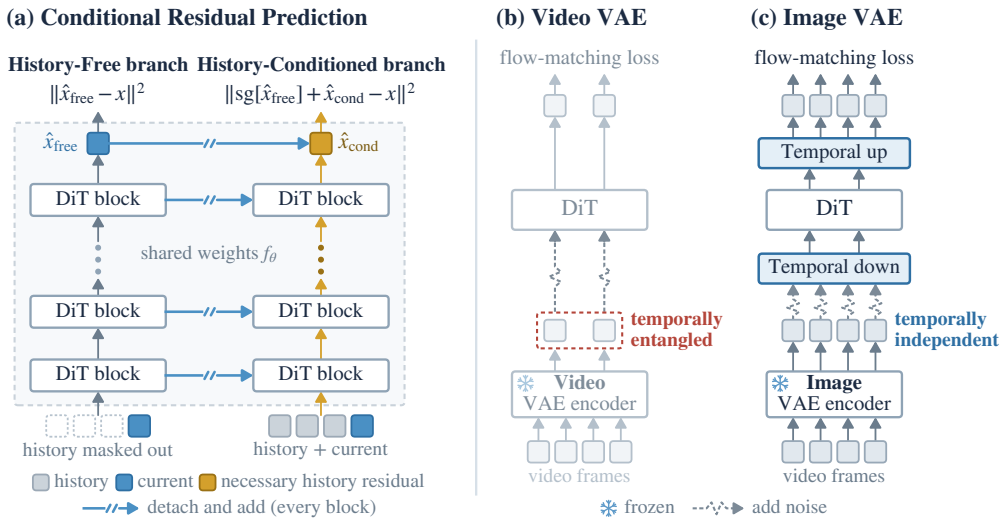

Figure 1: **Overview.** **(a)** Conditional Residual Prediction (CRP) applied to history. The same network f_θ is evaluated twice on the same input: the History-Free branch masks out the history and predicts $\hat{x}_{\text{free}}$, and the History-Conditioned branch sees the history and predicts only a residual $\hat{x}_{\text{cond}}$ on top of it. The History-Free hidden states are detached and added to the History-Conditioned branch after every block, never the reverse, so the History-Free branch stays free of history. **(b)** A video VAE encodes the first frame into its own latent and every later four frames into one, with causal temporal convolutions that entangle latents across time. **(c)** An image VAE encodes every frame independently into the same kind of latent; a temporal down/up layer keeps the cost of the DiT unchanged.

for ODE initialization or score distillation (Yin et al., 2025; Huang et al., 2025). Later methods still start from a bidirectional base but replace the ODE- or score-providing teacher with a causal one, at initialization or throughout scoring (Zhu et al., 2026; Zhao et al., 2026; Bandyopadhyay et al., 2026).

Compared with vanilla causal video models, these methods have achieved clear success, approaching or even surpassing their bidirectional teacher models. Yet their complex training and reliance on a pretrained teacher limit scaling: a stronger causal model first needs a stronger bidirectional one. More importantly, recent work finds that using a causal rather than bidirectional teacher for distillation yields a better causal student (Zhu et al., 2026; Bandyopadhyay et al., 2026): the mismatch between bidirectional and causal models limits post-distillation performance, and a bidirectional teacher can leak future information that the student cannot access at generation time. This line of work further motivates our objective. Directly training a multi-step causal video model without relying on a bidirectional model not only provides a simple and scalable multi-step generator in its own right, but may also enable distillation into a stronger, efficient few-step model.

Our approach centers on one idea: reducing over-reliance on history. We find that a causal model trained on a ground-truth prefix becomes strongly dependent on it, and we hypothesize that much of this dependence is unnecessary: it covers information that the current input already determines (Sec. 3.1). We propose Conditional Residual Prediction (CRP; Fig. 1(a)), which first predicts the target without a condition and lets the condition add only a residual on top of this prediction. Applied to history, we split the forward passes over the same input into a History-Free branch and a History-Conditioned branch and combine them to form the final output. To prevent the History-Conditioned branch from contaminating the History-Free branch, we detach the output of the History-Free branch with a stop-gradient, so that the History-Conditioned loss does not backpropagate through it; the shared parameters are updated by both losses. Under this design, the History-Conditioned branch effectively models the residual of the History-Free prediction. Intuitively, the History-Free branch models as much as possible without relying on history, while the History-Conditioned branch sup-

plies only the signals that necessarily depend on history. In our controlled experiments, CRP, together with the x_0-prediction it builds on, narrows the VBench Total gap between the causal and bidirectional models from 6.14 to 0.19, with CRP itself accounting for 4.68 points.

CRP acts within the DiT, but the video VAE introduces a second source of history dependence that no change to the DiT can remove: its causal temporal compression couples latent frames. We therefore also use an image VAE rather than a video VAE. On top of CRP, the image VAE further narrows the gap from 0.19 to -0.20. It also suits CRP: encoding every frame independently removes the encoder-induced distinction between single-frame and four-frame latents.

Overall, this paper demonstrates that a causal video model can achieve strong performance without relying on a bidirectional teacher: building on CRP, we train Optica, a 2B-parameter causal video model that reaches 82.78 on VBench. It also demonstrates the effectiveness of reducing history over-reliance and offers a simple recipe for native, scalable training of causal video models.

2 Background

Let the latent sequence of a video be $x = (x_1, \ldots, x_N)$, where the subscript denotes a position along the temporal dimension. Let y be the text condition and f_θ the DiT to be trained. We use this notation throughout the paper.

2.1 Bidirectional Video Diffusion Model

In image diffusion, attention is typically bidirectional (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Peebles & Xie, 2023). This carried over naturally to video diffusion, where it has been successful (Ho et al., 2022b; Singer et al., 2023; Ho et al., 2022a; Blattmann et al., 2023b; Guo et al., 2024; Blattmann et al., 2023a) and now dominates the field (Yang et al., 2025; Lin et al., 2024; Team Wan et al., 2025; Wu et al., 2025a; HaCohen et al., 2026; Team Seedance, 2026; MiniMax, 2026). Such methods draw one noise level t shared across all positions, form the interpolation $x^t = (1 - t)\, x + t\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and regress the corresponding velocity (Lipman et al., 2023; Liu et al., 2023). Attention is bidirectional along time, so every position attends to every other within the window, and generation integrates the ODE jointly over the whole window from $t = 1$, producing all positions at once.

2.2 Causal Video Diffusion Model

Causal video diffusion models (Weng et al., 2023; Valevski et al., 2025; Chen et al., 2024; Ruhe et al., 2024; Jin et al., 2025; Sun et al., 2025; Chen et al., 2025a; Sand.ai et al., 2025; Gu et al., 2025; Deng et al., 2025) differ from bidirectional models above all in how they treat time, factorizing it by the chain rule: $p(x \mid y) = \prod_{i=1}^{N} p(x_i \mid x_{<i}, y)$. A matching causal attention mask in the DiT keeps information flowing in one direction only. Some works apply causality at the chunk level (Yin et al., 2025; Huang et al., 2025; Yang et al., 2026b; Sand.ai et al., 2025) rather than the frame level (Weng et al., 2023; Valevski et al., 2025; Chen et al., 2024; Sun et al., 2025; Gu et al., 2025; Cheng et al., 2025; Bai et al., 2026), which places them between the two extremes; adjusting the chunk size turns such a model into either a bidirectional or a causal one (Ma et al., 2026). Throughout this paper we call a model causal when its chunk size is below two seconds and bidirectional when it is above.

Training normally uses teacher forcing (Williams & Zipser, 1989): the model is conditioned on the ground-truth prefix and noise is added only to the current position, so the noised query at position i attends to the clean prefix $x_{<i}$ and to itself: $\mathcal{L}_{\text{causal}} = \mathbb{E}[\sum_{i=1}^{N} \| f_\theta(x_i^{t_i},\, x_{<i},\, t_i,\, y) - (\epsilon_i - x_i) \|^2]$. Here, each position has an independent noise level t_i, while the conditioning prefix $x_{<i}$ comes from real data. Inference instead proceeds autoregressively, that is, on the model's own generated frames as history: for $i = 1, \ldots, N$, the model conditions on the generated prefix $\hat{x}_{<i}$ and integrates the ODE from $x_i^1 \sim \mathcal{N}(0, I)$ to obtain $\hat{x}_i$. Under autoregressive inference the generated history may differ from the real history, which brings a training–inference gap and hence a loss of quality. This is the central problem studied in this paper.

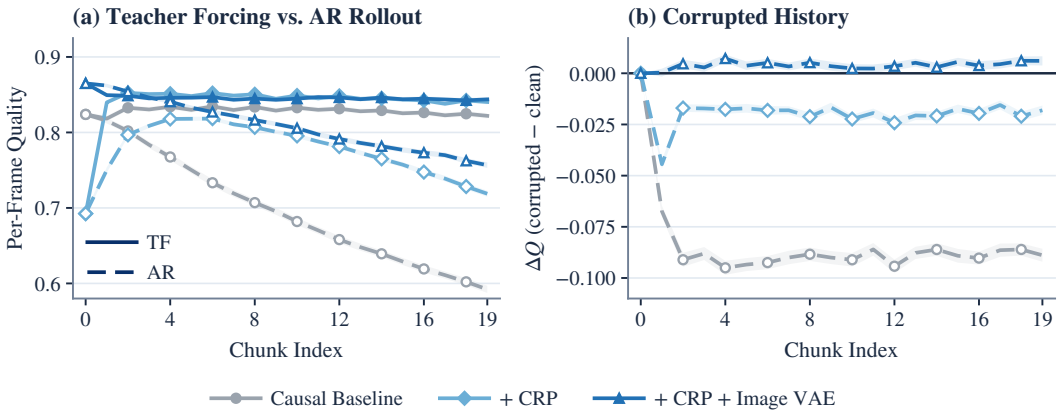

Figure 2: **Per-frame image quality degradation and comparison.** Per-frame no-reference quality is measured with Q-ReAlign Mini (Wu et al., 2024). **(a)** Teacher-forced (TF) generation remains approximately stable, whereas autoregressive (AR) quality decreases with chunk index. **(b)** Quality change under a corrupted history, $\Delta Q = Q_{\text{corrupted}} - Q_{\text{clean}}$, with each corruption type triggered independently per history frame with probability 0.3; negative values indicate degradation. Curves show means and shaded regions standard errors across clips. CRP and the image VAE progressively reduce the degradation. With the video VAE, + CRP falls below the baseline at chunk 0, which is generated without history; the image VAE removes this drop (Sec. 3.3).

2.3 Initialization and Distillation from a Bidirectional Teacher

Many works address the quality degradation of causal video models by exploiting a high-quality bidirectional model: some build on it by initialization or fine-tuning, as short-chunk causal models (Gao et al., 2025; Zhang et al., 2025) or clip- and shot-level extenders (Henschel et al., 2025; Guo et al., 2025; Li et al., 2026), while others distill from it, usually with Diff-Instruct (Luo et al., 2023) or DMD (Yin et al., 2024) objectives (Yin et al., 2025; Huang et al., 2025; Cui et al., 2026; Yang et al., 2026b; Liu et al., 2026; Chen et al., 2026b; Zheng et al., 2026; Bai et al., 2026), or post-train the initialized causal model adversarially (Lin et al., 2025). This line of work has been markedly successful, with the causal student approaching or even surpassing its bidirectional teacher.

Recent work further finds that the bidirectional teacher is itself a bottleneck, and that making each stage causal works better. Self Forcing (Huang et al., 2025) replaces the ground-truth prefix with the model's own rollout during training. Causal Forcing (Zhu et al., 2026) traces the problem to ODE initialization and uses a teacher-forced multi-step causal model as the ODE teacher. Context-Matched Distillation (Bandyopadhyay et al., 2026) makes the score-providing teacher causal as well, since a bidirectional one sees future frames the student cannot. Appendix A expands these comparisons.

3 Method

3.1 The History Over-Reliance Hypothesis

The most important difference between a causal video model and a bidirectional model is the former's autoregressive generation process: the model first generates the earlier portion and then uses it as the basis for generating later positions. A common source of poor quality during autoregressive generation is exposure bias (Bengio et al., 2015; Lamb et al., 2016; Ross et al., 2011; Ranzato et al., 2016). Errors in early positions are not only carried forward as fixed history and propagated to later positions, but also move the sequence away from the ground-truth history distribution seen during training, thereby degrading model performance. This train–test mismatch has also been observed directly in autoregressive video diffusion (Huang et al., 2025; Guo et al., 2026). We first quantify the mismatch through per-frame image quality, scored with Q-ReAlign Mini, a modern checkpoint following the discrete-level visual-scoring formulation of Q-Align (Wu et al., 2024), over 1,000 real clips (Appendix D.2). On our causal baseline (Sec. 4.2), quality under a ground-truth prefix is flat

across the clip, 0.824 in the first chunk and 0.822 in the twentieth, whereas the same model under an autoregressive rollout falls to 0.592, a mean gap of 0.138 ± 0.003 (Fig. 2(a)). Corrupting the ground-truth history rather than replacing it with generated frames reproduces the effect, costing 0.089 ± 0.001 (Fig. 2(b)). We use these two diagnostics throughout to test whether an intervention reduces history reliance, independently of VBench.

History dependence thus poses an apparent dilemma. The model must rely on the history to generate the current position, yet this reliance lets errors accumulate. Simply weakening or severing its access to the history reduces these errors but also removes useful signals, which hurts performance. We decompose this problem from an information-theoretic perspective. Information provided by history may be redundant with information at later positions; that is, some information can be recovered from the current-position signal alone, without relying on history.

More precisely, let the history be $H = x_{<i}$, the current input be $c = x_i^t$, the target be $x = x_i$, and the text input be y, where i denotes the temporal position and t the noise level in the diffusion or flow-matching formulation. By the chain rule for mutual information (Cover & Thomas, 2006),

$$I(x; c, H, y) = I(x; c, y) + I(x; H \mid c, y).$$

We call the first term the *history-free* information, which the current input and text determine on their own, and the second the *history-exclusive* information, which only the past can supply.

Hypothesis 1 (History Over-Reliance)**.** A causal video model trained on a ground-truth prefix relies on the history for history-free information instead of modeling it from the current input, and thereby acquires history dependence that the task does not require.

Why is this hypothesis plausible, and why would it cause severe exposure bias? During training, the model receives a ground-truth prefix, so relying on the ground-truth prefix for this History-Free signal yields a lower error than generating it independently. At inference, however, the history itself contains errors that accumulate during generation, exacerbating exposure bias. If the model generates this component by itself, the training loss may be higher, but at inference time the component no longer depends on history and therefore avoids interacting with its accumulated errors.

We therefore separate history-free and history-exclusive information explicitly: the former is modeled from the current input, and a History-Conditioned branch models only the remaining residual. We formalize this design as Conditional Residual Prediction in Sec. 3.2, and complement it in Sec. 3.3 by removing the history dependence that the video VAE introduces. CRP also differs from two standard techniques: training on corrupted history and on rollout history. Diffusion Forcing (Chen et al., 2024), the best-known example of the former, injects noise into the history during training. If inference uses a clean history, training and inference become inconsistent; if inference keeps the noise, it discards history information that the current position needs, by an amount set by a hand-crafted noise level that is hard to control. Training with rollout history, or scheduled sampling (Bengio et al., 2015), changes the training objective and can lead to a suboptimal solution (Huszár, 2015). See Appendix A for more discussion.

3.2 Conditional Residual Prediction

We first focus on the DiT, where history may only add a residual on top of a History-Free prediction.

Definition 1 (Conditional Residual Prediction)**.** Let a network f_θ predict a target x from an input c and a condition h. CRP evaluates the same network twice: a condition-free branch $\hat{x}_{\text{free}} = f_\theta(c)$ that never accesses h, and a residual branch $\hat{x}_{\text{cond}} = f_\theta(c, h)$ that may also receive the detached features of the condition-free branch but passes nothing back to it. With a loss ℓ, the two branches are trained with

$$\mathcal{L}_{\text{free}} = \ell(\hat{x}_{\text{free}}, x), \qquad \mathcal{L}_{\text{cond}} = \ell(\text{sg}[\hat{x}_{\text{free}}] + \hat{x}_{\text{cond}}, x),$$

where the stop-gradient $\text{sg}[\cdot]$, which we use throughout, keeps $\mathcal{L}_{\text{cond}}$ from backpropagating through the condition-free branch, and the prediction is $\hat{x}_{\text{free}} + \hat{x}_{\text{cond}}$, or $\hat{x}_{\text{free}}$ alone when h is unavailable. The condition-free branch thus carries no information about h in its features or output; everything that depends on h is carried by the residual branch, without additional parameters.

In this paper, ℓ is the squared error, and we instantiate Definition 1 with the noisy current chunk and the text as the input, $c = (x_i^t, y)$, the history as the condition, $h = H = x_{<i}$, and the clean chunk as the target, $x = x_i$.

Conditional residual branches have conceptual precedents in ControlNet and temporal-residual video diffusion (Zhang et al., 2023; 2024). ControlNet, however, injects an additional, separately parameterized condition branch into a frozen base, so that the features and output of the base carry the condition. CRP reverses this direction within a single network and keeps the condition-free branch clean. SCD separates temporal causal reasoning from frame-wise denoising (Bai et al., 2026). Appendix A further discusses History Guidance and remedies for over-reliance on history in imitation learning.

Table 1: Roadmap of the controlled ablation, trained on 4M videos at 256px and 5s. Each row adds one configuration change on top of the row above, and Δ is the resulting change in Total. The causal baseline is substantially worse than the bidirectional baseline, CRP nearly closes this gap, and the image VAE further brings it on par with the bidirectional model. All scores use the standard VBench prompt set and its fixed seeds; Appendix F compares reconstruction in image and video modes.

Configuration	Total	Quality	Semantic	Δ
Bidirectional reference	77.28	79.64	67.82	–
Causal baseline	71.14	75.13	55.16	–
+ x_0-prediction	72.41	74.96	62.22	+1.27
+ CRP	77.09	78.09	73.10	+4.68
+ Image VAE	77.48	78.43	73.72	+0.39

Specifically, for the same network and the same set of inputs, we compute two forward-pass results simultaneously (Fig. 1(a)): a History-Free branch with past frames masked out and a History-Conditioned residual branch with access to past frames. During sampling, the first chunk, which has no history, uses only $\hat{x}_{\text{free}}$. We pass the detached hidden state of the History-Free branch to the History-Conditioned branch after every block, allowing the latter to process History-Free signals in latent space rather than only in output space. Since the residual is most naturally defined on the clean target, CRP uses x_0-prediction.

3.3 From Video VAE to Image VAE

CRP acts within the DiT, but the latents it models come from the VAE. Like many common video VAEs, the Wan 2.1 VAE we adopt (Team Wan et al., 2025) is built from causal temporal convolutions: it encodes the first frame into its own latent and every later four frames into one, and each latent depends on the frames before it (Fig. 1(b)). This causes two problems.

First, the VAE introduces history dependence of its own, which no change to the DiT can remove: it learns to reconstruct later positions partly from preceding signals. We observe that, under the common setting of 5 seconds at 16 fps, an error in an early latent propagates all the way to the last frame (Appendix E). In autoregressive generation, an error in an earlier chunk is thus carried into every later one.

Second, the video VAE produces two types of latent, a single-frame latent for the first frame and a four-frame latent for each later group of four frames, and one model must generate both. The two are also inherently imbalanced: the single-frame latent appears only at the first position and receives a small share of training, which an extra image task can alleviate. In CRP, the imbalance becomes more pronounced: the History-Free branch predicts without history at every position, almost always on four-frame latents. Since the first chunk is generated by this branch alone, it falls below the baseline (Fig. 2(a), chunk 0). Rather than adding an image task, we unify the latent representation.

Motivated by these two issues, we replace the video VAE with an image VAE, which encodes every frame independently into the same kind of latent (Fig. 1(c)). To avoid introducing other changes, we keep the same Wan 2.1 VAE but use its image mode instead of its video mode, which loses no fidelity (1.11 dB higher paired PSNR; Appendix F). This removes both problems: a perturbation of one latent no longer spreads to later frames, and the chunk-0 drop disappears.

To keep the DiT's cost unchanged with four times as many latents, we add a lightweight temporal patch layer that merges every four adjacent latents at the input and splits them at the output (Fig. 1(c); Appendix B.2). Consistent with JiT (Li & He, 2026), we observe that x_0-prediction performs better than velocity prediction with these higher-dimensional tokens. We refer to this complete configuration as "+ Image VAE."

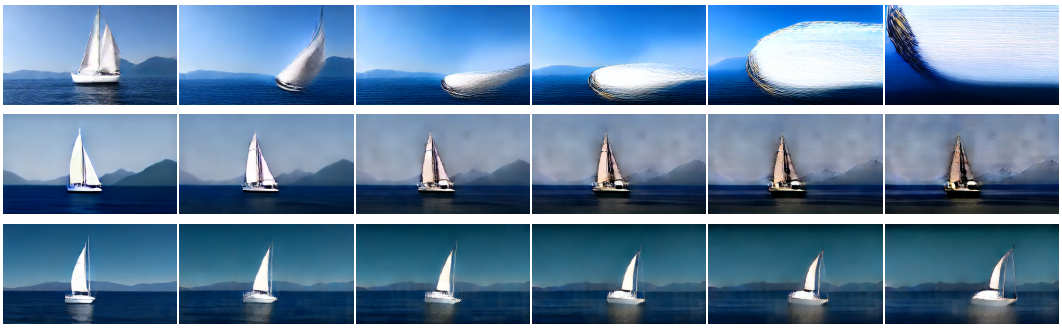

Figure 3: **Qualitative effect of reducing history over-reliance over an autoregressive rollout.** Rows, top to bottom: causal baseline; + CRP; + CRP + Image VAE. All three are the 4M-sample controlled-ablation models under the same sampling parameters, so the differences between them isolate the effect of each added component. The causal baseline deteriorates rapidly as the temporal index increases. With + CRP, the scene keeps its structure throughout, but later frames darken and develop speckle artifacts; adding the image VAE removes them.

4 Controlled Ablation Experiments

4.1 Setup

Data For the controlled ablation experiments, we use approximately 4M internal video clips at 256p resolution and a duration of 5 seconds. Further details on the data are provided in Appendix C.

Model Our model is initialized from the 1.6B-parameter SANA-1.5 T2I checkpoint (Xie et al., 2025a;b). Following SANA-Video (Chen et al., 2026a), we add a temporal convolutional layer to the MLP and adapt kernelized linear attention (Katharopoulos et al., 2020) for RoPE (Su et al., 2024). We replace 25% of the linear-attention layers with softmax attention. For the causal model, all components involving the temporal dimension, including temporal convolution, attention, and RoPE, are made strictly causal. The resulting model contains 2B parameters. We use the Wan 2.1 VAE. Since SANA-1.5 is pretrained in the DC-AE latent space (Chen et al., 2025b), we follow DC-VideoGen (Chen et al., 2025c) and adapt its input and output layers to the Wan 2.1 latent space before the main training (Appendix B.3).

4.2 Bidirectional and Causal Baseline

We first train bidirectional and causal baselines under this setup. For the baseline models, we use the standard velocity loss. For the causal baseline, we use teacher forcing rather than diffusion forcing, following prior work (Zhu et al., 2026). The causal baseline scores 71.14 VBench Total, well below the bidirectional baseline's 77.28. This is consistent with prior autoregressive video diffusion results (Yin et al., 2025; Huang et al., 2025). Qualitatively, we observe clear exposure bias: as generation proceeds, the causal baseline rapidly departs from the real-image distribution, producing distorted textures, incorrect structures, and artifacts (Fig. 3, first row).

4.3 Causal Model with CRP

Starting from the causal baseline, we first add CRP to the DiT, keeping the video VAE. Because CRP predicts a residual on top of an x_0 prediction, we first switch the causal baseline from velocity prediction to x_0-prediction, which on its own raises VBench Total from 71.14 to 72.41 (Table 1), and then add CRP (Sec. 3.2).

CRP raises VBench Total from 72.41 to 77.09, close to the bidirectional baseline of 77.28. CRP alone therefore accounts for 4.68 of the 6.14-point gap between the causal and bidirectional baselines. On the two diagnostics in Fig. 2, CRP reduces the quality gap between autoregressive generation and teacher forcing from 0.138 ± 0.003 to 0.066 ± 0.003, and the quality loss caused by

Table 2: Comparison with video generators on VBench. Total, Quality, and Semantic are short-video VBench-T2V scores (Huang et al., 2024). For our model, we report scores on 480p+5s videos generated autoregressively in 5 chunks. Appendix D.1 lists the sources of the quoted scores.

Method	Steps	#Params	Bi. source	Use	Total	Quality	Semantic
Fixed-window, non-causal							
Step-Video (Ma et al., 2025)	50	30B	–	–	81.83	84.46	71.28
Open-Sora-2.0 (Zheng et al., 2025)	50	11B	–	–	84.34	85.40	80.12
HunyuanVideo (Kong et al., 2024)	50	13B	–	–	83.24	85.09	75.82
Wan2.1 (Team Wan et al., 2025)	50	14B	–	–	83.69	85.59	76.11
CogVideoX1.5 (Yang et al., 2025)	50	5B	–	–	82.17	82.78	79.76
SANA-Video (Chen et al., 2026a)	50	2B	–	–	83.71	84.35	81.35
LTX-Video (HaCohen et al., 2024)	20	1.9B	–	–	80.00	82.30	70.79
Wan2.1 (Team Wan et al., 2025)	50	1.3B	–	–	84.26	85.30	80.09
Causal, distilled from a bidirectional teacher							
CausVid (Yin et al., 2025)	4	1.3B	Wan-1.3/14B	init.+dist.	81.20	84.05	69.80
Self Forcing (Huang et al., 2025)	4	1.3B	Wan-1.3/14B	init.+dist.	84.31	85.07	81.28
LongLive (Yang et al., 2026b)	4	1.3B	Wan-1.3/14B	init.+dist.	84.87	86.97	76.47
Rolling Forcing (Liu et al., 2026)	5	1.3B	Wan-1.3/14B	init.+dist.	81.22	84.08	69.78
Context Forcing (Chen et al., 2026b)	4	1.3B	Wan-1.3B	init.+dist.	83.44	84.98	77.29
Causal Forcing (Zhu et al., 2026)	4	1.3B	Wan-1.3/14B	init.+dist.	84.04	84.59	81.84
Causal-rCM (Zheng et al., 2026)	2	1.3B	Wan-1.3/14B	init.+dist.	84.63	85.46	81.31
SCD (Bai et al., 2026)	4	1.6B	Wan-1.3/14B	init.+dist.	84.03	85.14	79.60
Causal, bidirectional initialization only							
SkyReels-V2 (Chen et al., 2025a)	30	1.3B	own bi. model	init.	82.67	84.70	74.53
Causal, no bidirectional video model							
NOVA (Deng et al., 2025)	25	0.6B	–	–	80.12	80.39	79.05
Pyramid Flow (Jin et al., 2025)	20	2B	–	–	81.72	84.74	69.62
MAGI-1 (Sand.ai et al., 2025)	64	4.5B	–	–	79.18	82.04	67.74
Optica (ours)	50	2B	–	–	82.78	83.02	81.83

corrupted history from 0.089 ± 0.001 to 0.020 ± 0.001. In the rollout of Fig. 3, the scene keeps its structure throughout, although later frames darken and develop speckle artifacts.

4.4 Causal Model with Image VAE

We then replace the video VAE with the image VAE of Sec. 3.3, together with the temporal patch layer, which adds about 40M parameters, or 2% of the total (Fig. 1(c)).

The image VAE further raises VBench Total from 77.09 to 77.48, matching the bidirectional baseline.

Adding the image VAE further reduces the two diagnostics to 0.039 ± 0.002 and -0.004 ± 0.001, respectively; at this corruption strength, there is no longer any measurable degradation. It also removes the drop of CRP at chunk 0 (Fig. 2(a); Sec. 3.3) and the darkening and speckles in later frames (Fig. 3).

5 Scaling Experiments

Data Building on the controlled experiments, we continue training at 256p and 5 seconds up to 7M, and then switch to 480p and train up to 14.5M. Finally, we perform supervised fine-tuning (SFT) for 0.5M samples on internal high-quality data (Appendix C), for 15M samples in total.

Model The scaled-up model continues training from the controlled-experiment checkpoint. All controlled experiments use a chunk size of one latent position; for the scaled-up model, we increase the chunk size to four latent positions and fine-tune directly.

Results Optica achieves VBench Total/Quality/Semantic scores of 82.78/83.02/81.83. It is trained on only approximately 15M internal real video clips, contains 2B parameters, and uses no initialization, ODE sample, or score from a bidirectional teacher.

Among methods that do not rely on any bidirectional video model, Optica achieves the highest VBench Total score: 82.78, exceeding Pyramid Flow, NOVA, and MAGI-1 by 1.06, 2.66, and 3.60 points, respectively. This advantage comes primarily from Semantic: Optica scores 81.83, compared with 69.62, 79.05, and 67.74. Our Quality score remains below that of Pyramid Flow (83.02 vs. 84.74).

We observe a similar pattern when comparing against causal models distilled from bidirectional teachers. Optica's Semantic score is already comparable to those of the strongest models in this group. For example, Causal Forcing, Causal-rCM, and Self Forcing achieve Semantic scores of 81.84, 81.31, and 81.28, respectively, compared with 81.83 for Optica. On Quality, however, Optica remains 1.0–4.0 points lower. The same holds relative to the bidirectional SANA-Video 5-second checkpoint, the closest published model architecturally: their Semantic scores are nearly identical (81.83 vs. 81.35), and the primary gap lies in Quality (83.02 vs. 84.35).

The remaining gap should be read together with the training budget. Although many distilled causal models contain only 1.3B parameters, they are initialized from or distilled with pretrained Wan video models, commonly at 1.3B and in some cases with 14B teacher supervision (Yin et al., 2025; Huang et al., 2025; Yang et al., 2026b; Zheng et al., 2026; Bai et al., 2026). The cost of this upstream training is not reflected in their parameter counts. By contrast, Optica uses only about 15M video clips and starts from an image model. More importantly, in our controlled experiments with identical data, model sizes, and evaluation protocols, our causal model already matches the bidirectional reference (Table 1). This points to training budget, rather than the causal architecture, as the main source of the remaining gaps.

6 Conclusion

This paper studies history over-reliance in causal video models: teacher forcing supplies a consistently correct prefix, so the model comes to rely on history for signals it could model from the current position, and this reliance amplifies history errors during autoregressive inference. We address this with Conditional Residual Prediction, which splits the DiT into a History-Free branch and a History-Conditioned branch that predicts only a residual on top of it. We complement it with an image VAE that removes history dependence between latents. In controlled experiments, CRP narrows the 6.14-point VBench Total gap to the bidirectional reference to 0.19, and the image VAE narrows it further to -0.20, while substantially reducing quality degradation under generated and corrupted histories. The scaled 2B-parameter Optica uses no initialization, sample, or score from a bidirectional teacher and achieves 82.78 on VBench.

History remains necessary for modeling motion, identity, and temporal consistency; what CRP removes is the reliance on history for what the current position can model on its own. This also makes Optica complementary to distillation rather than an alternative to it. Recent work finds that a few-step causal model benefits from distilling a multi-step causal teacher rather than a bidirectional one (Zhu et al., 2026), and Optica offers a way to train such a teacher natively, keeping the causal factorization consistent between the pretrained model and the downstream student.

The same decomposition also suggests a target for compressing long context or memory in causal video models: keep the history-exclusive information, which the current position cannot infer on its own, and discard the rest. This gives an information-theoretic rather than heuristic criterion for which parts of the history are useful, and can guide the design of concrete compression algorithms.

Our experiments have several limitations. All controlled and scaling runs evaluate only 5-second videos, so we have not verified that CRP keeps suppressing error accumulation over longer autoregressive rollouts. We also instantiate CRP only with the history as the condition; other conditions that generative models are known to over-rely on, such as the conditioning image in image-to-video generation (Zhao et al., 2024a), are left for future work. This version of Optica also does not yet support image-to-video generation. The image VAE gives up temporal compression and the History-Free branch adds a forward pass, so the current 50-step model does not demonstrate real-time efficiency, and we have not yet distilled Optica into a few-step model.

AI USE STATEMENT

LLMs were used for language polishing, coding assistance, reviewing mathematical formulas, and searching for related literature. The core method and research were developed without AI assistance. The authors take full responsibility for the final content of this paper.

REFERENCES

Xingjian Bai, Guande He, Zhengqi Li, Eli Shechtman, Xun Huang, and Zongze Wu. Causality in video diffusers is separable from denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43373–43384, 2026. URL https://arxiv.org/abs/2602.10095.

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation. *arXiv preprint arXiv:2608.13391*, 2026. URL https://arxiv.org/abs/2608.13391.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.03099.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b. URL https://arxiv.org/abs/2304.08818.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024. URL https://arxiv.org/abs/2407.01392.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a. URL https://arxiv.org/abs/2504.13074.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. SANA-Video: Efficient video generation with block linear diffusion transformer. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2509.24695.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2410.10733.

Junyu Chen, Wenkun He, Yuchao Gu, Yuyang Zhao, Jincheng Yu, Junsong Chen, Dongyun Zou, Yujun Lin, Zhekai Zhang, Muyang Li, Haocheng Xi, Ligeng Zhu, Enze Xie, Song Han, and Han Cai. DC-VideoGen: Efficient video generation with deep compression video autoencoder. *arXiv preprint arXiv:2509.25182*, 2025c. URL https://arxiv.org/abs/2509.25182.

Shuo Chen, Cong Wei, Sun Sun, Tiancheng Shen, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2602.06028.

Xinle Cheng, Tianyu He, Jiayi Xu, Junliang Guo, Di He, and Jiang Bian. Playing with transformer at 30+ FPS via next-frame diffusion. *arXiv preprint arXiv:2506.01380*, 2025. URL https://arxiv.org/abs/2506.01380.

Chia-Chi Chuang, Donglin Yang, Chuan Wen, and Yang Gao. Resolving copycat problems in visual imitation learning via residual action prediction. In *European Conference on Computer Vision*, volume 13699 of *Lecture Notes in Computer Science*, pp. 392–409. Springer, 2022. doi: 10.1007/978-3-031-19842-7_23. URL https://arxiv.org/abs/2207.09705.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X. URL https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.02283.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.11979.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.14169.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-VDM: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *International Conference on Machine Learning*, pp. 18550–18565, 2025. URL https://arxiv.org/abs/2411.16375.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022. doi: 10.1007/978-3-031-19790-1_7. URL https://arxiv.org/abs/2204.03638.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. URL https://arxiv.org/abs/2503.19325.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.04725.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17281–17291, 2025. URL https://arxiv.org/abs/2503.10589.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. In *European Conference on Computer Vision*, volume 17021 of *Lecture Notes in Computer Science*, pp. 324–344. Springer, 2026. doi: 10.1007/978-3-032-37574-2_18. URL https://arxiv.org/abs/2512.15702.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11495.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2568–2577, 2025. URL https://arxiv.org/abs/2403.14773.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a. URL https://arxiv.org/abs/2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022b. URL https://arxiv.org/abs/2204.03458.

Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.07696.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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. doi: 10.1109/CVPR52733.2024.02060. URL https://arxiv.org/abs/2311.17982.

Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015. URL https://arxiv.org/abs/1511.05101.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05954.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. A study of BFLOAT16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019. URL https://arxiv.org/abs/1905.12322.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020. URL https://arxiv.org/abs/2006.16236.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1610.09038.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026. URL https://arxiv.org/abs/2511.13720.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.09212.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26069–26080, 2025. URL https://arxiv.org/abs/2501.18590.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-Sora Plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. URL https://arxiv.org/abs/2412.00131.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.09350.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25161.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/s11633-025-1562-4. URL https://arxiv.org/abs/2211.01095.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.18455.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-Video-T2V technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. URL https://arxiv.org/abs/2502.10248.

Xinyin Ma, Julius Berner, Chao Liu, Arash Vahdat, Weili Nie, and Xinchao Wang. Flex-Forcing: Towards a unified autoregressive and bidirectional video diffusion model. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2607.03509.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. URL https://arxiv.org/abs/1907.07484.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.03740.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research Blog, 2026. URL https://www.minimax.io/blog/minimax-h3.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://arxiv.org/abs/2212.09748.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06732.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://arxiv.org/abs/1011.0686.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42818–42835. PMLR, 2024. URL https://proceedings.mlr.press/v235/ruhe24a.html.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. URL https://arxiv.org/abs/2505.13211.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14792.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56242–56280. PMLR, 2025. URL https://proceedings.mlr.press/v267/song25b.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://arxiv.org/abs/2104.09864.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. AR-Diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7364–7373, 2025. URL https://arxiv.org/abs/2503.07418.

Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.14837.

VBench Team. VBench leaderboard. https://huggingface.co/spaces/Vchitect/VBench_Leaderboard, 2026. Accessed September 2026.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.09853.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2010.14876.

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. ART•V: Auto-regressive text-to-video generation with diffusion models. *arXiv preprint arXiv:2311.18834*, 2023. URL https://arxiv.org/abs/2311.18834.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a. URL https://arxiv.org/abs/2511.18870.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video VAE for latent video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18124–18133, 2025b. doi: 10.1109/CVPR52734.2025.01689. URL https://arxiv.org/abs/2411.06449.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.10629.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68578–68598. PMLR, 2025b. URL https://proceedings.mlr.press/v267/xie25b.html.

Lingxiao Yang, Liu Liu, Moran Li, Han Feng, Wenjian Cao, Jiangning Zhang, and Ye Shi. In-context forcing: Uncovering context effects in autoregressive video diffusion. *arXiv preprint arXiv:2608.05237*, 2026a. URL https://arxiv.org/abs/2608.05237.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2509.22622.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06072.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024. URL https://arxiv.org/abs/2311.18828.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. 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. URL https://arxiv.org/abs/2412.07772.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023. URL https://arxiv.org/abs/2212.05199.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion—tokenizer is key to visual generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.05737.

Lin Zhang, Sicheng Mo, Zefan Cai, Jinhong Lin, Zihao Lin, Jiuxiang Gu, Krishna Kumar Singh, Yuheng Li, and Yin Li. UniTemp: Unlocking video generation in any temporal order via bidirectional distillation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.18702.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://arxiv.org/abs/2302.05543.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.12626. Spotlight.

Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. TRIP: Temporal residual learning with image noise prior for image-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8671–8681, 2024. URL https://arxiv.org/abs/2403.17005.

Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://arxiv.org/abs/2406.15735.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026. URL https://arxiv.org/abs/2605.15141.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. CV-VAE: A compatible video VAE for latent generative video models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12847–12871, 2024b. URL https://arxiv.org/abs/2405.20279.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026. URL https://arxiv.org/abs/2606.25473.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-Sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025. URL https://arxiv.org/abs/2503.09642.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02214.

A Extended Related Work

Masked-Conditional Video Diffusion MCVD (Voleti et al., 2022), FDM (Harvey et al., 2022), and RaMViD (Höppe et al., 2022) can also be rolled out autoregressively by conditioning on previously generated frames, but they are trained to condition on arbitrary frames, including future ones, and are therefore not causal in the sense of Sec. 2.2.

Causal Distillation These works extend the pipeline of Sec. 2.3. Self-Forcing++ (Cui et al., 2026) re-noises long rollouts to push self-rollout training past the teacher's window. LongLive (Yang et al., 2026b) adds KV recaching at prompt switches and a frame-level attention sink for real-time interactive streaming. Rolling Forcing (Liu et al., 2026) jointly denoises a rolling window with bidirectional intra-window attention while emitting frames sequentially. Context Forcing (Chen et al., 2026b) distills from a long-context teacher and restructures the KV cache. Causal-rCM (Zheng et al., 2026) combines teacher-forcing consistency distillation with self-forcing DMD under a forward/reverse divergence view. SCD (Bai et al., 2026) decouples once-per-frame causal reasoning from frame-wise

denoising. Causal Forcing++ (Zhao et al., 2026) inherits Causal Forcing's teacher-forcing and DMD stages, replaces its causal ODE stage with causal consistency distillation, and extends to frame-wise one- and two-step generation.

Training with Perturbed History Diffusion Forcing (Chen et al., 2024) samples each noise level independently, so that 16 frames with 1000 timesteps each already admit $O(10^{48})$ noise-level combinations (Sun et al., 2025). Rolling Diffusion (Ruhe et al., 2024), AR-Diffusion (Sun et al., 2025), SkyReels-V2 (Chen et al., 2025a) and MAGI-1 (Sand.ai et al., 2025) restrict the schedule to be non-decreasing toward the future. The clean prefix can also be corrupted on top of a teacher-forced schedule. MAGI-1 injects up to 5% noise into its clean chunks (Sand.ai et al., 2025). SkyReels-V2 adds noise to the clean condition at inference to stabilize long rollouts (Chen et al., 2025a). Pyramid Flow noises the history during training while conditioning on clean frames at inference (Jin et al., 2025). GameNGen finds corrupting context frames critical for long-horizon visual stability (Valevski et al., 2025). In-Context Forcing (Yang et al., 2026a) argues that clean context frames leak local detail and invite shortcuts; it adds more noise to adjacent frames and less to distant ones while leaving the prediction target fixed, partly in service of cross-frame parallel denoising, and studies few-step models downstream of a distillation pipeline. Such corruption is not selective: it weakens the history-exclusive information that the current position needs along with the redundant history-free information. It also leaves a dilemma between training and inference. With noise augmentation as in Diffusion Forcing, noise applied only in training leaves inference with a clean history that training rarely provides, whereas noise applied at inference as well discards information by an amount set by the noise level rather than by what the current position needs. Frame-, block-, and patch-level corruption or dropout face the same trade-off. Our runs condition on a clean prefix throughout, following Zhu et al. (2026), who report that plain teacher forcing usually outperforms diffusion forcing, a finding our internal experiments reproduce.

Training with Rollout History Self Forcing (Huang et al., 2025) unrolls the student with KV caching inside the training loop and supervises the rollout with a distribution-matching objective; the extensions above (Cui et al., 2026; Yang et al., 2026b; Liu et al., 2026; Zheng et al., 2026) keep this structure. A rollout carries no ground-truth target, so all of them draw one from the distillation pipeline, and hence from a teacher. Resampling Forcing (Guo et al., 2026) does without one and is the closest teacher-free work to ours: it shares our target of exposure bias under teacher forcing, our stop-gradient, and our avoidance of any teacher or discriminator. It autoregressively resamples the history with the online weights, detaches that rollout, and asks the model to predict a clean target from the degraded prefix, so it learns to *rectify* history errors. Supervising a self-generated history with the ground-truth continuation, as in scheduled sampling (Bengio et al., 2015), is an improper objective in general (Huszár, 2015): once the rollout drifts or collapses, the same generated history is consistent with many real continuations, so the regression target either averages over them or asks the model to recover information that the history no longer contains. Resampling Forcing initializes from pretrained Wan2.1-1.3B weights rather than from an image model, and its history routing targets long-horizon attention cost, which we do not address. Both families change the conditioning signal; CRP changes the prediction target instead, at the price of one extra forward pass.

Controlling the History Dependency History Guidance (Song et al., 2025) uses a CFG-style mechanism to inject the history signal at inference time, allowing the strength of the history signal to be controlled. It evaluates a single network with and without history, but trains it to predict the full target in every case and combines the predictions only at sampling time. MAGI-1 (Sand.ai et al., 2025) sets a CFG guidance coefficient for the history separately from the text one, and schedules it over the denoising steps. SCD (Bai et al., 2026) separates the history-information extractor from the frame-wise denoiser, making the denoiser depend on a compressed history and thereby providing a bottleneck on the history signal. In imitation learning, policies conditioned on observation histories are known to over-rely on the history. This is addressed with targeted interventions (de Haan et al., 2019), with an adversarial objective that removes information about the previous action (Wen et al., 2020), or by training a separate history module to predict the change of action and feeding its stop-gradiented features to a policy that predicts the full action (Chuang et al., 2022). By contrast, the residual in CRP is added to a stop-gradiented History-Free prediction of the same network.

Image-initialized causal video models Pyramid Flow (Jin et al., 2025) initializes from SD3 Medium, NOVA (Deng et al., 2025) from its own from-scratch text-to-image model, and GameNGen (Valevski et al., 2025) and ART·V (Weng et al., 2023) from Stable Diffusion. Our setting belongs to this group, and Pyramid Flow is the closest published point of comparison in Table 2.

B Implementation Details

Optimization All models are trained with AdamW (Loshchilov & Hutter, 2019) (β_1=0.9, β_2=0.999, ϵ=10^{-10}, weight decay 0) under a constant learning-rate schedule with a linear warmup of 500 steps per data-parallel replica. The learning rate is 3.5×10^{-5}, except for the final SFT stage of the scaled model (Sec. 5), which uses 1×10^{-5} with all other settings unchanged. Gradients are clipped to a norm of 0.5. Training uses bf16 mixed precision (Kalamkar et al., 2019; Micikevicius et al., 2018) on an H20 cluster with a global batch size of 128. We do not use an EMA of the weights.

Objective We use flow matching with a linear schedule (Lipman et al., 2023; Liu et al., 2023). Timesteps are drawn from a logit-normal distribution (μ=0, σ=1) and warped by a training shift of 3.0. Each latent position receives an independently sampled timestep and independent noise; when the chunk size exceeds one, a single timestep is shared within a chunk and sampled independently across chunks. Text conditions are dropped with probability 0.1 during training to enable classifier-free guidance at inference (Ho & Salimans, 2022).

Sampling We sample with a flow-adapted DPM-Solver++ multistep implementation (Lu et al., 2025), using 50 steps and an inference timestep shift of 7.0 (versus 3.0 during training) in all experiments. Classifier-free guidance uses an empty-string negative prompt and no separate guidance on the history stream; its text scale is 6.5 for all controlled ablations, diagnostics, and visualizations (Sec. 4.1–4.4 and the corresponding appendices), and 9.5 for the scaling experiments (Sec. 5). Videos have 81 frames with the video VAE and 80 frames with the image VAE, all at 16 fps, and are generated at 256p in the controlled experiments and at 480×832 in the scaling experiments. Each latent position covers four frames: it holds one latent of the video VAE (except the first, which covers a single frame) or four per-frame latents of the image VAE merged by the temporal down layer. The controlled causal models generate 20 chunks, with an extra first-frame chunk for the video VAE; the scaled model generates 5 chunks of four latent positions each.

B.1 Architecture

The DiT is warm-started from an adapted 1.6B-parameter SANA-1.5 checkpoint (Xie et al., 2025b) and keeps its width and depth: 20 blocks at hidden width 2240, 20 attention heads of dimension 112. The input layer is a 3D patch embedding with patch size $(1, 2, 2)$ over the latent, i.e. no temporal patching at this stage; temporal compression is handled separately (Appendix B.2). Captions are encoded by a frozen Gemma-2-2B-IT decoder (Gemma Team et al., 2024) at a maximum length of 512 tokens.

Temporal conv in the MLP Following SANA-Video (Chen et al., 2026a), each block's MLP carries a zero-initialized temporal convolution with kernel size 3 and a shortcut after the spatial GLU path. In our implementation this is a full (not depthwise) $\mathrm{Conv2d}$ with kernel $(3, 1)$ over the (T, HW) layout.

Attention Following SANA-Video (Chen et al., 2026a), we apply a ReLU feature map to q and k and only then apply RoPE (Su et al., 2024) to the features, making the two compatible. Positions use axis-wise video RoPE; we set θ=10^4 along time and θ=2×10^3 along height and width. We replace 25% of the blocks with softmax attention; with depth 20 the rule places softmax at blocks $\{3, 7, 11, 15, 19\}$, i.e. every fourth block ending at the last.

B.2 Temporal Patch Layer

With the image VAE every latent frame corresponds to one video frame, so we restore the token budget inside the DiT rather than in the VAE. Immediately after the patch embedding, a down-projection maps each group of p=4 adjacent latent frames to a single token group, and a matching

up-projection inverts it before the output layer; the transformer blocks therefore operate on $80/4 = 20$ temporal positions.

Both projections are implemented as a strided convolution along time with kernel size equal to stride, which is exactly a bias-free linear map on the p concatenated frames. The pair costs 40M parameters, about 2% of the model; Fig. 1(c) shows where they sit.

B.3 Adapting SANA-1.5 to the Wan VAE Latent Space

SANA-1.5 is pretrained in the DC-AE latent space (Chen et al., 2025b), whose $32\times$ spatial compression and 32 latent channels differ from the Wan 2.1 VAE's $8\times$ and 16, so the pretrained input and output layers cannot be reused as they are. We adapt them in two stages, following the embedding-alignment recipe of DC-VideoGen (Chen et al., 2025c).

The first stage aligns the two embedding spaces. The new input layer is a randomly initialized 3D patch embedding with stride $(1, 2, 2)$ over the Wan latent, and its output tokens are average-pooled 2×2 so that the grid matches the one the original SANA-1.5 input layer produces from the DC-AE latent of the same image. With everything else frozen, we train the 146K parameters of this layer to minimize the MSE between the two embeddings. It converges quickly: the cosine similarity between them exceeds 0.4 within 500 steps and then climbs to about 0.7. The second stage freezes the DiT blocks and trains only the input and output layers with the flow-matching loss on Wan VAE latents, under the same prediction parameterization as the run it initializes. This stage is image-only and consumes on the order of 1M samples.

The DC-VideoGen code had not yet been released at the time of writing, so the procedure above is our own reimplementation from the paper's description. It sped up convergence of the subsequent video training in our runs.

C Training Data

Our training data come from an internal video corpus. Raw videos are first split into single-shot clips with a shot-boundary detector, and the starting frame of each clip is chosen for sharpness; the clips are then deduplicated and concept-balanced using video-level embeddings. A cascade of filters discards clips with low aesthetic or technical quality, visual blur, and little or no motion as measured by optical flow. The experiments in this paper use a subset of about 20M clips drawn from the resulting pool. Each clip is 5–6 seconds long, and during training we uniformly sample a 5-second window at 16 fps.

Captions are generated for whole clips by a vision-language model and come in short, medium, and long variants, which are sampled during training with probabilities of 10%, 30%, and 60%, respectively. The controlled experiments and the subsequent 256p/480p scaling experiments use the same dataset and differ only in training resolution. After these stages, the scaled model is further fine-tuned (SFT) for 0.5M samples on internal high-quality video data; the controlled experiments do not include this stage.

D Evaluation Protocol

D.1 VBench

We use the official VBench T2V evaluation (Huang et al., 2024). For the scaled model, following the evaluation protocol in the official VBench repository, we generate videos with the augmented prompts from Huang et al. (2025) but compute scores using the original short VBench prompts. The results for Optica are measured at 480×832, 80 frames, and 16 fps. We sample for 50 steps with a flow-adapted DPM-Solver++ implementation (Lu et al., 2025), using a text CFG scale of 9.5 and an inference flow shift of 7.0. The controlled ablations (Table 1) instead generate 256p videos directly from the standard short VBench prompts, with 50 steps, a flow shift of 7.0, and a text CFG scale of 6.5 (Appendix B).

In Table 2, the scores of Step-Video, Open-Sora-2.0, HunyuanVideo, Wan2.1-14B, CogVideoX1.5, and LTX-Video are taken from the public VBench leaderboard (VBench Team, 2026); those of

Wan2.1-1.3B, SkyReels-V2, MAGI-1, and CausVid from Table 1 of Huang et al. (2025), which evaluates CausVid through its official Wan2.1-1.3B implementation; and all other rows from their own papers. The Rolling Forcing Total is taken from Chen et al. (2026b), since Liu et al. (2026) reports only Quality and Semantic. Evaluation protocols, including prompt rewriting, resolution, and sampling, differ across rows.

D.2 Per-Frame Quality and the History-Corruption Diagnostic

To quantify quality degradation during autoregressive rollouts, we compare the per-frame image quality of teacher forcing and autoregressive generation on 1,000 real video clips. The autoregressive setting performs a complete autoregressive rollout. The teacher-forcing setting instead uses real latents as history and generates each chunk sequentially. Both settings use a flow shift of 7.0 and a text CFG scale of 6.5. For the video VAE, 81 frames correspond to 21 latent positions; Fig. 2 reports the first 20 positions (0–19), including the initial position.

We use the public Q-ReAlign Mini checkpoint to evaluate the image quality of every frame. Q-ReAlign reuses the discrete text-defined rating formulation introduced by Q-Align (Wu et al., 2024). Specifically, the model predicts probabilities for five ratings (excellent, good, fair, poor, and bad), which are weighted by 1.00, 0.75, 0.50, 0.25, and 0.00, respectively, to obtain the final score. We first average all frames within each chunk, and then use each video clip as the statistical unit when computing the mean and standard error.

To more directly measure model sensitivity to erroneous history, we corrupt only the ground-truth history used as the condition while leaving the current chunk to be generated unchanged. We use five corruption types: brightness, contrast, saturation, elastic transform, and smooth grid warp. The magnitude of each of the first three corruptions is 0.20. Elastic transform uses severity 3 from the imagecorruptions library (Michaelis et al., 2019), which implements the corruptions of Hendrycks & Dietterich (2019), and the maximum displacement of smooth grid warp is 0.03 times the shorter image side. Each corruption is triggered independently on every frame with probability 0.3, so multiple corruptions may be applied to the same frame, while a frame on which none is triggered remains unchanged. In Fig. 2, ΔQ denotes $Q_{\text{corrupted}} - Q_{\text{clean}}$ for the same video clip and chunk. All curves report the mean and standard error over 1,000 clips.

D.3 VAE Decoder Perturbation Analysis

To analyze the history dependence introduced by the video VAE, we perform latent-perturbation experiments directly, without involving the DiT. We first encode 1,000 real video clips using the Wan 2.1 VAE, then add Gaussian perturbations to one latent frame:

$$z_i' = z_i + 3\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

while leaving all other latent frames unchanged. We then decode the original and perturbed latents separately and compute the pixel-space RMS for every output frame. This metric first computes the root mean square over the spatial dimensions and then averages over channels. Figure 4(a) reports the mean over all clips, with shading indicating the 95% confidence interval computed by clip-level bootstrap.

For Fig. 4(b), at each latent position we pool all clips and spatial positions, estimate the mean and variance of each channel, and compute $\mathrm{KL}(q(z_t) \,\|\, \mathcal{N}(0, I))$ using a moment-matched Gaussian. We finally average over channels to obtain the KL per latent dimension.

E History Dependence in the Video VAE

Like many common video VAEs, the Wan 2.1 VAE we adopt (Team Wan et al., 2025) uses an architecture based on causal temporal convolutions; related temporally compressive representations are used throughout latent video generation (Yang et al., 2025; Kong et al., 2024; HaCohen et al., 2024; Yu et al., 2023; 2024; Zhao et al., 2024b; Ge et al., 2022). Related consequences of this design have been noted before, though not for autoregressive rollouts: IV-VAE (Wu et al., 2025b) shows that causal temporal convolution gives frames unequal information access and unbalanced reconstruction quality, and UniTemp (Zhang et al., 2026) identifies the same temporal bias as an obstacle to

generating in reverse temporal order. Neither concerns exposure bias or error accumulation during autoregressive generation, which is the question our analysis targets. We find that, when paired with the VAE training objective, the causal architecture creates dependencies between earlier and later frames. Specifically, perturbing an earlier frame affects the decoded results at later positions. Per-frame latents, which avoid this coupling, have precedent: early latent video diffusion models operated on them (Blattmann et al., 2023b;a), and FAR (Gu et al., 2025) adopts an image VAE for frame-level autoregression, though without arguing for the choice.

As shown in Fig. 4(a), an error introduced at one position propagates to subsequent positions. Under the common setting of 5 seconds at 16 fps, an error introduced in an early latent propagates all the way to the last frame. We also examine the KL divergence between per-channel moment-matched Gaussian approximations of the Wan VAE latents and the standard Gaussian prior (Fig. 4(b)).

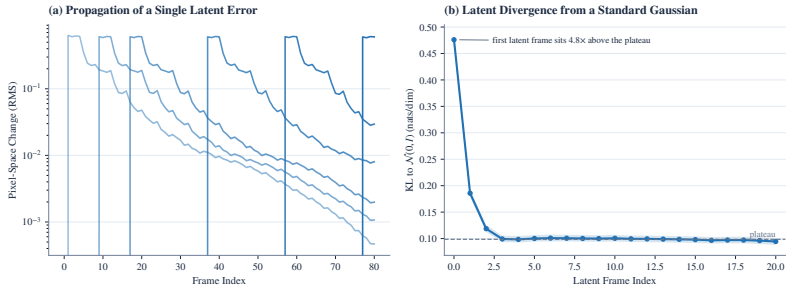

Figure 4: **The video VAE couples latent frames. (a)** Pixel-space RMS change over output-frame index after perturbing a single latent frame and decoding the same sequence, for several perturbation locations; every perturbation produces a persistent tail that reaches the end of the clip. **(b)** KL between per-channel moment-matched Gaussian approximations and the standard Gaussian prior, averaged over channels at each latent position. The KL is largest at the first latent frame and lower from latent frame 3 onward.

Under this moment-matched Gaussian approximation, the first latent frame has a substantially larger KL to the prior than later frames. From latent frame 3 onward, the KL stays on a plateau $4.8\times$ below that of the first latent frame. The first latent thus has different statistics from the later ones, matching the two latent types of Sec. 3.3. For causal video, the propagation in Fig. 4(a) means that once a signal at an earlier position is generated incorrectly, the VAE continually carries it into all subsequent positions.

Note that we remove all history dependence from the VAE, not only the unnecessary part, and leave history modeling entirely to the DiT, where CRP handles it. Another possible path is to retain necessary dependence on history in the VAE while removing only unnecessary dependence. This alternative could further compress the signal and reduce the modeling difficulty of the DiT, thereby improving overall performance. We discuss this possibility in the hope of motivating further exploration by the community.

F VAE Reconstruction: Image vs. Video Mode

We verify that changing the temporal mode of the VAE does not improve the downstream model by sacrificing reconstruction fidelity. Both settings use the same pretrained Wan 2.1 VAE checkpoint; video mode retains its causal temporal convolutions, whereas image mode encodes and decodes every frame independently. We evaluate both modes on the same 1,000 real-video clips, each containing 81 frames at 192×336 resolution. For each mode, we encode and immediately decode the clip without involving the DiT.

We clamp both the input and reconstruction to $[0, 1]$, compute PSNR for every frame, average frames within each clip, and finally average over clips. Uncertainty is computed with clips as the statistical unit.

Image mode improves paired PSNR by 1.11 ± 0.04 dB. Figure 5 shows the corresponding qualitative comparison, and additionally visualizes in pixel space the latent coupling measured in Fig. 4(a).

Table 3: Reconstruction fidelity of the same Wan 2.1 VAE checkpoint under video-mode and image-mode encoding/decoding. Values are mean per-frame PSNR $\pm$ standard error over 1,000 paired real-video clips. Encoding latency is the synchronized batch-one mean for an 81-frame clip.

VAE mode	PSNR (dB)	Encode (ms/clip)
Video mode (causal temporal)	27.93 ± 0.13	754
Image mode (per-frame)	29.04 ± 0.15	909

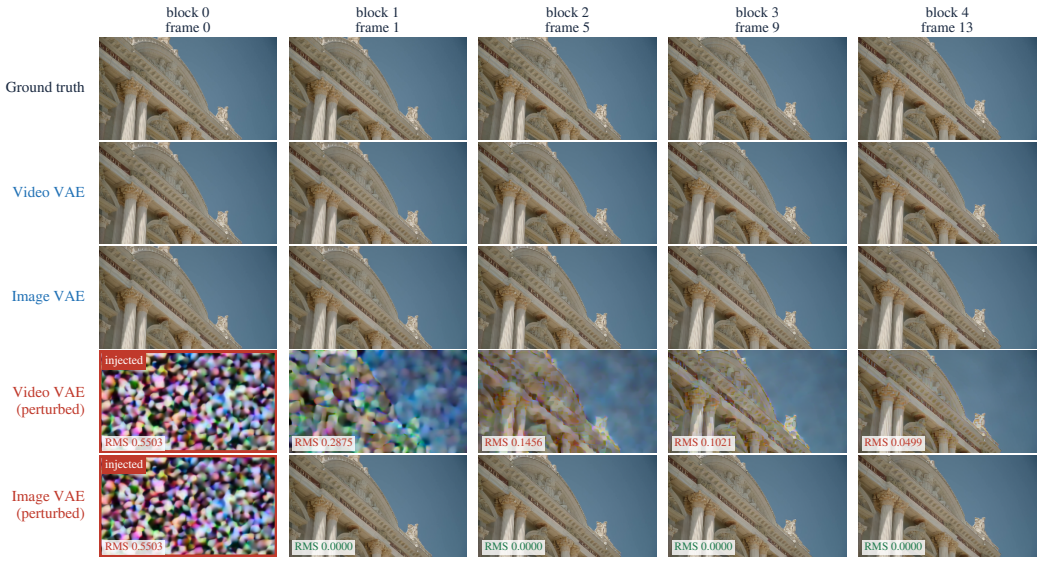

Figure 5: **Reconstruction quality is comparable between the two VAE modes, but only the video VAE propagates a latent error across frames.** No DiT is involved. Columns are the pixel frames $0, 1, 5, 9, 13$. Rows 2–3 are ordinary reconstructions. Rows 4–5 repeat them after adding 3ϵ, $\epsilon \sim \mathcal{N}(0, I)$, only to the latent frame that owns pixel frame 0; the red box marks that injected column, which therefore changes in both modes. Insets give the pixel-space RMS change from the corresponding clean reconstruction. Past the injected column the video VAE still drifts by 0.05–0.29 RMS, while the image VAE is exactly 0.

Clean reconstructions from the two modes are visually close, consistent with the small paired PSNR gap in Table 3. The perturbed rows behave very differently. Following the protocol of Appendix D.3, we perturb the single latent frame that owns pixel frame 0 and decode the sequence again. In video mode, the corruption remains plainly visible at every subsequent displayed frame. In image mode, the same latent frame owns pixel frame 0 alone, and every later frame is recovered bit-for-bit.