ON-POLICY FORCING: AUTOREGRESSIVE VIDEO GENERATION WITHOUT BIDIRECTIONAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent autoregressive (AR) video generators often rely on a separate bidirectional teacher for few-step distillation, limiting the role of AR training as a native foundation for the full generation pipeline. In this work, we show that teacher-forcing AR diffusion training can produce strong many-step video generators. We then introduce *On-Policy Forcing*, which directly distills the trained AR model into a few-step student while preserving the same causal factorization from AR training through few-step distillation. By using AR models for both real- and fake-score estimation during on-policy distillation, the method removes the need for an additional bidirectional teacher and the conditioning mismatch of asymmetric distillation, in which bidirectional score models see future context that the AR student cannot. With all methods evaluated by our VBench pipeline, our four-step student distilled with 1.3B AR score models reaches 84.08 Total, outperforming Self Forcing and Causal Forcing distilled with 1.3B bidirectional score models by 0.86 and 0.79 points, respectively. At an equal training budget, On-Policy Forcing improves Dynamic Degree on the official VBench subset by 28.33 points over distillation with bidirectional real- and fake-score models, while a conditioning mismatch in the fake-score model alone leads to severe degradation. These results validate AR training as a viable foundation for video generation and establish a fully causal training-to-distillation recipe for AR video models.

1 INTRODUCTION

Autoregressive (AR) video generation models provide a foundation for applications that require continuous prediction and interaction, including real-time video synthesis, interactive world modeling, and embodied intelligence. These applications require models that anticipate how a scene evolves and respond as new observations and actions arrive. Next-frame prediction captures this requirement by training models to predict the next observation from the available history and current inputs. Like next-token prediction in GPT (Radford et al., 2018; Brown et al., 2020), this objective connects learning from sequential data with generating its continuation, providing a foundation for video models that can interact with their environment.

Many current methods obtain AR video generators through asymmetric distillation (Yin et al., 2025; Huang et al., 2025). However, these methods have two inherent limitations. First, a bidirectional (BI) teacher can use future context, so its predictions generally differ from the past-conditioned predictions learned through teacher forcing under AR maximum likelihood. Second, the student learns the data distribution indirectly through the teacher, rather than directly from video data. Causal Forcing and its successors (Zhu et al., 2026; Zhao et al., 2026; Zheng et al., 2026) introduce an AR teacher for initialization but retain BI supervision for refinement, leaving the dependence on a separate BI teacher in place.

This remaining BI supervision matters because DMD updates the student with the difference between a real and a fake score evaluated on the same noisy block. The AR student generates each block from its own prefix, whereas a BI score model conditions the block on the whole noisy video, including future blocks. We call this a *conditioning mismatch*. Causal Forcing removes the corresponding mismatch from initialization by distilling from an AR teacher (Zhu et al., 2026). We

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.

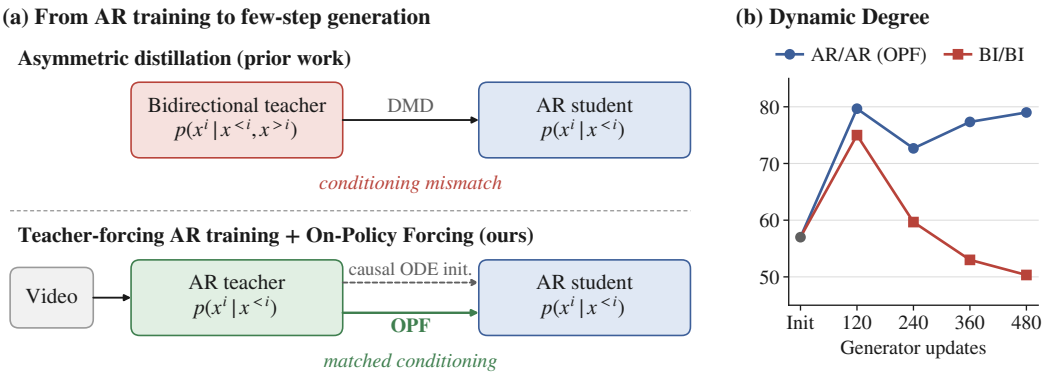

Figure 1: Autoregressive generation without bidirectional distillation. **(a)** Top: prior few-step AR video models (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026) are distilled asymmetrically, with bidirectional (BI) score models that condition each block on both past and future context, whereas the student conditions only on the past, a conditioning mismatch. Bottom: On-Policy Forcing (OPF) trains an AR teacher on video with teacher forcing and uses it both for causal ODE initialization, as in Causal Forcing, and for the final distillation stage. **(b)** VBench Dynamic Degree on 100 rich-motion prompts when both score models are AR (AR/AR, i.e., OPF) or both are BI (BI/BI), starting from the same causal ODE initialization (Init) with 1.3B score models. AR/AR stays above BI/BI at every checkpoint, whereas BI/BI declines below Init as training continues. Points show means over three training seeds (Table 10).

remove it from the final stage. We can drop the BI teacher because teacher forcing alone already trains a strong AR model. Self Forcing and Causal Forcing find it the stronger AR training scheme (Huang et al., 2025; Zhu et al., 2026). Its weak few-step results in Self Forcing come from students trained on ground-truth context, an exposure bias that on-policy distillation removes. *On-Policy Forcing* (OPF) keeps teacher-forcing AR training and causal ODE initialization, and replaces both BI score models with AR ones: the AR teacher serves as the frozen real-score model, paired with an online AR fake-score model, and both condition on the student's generated prefix. The AR teacher, trained directly on video, thus supervises every stage from initialization to distillation.

Teacher forcing suffices as the only source of supervision: starting from Wan2.1-T2V-1.3B, the AR teacher reaches 82.90 VBench Total with 50 denoising steps per block. The four-step OPF student reaches 84.08, above its teacher and above Causal Forcing and Self Forcing distilled with 1.3B BI score models under the same evaluation, and its three-seed mean stays above both baselines at every evaluated checkpoint. Changing only how the fake-score model conditions exposes the mismatch directly: from the same pretrained Wan weights, a causal fake-score model reaches 83.19 Total, whereas a bidirectional one collapses to 63.22. With BI score models on both sides, motion decays below its initialization during training, while OPF retains it (Figure 1b). In summary, our contributions are:

- We validate teacher-forcing AR training as a foundation for both many-step generation and few-step distillation.
- We introduce On-Policy Forcing, which distills the AR teacher on student rollouts with consistent causal conditioning and no additional BI teacher.
- We show that OPF retains quality and motion gains across evaluated checkpoints, while a conditioning mismatch in either score model can lead to severe degradation.

2 Related Work

Video diffusion models. Text-to-video diffusion models typically denoise all frames of a clip jointly with bidirectional attention (Team Wan et al., 2025; HaCohen et al., 2024). They produce high-quality short videos, but no frame can be shown before the whole clip is denoised, which rules out streaming and interaction. Autoregressive variants instead generate frames or chunks in sequence, conditioning each on its predecessors (Jin et al., 2025; Sun et al., 2025; Xie et al., 2025; Sand AI et al., 2025). OPF starts from a bidirectional backbone and keeps every subsequent training stage autoregressive.

Autoregressive video training. Teacher forcing learns next-block distributions from clean data histories, whereas diffusion forcing varies noise levels across video tokens, and progressive chunk-wise schedules let later chunks begin denoising before earlier ones finish (Chen et al., 2024; Sand AI et al., 2025). Self Forcing and Causal Forcing compare the two and find teacher forcing the stronger scheme for AR diffusion training (Huang et al., 2025; Zhu et al., 2026). Both clean and corrupted data histories differ from model-generated histories, so exposure bias remains (Huang et al., 2025). History-guided training, self-resampling, and context packing further explore how to construct histories and mitigate autoregressive drift (Song et al., 2025; Guo et al., 2025; Zhang et al., 2025), and rolling, hierarchical, or long-context designs extend generation to long videos (Liu et al., 2025; Yang et al., 2025; Cui et al., 2025; Zou et al., 2026; Chen et al., 2026). We train the teacher with teacher forcing, as in Causal Forcing (Zhu et al., 2026), and address generated histories in the on-policy distillation stage.

Few-step autoregressive video distillation. Few-step video generation reduces sampling cost through consistency, reward, adversarial, or distribution-matching objectives (Ding et al., 2024; Li et al., 2024; Zhang et al., 2024). Few-step AR video generators are mostly distilled from BI teachers. CausVid (Yin et al., 2025) supervises a causal student with a BI teacher through DMD (Yin et al., 2024b). Self Forcing (Huang et al., 2025) trains on the student's own rollouts to close the train–test gap. Both initialize the student from the BI teacher. CausVid also reports that distilling from a causal teacher underperforms distilling from a BI one. Its causal teacher, however, assigns independent noise levels to blocks as in diffusion forcing, which Causal Forcing shows mismatches the AR conditionals (Zhu et al., 2026), whereas our teacher is trained with teacher forcing. Causal Forcing and its successors (Zhu et al., 2026; Zhao et al., 2026; Zheng et al., 2026) instead train an AR teacher with teacher forcing to make initialization causal, but keep BI score models for the final stage. OPF keeps the rollouts and the causal initialization and makes the final supervision causal as well. Adversarial self-distillation and post-training offer alternative routes (Yang et al., 2026; Lin et al., 2025; Li et al., 2026), while OPSD-V and AnyFlow (Liu et al., 2026; Gu et al., 2026) extend on-policy training through privileged cache targets or flow-map distillation.

3 METHOD

3.1 PRELIMINARIES

Autoregressive video diffusion. Let a video be partitioned into N temporal blocks $\boldsymbol{x}_0^{1:N} = (\boldsymbol{x}_0^1, \ldots, \boldsymbol{x}_0^N)$, and let $\boldsymbol{c}$ denote the text condition. An autoregressive (AR) video model factorizes the data distribution as

$$p^{\mathrm{AR}}(\boldsymbol{x}_0^{1:N} \mid \boldsymbol{c}) = \prod_{i=1}^{N} p(\boldsymbol{x}_0^i \mid \boldsymbol{x}_0^{<i}, \boldsymbol{c}), \tag{1}$$

where $\boldsymbol{x}_0^i$ is the clean latent of block i and $\boldsymbol{x}_0^{<i}$ contains only its preceding blocks. An AR video diffusion model learns each conditional with flow matching: for a noise level $t \in (0, 1]$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, it forms $\boldsymbol{x}_t^i = (1-t)\boldsymbol{x}_0^i + t\boldsymbol{\epsilon}$ and predicts the velocity $\boldsymbol{\epsilon} - \boldsymbol{x}_0^i$ from $\boldsymbol{x}_t^i$, t, the preceding blocks, and $\boldsymbol{c}$. Teacher forcing conditions this prediction on clean preceding blocks from the training data, whereas diffusion forcing assigns independent noise levels to all blocks (Chen et al., 2024).

Distribution matching distillation. DMD (Yin et al., 2024b;a) turns a many-step diffusion model into a few-step generator G_θ by minimizing the reverse KL divergence between the generator's output distribution and the data distribution, averaged over noise levels. For a generated sample $\tilde{\boldsymbol{x}}_0$ and its noised version $\tilde{\boldsymbol{x}}_t$, the gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = -\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t \left(s_{\mathrm{real}}(\tilde{\boldsymbol{x}}_t, t) - s_{\mathrm{fake}}(\tilde{\boldsymbol{x}}_t, t) \right)^\top \frac{\partial \tilde{\boldsymbol{x}}_0}{\partial \theta} \right], \tag{2}$$

where w_t is a time-dependent weight, s_{real} is the score of a frozen diffusion model trained on real data, and s_{fake} is the score of an online diffusion model trained on the generator's samples. In prior AR video distillation, G_θ is causal while both score models are bidirectional video diffusion models (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026).

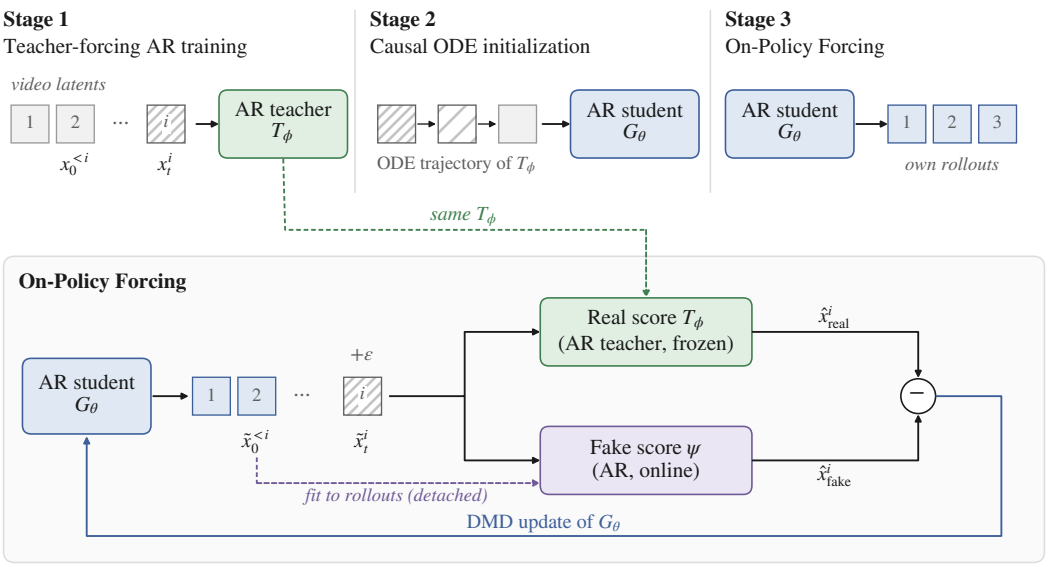

Figure 2: AR training and On-Policy Forcing. Top: Stage 1 trains a many-step AR teacher on video latents with teacher forcing. Stage 2 initializes a four-step AR student from the teacher's causal ODE trajectories. Stage 3 distills the student on its own rollouts. Stages 1 and 2 follow Causal Forcing. Bottom: in Stage 3, the frozen AR teacher and an online AR model serve as the real- and fake-score models and receive the same generated prefix and noisy block, so neither introduces a conditioning mismatch. Their score difference updates the student through DMD, and the fake-score model is trained on detached rollouts.

3.2 Overview

Figure 2 shows the three training stages, which learn the conditionals in equation (1) from video and keep both many-step and few-step generation autoregressive. Stage 1 (AR diffusion training) initializes an AR diffusion model T_ϕ from a pretrained video diffusion backbone and trains it with teacher forcing (Sec. 3.3). Stage 2 (causal ODE initialization) initializes a K-step AR generator G_θ from causal ODE trajectories of T_ϕ (Sec. 3.4). Stage 3 (On-Policy Forcing) distills G_θ on its own rollouts with T_ϕ as the frozen real-score model and an online AR fake-score model, both conditioned on the same generated prefix (Sec. 3.5), using standard DMD.

Stages 1 and 2 follow Causal Forcing (Zhu et al., 2026), whose Stage 3 is asymmetric DMD with bidirectional score models. OPF keeps both score models autoregressive, so all three stages share the factorization in equation (1). It changes which models supply the scores, not the matching objective.

3.3 Teacher-Forcing AR Diffusion Training

Following Causal Forcing, we train the teacher to predict each video block from its past, so that it can support both AR generation and subsequent distillation. We therefore use block-causal attention during training: each noisy block can attend to its clean preceding blocks and the text condition, but not to future video blocks. The starting weights come from a pretrained video diffusion backbone. This stage learns the AR conditionals in equation (1).

With the noised block $\boldsymbol{x}_t^i$ defined in Sec. 3.1, we train T_ϕ with causal attention while exposing the clean preceding blocks:

$$\mathcal{L}_{\mathrm{AR}}(\phi) = \mathbb{E}_{i,t,\boldsymbol{\epsilon}}\left[\lambda(t)\left\|v_\phi(\boldsymbol{x}_t^i, t \mid \boldsymbol{x}_0^{<i}, \boldsymbol{c}) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0^i)\right\|_2^2\right]. \tag{3}$$

Here, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and v_ϕ is the teacher's velocity prediction. Each block draws its own noise level: u is sampled uniformly from a 1,000-point grid on $(0, 1]$ and mapped to the shifted level $t = su/(1 + (s-1)u)$, with shift s given in Sec. 4. The fixed weight $\lambda(t)$ of the flow-matching schedule emphasizes intermediate noise levels. The teacher uses many-step denoising within each block. The later stages reduce the number of steps while keeping the same causal factorization.

Teacher forcing trains on clean data prefixes, whereas generation conditions on the model's own outputs. The final stage closes this gap by evaluating both score models on the student's generated prefixes.

3.4 Causal ODE Initialization

Before distribution matching, we initialize G_θ by regressing the endpoints of causal ODE trajectories from the AR teacher (Zhu et al., 2026). We sample a noisy block from a stored trajectory and predict its endpoint conditioned on the clean preceding blocks:

$$\mathcal{L}_{\text{init}}(\theta) = \mathbb{E}\left[\left\|G_\theta(\boldsymbol{x}_t^i, t \mid \boldsymbol{x}_0^{<i}, \boldsymbol{c}) - \hat{\boldsymbol{x}}_{0,\phi}^i\right\|_2^2\right]. \tag{4}$$

Here, G_θ predicts the clean block in a single network evaluation, and $\hat{\boldsymbol{x}}_{0,\phi}^i$ is the endpoint produced by the AR teacher. At inference, the student uses a K-step denoising schedule.

3.5 On-Policy Distillation

We use *on-policy distillation* to mean that the score models are evaluated on videos generated by the current student. The DMD gradient in equation (2) subtracts a real and a fake score estimated on the same perturbed sample. The AR student generates each block from its own prefix, whereas a BI score model conditions the block on the whole noisy video, including future blocks. When either score model conditions differently from the student, a *conditioning mismatch*, the score difference also reflects different conditional distributions. A controlled comparison isolates this effect (Sec. 4.3): with the AR real-score model fixed and the fake-score model initialized from the same pretrained Wan weights, a causal fake-score model trains normally and reaches 83.19 VBench Total, whereas a bidirectional one collapses to 63.22. OPF gives both score models the student's AR factorization and the same generated prefix, so the subtraction is taken between two estimates over the same conditional family. The AR teacher already estimates this family for real data, which is why it serves as the real-score model.

The student first produces a full autoregressive rollout $\tilde{\boldsymbol{x}}_0^{1:N} \sim p_\theta^{\text{AR}}$ with its inference-time cache. We draw one noise level t per rollout, shared by all blocks, by applying the same shift map to u uniform on a 1,000-point grid and clipping t to $[0.02, 0.98]$. At block i, we form $\tilde{\boldsymbol{x}}_t^i = (1-t)\tilde{\boldsymbol{x}}_0^i + t\boldsymbol{\epsilon}$ and evaluate

$$\hat{\boldsymbol{x}}_{\text{real}}^i = \hat{\boldsymbol{x}}_\phi(\tilde{\boldsymbol{x}}_t^i, t \mid \tilde{\boldsymbol{x}}_0^{<i}, \boldsymbol{c}), \tag{5}$$

$$\hat{\boldsymbol{x}}_{\text{fake}}^i = \hat{\boldsymbol{x}}_\psi(\tilde{\boldsymbol{x}}_t^i, t \mid \tilde{\boldsymbol{x}}_0^{<i}, \boldsymbol{c}). \tag{6}$$

Both score models use a flow-matching parameterization and return clean-sample predictions. The frozen teacher $\hat{\boldsymbol{x}}_\phi$ represents the real conditional distribution, and the online model $\hat{\boldsymbol{x}}_\psi$ represents the student's conditional distribution. Both are AR models: causal attention exposes the same generated past $\tilde{\boldsymbol{x}}_0^{<i}$ and hides future blocks. The real prediction applies CFG against the negative prompt, whereas the fake prediction uses $\boldsymbol{c}$ only. Under this parameterization, a clean-sample prediction $\hat{\boldsymbol{x}}_0$ corresponds to the score $\big((1-t)\hat{\boldsymbol{x}}_0 - \tilde{\boldsymbol{x}}_t^i\big)/t^2$, so the real–fake score difference is proportional to $\hat{\boldsymbol{x}}_{\text{real}}^i - \hat{\boldsymbol{x}}_{\text{fake}}^i$. The student receives the standard DMD gradient, up to a constant factor,

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\sum_{i=1}^{N} \frac{1}{Z}\left(\hat{\boldsymbol{x}}_{\text{real}}^i - \hat{\boldsymbol{x}}_{\text{fake}}^i\right)^\top \frac{\partial \tilde{\boldsymbol{x}}_0^i}{\partial \theta}\right], \quad Z = \text{mean}\big|\tilde{\boldsymbol{x}}_0^{1:N} - \hat{\boldsymbol{x}}_{\text{real}}^{1:N}\big|. \tag{7}$$

Following DMD, the time-dependent factor of the score difference is replaced by the per-sample normalizer Z, computed without gradient as the mean absolute difference between the rollout and the real prediction over all latent elements. The fake-score model is trained on detached samples from the same student rollout, with t drawn as above:

$$\mathcal{L}_{\text{fake}}(\psi) = \mathbb{E}_{i,t,\boldsymbol{\epsilon}}\left[\left\|v_\psi(\tilde{\boldsymbol{x}}_t^i, t \mid \tilde{\boldsymbol{x}}_0^{<i}, \boldsymbol{c}) - (\boldsymbol{\epsilon} - \tilde{\boldsymbol{x}}_0^i)\right\|_2^2\right]. \tag{8}$$

Here, v_ψ is the velocity prediction corresponding to the online AR fake-score model. Unlike equation (3), this loss uses no timestep weight, and its gradients do not pass through the student samples. The two updates alternate: the fake-score model is optimized r times per student update ($r = 5$ in our experiments), and no forward-KL anchor, real-data regression loss, adversarial loss, or temporal regularizer is added.

4 EXPERIMENTS

We evaluate the AR teacher as a many-step generator and as the teacher for OPF, then compare score configurations at an equal training budget.

4.1 EXPERIMENTAL SETUP

Models and training stages. Our teacher, student, and controlled score-configuration experiments use the Wan2.1-T2V-1.3B backbone, and every real-score model we train with is also 1.3B. The Self Forcing and Causal Forcing baselines we train are marked † in Table 1. Cited rows list the real-score model sizes reported by their sources. Starting from the pretrained 1.3B weights, we perform teacher-forcing AR diffusion training on real videos from OpenVid (Nan et al., 2025), using 64 GPUs, a per-GPU batch size of 4, and gradient accumulation 8. We use the checkpoint after 3,900 training steps as the AR teacher. The teacher uses timestep shift 8 and generates each temporal block with 50 denoising steps. Causal Forcing instead trains its AR diffusion model for 2K steps on 3K videos synthesized by the bidirectional base model (Zhu et al., 2026). In preliminary experiments, longer training and a moderately larger timestep shift improved the teacher, whereas adding noise to the clean context blocks degraded it (Appendix E). We then use this teacher for causal ODE initialization of a four-step AR student. The resulting checkpoint is shared by every controlled on-policy distillation run.

Our method uses the same AR teacher as the frozen real-score model and initializes the online AR fake-score model from the same checkpoint. We denote the score configurations by AR/AR, AR/BI, BI/AR, and BI/BI, with the real-score model listed first. AR/AR is On-Policy Forcing. The controlled runs use 16 GPUs and share the generator initialization, four-step schedule $[1000, 750, 500, 250]$, timestep shift 8, training prompts, classifier-free guidance (CFG), negative prompt, effective batch size 32, and a ratio of five fake-score updates per generator update. Each of the four configurations is trained once, and AR/AR and BI/BI are repeated with two additional training seeds. We measure training progress in generator updates and evaluate checkpoints after 120, 240, and 360 updates.

VBench evaluation. Every model we train is evaluated with the same VBench protocol with extended prompts (Huang et al., 2024). It contains 946 prompts and produces 6,230 videos per model: five samples for ordinary prompts and 25 samples for the Temporal Flickering prompts. Generation uses seed 0, 81 frames, 16 frames per second, 832×480 resolution, CFG 3, and the same negative prompt. We evaluate all 16 VBench dimensions using the official aggregation procedure. Scores are reported in percentage points. For direct paired comparisons, we report 95% confidence intervals from 10,000 paired bootstrap resamples over prompts (Table 6) and, for repeated runs, the mean and sample standard deviation over three training seeds.

4.2 MAIN RESULTS ON VBENCH

The AR teacher is already a capable many-step generator: with 50 denoising steps per block, it reaches 82.90 Total (Table 1). Teacher-forcing models in prior work mainly serve as initialization or as baselines that trail the distilled students (Huang et al., 2025; Zhu et al., 2026). We find that the quality of a teacher-forcing model depends on its training recipe, and Appendix E compares the choices of Section 4.1 on fixed prompts. Distilling it into four steps with OPF raises Total to 84.08 and Quality to 85.36, so the student surpasses its own teacher while generating more motion (70.83 vs. 55.56 Dynamic Degree; Appendix B). OPF also outperforms Self Forcing and Causal Forcing distilled from 1.3B bidirectional real-score models, by 0.86 and 0.79 Total points, and falls within the range of recent four-step methods that distill from 14B bidirectional real-score models.

Relative to the shared causal ODE initialization, OPF improves Total, Quality, and Semantic by 2.2–2.4 points each and Dynamic Degree by 10.3 points. Paired bootstrap intervals exclude zero for every gain over the initialization and the teacher, except for the small Semantic difference to the teacher (Table 6). Table 1 reports one training run at its highest-Total checkpoint, reached after 240 generator updates. Averaged over three training seeds, OPF stays between 83.69 and 83.78 Total at every evaluated checkpoint, above both baselines (Sec. 4.4).

Table 1: Comparison with existing methods on VBench. With a 1.3B AR real-score model, On-Policy Forcing exceeds Self Forcing and Causal Forcing distilled from 1.3B BI real-score models under the same evaluation, and falls within the range reported for four-step methods that distill from 14B BI real-score models. Cited rows are reported values. LTX-1.9B, Wan2.1-1.3B, CausVid, Self Forcing, and Causal Forcing are transcribed from Causal Forcing (Zhu et al., 2026), the others from the cited papers. Reported Dynamic values use rich-motion prompts. For the models we evaluate, Dynamic Degree on the official subset is reported in Appendix B. Bold marks the best Total, Quality, and Semantic among the models we evaluate. Dashes denote inapplicable or unreported entries.

Model	Generation unit	Denoising steps	Real-score model	Total	Quality	Semantic	Dynamic
Bidirectional Video Diffusion Models							
LTX-1.9B (HaCohen et al., 2024)	Full video	50	–	79.83	81.88	71.62	46
Wan2.1-1.3B (Team Wan et al., 2025)	Full video	50	–	83.37	84.30	79.65	61
Autoregressive Video Diffusion Models							
AR teacher†	Chunk	50	–	82.90	83.75	**79.52**	–
Distilled Autoregressive Video Models							
CausVid (Yin et al., 2025)	Chunk	4	BI, 14B	81.33	83.98	70.72	62
Self Forcing (Huang et al., 2025)	Chunk	4	BI, 14B	83.74	84.48	80.77	57
Causal Forcing (Zhu et al., 2026)	Chunk	4	BI, 14B	84.04	84.59	81.84	68
Adversarial Self-Distillation (Yang et al., 2026)	Chunk	4	BI, 14B	84.38	85.16	81.25	–
Causal Forcing++ (Zhao et al., 2026)	Frame	4	BI, 14B	84.10	84.94	80.75	71
Causal-rCM (Zheng et al., 2026)	Chunk	4	BI, 14B	84.37	85.02	81.73	–
Self Forcing†	Chunk	4	BI, 1.3B	83.22	84.19	79.32	–
Causal Forcing†	Chunk	4	BI, 1.3B	83.29	84.43	78.70	–
On-Policy Forcing **(Ours)**	Chunk	4	AR, 1.3B	**84.08**	**85.36**	78.95	–

† Trained and evaluated by us under the same protocol as On-Policy Forcing.

Table 2: Real- and fake-score configurations. A conditioning mismatch in either score model lowers Total, and in the fake-score model it collapses training. All runs share the student initialization, 360 generator updates, and training seed. All score models use the Wan2.1-T2V-1.3B backbone: AR score models start from the AR teacher and BI score models from pretrained Wan2.1 weights. The first row is On-Policy Forcing.

Real score	Fake score	Mismatch	Total	Quality	Semantic	Dynamic
AR	AR	None	**83.98**	**85.25**	78.91	71.67
AR	BI	Fake	63.22	67.55	45.91	0.00
BI	AR	Real	80.97	82.29	75.66	**89.17**
BI	BI	Both	82.78	83.70	**79.11**	43.33

4.3 Effect of the Real- and Fake-Score Configurations

Table 2 tests whether both score models must share the student's conditioning. Each asymmetric configuration places the conditioning mismatch differently: AR/BI in the fake-score model, BI/AR in the real-score model, and BI/BI in both. All runs start from the shared ODE initialization and stop after 360 generator updates with the same training seed, so this comparison fixes the training budget.

A conditioning mismatch lowers Total wherever it appears. Keeping the AR real-score model but making the fake-score model bidirectional collapses training: AR/BI falls to 63.22 Total and 0.00 Dynamic Degree, 20.8 Total points below AR/AR. With bidirectional score models on both sides, BI/BI does not collapse but trails AR/AR by 1.2 Total, 1.6 Quality, and 28.3 Dynamic Degree points, with a comparable Semantic score. BI/AR reaches the highest Dynamic Degree (89.17) at a lower Total (80.97). Table 6 gives paired bootstrap intervals for both comparisons with AR/AR.

As discussed in Sec. 3.5, the AR/BI collapse comes from the fake score's conditioning rather than its weights: a causal fake-score model initialized from the same pretrained Wan weights reaches 83.19 Total (Appendix B.3). Successful distillation therefore does not require the fake-score model to inherit AR-trained weights.

4.4 On-Policy Distillation Dynamics

We next examine how AR/AR and BI/BI evolve during training. Figure 3 reports Total and Quality as means and sample standard deviations over three training seeds from their shared initialization.

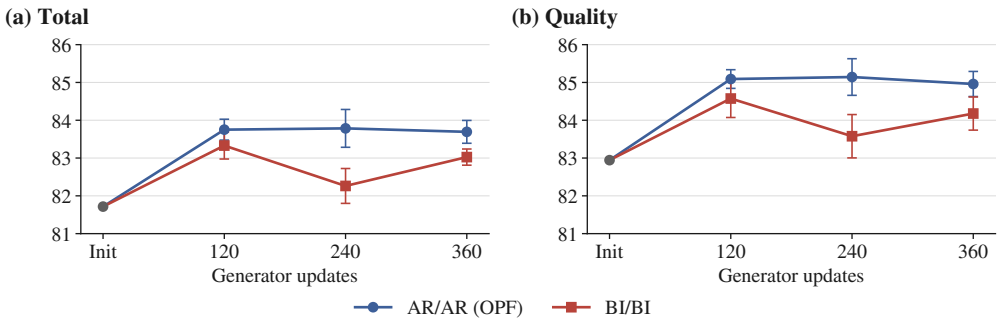

Figure 3: VBench Total and Quality during on-policy distillation. AR/AR (OPF) has a higher mean than BI/BI on both metrics at every checkpoint. Both configurations start from the same causal ODE initialization (Init), which is evaluated once. Points after Init show the mean ± one sample standard deviation over three training seeds.

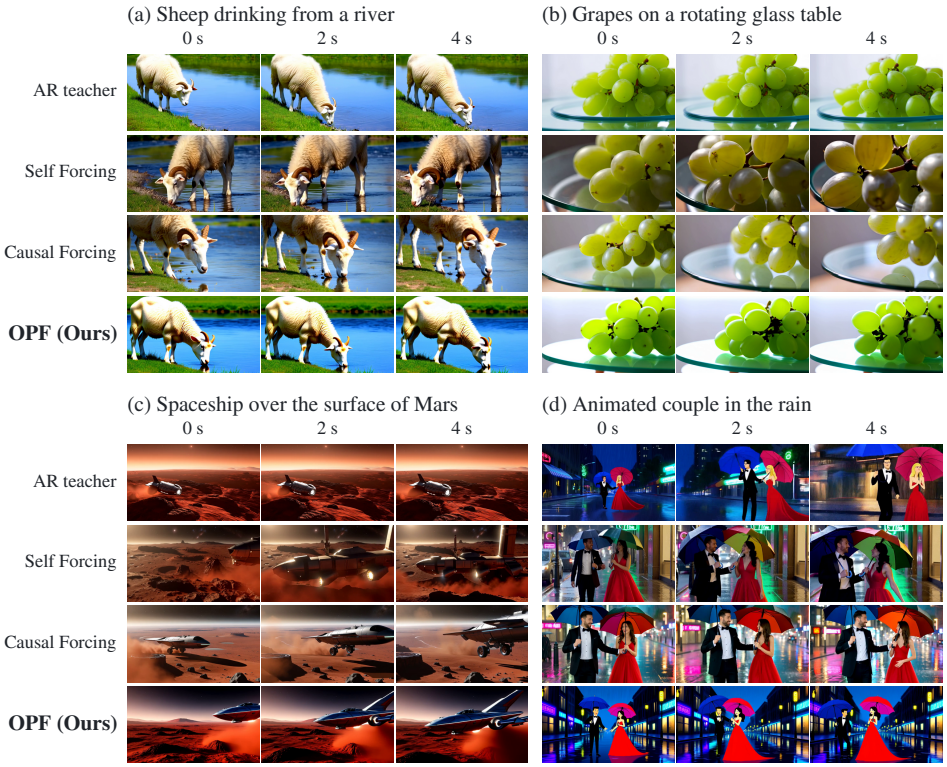

Figure 4: Qualitative comparison on VBench prompts. Frames at 0, 2, and 4 seconds from the AR teacher (50 denoising steps per block) and from Self Forcing (Huang et al., 2025), Causal Forcing (Zhu et al., 2026), and OPF (four steps), generated with the same prompt and sample index. Self Forcing and Causal Forcing are distilled from 1.3B BI real-score models. Under both baselines, object attributes change within the clip, such as the umbrella colors in (d). OPF keeps them stable.

Across the three training seeds, AR/AR keeps both Total and Dynamic Degree above the shared initialization at every checkpoint and has the higher mean Total throughout. BI/BI initially has higher Dynamic Degree, but its mean falls below the initialization after 360 updates (Table 7). BI/BI degrades further with longer training: in independent three-seed runs of both configurations trained for 480 generator updates, its mean Total falls to 81.97, close to the initialization at 81.72, while AR/AR reaches 83.84 (Table 9).

Trained once each, AR/BI and BI/AR follow different trajectories (Table 8): AR/BI collapses from the first checkpoint, with Dynamic Degree near zero, whereas BI/AR recovers gradually but stays below both the shared initialization and AR/AR in Total. Each placement of the conditioning mismatch impairs training, but along a different trajectory.

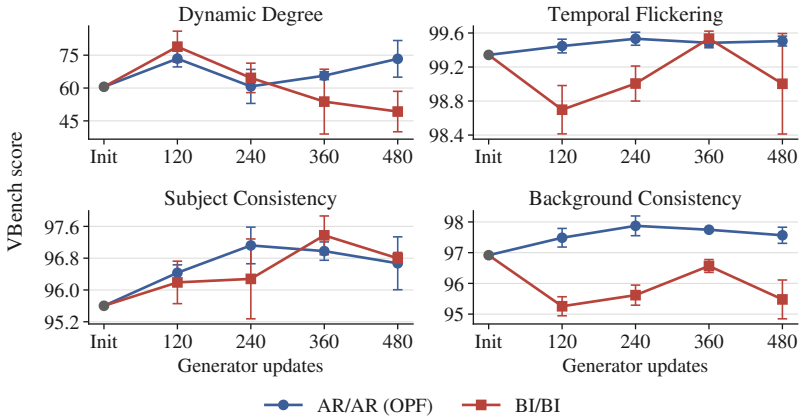

Figure 5: Motion-related VBench dimensions during on-policy distillation. AR/AR (OPF) keeps its Dynamic Degree above Init and stays above BI/BI on Background Consistency at every checkpoint. BI/BI starts with a higher Dynamic Degree but lower Temporal Flickering and Background Consistency, and its Dynamic Degree later falls below Init. Points after Init show the mean ± one sample standard deviation over three training seeds on the official VBench subset. Init is evaluated once. Each panel uses its own y-axis range.

4.5 Motion Metrics

Dynamic Degree alone cannot distinguish intended motion from flicker or structural artifacts. Figure 5 reports it alongside Temporal Flickering, Subject Consistency, and Background Consistency for the longer runs of Section 4.4 on the official VBench subset.

Early in training, BI/BI has the higher Dynamic Degree but lower Temporal Flickering and Background Consistency than both AR/AR and Init, so part of its early motion comes with flicker and background changes. Its Dynamic Degree then falls below Init. AR/AR keeps its Dynamic Degree above Init and stays above BI/BI on Background Consistency at every checkpoint and on Temporal Flickering at three of the four checkpoints. On the official VBench subset, OPF therefore gains motion without giving up temporal consistency (all dimensions in Appendix B).

On 100 rich-motion prompts, the same runs show a clearer gap. AR/AR stays between 72.7 and 79.7 Dynamic Degree, above BI/BI at every checkpoint, whereas BI/BI declines from 75.0 to 50.3, below the initialization at 57.0 (Figure 1b; Table 10).

4.6 Qualitative Comparisons

Figure 4 compares the OPF student from Table 1 with the Self Forcing and Causal Forcing models using 1.3B BI real-score models, with the AR teacher as a reference. Under both baselines, object attributes drift within the clip: an umbrella in (d) changes color between frames, and under Self Forcing the grapes in (b) darken and the spaceship in (c) merges with a tower-like structure. OPF keeps the umbrella colors and the grapes stable across the clip, in line with its higher Subject and Background Consistency (Table 4). OPF also follows the prompt as the AR teacher does: in (d), both render the requested animated style, whereas both baselines produce photorealistic people. In (c), the OPF spaceship moves with visible exhaust, whereas the AR teacher's stays nearly in place (sample details in Appendix C).

5 Conclusion

AR training supports both many-step generation and few-step distillation without an additional bidirectional teacher. Removing the conditioning mismatch from the final stage lets on-policy distillation improve the student over its AR teacher and over Self Forcing and Causal Forcing distilled from 1.3B bidirectional score models, while keeping the teacher's causal factorization. A conditioning mismatch in either score model degrades distillation. In the fake-score model it causes collapse even from the same initialization.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, including revising text for clarity and consistency and refining the presentation of tables and figures. Reported measurements come from evaluation outputs or the cited papers, not AI-generated estimates. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, results, and artifacts.

ETHICS STATEMENT

This work trains and distills an autoregressive text-to-video generator from an existing pretrained backbone using publicly available video and prompt datasets. It does not recruit human participants or collect new video data. Public availability of the source videos does not rule out privacy or consent concerns. Faster autoregressive video generation lowers the cost of producing synthetic video, which carries the misuse risks common to generative video models, including misleading or non-consensual synthetic media. Use of the pretrained backbone and datasets remains subject to their respective licenses and usage restrictions.

REPRODUCIBILITY STATEMENT

Section 4 describes the training stages, checkpoint selection, and evaluation protocol. Appendix A gives the model initializations, hyperparameters, hardware settings, and bootstrap procedure. Appendix B reports all VBench dimensions and evaluated training checkpoints. Appendix C gives the sample identities and generation settings for the qualitative comparisons. Results transcribed from prior papers are identified separately in Table 1.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In *Proceedings of NeurIPS*, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Boyuan Chen, Diego Marti Monso, 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, 2024.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026. URL https://arxiv.org/abs/2602.06028.

Justin 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. *arXiv preprint arXiv:2510.02283*, 2025. URL https://arxiv.org/abs/2510.02283.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. DOLLAR: Few-step video generation via distillation and latent reward optimization. *arXiv preprint arXiv:2412.15689*, 2024. URL https://arxiv.org/abs/2412.15689.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. AnyFlow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026. URL https://arxiv.org/abs/2605.13724.

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. *arXiv preprint arXiv:2512.15702*, 2025. URL https://arxiv.org/abs/2512.15702.

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.

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*, 2025.

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*, 2024.

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.

Haobo Li, Yanhong Zeng, Yunhong Lu, Jiapeng Zhu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yujun Shen, and Zhipeng Zhang. AAD-1: Asymmetric adversarial distillation for one-step autoregressive video generation. In *International Conference on Machine Learning*, 2026.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. OPSD-V: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. URL https://arxiv.org/abs/2509.25161.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical report*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Sand AI, Hansi Teng, Hongyu Jia, Lei Sun, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. URL https://arxiv.org/abs/2505.13211.

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, 2025.

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.

Team Wan, Ang Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6376–6386, 2025.

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. *arXiv preprint arXiv:2509.22622*, 2025. URL https://arxiv.org/abs/2509.22622.

Yongqi Yang, Huayang Huang, Xu Peng, Xiaobin Hu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yu Wu. Towards one-step causal video generation via adversarial self-distillation. In *International Conference on Learning Representations*, 2026.

Tianwei Yin, Michael Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Fredo 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, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo 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*, 2025.

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

Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. SF-V: Single forward video generation model. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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.

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.

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.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. HiAR: Efficient autoregressive long video generation via hierarchical denoising. *arXiv preprint arXiv:2603.08703*, 2026. URL https://arxiv.org/abs/2603.08703.

A Additional Implementation Details

Model initialization. Wan2.1-T2V-1.3B weights initialize the teacher-forcing AR diffusion run on OpenVid, and the checkpoint after 3,900 training steps is frozen as the AR teacher. Its causal ODE trajectories train the four-step causal ODE initialization for 2,500 training steps. The same AR teacher checkpoint is then used as the frozen real-score model and as the initialization of the online AR fake-score model. All OPF checkpoints start from this causal ODE initialization.

Table 3: Score-model initialization. Every configuration uses the same causal ODE initialization and Stage 3 training settings, and the real-score model is frozen. AR/AR is On-Policy Forcing. Wan2.1 denotes the pretrained bidirectional Wan2.1-T2V-1.3B weights.

Configuration	Real-score model	Fake-score initialization
AR/AR	AR teacher	AR teacher
AR/BI	AR teacher	Wan2.1
BI/AR	Wan2.1	AR teacher
BI/BI	Wan2.1	Wan2.1

Teacher training. The AR teacher is trained with causal attention and clean preceding blocks using 64 GPUs, a per-GPU batch size of 4, and gradient accumulation 8. We use timestep shift 8, three latent frames per temporal block, flow-matching loss, and AdamW with learning rate 1×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.95)$, and weight decay 0.1. The teacher checkpoint is evaluated without EMA and uses 50 denoising steps per block.

On-policy distillation. All score-configuration runs use 16 GPUs with a per-GPU batch size of 1 and gradient accumulation 2, giving effective batch size 32. They use 21 latent frames, seven blocks of three frames, the four-step schedule $[1000, 750, 500, 250]$, timestep shift 8, CFG 3, and the same negative prompt. Each of the four score configurations is trained once with the same training seed, and AR/AR and BI/BI are repeated with two additional training seeds. AR/AR and BI/BI are also trained for 480 generator updates in an independent set of three-seed runs with the same configuration (Appendix B.5). The student and fake-score learning rates are 2×10^{-6} and 4×10^{-7}, respectively, both with $(\beta_1, \beta_2) = (0, 0.999)$. The fake-score model is updated five times per generator update. EMA is disabled, so all evaluated student checkpoints use the raw generator weights. Stage 3 uses text prompts but no real-video latents. Both updates use samples from the current student: the student update backpropagates through them, and the fake-score update uses them detached.

The four score-configuration runs use the same causal ODE initialization, prompt file, training seed, schedule, shift, CFG, negative prompt, effective batch size, and update budget. Table 3 lists the initialization checkpoints used for each score model.

Evaluation and statistics. Each full VBench evaluation uses 946 prompts and 6,230 videos at 81 decoded frames, 16 FPS, and 832×480 resolution. The evaluator is run on all 16 VBench dimensions. Paired intervals use 10,000 prompt-level bootstrap resamples with seed 0. Each resample retains all videos of a selected prompt and uses the same prompt draws for both models. The intervals capture variation over evaluation prompts, and variation over training seeds is reported separately. For repeated AR/AR and BI/BI runs, we report the mean and sample standard deviation ($n = 3$, denominator $n - 1$) at each fixed checkpoint. All evaluation generation uses seed 0. The causal ODE initialization is a single checkpoint and is therefore reported without a standard deviation.

Training profile. We profile one generator update of the 16-GPU configuration on two nodes with eight H200 GPUs per node. With effective batch size 32, this covers one student update and five fake-score updates, including forward, backward, and optimizer steps. The peak allocated and reserved memory on the most loaded GPU are 56.22 and 101.09 GiB. One generator update takes 208.88 seconds: 97.34 seconds for generator forward/backward, 105.94 seconds for fake-score forward/backward, and 4.49 seconds for text encoding.

B Additional Quantitative Results

B.1 VBench Scores Across All Dimensions

Tables 4 and 5 report every VBench dimension for the models in Table 1, with On-Policy Forcing after 240 generator updates, and for the causal ODE initialization and the four score configurations of Table 2 after 360 generator updates. Values are percentage points. Dynamic Degree uses 360 videos and Temporal Flickering uses 1,875 videos.

Table 4: Quality-related VBench dimensions. Among the models in Table 1, On-Policy Forcing is highest on five of the seven dimensions, all except Motion Smoothness and Imaging Quality. †Trained and evaluated by us, as in Table 1. Column abbreviations: Subj. (Subject Consistency), Backg. (Background Consistency), Flicker (Temporal Flickering), Smooth (Motion Smoothness), Dynamic (Dynamic Degree), Aesth. (Aesthetic Quality), and Imaging (Imaging Quality).

Model	Subj.	Backg.	Flicker	Smooth	Dynamic	Aesth.	Imaging
Models in Table 1							
AR teacher†	96.64	97.31	99.43	98.43	55.56	64.70	65.57
Self Forcing†	94.48	95.57	99.03	98.49	69.72	64.54	67.24
Causal Forcing†	96.58	95.21	99.05	98.39	64.72	65.24	68.91
On-Policy Forcing **(Ours)**	97.01	97.86	99.55	98.34	70.83	65.58	66.29
Causal ODE initialization and score configurations after 360 generator updates							
Causal ODE initialization	95.60	96.92	99.34	98.59	60.56	61.81	62.12
AR/AR	96.95	97.71	99.42	98.06	71.67	65.58	66.74
AR/BI	77.36	91.64	98.03	97.85	0.00	42.35	46.39
BI/AR	93.11	93.23	98.98	97.93	89.17	59.00	57.53
BI/BI	97.52	96.50	99.54	98.88	43.33	64.63	69.65

Table 5: Semantic-related VBench dimensions. The models in Table 1 differ by less than 4.5 points on every dimension, whereas AR/BI falls below the other configurations on every dimension except Appearance Style. Groups and † as in Table 4. Column abbreviations: Object (Object Class), Multi (Multiple Objects), Action (Human Action), Spatial (Spatial Relationship), T. Style (Temporal Style), A. Style (Appearance Style), and Overall (Overall Consistency).

Model	Object	Multi	Action	Color	Spatial	Scene	T. Style	A. Style	Overall
Models in Table 1									
AR teacher†	94.18	83.20	96.40	81.41	74.57	59.01	24.33	21.19	26.63
Self Forcing†	94.64	82.97	97.00	85.23	74.47	56.15	24.38	20.43	26.51
Causal Forcing†	94.26	85.63	95.20	83.74	73.82	54.56	24.09	20.34	26.19
On-Policy Forcing **(Ours)**	93.31	85.18	95.80	83.48	73.34	56.51	23.66	21.03	26.24
Causal ODE initialization and score configurations after 360 generator updates									
Causal ODE initialization	93.54	73.72	96.20	83.09	69.11	52.69	23.92	21.12	26.15
AR/AR	93.37	85.85	96.20	83.80	73.35	55.04	23.67	21.08	26.16
AR/BI	48.61	17.64	53.80	67.00	26.79	13.78	18.72	22.35	19.27
BI/AR	92.44	69.13	95.40	89.26	69.36	47.69	24.06	19.66	26.37
BI/BI	95.24	86.34	93.80	83.69	73.06	56.02	24.15	20.66	26.59

B.2 Paired Bootstrap Intervals

Table 6 compares the On-Policy Forcing run of Table 1 with the AR teacher and the causal ODE initialization after 240 generator updates, and with AR/BI and BI/BI after 360 generator updates, using the paired bootstrap of Appendix A.

B.3 AR Fake-Score Model Initialized from Wan2.1

We initialize an AR fake-score model from pretrained Wan2.1 weights while retaining the AR real-score model and the causal ODE initialization. After 360 generator updates this configuration reaches 83.19 Total, 84.21 Quality, 79.12 Semantic, and 65.56 Dynamic Degree. In this setting, successful distillation does not require initializing the fake-score model from the AR teacher. Compared with AR/BI, both fake-score models start from the same pretrained Wan2.1 weights, and the

Table 6: Paired bootstrap comparisons. Against every compared model, the Total, Quality, and Dynamic Degree intervals of On-Policy Forcing lie above zero, while the Semantic intervals against the AR teacher and BI/BI include zero. Entries are On-Policy Forcing minus the compared model, in percentage points, with 95% confidence intervals from 10,000 prompt-level paired bootstrap resamples.

Compared model	Total	Quality	Semantic	Dynamic
After 240 generator updates (Table 1)				
AR teacher	$+1.17$ [0.68, 1.68]	$+1.61$ [1.06, 2.18]	-0.57 [-1.64, 0.48]	$+15.28$ [8.89, 22.22]
Causal ODE initialization	$+2.36$ [1.90, 2.86]	$+2.41$ [1.90, 2.96]	$+2.15$ [1.04, 3.28]	$+10.28$ [4.72, 16.39]
After 360 generator updates (Table 2)				
AR/BI	$+20.76$ [19.99, 21.52]	$+17.70$ [16.87, 18.48]	$+33.00$ [30.91, 35.12]	$+71.67$ [62.78, 80.00]
BI/BI	$+1.20$ [0.54, 1.86]	$+1.55$ [0.77, 2.33]	-0.20 [-1.35, 0.96]	$+28.33$ [19.17, 37.78]

two runs differ only in whether the fake-score model conditions each block causally, so this pair isolates the conditioning mismatch of the fake-score model.

B.4 Training Trajectories

Table 7 reports the three-seed means underlying Figure 3. Table 8 reports all four score configurations, each trained once with the same training seed.

Table 7: Three-seed VBench scores of AR/AR and BI/BI. AR/AR (On-Policy Forcing) has the higher mean Total at every checkpoint, while BI/BI loses Dynamic Degree after 120 updates. Entries are mean $\pm$ sample standard deviation over three training seeds, in percentage points. Dynamic uses the official VBench subset.

Configuration	Generator updates	Total	Quality	Semantic	Dynamic
AR/AR	120	83.75 ± 0.28	85.09 ± 0.25	78.40 ± 0.42	68.24 ± 6.92
AR/AR	240	83.78 ± 0.50	85.14 ± 0.48	78.35 ± 0.64	68.33 ± 9.55
AR/AR	360	83.69 ± 0.30	84.96 ± 0.33	78.63 ± 0.24	69.63 ± 9.06
BI/BI	120	83.33 ± 0.36	84.57 ± 0.50	78.35 ± 0.49	90.19 ± 2.80
BI/BI	240	82.26 ± 0.46	83.58 ± 0.57	77.00 ± 0.35	63.15 ± 16.04
BI/BI	360	83.03 ± 0.21	84.18 ± 0.44	78.42 ± 0.72	56.48 ± 11.79

Table 8: Training trajectories of the four score configurations. AR/BI collapses from the first checkpoint, and AR/AR (On-Policy Forcing) has the highest Total at every checkpoint. Each configuration is trained once. Generator updates count student optimizer updates, each paired with five fake-score updates. 120, 240, and 360 updates correspond to 125.8, 251.6, and 377.4 million transformer tokens.

Configuration	Generator updates	Total	Quality	Semantic	Dynamic
Causal ODE initialization	0	81.72	82.94	76.80	60.56
AR/AR	120	83.81	85.10	78.61	71.67
AR/AR	240	84.08	85.36	78.95	70.83
AR/AR	360	83.98	85.25	78.91	71.67
AR/BI	120	64.34	68.60	47.32	2.50
AR/BI	240	60.18	66.60	34.54	0.00
AR/BI	360	63.22	67.55	45.91	0.00
BI/AR	120	77.93	79.56	71.40	100.00
BI/AR	240	80.14	81.69	73.93	92.50
BI/AR	360	80.97	82.29	75.66	89.17
BI/BI	120	83.69	85.01	78.43	87.22
BI/BI	240	82.55	83.85	77.36	76.94
BI/BI	360	82.78	83.70	79.11	43.33

B.5 Longer Training

We train AR/AR and BI/BI for 480 generator updates in an independent set of runs with three training seeds, starting from the causal ODE initialization with the settings of the runs in Figure 3, and evaluate every 120 updates on the official VBench subset (Table 9). Table 10 reports the same runs on rich-motion prompts.

Table 9: VBench scores of the longer runs. After 480 generator updates, BI/BI falls to 81.97 Total, close to the initialization, while AR/AR reaches 83.84. Entries are mean ± sample standard deviation over three training seeds, in percentage points. The initialization is evaluated once. Dynamic uses the official VBench subset.

Configuration	Generator updates	Total	Quality	Semantic	Dynamic
Causal ODE initialization	0	81.72	82.94	76.80	60.56
AR/AR	120	83.72 ± 0.19	85.08 ± 0.22	78.28 ± 0.05	73.43 ± 3.78
AR/AR	240	83.46 ± 0.10	84.72 ± 0.13	78.45 ± 0.17	60.74 ± 7.77
AR/AR	360	83.65 ± 0.17	84.91 ± 0.16	78.63 ± 0.25	65.65 ± 1.79
AR/AR	480	83.84 ± 0.18	85.07 ± 0.27	78.94 ± 0.24	73.33 ± 8.37
BI/BI	120	83.79 ± 0.21	85.07 ± 0.28	78.65 ± 0.22	78.89 ± 7.09
BI/BI	240	82.58 ± 0.50	84.07 ± 0.24	76.63 ± 1.66	64.63 ± 6.73
BI/BI	360	83.23 ± 0.85	84.33 ± 0.98	78.84 ± 0.71	53.80 ± 14.79
BI/BI	480	81.97 ± 0.87	83.00 ± 0.96	77.83 ± 0.94	49.26 ± 9.23

B.6 Dynamic Degree on Rich-Motion Prompts

We evaluate the causal ODE initialization and all three training seeds of AR/AR and BI/BI on the same 100-prompt rich-motion list: the longer runs after 120, 240, 360, and 480 generator updates, and the runs of Figure 3 after 240 and 360 updates. Each checkpoint generates one video per prompt with generation seed 0, 81 frames, 16 FPS, and 832×480 resolution. We use the rich-motion prompt list from Causal Forcing (Zhu et al., 2026) and run VBench's Dynamic Degree evaluator in custom-input mode on 100 videos per checkpoint.

Table 10: Dynamic Degree on 100 rich-motion prompts. AR/AR stays above BI/BI at every checkpoint, and BI/BI falls below the initialization by 360 generator updates. Init is the causal ODE initialization, evaluated once. Entries are mean ± sample standard deviation over three training seeds, in percentage points, with one video per prompt. This prompt set differs from the official VBench Dynamic Degree subset.

Configuration	Init	120 updates	240 updates	360 updates	480 updates
Longer runs (Figure 1b)					
AR/AR	57.00	79.67 ± 8.39	72.67 ± 5.51	77.33 ± 2.89	79.00 ± 5.57
BI/BI	57.00	75.00 ± 11.27	59.67 ± 16.92	53.00 ± 5.57	50.33 ± 3.21
Runs of Figure 3					
AR/AR	57.00	–	68.67 ± 5.69	76.33 ± 12.10	–
BI/BI	57.00	–	66.67 ± 13.65	48.67 ± 19.66	–

C Qualitative Comparison Details

Panels (a)–(d) of Figure 4 use VBench prompts 310, 683, 772, and 562 with sample indices 2, 3, 2, and 2, respectively. For each panel, the AR teacher, Self Forcing, Causal Forcing, and On-Policy Forcing videos use the same prompt and sample index. Three frames are shown at 0, 2, and 4 seconds. Generation follows the VBench evaluation settings: seed 0, 81 frames, 16 FPS, 832×480 resolution, CFG 3, and the same negative prompt. Both few-step baselines use 1.3B BI real-score models. The AR teacher uses 50 denoising steps per block and is shown as a reference.

D Limitations

While On-Policy Forcing removes the bidirectional teacher from few-step distillation, several limitations remain. First, all experiments build on a single 1.3B backbone and a single training set, so

how the AR teacher and on-policy distillation scale with model size and data remains to be examined. Second, the asymmetric configurations AR/BI and BI/AR are each trained with a single seed, and their trajectories may vary across seeds. Third, in the real-score comparison of Section 4.3, the bidirectional and AR real-score models differ in both factorization and weights, so this comparison does not isolate factorization as the fake-score comparison in Section 3.5 does. Finally, all generated videos contain 81 frames, and extending OPF to longer videos with streaming generation is a natural direction for future work.

E Teacher-Forcing Training Choices

Section 4.1 names three choices that shaped the AR teacher, namely clean context blocks, timestep shift 8, and longer training. Figures 6, 7, and 8 compare them on ten fixed MovieGenBench prompts. Every variant generates one video per prompt with seed 0 and the same initial noise, using 50 denoising steps per block and its own training timestep shift. Each panel shows the frame at 4 seconds. The context-noise and shift variants come from shorter exploratory runs with a smaller batch and are compared after 1,200 training steps. The training-length variants are checkpoints of the AR teacher run.

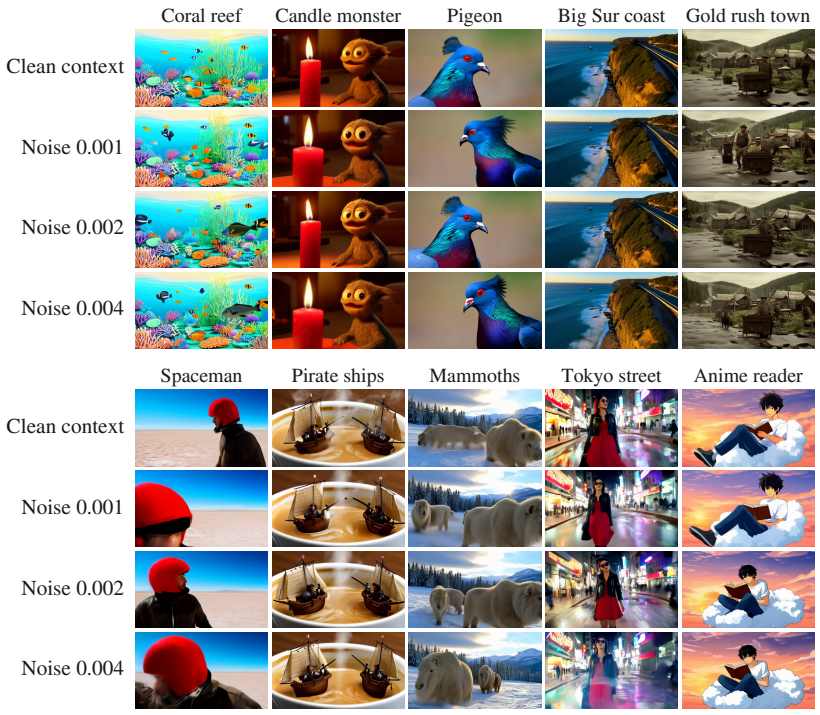

Figure 6: Noise on the context blocks. With context noise ratios of 0.001, 0.002, and 0.004, the Spaceman scene drifts into a close-up by 4 seconds, and the pigeon and mammoth scenes drift at some ratios. Clean context keeps the composition. All variants use timestep shift 8 and 1,200 training steps.



Figure 7: Timestep shift. Shifts 12 and 20 turn the Spaceman scene into a close-up, and shift 12 blurs the pirate ships. Shifts 5 and 8 keep both compositions. All variants use clean context and 1,200 training steps.



Figure 8: Training length. Checkpoints of the AR teacher run after 1.2K, 2.4K, and 3.9K steps. With longer training, the mammoth scene changes from dark bears to woolly mammoths and the Spaceman pose becomes coherent. The 3.9K checkpoint is the AR teacher used throughout the paper.