ONE-FORCING: TOWARDS STABLE ONE-STEP AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in autoregressive diffusion-based video generation have substantially improved the quality of real-time video synthesis. However, most existing methods still require multiple denoising steps, while reducing sampling to a single step often leads to severe quality degradation: trajectory-based consistency distillation methods often produce videos with weak dynamics, whereas DMD-based methods, such as Self-Forcing, tend to generate blurry frames. We attribute this limitation to the teacher trajectories exhibiting highly concentrated curvature near the high-noise endpoint, which poses a fundamental geometric challenge for one-step distillation under the consistency distillation framework. To address these limitations, we propose One-Forcing, a simple yet effective approach that augments the DMD objective with an auxiliary GAN loss for high-quality and efficient one-step video generation. We further find that framewise autoregression stabilizes adversarial training, enabling higher-quality generation with substantially fewer training iterations than chunkwise autoregression. Experiments on VBench show that One-Forcing achieves a total score of 83.76, establishing state-of-the-art performance among one-step causal video generation methods while remaining competitive with strong multi-step approaches. Project page is available here.

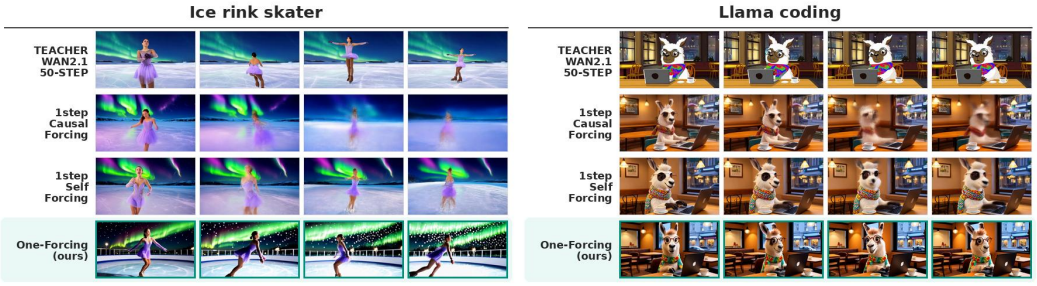

Figure 1: Visualization results on two prompts: Wan2.1 teacher uses 50 denoising steps, Causal Forcing, Self Forcing, and One-Forcing use one-step autoregressive sampling. Our method exhibits excellent dynamics and visual quality.

1 INTRODUCTION

Diffusion-based video generation has achieved striking visual fidelity and complex spatiotemporal dynamics, as demonstrated by bidirectional models such as Sora (Brooks et al., 2024), Veo (Google DeepMind, 2025), Wan (Team Wan et al., 2025), HunyuanVideo (Kong et al., 2024), and Seedance (Team Seedance et al., 2026). However, jointly denoising entire sequences incurs substantial computational costs and limits real-time interaction. Autoregressive video generators address this limitation by producing frames or short temporal blocks sequentially (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026a), enabling streaming applications such as world simulation (Ha & Schmidhuber, 2018; Hafner et al., 2025; Parker-Holder et al., 2024; Zhu et al., 2026b) and interactive game engines (Bruce et al., 2024; Valevski et al., 2025). Yet most causal video systems still require multiple denoising

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.

steps per block, leaving the sampling budget as a key bottleneck for end-to-end latency. The central question of this paper is therefore whether a causal video generator can preserve strong visual quality and motion dynamics in the one-step regime.

Existing fast distillation objectives leave a gap in this regime. Consistency-style methods can produce high-quality videos with a few denoising steps, but often lose motion and structure when reduced to one step. As shown in Figure 3, Wan (Team Wan et al., 2025) video trajectories exhibit sharply concentrated curvature near the high-noise endpoint: their strongest departure from a straight noise-to-data path occurs near the starting point of generation. This suggests a geometric challenge for predicting the final video directly from the highest-noise input, whereas multiple sampling steps allow subsequent predictions to refine the initial estimate. Distribution Matching Distillation (DMD) offers an alternative by matching output distributions without following a prescribed teacher trajectory (Yin et al., 2024b). However, in our one-step evaluations, Self Forcing still produces blurry outputs with missing fine detail. Recent analysis shows that low-frequency components can dominate DMD updates and delay the recovery of sharp boundaries and textures, offering a possible explanation for this behavior (Zhou et al., 2026). Adversarial supervision can provide complementary feedback on fine details, as demonstrated at low noise levels in SDXL-Lightning (Lin et al., 2024). These observations motivate augmenting DMD with a discriminator trained against real videos to improve visual realism. In autoregressive generation, this supervision must remain effective as the model's own predictions become the context for subsequent frames.

To address these challenges, we propose One-Forcing, a framework for high-quality one-step autoregressive video generation. Building on the adversarial distribution matching formulation of DMD2 (Yin et al., 2024a), we complement teacher guidance with direct supervision from real videos. Making this supervision effective under self-generated context also requires stable adversarial learning. Here, we find that framewise autoregression plays a key role, enabling stable and fast adversarial learning while improving generation quality over its chunkwise counterpart. We hypothesize that finer temporal credit assignment helps adversarial feedback guide learning as the generated context evolves. This finding highlights the importance of considering adversarial supervision and autoregressive granularity together when distilling video generators to a single step. On VBench, One-Forcing achieves a total score of 83.76, outperforming the evaluated one-step causal video baselines and remaining competitive with strong many-step approaches. In summary, our contributions are:

- We identify a geometric obstacle to one-step video distillation: video teacher trajectories exhibit sharply concentrated curvature near the high-noise endpoint, unlike image teachers commonly used in consistency distillation. This provides an explanation for why trajectory-based objectives degrade sharply when compressed to a single video generation step.
- We propose One-Forcing, which combines DMD and adversarial learning for one-step autoregressive video generation. By reusing the fake-score network as a shared critic, our method introduces supervision from real videos without requiring a separate discriminator backbone.
- We demonstrate that framewise autoregression supports stable and fast adversarial learning and improves generation quality over its chunkwise counterpart, making it an important component of effective one-step video distillation.
- On VBench, One-Forcing achieves *state-of-the-art* one-step causal video generation (83.76 total) and remains competitive with strong many-step approaches.

2 Related Works

2.1 Bidirectional and Autoregressive Video Generation

Current video diffusion models fall into two paradigms. *Bidirectional* models denoise an entire clip with full spatiotemporal attention, achieving strong coherence at quadratic attention cost (Ho et al., 2022b;a; Yang et al., 2025; Kong et al., 2024; Team Wan et al., 2025). Full-clip processing limits their suitability for streaming.

Autoregressive (causal) video generators condition each new block on previously generated outputs through a KV cache, enabling real-time streaming. Self Forcing (Huang et al., 2025) first demonstrated

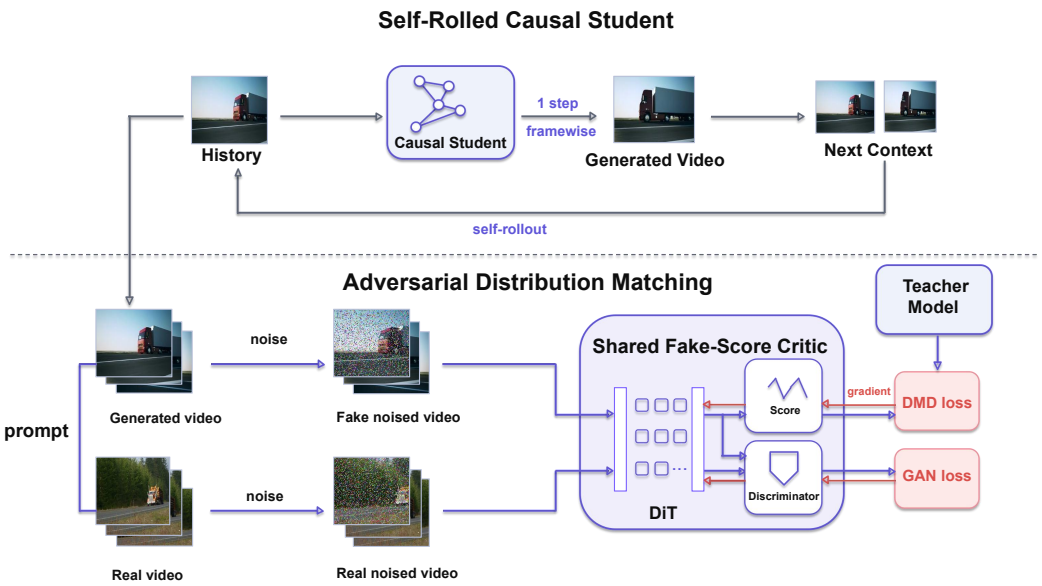

Figure 2: **One-Forcing training framework.** One-Forcing trains on one-step framewise autoregressive rollouts, combining teacher guidance through DMD and adversarial supervision from real videos with a shared fake-score backbone.

that training on self-generated context with a holistic video-level loss can close the train–test gap in causal video diffusion. Causal Forcing (Zhu et al., 2026a) uses an autoregressive teacher for ODE initialization to bridge the architectural gap between bidirectional and causal models, yielding improvements in dynamics and instruction following. Causal Forcing++ (Zhao et al., 2026) replaces causal ODE initialization with causal consistency distillation, reducing preparation cost and enabling framewise two-step generation. These initialization strategies complement One-Forcing's focus on one-step distribution matching. Other notable systems include CausVid (Yin et al., 2025), MAGI-1 (Sand.ai et al., 2025), LongLive (Yang et al., 2026a), Rolling Forcing (Liu et al., 2026b), Infinity-RoPE (Yesiltepe et al., 2025), and Self-Forcing++ (Cui et al., 2026). Despite these advances, reducing sampling to one step still causes pronounced quality degradation.

2.2 Diffusion Distillation

Two complementary approaches exist for compressing multi-step diffusion or flow models into fewer steps. One line relies on continuous-time transport trajectories between noise and data. Flow matching (Lipman et al., 2023) and rectified flow transformers (Esser et al., 2024) learn velocity fields that parameterize such transport paths, while consistency distillation enforces that a student's prediction remains invariant along a teacher trajectory, typically a PF-ODE, enabling few-step or one-step generation (Song et al., 2023; Song & Dhariwal, 2024; Lu & Song, 2025). Consistency-style and related trajectory-compression methods have scaled well for images (Luo et al., 2023; Frans et al., 2025; Zheng et al., 2026) and been extended to video (Lv et al., 2025), but they implicitly assume trajectories that are smooth enough to be faithfully compressed.

Distribution Matching Distillation (DMD) (Yin et al., 2024b) instead estimates a reverse-KL gradient from the difference between teacher and generated-distribution scores, matching distributions without following a specific teacher trajectory. DMD2 (Yin et al., 2024a) further incorporates adversarial supervision from real data. Recent video extensions use windowed autoregressive rollouts (Huang et al., 2025; Ge et al., 2026), reward-weighted distribution matching (Lu et al., 2025), and diagonal multi-step scheduling (Liu et al., 2026a).

2.3 Adversarial Training for Video Generation

Generative adversarial networks (Goodfellow et al., 2014) offer single-pass generation by training a discriminator to separate real and generated samples. Early video GANs used 3D convolutions or motion–appearance decompositions (Vondrick et al., 2016; Tulyakov et al., 2018), while later work improved temporal fidelity with continuous-time generators (Skorokhodov et al., 2022). Recent methods use adversarial learning as a post-training or distillation signal for diffusion models. Adversarial Diffusion Distillation (ADD) (Sauer et al., 2024) sharpens one-step image outputs. Adversarial Post-Training (APT) (Lin et al., 2025a) extends this approach to one-step text-to-video generation. Autoregressive APT (AAPT) (Lin et al., 2025b) combines adversarial training with student-forcing in a causal KV-cache architecture, generating one latent frame per forward pass for real-time streaming. Adversarial Self-Distillation (Yang et al., 2026b) and Phased One-Step Adversarial Equilibrium (Cheng et al., 2026) similarly apply adversarial objectives to few-step video generation. Overall, these advances demonstrate that adversarial training can improve video quality while reducing generation to one or a few sampling steps.

3 Method

3.1 Preliminaries

Consistency distillation. Consistency distillation trains an endpoint predictor $f_\theta(x_t, t, c)$ to produce consistent clean predictions from different states along a teacher trajectory, conditioned on a text prompt c. Predictions at adjacent teacher states are encouraged to agree through

$$\mathcal{L}_{\mathrm{CD}}(\theta) = \mathbb{E}_{x_t,t,c}\left[\mathrm{MSE}\left(f_\theta(x_t,t,c), \mathrm{sg}\left[f_{\bar{\theta}}\left(\Phi_{\Delta t}(x_t,t,c), t-\Delta t, c\right)\right]\right)\right], \tag{1}$$

where $\Phi_{\Delta t}$ advances the teacher trajectory from t to $t-\Delta t$, $f_{\bar{\theta}}$ is an exponential moving average of the student, and sg denotes stop-gradient. MSE denotes mean squared error. The resulting predictor supports direct noise-to-data generation and refinement through additional sampling steps (Song et al., 2023).

Self Forcing. Self Forcing trains an autoregressive generator on its own rollouts, with each prediction conditioned on previously generated outputs through the KV cache (Huang et al., 2025). Given a generated sequence $x_\theta = G_\theta(z, c)$, DMD supervises its distribution through a reverse-KL gradient (Yin et al., 2024b):

$$\nabla_\theta \mathrm{KL}(p_\theta \| p_{\mathrm{data}}) = \mathbb{E}_{x_\theta}\left[J_\theta(z,c)^\top \left(\nabla_x \log p_\theta(x_\theta \mid c) - \nabla_x \log p_{\mathrm{data}}(x_\theta \mid c)\right)\right], \tag{2}$$

where p_θ and p_{data} denote the generated and target distributions, respectively, and $J_\theta(z,c)$ is the generator Jacobian with respect to θ. DMD estimates the score difference using a trainable fake-score model and a frozen teacher on noised rollouts. This objective matches output distributions without prescribing a teacher endpoint for each input noise sample.

3.2 Challenges in One-Step Autoregressive Video Distillation

Trajectory geometry challenges consistency distillation. Consistency distillation can produce high-quality videos with four denoising steps, but generation quality drops noticeably when sampling is reduced to one step. This behavior is also observed for rCM (Nie et al., 2026). To investigate this gap, we examine the geometry of teacher denoising trajectories and its effect on distillation.

Following the trajectory analyses in TMD and rectified flow (Nie et al., 2026; Liu et al., 2023), we compare the local teacher velocity with the constant velocity connecting its noise and data endpoints. Specifically, we compute

$$C(t_i) = \frac{1}{d}\left\| \frac{x_{t_i} - x_{t_{i-1}}}{t_i - t_{i-1}} - (x_1 - x_0) \right\|_2^2, \tag{3}$$

where $t = 0$ and $t = 1$ denote the data and highest-noise endpoints, respectively, and d is the number of latent coordinates. This quantity serves as a curvature proxy, measuring how much local trajectory velocity departs from linear evolution between the endpoints.

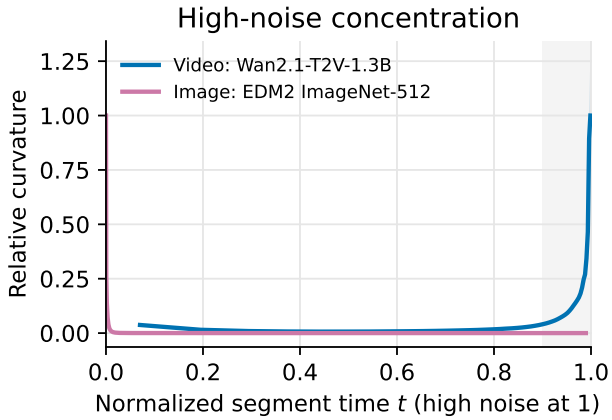

Figure 3: Relative trajectory-curvature profiles show high-noise concentration for Wan video generation but not for EDM2 ImageNet-512 image generation. Each curve is normalized by its own peak.

Table 1: VBench results for four-step consistency distillation with original and rectified teacher trajectories after 300 training steps.

Teacher trajectory	Total	Quality	Semantic
Original curved	61.16	63.92	50.11
Rectified	**71.11**	**75.37**	**54.05**

Figure 3 shows that Wan (Team Wan et al., 2025) video trajectories concentrate 92.5% of their measured curvature mass in the high-noise region $t \geq 0.9$, whereas the EDM2 ImageNet-512 (Karras et al., 2024) reference exhibits no comparable concentration. As shown in Appendix B, this pattern is consistent across prompts and persists when the measurement emphasizes temporal variation. The strongest departure from linear evolution therefore occurs near the starting point of video generation. This suggests a geometric challenge for one-step consistency distillation: four-step sampling allows later predictions to refine the initial estimate at lower noise levels, while one-step generation depends entirely on the endpoint prediction made from the highest-noise input.

To test whether trajectory geometry affects distillation quality, we preserve the teacher's noise and output endpoints and replace only its three intermediate states with time-aware linear interpolation. Table 1 shows that, with identical initialization, sample order, and optimization settings, rectification improves four-step VBench from 61.16 to 71.11 after 300 training iterations. Because the endpoint pairs remain unchanged, this improvement shows that simplifying the intermediate trajectory helps the student learn the same endpoint mapping under a fixed training budget.

Together, the trajectory analysis and rectification experiment support a geometric explanation for the difficulty of compressing video consistency models to one step. They motivate direct output distribution matching, which allows the student to learn a noise-to-data mapping without reproducing the teacher's prescribed endpoint correspondence.

Fine-detail loss in one-step Self Forcing. Self Forcing provides this distributional alternative, but a substantial quality gap remains in the one-step setting. In our evaluation, its one-step outputs exhibit visible blur and loss of fine detail despite being trained on self-generated context, as shown in Figure 1. These imperfections also enter the conditioning context for subsequent predictions.

Recent analysis offers a possible explanation for this behavior. SAP-DMD finds that low spatial frequencies can dominate the DMD update direction, favoring coarse structure while delaying the learning of sharp boundaries and fine textures (Zhou et al., 2026). SDXL-Lightning further reports that its adversarial loss emphasizes high-frequency details at low noise levels (Lin et al., 2024).

These observations motivate complementing teacher-derived distributional guidance with direct supervision from real videos. An adversarial objective provides an additional learning signal for

visual realism as the student distribution evolves. One-Forcing combines this supervision with DMD and uses framewise autoregression to support stable learning from the resulting self-generated rollouts.

3.3 ONE-FORCING

Figure 2 summarizes One-Forcing. Starting from Self Forcing, we introduce real-video adversarial supervision alongside DMD and use framewise prediction to improve one-step generation. Both objectives supervise the videos produced under the student's own generated context.

Adversarial supervision with a shared backbone. We add a discriminator head D_ϕ to the trainable fake-score network, sharing its backbone for denoising and sequence-level real/fake discrimination. The head aggregates features from selected transformer layers into a scalar logit for the entire sequence. The bidirectional critic can evaluate both frame appearance and relationships across frames.

The discriminator compares real video latents with generated rollouts at matched noise levels. Real videos provide a direct reference for visual realism as the student distribution evolves. Denoting the noised real and generated sequences by $x_{\mathrm{real},t}$ and $x_{\theta,t}$, the adversarial losses are

$$\mathcal{L}_G^{\mathrm{adv}} = \mathbb{E}\left[\mathrm{softplus}\left(-D_\phi(x_{\theta,t}, t, c)\right)\right], \tag{4}$$

$$\mathcal{L}_D^{\mathrm{adv}} = \mathbb{E}\left[\mathrm{softplus}\left(-D_\phi(x_{\mathrm{real},t}, t, c)\right) + \mathrm{softplus}\left(D_\phi(x_{\theta,t}, t, c)\right)\right], \tag{5}$$

where $\mathrm{softplus}(u) = \log(1 + \exp(u))$. Generated samples are detached when updating the discriminator.

Framewise autoregressive rollouts. To support stable learning with this joint supervision, we use framewise autoregressive rollouts. The generator predicts one latent frame at a time, conditioned on previously generated frames. Each prediction uses one denoising step and updates the KV cache before the next frame is generated. The student therefore learns with the evolving context produced by its own predictions.

Framewise autoregression decomposes generation into smaller conditional prediction problems and incorporates the newest generated content after every latent frame. Chunkwise generation jointly predicts several new frames before updating this context. This distinction becomes particularly relevant when each prediction has only one denoising step.

The adversarial objective evaluates the generated sequence jointly. We believe that this finer temporal decomposition helps assign sequence-level adversarial feedback to individual causal predictions, making learning more effective as the generated context evolves. Our experiments show that framewise generation achieves higher quality and converges in fewer training iterations than the chunkwise setting.

Joint objective and training. The generator combines DMD guidance with adversarial supervision:

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{DMD}} + \lambda_G \mathcal{L}_G^{\mathrm{adv}}. \tag{6}$$

The shared critic learns to estimate the generated-distribution score and discriminate real from generated videos:

$$\mathcal{L}_{\mathrm{critic}} = \mathcal{L}_{\mathrm{fake}} + \lambda_D \mathcal{L}_D^{\mathrm{adv}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{fake}}$ denotes the fake-score denoising loss, and λ_G and λ_D control the adversarial contributions. DMD supplies teacher-derived distributional guidance, while the adversarial objective supplies direct supervision from real videos. Both act on framewise autoregressive rollouts.

We initialize the generator from the ODE-distilled causal checkpoint provided by Causal Forcing (Zhu et al., 2026a). Each training iteration updates the critic, and the generator is updated every five iterations using a separate minibatch. During generator updates, critic parameters are frozen while gradients propagate through its input. Generator gradients are restricted to the selected rollout window, and KV-cache updates are detached.

Table 2: VBench results for one-step and many-step video generation.

Model	#Params	Resolution	NFE	VBench Scores ↑		
				Total	Quality	Semantic
Multiple steps						
MAGI-1 (Sand.ai et al., 2025)	4.5B	832×480	64	79.18	82.04	67.74
Wan2.1 (Team Wan et al., 2025)	1.3B	832×480	50	**84.26**	**85.30**	**80.09**
SkyReels-V2 (Chen et al., 2025)	1.3B	960×540	30	82.67	84.70	74.53
NOVA (Deng et al., 2025)	0.6B	768×480	25	80.12	80.39	79.05
LTX-Video (HaCohen et al., 2024)	1.9B	768×512	20	80.00	82.30	70.79
Pyramid Flow (Jin et al., 2025)	2B	640×384	20	81.72	84.74	69.62
CausVid (Yin et al., 2025)	1.3B	832×480	4	81.18	84.41	68.30
Self Forcing (Huang et al., 2025)	1.3B	832×480	4	83.46	84.77	78.24
Single step						
Self Forcing (Huang et al., 2025)	1.3B	832×480	1	77.18	79.40	68.34
Causal-Forcing (Zhu et al., 2026a)	1.3B	832×480	1	78.39	80.67	69.25
ASD (Yang et al., 2026b)	1.3B	832×480	1	79.12	81.35	70.19
Ours(chunkwise)	1.3B	832×480	1	81.60	83.65	73.41
Ours(framewise)	1.3B	832×480	1	**83.76**	**85.22**	**77.91**

4 Experiments

4.1 Implementation Details

Following (Zhu et al., 2026a; Huang et al., 2025), the generator is initialized from an ODE initialized checkpoint (Zhu et al., 2026a) and uses one denoising timestep per autoregressive block. The trainable fake-score critic uses Wan2.1-T2V-1.3B, while the frozen real-score teacher uses Wan2.1-T2V-14B. Unless otherwise specified, the reported configuration uses 21 latent frames with 1 latent frame per autoregressive block for the framewise model and 3 latent frames per block for the chunkwise model, a generator update every five critic updates, and non-relativistic adversarial losses. We use $8\times$H100 GPUs for training. Additional implementation details from the final training configuration are provided in Appendix A.

4.2 Evaluation

We evaluate visual quality and text-video alignment using VBench (Huang et al., 2024), reporting the official normalized total, quality, and semantic scores. Unless otherwise specified, our model generates videos at 832×480 resolution and 16 FPS, with one latent frame per autoregressive block. For inference, following ASD (Yang et al., 2026b), we adopt First-Frame Enhancement (FFE) to improve the quality of the first frame. Inference requires a single H100 GPU. Additional evaluation settings and prompt-processing details are provided in Appendix A.

Baselines. Our primary comparisons are with one-step autoregressive video generators, including Self Forcing (Huang et al., 2025), Causal-Forcing (Zhu et al., 2026a), and ASD (Yang et al., 2026b). These methods use the same Wan2.1-1.3B generator backbone as One-Forcing, allowing comparison at a common model scale. To assess how closely one-step generation approaches multi-step quality, we also include Wan2.1 (Team Wan et al., 2025), LTX-Video (HaCohen et al., 2024), SkyReels-V2 (Chen et al., 2025), NOVA (Deng et al., 2025), Pyramid Flow (Jin et al., 2025), MAGI-1 (Sand.ai et al., 2025), CausVid (Yin et al., 2025), and four-step Self Forcing. Model sizes, output resolutions, and sampling steps are summarized in Table 2.

4.3 Main Results

Table 2 summarizes the results. In the one-step setting, One-Forcing (framewise) achieves a total score of 83.76 with quality 85.22 and semantic 77.91, outperforming all prior one-step causal methods including Self Forcing, Causal-Forcing, and ASD by 4–7 points in total score. Notably, with only a

single denoising step, One-Forcing *surpasses most many-step baselines* that use 4 to 25 denoising steps, including Self Forcing, CausVid, LTX-Video, Pyramid Flow, and NOVA. The remaining gap to the 50-step Wan2.1-1.3B baseline is marginal, showing that one-step generation can approach multi-step quality given an effective distillation objective. The framewise model, which generates one latent frame per autoregressive block, outperforms the chunkwise variant with three-frame blocks, achieving a VBench total score of 83.76 versus 81.60. Beyond improving generation quality, framewise autoregression also simplifies adversarial training. In our experiments, the framewise configuration trains stably without R1/R2 regularization, whereas the chunkwise counterpart requires these penalties to maintain stability. This avoids the additional computational overhead associated with discriminator regularization. Under their respective stable training configurations, the framewise model reaches its reported performance after 200 distillation iterations, compared with 750 iterations for the chunkwise model. Our chunkwise model (81.60) nonetheless surpasses all existing one-step methods, confirming that the proposed objective is effective across different block sizes.

Long video generation. We compare one-step models on 20-second videos to assess longer autoregressive rollouts. As shown in Table 3, One-Forcing outperforms Self Forcing on VBench, with improvements in quality and semantic alignment. These results extend its benefits beyond short-clip generation.

Table 3: VBench results for 20-second video generation. Both models use one-step inference.

Model	Total	Quality	Semantic	Multiple objects
Self Forcing	75.30	77.47	66.61	39.33
One-Forcing	**78.52**	**79.38**	**75.11**	**67.91**

4.4 Ablation

In this section, we study the effects of adversarial supervision, FFE, and initialization on generation quality, and examine discriminator behavior during training.

Table 4: VBench ablations of adversarial training and FFE. More details are in Appendix A.

(a) Adversarial objective

Configuration	Total	Quality	Semantic
DMD-only	75.27	78.29	63.19
DMD+GAN(Ours)	**80.36**	**83.22**	**68.92**

(b) First-Frame Enhancement (FFE)

Configuration	Total	Quality	Semantic
Without FFE	80.69	81.15	**78.84**
With FFE (Ours)	**83.30**	**84.65**	77.93

Adversarial objective and FFE. As shown in Table 4a, removing the adversarial branch while retaining DMD and FFE degrades generation quality. This supports our central design of complementing score matching with real-data discrimination, and shows that the gain from adversarial training is not attributable to FFE. The FFE ablation in Table 4b shows that our method outperforms the evaluated prior one-step baselines even without FFE. Consistent with ASD (Yang et al., 2026b), enhancing the first block, which conditions subsequent generation, provides an inference-time complement to our training objective.

Initialization strategy. We compare two framewise configurations that differ in initialization: causal ODE initialization (row 1) and causal Consistency Distillation initialization (row 2). Both use 1-frame blocks. The ODE-initialized model achieves a higher total score (83.76 vs. 82.36) and substantially stronger dynamic degree (52.76 vs. 23.61), while the CD-initialized variant obtains better semantic scores (80.50 vs. 77.91). We attribute the dynamic advantage of ODE initialization to its training data containing richer motion information from the multi-step ODE trajectory, which provides the generator with a stronger motion prior during distillation. The CD initialization, on the other hand, starts from a model already trained for consistency, which benefits semantic alignment but tends to suppress large motions.

Table 5: Ablation study of Causal ODE Initialization and Consistency Distillation Initialization on VBench.

Configuration	Total	Quality	Semantic
Causal ODE Initialization(Ours)	**83.76**	**85.22**	**77.91**
Consistency Distillation Initialization	82.36	82.82	80.50

Discriminator effectiveness. We compare the adversarial training dynamics of One-Forcing against ASD (Yang et al., 2026b). Both methods attach a classification branch to the fake-score backbone operating in noised latent space, but they differ fundamentally in what constitutes the "real" distribution for the discriminator. One-Forcing trains the discriminator on *real data*: noised samples from actual videos in the training set, providing a fixed, high-quality target distribution. ASD instead uses a *self-distillation* target, where the output of an $(n{+}1)$-step model serves as "real" for the n-step student, meaning the discriminator must distinguish between two imperfect model outputs rather than between generated and genuine data.

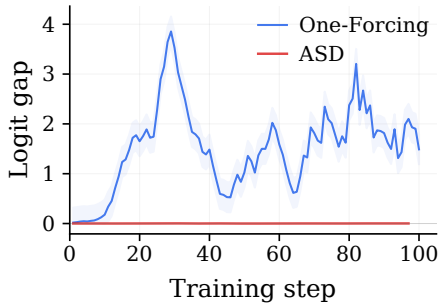

Figure 4: Discriminator logit gap between reference and generated videos during training. One-Forcing maintains a clear gap, while ASD stays near zero.

Figure 4 plots the discriminator logit gap $|l_r - l_f|$ during training, where l_r and l_f denote the discriminator logits for reference and generated videos, respectively. One-Forcing maintains a large, actively varying gap (μ=1.53, σ=1.20): the distributional distance between generated latents and real data is substantial and evolves as the generator improves, indicating a healthy adversarial dynamic. In contrast, ASD's logit gap stays near zero (μ=0.001, $\max$ <0.006) from the very first steps. Because both sides of ASD's comparison are model-generated latents with minimal distributional difference, the discriminator never receives a meaningful learning signal and effectively collapses. This confirms that grounding the adversarial signal in *real data* is critical for effective GAN-based video distillation.

5 Conclusion

We presented One-Forcing, a framework for high-quality one-step autoregressive video generation that combines DMD with adversarial supervision from real videos. Our trajectory analysis reveals concentrated curvature near the high-noise endpoint, providing a geometric perspective on the difficulty of one-step consistency distillation. Within this framework, we identify framewise autoregression as an important ingredient for stable and fast adversarial learning, achieving higher generation quality than its chunkwise counterpart. On VBench, One-Forcing achieves a total score of 83.76, outperforming the evaluated one-step causal video baselines and remaining competitive with strong many-step approaches.

6 Limitations and Future Work

One-Forcing requires real data as the "real" distribution for the discriminator, which is different from data-free methods such as Self Forcing (Huang et al., 2025) and ASD (Yang et al., 2026b). However, these data are already available in standard forcing-like distillation settings where training videos or their precomputed representations are used. Future work will explore larger backbones, efficient attention, and long-video generation frameworks (Yang et al., 2026a; Liu et al., 2026b; Yesiltepe et al., 2025; Cui et al., 2026) for higher-resolution and longer-duration generation. Other directions include adaptive step scheduling that allocates more denoising steps to perceptually complex segments, and action-conditioned generation for interactive world modeling (Ball et al., 2025; Bruce et al., 2024; Sun et al., 2025; Zhang et al., 2025).

AI USE STATEMENT

We used generative AI tools to assist with literature review, manuscript drafting and editing, refinement of research hypotheses, interpretation of experimental results, code implementation, LaTeX formatting, and figure preparation. We also used Qwen2.5-7B-Instruct to rewrite evaluation prompts, as described in Appendix A. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

Efficient autoregressive video generation can support creative content production, accessibility tools, rapid prototyping, simulation, and interactive world modeling. However, it also lowers barriers to creating fabricated media, spreading disinformation, impersonating individuals, and reproducing harmful stereotypes or biases in training data. Responsible deployment should therefore include dataset and prompt filtering, provenance tracking, watermarking or content credentials, synthetic-media detection, clear disclosure of generated content, and usage restrictions in sensitive domains. These safeguards require dedicated evaluation alongside improvements in generation quality and inference speed.

REPRODUCIBILITY STATEMENT

Appendix A describes the model initialization, noise schedule, training objectives, hyperparameters, and inference procedure. Appendix B provides the trajectory-curvature sampling and robustness details, and Appendices C and E report training curves and detailed evaluation results.

REFERENCES

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. Google DeepMind blog, 2025. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/. Accessed: 2026-05-07.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang,

Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length film generative model, 2025. URL https://arxiv.org/abs/2504.13074.

Jiaxiang Cheng, Bing Ma, Xuhua Ren, Hongyi Henry Jin, Kai Yu, Peng Zhang, Wenyue Li, Yuan Zhou, Tianxiang Zheng, and Qinglin Lu. Phased one-step adversarial equilibrium for video diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3237–3245, March 2026. doi: 10.1609/aaai.v40i5.37318. URL https://ojs.aaai.org/index.php/AAAI/article/view/37318.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DzvPiqh23f.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JE9tCwe3lp.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Xingtong Ge, Yi Zhang, Yushi Huang, Dailan He, Xiahong Wang, Bingqi Ma, Guanglu Song, Yu Liu, and Jun Zhang. Salt: Self-consistent distribution matching with cache-aware training for fast video generation, 2026. URL https://arxiv.org/abs/2604.03118.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Google DeepMind. Veo: a text-to-video generation system. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2450–2462. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

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, 2024. URL https://arxiv.org/abs/2501.00103.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

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, 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f4823f831af67a3ef15e41a85434422a-Paper-Conference.pdf.

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 (CVPR)*, pp. 21807–21818, June 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, June 2024. doi: 10.1109/CVPR52733.2024.02282. URL https://openaccess.thecvf.com/content/CVPR2024/html/Karras_Analyzing_and_Improving_the_Training_Dynamics_of_Diffusion_Models_CVPR_2024_paper.html.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2402.13929.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37959–37974. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/lin25m.html.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 41061–41086. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a9468a918fc65dc9ce7b7bd99f4f0ef-Paper-Conference.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=X7YW6STzeL.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=IAyzXjbfwo.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LyJi5ugyJx.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation, 2025. URL https://arxiv.org/abs/2512.04678.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. URL https://arxiv.org/abs/2310.04378.

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14983–14993, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lv_Dual-Expert_Consistency_Model_for_Efficient_and_High-Quality_Video_Generation_ICCV_2025_paper.html.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition matching distillation for fast video generation, 2026. URL https://arxiv.org/abs/2601.09881.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. Google DeepMind blog, 2024. URL https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/. Accessed: 2026-05-07.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. MAGI-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision – ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6. URL https://doi.org/10.1007/978-3-031-73016-0_6.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3636, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Skorokhodov_StyleGAN-V_A_Continuous_Video_Generator_With_the_Price_Image_Quality_CVPR_2022_paper.html.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling, 2025. URL https://arxiv.org/abs/2512.14614.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, et al. Seedance 2.0: Advancing video generation for world complexity, 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, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1526–1535, June 2018. doi: 10.1109/CVPR.2018.00165. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Tulyakov_MoCoGAN_Decomposing_Motion_CVPR_2018_paper.html.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, volume 29, pp. 613–621. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/04025959b191f8f9de3f924f0940515f-Paper.pdf.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=nCAODkpsPJ.

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 *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=P3O0fNmnWa.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with

an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout, 2025. URL https://arxiv.org/abs/2511.20649. CVPR 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54dcf25318f9de5a7a01f0a4125c541e-Paper-Conference.pdf.

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 (CVPR)*, pp. 6613–6623, June 2024b. doi: 10.1109/CVPR52733.2024.00632. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

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 (CVPR)*, pp. 22963–22974, June 2025. doi: 10.1109/CVPR52734.2025.02138. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model, 2025. URL https://arxiv.org/abs/2506.18701.

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, 2026. URL https://arxiv.org/abs/2605.15141.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2uNlM353RI.

Zhenyu Zhou, Can Wang, Chun Chen, Zeyu Zheng, and Defang Chen. Spectral amplitude purification in distribution matching for diffusion distillation, 2026. URL https://arxiv.org/abs/2609.29116.

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, 2026a. URL https://arxiv.org/abs/2602.02214.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=8UZpmrxoLG.

A Details of Implementations

Our implementation is based on the Causal Forcing codebase (Zhu et al., 2026a) and the Wan2.1 model family (Team Wan et al., 2025). The reported framewise One-Forcing model is initialized from an ODE-trained causal model. The real-score network is a frozen bidirectional Wan2.1-T2V-14B model, and the trainable fake-score network is initialized from Wan2.1-T2V-1.3B. We reuse the fake-score backbone as the adversarial discriminator by adding register tokens, lightweight attention blocks, and a classification head to selected transformer layers. No separate pixel-space discriminator is used.

Noise schedule and model parameterization. Following Wan2.1, we use a flow-matching scheduler. For a sampled timestep $t \in [0, 1000]$, the shifted noise level is

$$\sigma_t = \frac{k(t/1000)}{1 + (k-1)(t/1000)},$$

with shift factor $k = 5$ for generator rollouts and for the DMD/GAN critic timestep sampling. The forward corruption process is

$$x_t = (1 - \sigma_t)x_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

and the flow-prediction target is $\epsilon - x_0$. During generation, the model predicts $v_\theta(x_t, t, c)$ and converts it to a clean latent estimate by $\hat{x}_0 = x_t - \sigma_t v_\theta(x_t, t, c)$. Training rollouts use a single denoising timestep per autoregressive block.

Data and prompt processing. The distillation stage uses precomputed training data. Each training example contains a text prompt and the corresponding real data sample; raw videos are not decoded or reloaded during this stage. The adversarial real samples are drawn from this dataset, while fake samples are produced by the current one-step causal generator. The reported framewise model is trained on 21 latent frames with spatial latent size 60×104 and 16 latent channels, corresponding to 832×480 video generation. Text embeddings are computed with the frozen Wan text encoder, and classifier-free guidance uses the standard Wan negative prompt. For VBench evaluation, we rewrite the test prompts using Qwen/Qwen2.5-7B-Instruct following previous works (Huang et al., 2025; Cui et al., 2026), except for the trajectory-rectification experiment in Table 1, the GAN-objective experiment in Table 4a, and the 20-second video evaluation, which use original prompts. Rewriting is held fixed within each paired comparison, including both FFE rows in Table 4b.

GAN and FFE evaluation. The DMD-only and DMD+GAN comparison in Table 4a uses step-200 raw generator weights without EMA, matched initialization, framewise training, and the same FFE inference schedule. Both are evaluated on the same 944 original prompts and generation seeds, with one generated sample per prompt and all 16 VBench dimensions using the normalized scoring formula. These scores use a one-sample protocol rather than the five-sample-per-prompt leaderboard protocol. The FFE comparison in Table 4b keeps the same step-200 checkpoint and prompt set, changing only the inference schedule.

Training details. We train the fake-score critic with the flow denoising objective on generated latents and train the adversarial branch to distinguish noised real data from noised generated samples. Each training iteration performs one critic update; every five iterations, we additionally update the generator on a separately sampled minibatch using the DMD surrogate and the adversarial generator loss. For DMD, the real-score model is evaluated with classifier-free guidance scale 5.0, while the fake-score model is evaluated without classifier-free guidance. The DMD gradient is normalized by the average absolute real-score residual. We use AdamW for both generator and critic, mixed precision, gradient checkpointing, and full-shard FSDP. The reported framewise training run uses 8 NVIDIA H100 GPUs with a per-GPU batch size of 1.

Inference details. The reported framewise model generates one latent frame per autoregressive block. For the one-step setting, subsequent autoregressive blocks use one denoising update. Following (Yang et al., 2026b), the first block is generated with a short four-step warm-up schedule to initialize the KV cache before one-step streaming continues. Unless otherwise specified, videos are decoded at 832×480 resolution and 16 FPS.

Algorithm 1 One-Forcing Training

Require: Generator G_θ, fake-score network v_ϕ with discriminator head D_ϕ, frozen teacher v_T, Dataset $\mathcal{D}$, generator interval K, weights λ_G, λ_D

for training iteration $i = 0, 1, \ldots$ **do**
  if $i \bmod K = 0$ **then**
    Sample prompt c_G from $\mathcal{D}$
    Generate fake samples $x_\theta \leftarrow G_\theta(\epsilon_G, c_G)$ with one-step framewise autoregressive rollout
    Compute $\mathcal{L}_{\mathrm{DMD}}$ from teacher and fake-score predictions on noised rollouts
    Compute $\mathcal{L}_G^{\mathrm{adv}}$ with D_ϕ on independently noised fake samples
    Update θ: $\mathcal{L}_G = \mathcal{L}_{\mathrm{DMD}} + \lambda_G \mathcal{L}_G^{\mathrm{adv}}$
  end if
  Sample a critic minibatch with prompt c_ϕ and real data x_{real} from $\mathcal{D}$
  Generate fake samples $\tilde{x}_\theta \leftarrow G_\theta(\tilde{\epsilon}, c_\phi)$ without generator gradients
  Train v_ϕ to denoise noised generated samples, giving $\mathcal{L}_{\mathrm{fake}}$
  Train D_ϕ to classify noised real data as real and noised generated samples as fake, giving $\mathcal{L}_D^{\mathrm{adv}}$
  Update ϕ: $\mathcal{L}_{\mathrm{critic}} = \mathcal{L}_{\mathrm{fake}} + \lambda_D \mathcal{L}_D^{\mathrm{adv}}$
end for

Table 6: Training hyperparameters for the reported framewise One-Forcing configuration.

Hyperparameter	**Value**
Generator initialization	ODE-trained Causal model
Generator / real score / fake score	Wan2.1-T2V-1.3B / Wan2.1-T2V-14B / Wan2.1-T2V-1.3B
Objective	Flow denoising for fake score; DMD + noised-latent GAN for generator
Training frames	21 latent frames, $16 \times 60 \times 104$ per frame
Frames per autoregressive block	1
Training rollout steps per block	1
Guidance scale	5.0 for real-score guidance; no CFG for generator or fake score
Timestep range and shift	1000 training timesteps; shift factor 5.0 for sampled DMD/GAN timesteps
Update schedule	One critic update per iteration; one generator update every five iterations
Optimizer	AdamW, $\beta_1 = 0$, $\beta_2 = 0.999$, weight decay 0.01
Learning rates	1.0×10^{-5} for generator and fake-score critic
Batch size	1 per GPU on 8 GPUs
EMA	Decay 0.99, starting after 50 iterations
GAN branch	Layers $\{21, 29\}$, 2 registers, 1536 feature dim, 2048 FFN dim, 12 heads
GAN loss	Non-relativistic logistic loss, $\lambda_G = \lambda_D = 0.03$
Discriminator regularization	None
Systems	Mixed precision, gradient checkpointing, full-shard FSDP, activation CPU offload
Convergence steps	200 iterations

B Trajectory Curvature Analysis Details

This appendix provides the sampling and robustness details for the trajectory-curvature calculation in Equation 3. The video comparison uses 100 50-step Wan2.1-T2V-1.3B (Team Wan et al., 2025) trajectories with diverse motion and scene prompts, shift 8, classifier-free guidance 6, and one deterministic seed per prompt. The image-domain comparison uses eight 256-step trajectories from the official EDM2 (Karras et al., 2024) ImageNet-512 teacher, matching the teacher family used by scalable consistency models (Lu & Song, 2025). EDM2 noise levels are normalized so $t = 1$ is the highest-noise endpoint.

The high-noise concentration is stable across prompts: per-prompt estimates give $92.49\% \pm 0.13\%$ curvature mass at $t \geq 0.9$ (mean $\pm$ SEM; 95% bootstrap CI $[92.24, 92.73]\%$) and a high-noise/mid-noise ratio of 33.1 ± 0.7 (95% bootstrap CI $[31.8, 34.4]$). A temporal-difference version of the same metric, which removes static appearance and emphasizes motion structure, still places 88.6% of the curvature mass at $t \geq 0.9$ with a high/mid ratio of 19.3.

The additional GAN-objective and trajectory-rectification training runs use 8×H200 GPUs and a single training seed. Their paired evaluations use one generated sample per prompt; they do not estimate variation across training seeds.

Four-step rectification evaluation. The experiment in Table 1 uses two 300-step checkpoints trained on the same teacher-trajectory dataset. Only the three intermediate states differ; endpoints, initialization, sample order, and optimization are fixed. Both checkpoints are evaluated with four-step sampling, using raw generator weights without EMA on the same 944 original VBench prompts and generation seeds, with one sample per prompt and all 16 dimensions. This uses the normalized VBench scoring formula, but is not the five-sample-per-prompt leaderboard protocol. Rectification improves total score from 61.16 to 71.11, quality from 63.92 to 75.37, and semantic score from 50.11 to 54.05.

C Training Loss Curves

Figure 5 compares the training loss curves of our One-Forcing and ASD (Yang et al., 2026b) over the first 100 steps. Both methods start from an ODE-initialized generator checkpoint. Panel (a) shows the DMD loss, which drives the score-matching component; both methods exhibit similar initial magnitudes, though One-Forcing stabilizes at a lower level. Panel (b) reveals the generator GAN loss: One-Forcing's loss varies actively as the discriminator provides meaningful gradients, whereas ASD's GAN loss flatlines at $\ln 2 \cdot 0.01 \approx 0.0069$ throughout training. Panels (c) and (d) show the critic and discriminator losses, respectively; One-Forcing's discriminator loss decreases over training as it learns to distinguish real from fake, while ASD's discriminator loss remains constant.

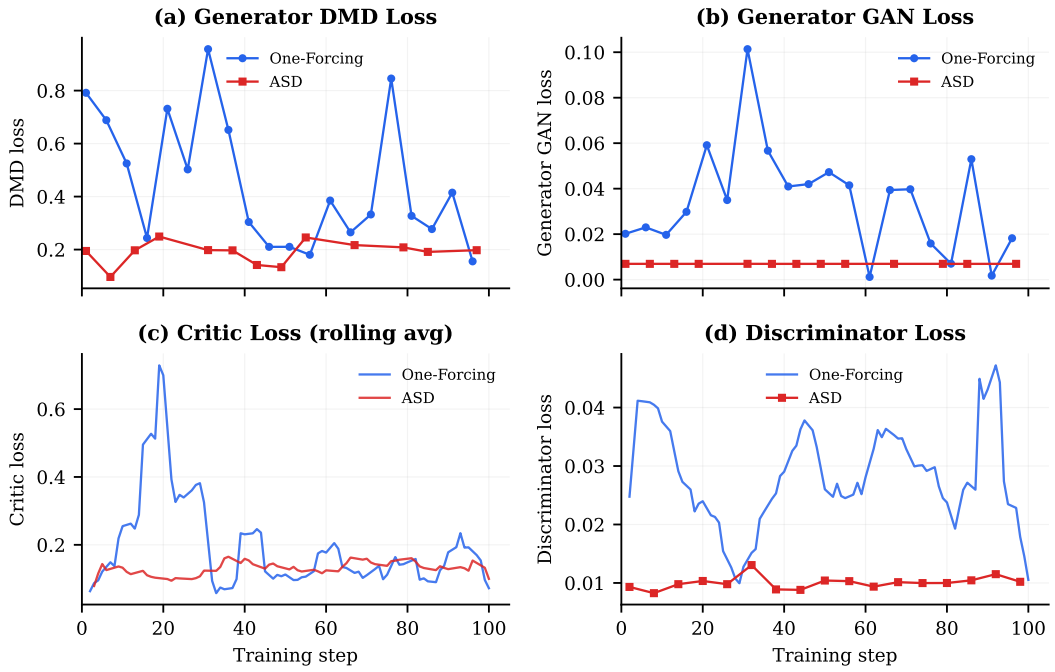

Figure 5: Training loss curves for One-Forcing (blue) and ASD (red) over the first 100 steps. (a) DMD loss. (b) Generator GAN loss. (c) Critic loss (rolling average). (d) Discriminator loss.

D Training Beyond 200 Steps

To assess stability under extended training, we compare the main 200-step framewise One-Forcing result from Table 2 with an additional evaluation at step 400. This evaluation uses Qwen-rewritten conditioning prompts, with five samples per prompt and all 16 VBench dimensions evaluated. As

shown in Table 7, the total score changes only slightly from 83.76 at step 200 to 83.93 at step 400, while the quality and semantic scores remain comparable. These results show that One-Forcing maintains stable generation quality even when we double the training budget.

Table 7: VBench scores during extended framewise training.

Training step	Total	Quality	Semantic
200	83.76	85.22	77.91
400	83.93	85.41	78.04

E VBench Scores Across All Dimensions

Following the evaluation visualization style of Self Forcing, Figure 6 expands the full 16-dimensional VBench profile for the Table 2 entries with available per-dimension records: One-Forcing (framewise), Causal-Forcing 1-step (Zhu et al., 2026a), CausVID 4-step (Yin et al., 2025), ASD (Yang et al., 2026b), Self Forcing DMD 1-step (Huang et al., 2025), and Self Forcing DMD 4-step. One-Forcing improves over the one-step causal baselines in the aggregate score and shows stronger object, spatial-relation, scene, and dynamic-degree performance than ASD, Causal-Forcing, and one-step Self Forcing, while keeping high temporal smoothness. Compared with four-step Self Forcing, One-Forcing has a higher normalized VBench total and quality score, with gains in dynamic degree and several object/action/scene dimensions, while remaining slightly lower on color, imaging quality, and some consistency-style dimensions.

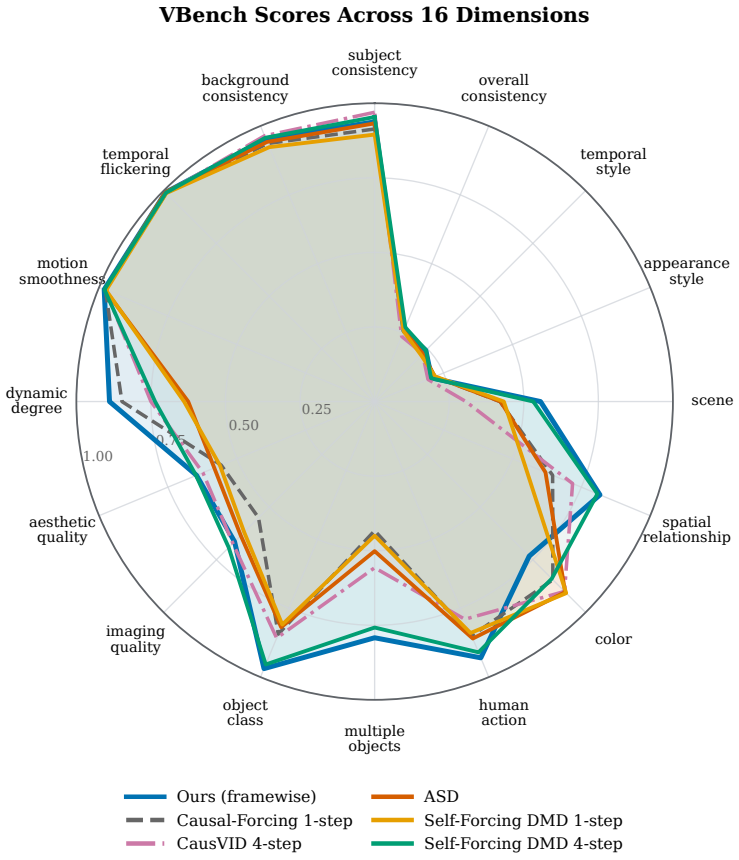

Figure 6: VBench scores across all 16 dimensions for selected Table 2 entries. Higher radial values indicate better normalized VBench sub-metric scores.

F Additional VBench Ablation Dimensions

Table 8 expands the trajectory-rectification, adversarial-objective, and FFE comparisons in Tables 1 and 4 with additional VBench dimensions. Trajectory rectification improves subject consistency, motion smoothness, and imaging quality, connecting the geometric intervention to both temporal and visual quality. The adversarial objective improves subject consistency and imaging quality alongside multiple-object composition, while FFE improves subject consistency and multiple-object composition. These measurements clarify which aspects of generation contribute to the aggregate gains. They use the same paired settings as their corresponding main-text ablations: original prompts for rectification and the GAN comparison, and rewritten prompts for the FFE comparison.

Table 8: Additional VBench dimensions for the main-text ablations. The baseline/variant pairs are original/rectified trajectories, DMD-only/DMD+GAN, and without/with FFE, respectively.

Ablation	Metric	Baseline	Variant
Trajectory rectification	Subject consistency	71.18	**81.37**
	Motion smoothness	86.00	**94.09**
	Imaging quality	47.78	**53.80**
Adversarial objective	Multiple objects	39.41	**65.24**
	Subject consistency	81.54	**92.93**
	Imaging quality	58.19	**69.48**
FFE	Subject consistency	91.05	**96.42**
	Multiple objects	76.14	**82.39**

G Forward-KL Regularization Details

We also test a forward-KL-style distillation regularizer that matches the one-step generator output to the teacher ODE endpoint conditioned on the same noisy latent. The probabilistic objective this regularizer approximates is

$$\mathcal{L}_{\text{fkl}} = \mathbb{E}_{x_{t_0}^{\text{ode}},c}\left[D_{\text{KL}}\big(q_{\text{ODE}}(x_0 \mid x_{t_0}^{\text{ode}}, c) \,\|\, p_\theta(x_0 \mid x_{t_0}^{\text{ode}}, t_0, c)\big)\right], \tag{8}$$

where q_{ODE} is represented by the saved ODE trajectory. Our implementation does not estimate this KL directly. Instead, it optimizes a deterministic squared-error surrogate: the generator prediction is regressed to the stored clean endpoint $x_{\text{tar}}^{\text{ode}}$,

$$\widehat{\mathcal{L}}_{\text{fkl}} = \mathbb{E}_{x_{t_0}^{\text{ode}},x_{\text{tar}}^{\text{ode}}}\left\|G_\theta(x_{t_0}^{\text{ode}}, t_0, c) - x_{\text{tar}}^{\text{ode}}\right\|_2^2. \tag{9}$$

This squared-error surrogate is added to the DMD and adversarial generator objectives with weight λ_{fkl}. Adding λ_{fkl}=1 substantially hurts performance: quality drops to 75.03, total score to 74.83, and dynamic degree to 1.30. Relative to the chunkwise baseline (total 81.60), the total score drops by nearly 7 points and dynamics almost vanish. These results suggest that the deterministic squared-error surrogate for forward-KL regularization is poorly aligned with the distributional objectives used by One-Forcing in the one-step setting. The objectives therefore conflict, and the extra anchor suppresses motion rather than improving fidelity.

H Diversity and Distributional Coverage

We evaluate whether adding the adversarial objective reduces diversity using two complementary comparisons. First, we measure LPIPS-VGG variation between generated samples for 100 randomly sampled prompts, with four generation seeds per prompt, comparing DMD-only and DMD+GAN at training step 200. Second, we evaluate recall using 256 videos randomly sampled from our dataset. These two protocols are reported together in Table 9, but use different evaluation samples.

LPIPS-VGG is nearly unchanged (0.6437 versus 0.6440), indicating comparable within-prompt perceptual variation. Recall increases from 0.457 to 0.711, providing evidence of improved distributional coverage in this evaluation. These measurements complement the VBench quality gains by assessing within-prompt variation and coverage.

Table 9: Diversity and coverage with and without the adversarial objective. LPIPS-VGG uses 100 prompts with four generation seeds each; recall uses a separate 256-video evaluation.

Objective	LPIPS-VGG	Recall ↑
DMD-only	0.6437	0.457
DMD+GAN	0.6440	**0.711**

I Four-Step Generation

Although our main focus is one-step generation, the joint DMD and adversarial objective also applies to a larger denoising budget. We train a four-step One-Forcing model and compare it with four-step Self Forcing on 944 VBench prompts, using Qwen prompt rewriting and the same backbone and inference-step budget. Table 10 shows higher total and quality scores for One-Forcing, together with improvements in imaging quality and multiple-object composition; the semantic score is slightly lower. The gain at this operating point shows that the benefit of the training objective is not restricted to one-step generation.

Table 10: VBench comparison of models trained for four-step generation.

Model	Total	Quality	Semantic	Imaging quality	Multiple objects
Self Forcing	83.46	84.77	**78.24**	69.19	75.76
One-Forcing	**83.85**	**85.35**	77.82	**71.12**	**77.74**

J Inference Latency and Training Cost

We benchmark three inference schedules for 21 latent frames over 20 trials on a single H200 GPU. The all-one-step schedule uses one denoising step for every block; FFE uses four steps for the first block and one for each subsequent block; the all-four-step schedule uses four steps for every block. End-to-end timing includes VAE decoding. As shown in Table 11, FFE adds only 0.06 s to the end-to-end time relative to all-one-step generation, while retaining a similar steady-state block latency. Relative to all-four-step sampling, FFE reduces end-to-end time from 7.72 s to 4.19 s, a $1.84\times$ speedup. These measurements complement the FFE quality ablation: the extra denoising budget is confined to the first block rather than repeated throughout the rollout.

Table 11: Inference timing on one H200 GPU for 21 latent frames and 20 trials. Steady-state block latency excludes the first block; end-to-end time includes VAE decoding.

Schedule	Steady block (ms)	Diffusion (s)	End-to-end (s)
All one-step	117.8	2.46	4.13
FFE	120.8	2.53	4.19
All four-step	290.9	6.07	7.72

Training FLOPs. As detailed in Appendix A, the reported training configuration uses no R1/R2 gradient penalties. Omitting this regularization reduces training FLOPs by 28.59%. This saving concerns arithmetic operations during training, not a measured reduction in training time, and is separate from the inference timings in Table 11.