ENHANCING AUTOREGRESSIVE VIDEO GENERATION VIA REPRESENTATION ADVERSARIAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step autoregressive video generation enables efficient streaming synthesis, but errors introduced in early temporal blocks are reused as context and can propagate through subsequent rollouts, leading to detail degradation, structural drift, and unstable motion. Existing distribution matching distillation (DMD) primarily aligns student and teacher distributions in diffusion latent space, but provides no direct supervision over the perceptual quality of decoded videos. We introduce **Radian**, a representation-space adversarial distillation framework that complements on-policy DMD with real-data adversarial supervision in the feature space defined by a frozen visual foundation model (VFM). During training, Radian sparsely decodes frames from autoregressive student rollouts, extracts multi-level visual representations, and applies lightweight discriminator heads to distinguish generated outputs from real video frames. The DMD objective anchors the student to the pretrained teacher, while the representation-space adversarial objective supplies complementary perceptual and semantic gradients that promote high-quality modes. These additional components are discarded after training, leaving the generator architecture and inference-time denoising budget unchanged. Experiments on Wan2.1-1.3B cover four-step chunk-wise, one-step frame-wise, and minute-long autoregressive generation. Our method achieves a VBench Total of 0.8444 and a VideoAlign Total of 0.8033 under four-step generation, and improves VBench-Long from 0.7805 to 0.8041 over Rolling Forcing while using fewer denoising steps. Controlled comparisons across image, video, and diffusion representations further indicate that the choice of representation spaces induces distinct adversarial signals, and external VFM gradients complement DMD more effectively than adversarial supervision derived from diffusion-internal features.

1 INTRODUCTION

Recent video generation models have substantially advanced visual fidelity, motion realism, and prompt alignment (Wan et al., 2025; Wiedemer et al., 2025; Kong et al., 2025; Seedance & et al., 2026; HaCohen et al., 2026). However, typical video generators synthesize fixed-length clips through bidirectional temporal attention, incurring prohibitive latency, computational costs and incompatibility with causal streaming generation. These limitations conflict with today's urgent demand for interactive world models, real-time audio-visual avatars, and online video editing, which require continuous generation and immediate response (Zhao et al., 2026b; Sun et al., 2025; Huang et al., 2025b). The central challenge is to retain the quality of a bidirectional teacher under a small sampling budget and long autoregressive rollouts.

Most previous works address the first challenge through few-step diffusion distillation. Progressive and consistency-based methods compress iterative denoising by direct mappings between noise levels (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Lu & Song, 2025; Zheng et al., 2026c), while Distribution Matching Distillation (DMD) trains a student to reproduce the output distribution of a pretrained diffusion teacher (Yin et al., 2024b;a). Recent video generation methods like Self Forcing further combine distillation with block-wise causal generation and KV caching, and train the student on its own generated histories to reduce the train-test gap (Yin et al., 2025; Huang et al., 2025a; Zhu et al., 2026; Gu et al., 2026; Ge et al., 2026; Liu et al., 2026; Cui et al., 2026). Nevertheless, early generated blocks with errors become the context for subsequent blocks, and these errors then propagate and accumulate over generation, leading to visual degradation, de-

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.

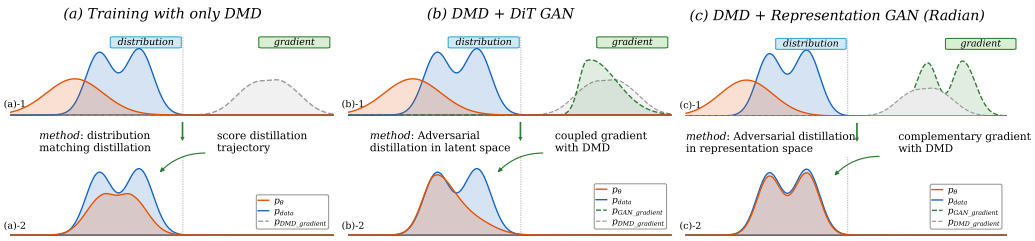


Figure 1: **Complementary Representation-Space Supervision. Radian** complements DMD with perceptual gradients, guiding the student toward a better-matched data distribution.

tail loss, and semantic drift. Moreover, methods like DMD transfer the teacher distribution through score differences in the diffusion latent space, failing to align decoded outputs with real videos.

Another line of works improve few-step generators through adversarial distillation (Sauer et al., 2024a; Lin et al., 2025a;b; Feng et al., 2026a; Li et al., 2026; Lin et al., 2024; Cheng et al., 2026). When real samples are used as the positive reference, the discriminator provides a measure of data realism that teacher-based score matching does not offer. ADD exploits this principle for image generation in the representation-space of a frozen DINOv2 encoder (Sauer et al., 2024b). When it comes to video generation, in contrast, discrimination is performed in clean or noise-corrupted VAE latents or internal features from DiT backbones. However, such discrimination may be sub-optimal for improving the video quality, as VAE latents are optimized for compression and reconstruction, whereas diffusion-backbone features are learned for denoising or velocity prediction. The diffusion space used by DMD is close to them, causing adversarial gradients to overlap with score distillation. Moreover, without RGB decoding, the discriminator cannot detect texture blur, artifacts, or VAE decoding errors, making such features suboptimal for adversarial supervision.

In this work, we present **Radian**, a method that **performs adversarial distillation in representation space** to enhance few-step autoregressive video generation. Radian augments on-policy DMD with an adversarial objective over features from a frozen visual foundation model (VFM). During training, selected outputs from causal autoregressive rollouts are decoded into RGB frames and mapped to multi-level VFM features, which are distinguished from those of real videos using lightweight discriminator heads. The VFM and discriminator heads are removed after training, leaving the generator architecture and inference cost unchanged. DMD anchors the student to the teacher distribution, while representation-space adversarial supervision uses real data to suppress perceptually and semantically degraded outputs and reshape the learned distribution toward higher-quality modes.

Our contributions are as follows. We introduce a new representation adversarial distillation framework that complements on-policy DMD with real-data supervision over decoded outputs while preserving the original inference architecture and cost. Beyond the proposed framework, we systematically study adversarial representation spaces spanning image VFMs including DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025), and SigLIP2 (Tschannen et al., 2025), video encoders including V-JEPA 2.1 (Mur-Labadia et al., 2026) and VideoMAE (Tong et al., 2022), as well as diffusion-internal features. Our analysis further shows that different image representations induce distinct adversarial signals, video representations provide complementary temporal regularization, and external VFM gradients are more complementary to DMD than diffusion-internal supervision as conceptually illustrated in Fig. 1. Finally, experiments on causal generators built upon Wan2.1 1.3B demonstrate consistent improvements in visual quality, motion quality, and prompt alignment over strong Forcing-series baselines, achieving state-of-the-art performance on VBench (Huang et al., 2023) and VBench-Long (Huang et al., 2024) across 4-step, 1-step, and long-horizon generation.

2 Related Work

Few-Step and Autoregressive Video Generation. Few-step diffusion distillation generally follows trajectory- or distribution-based paradigms. Trajectory-based methods compress teacher sampling trajectories through progressive or consistency objectives (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Lu & Song, 2025; Zheng et al., 2026c;b), whereas distribution-based methods directly align student and teacher output distributions. DMD estimates a reverse Kullback–Leibler (KL) gradient from teacher and student score functions, and DMD2 further improves score esti-

mation and training stability (Yin et al., 2024b;a; Ge et al., 2025). Related variants further adapt distribution matching to efficient few-step video generation (Gu et al., 2026; Ge et al., 2026). For streaming generation, autoregressive video models factorize videos into causal temporal blocks and reuse previous outputs through KV caching. CausVid (Yin et al., 2025) distills bidirectional video diffusion into a few-step causal generator, while Self Forcing (Huang et al., 2025a) reduces the training–inference gap by training on the student's own rollouts. Causal Forcing and Causal Forcing++ further improve initialization and scalable few-step frame-wise generation (Zhu et al., 2026; Zhao et al., 2026c), while Rolling Forcing and Self Forcing++ extend self-rollout training to long-horizon generation (Liu et al., 2026; Cui et al., 2026).

Adversarial Training for Diffusion Models. Adversarial objectives have been widely used to preserve sample quality under small inference budgets. In image generation, ADD combines diffusion score distillation with a discriminator on frozen DINOv2 features (Sauer et al., 2024b), while related methods also exploit internal diffusion representations (Sauer et al., 2024a). Extending to video generation, APT applies adversarial post-training to one-step synthesis (Lin et al., 2025a), and AAPT further adapts it to autoregressive video generation (Lin et al., 2025b). Subsequent work explores adversarial distribution matching, cross-step self-distillation, phased training, and other autoregressive adversarial formulations (Lu et al., 2025; Yang et al., 2026; Cheng et al., 2026; Li et al., 2026; Feng et al., 2026a; Luo et al., 2026). Nevertheless, most existing methods discriminate with diffusion-internal features, which may provide redundant or conflicting supervision with DMD and remain weakly aligned with decoded videos, limiting sensitivity to pixel-space artifacts.

Pretrained Representations for Generative Modeling. Pretrained representation spaces are increasingly used in generative modeling (Oquab et al., 2023; Siméoni et al., 2025; Assran et al., 2023; Mur-Labadia et al., 2026; Tong et al., 2022; Wang et al., 2026). REPA (Yu et al., 2024) aligns intermediate DiT states with clean VFM features, while RAE (Zheng et al., 2026a) adopts frozen VFM encoders as generative latent spaces. Subsequent works further explore structured representations and representation-space generative supervision (Shi et al., 2026; Wang et al., 2026; Zhao et al., 2026a; Feng et al., 2026b). In contrast, our proposed method targets an already distilled autoregressive generator and adversarially distinguishes decoded real and generated videos using frozen VFM features, rather than aligning features or redesigning the latent space. This provides real-data supervision complementary to DMD without additional inference cost.

3 METHOD

3.1 PRELIMINARIES

Flow-Matching Models and Causal Video Generation. Let $\mathbf{x}_0$ denote a clean video latent and $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$ its noisy state at timestep t, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and α_t and σ_t denote the signal and noise scaling coefficients determined by the diffusion noise schedule, respectively. A conditional diffusion or flow-matching model learns a time-dependent vector field, $\mathbf{v}(\mathbf{x}_t, t, c)$, that transports noise toward the data distribution through the corresponding probability-flow ODE. For causal generation, the video is partitioned into B temporal blocks, $\mathbf{x}_0 = (\mathbf{x}_0^1, \ldots, \mathbf{x}_0^B)$, and its distribution is factorized as

$$p_\theta(\mathbf{x}_0 \mid c) = \prod_{b=1}^{B} p_\theta\big(\mathbf{x}_0^b \mid \mathbf{x}_0^{<b}, c\big)\,. \tag{1}$$

Each block is generated with a few denoising steps while previously computed context is reused through the KV cache. During on-policy training, the student conditions on its own generated history $\hat{\mathbf{x}}_0^{<b}$, matching context at inference time (Lipman et al., 2023; Yin et al., 2025; Huang et al., 2025a).

Distribution Matching Distillation. DMD trains a few-step generator by minimizing the reverse KL divergence between the student distribution and a reference distribution defined by a pretrained diffusion teacher (Yin et al., 2024b;a). Given an autoregressive student rollout $\hat{\mathbf{x}}_0 = G_\theta(\mathbf{z}, c)$, we obtain $\hat{\mathbf{x}}_t = \alpha_t \hat{\mathbf{x}}_0 + \sigma_t \boldsymbol{\epsilon}$ through forward diffusion. At a randomly sampled timestep t, gradient of the reverse KL divergence is approximated by the teacher-student score functions. Specifically, let $\mathbf{s}_r(\hat{\mathbf{x}}_t, t, c)$ denote the frozen teacher score and $\mathbf{s}_\phi(\hat{\mathbf{x}}_t, t, c)$ be the score of the current student distribution, estimated by a trainable fake-score model. The generator update is

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_{t,\mathbf{z},\boldsymbol{\epsilon}}\left[\big(\mathbf{s}_\phi(\hat{\mathbf{x}}_t, t, c) - \mathbf{s}_r(\hat{\mathbf{x}}_t, t, c)\big)\,\frac{\partial \hat{\mathbf{x}}_t}{\partial \theta}\right]. \tag{2}$$

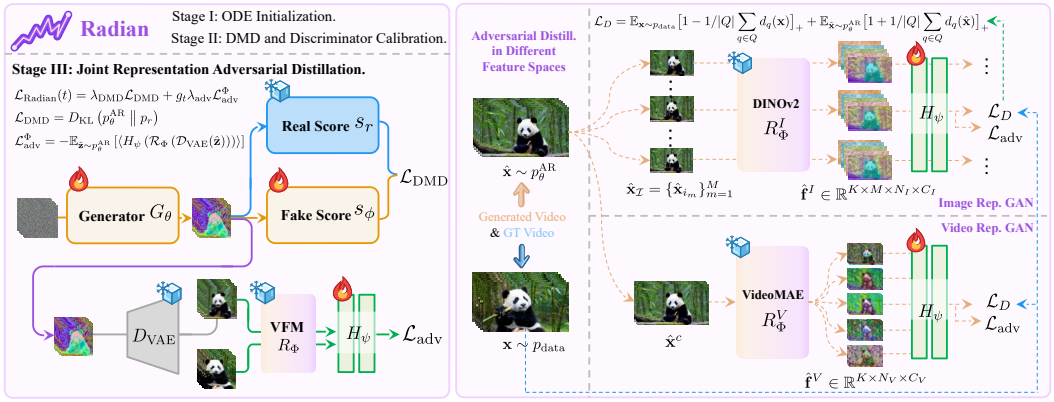


Figure 2: **Overview of the Three-stage Training Pipeline.**

The fake-score model and generator are updated alternately, allowing DMD to match the student rollout distribution to the teacher under a small sampling budget.

Adversarial Distillation. Adversarial distillation introduces supervision from real data by jointly training a generator G_θ and a discriminator D_ψ that distinguishes generated samples from real ones (Sauer et al., 2024b;a; Lin et al., 2025a;b; Li et al., 2026). Given noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and conditioning c, the standard adversarial objective is

$$\min_\theta \max_\psi \; \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}} \left[\log D_\psi(\mathbf{x}, c)\right] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\log(1 - D_\psi(G_\theta(\mathbf{z}, c), c))\right]. \tag{3}$$

Here, D_ψ denotes the complete discriminator mapping, including any input transformation or feature extractor. For autoregressive video generation, $G_\theta(\mathbf{z}, c)$ represents a complete causal rollout. In practice, the logistic objective may be replaced by hinge or related adversarial losses. Unlike teacher-based distribution matching, adversarial distillation directly compares generated samples against real data and therefore provides complementary signals toward the real-data distribution.

3.2 Representation Adversarial Distillation for Causal Video Generation

We consider a causal student generator G_θ that synthesizes a video as a sequence of B consecutive latent blocks, each generated with K denoising steps. We denote the resulting latent rollout by $\hat{\mathbf{z}} = \hat{\mathbf{z}}^{(1:B)} \sim p_\theta^{\text{AR}}$. Since the rollout reuses previously computed context through the KV cache, it follows the same causal computation at inference and exposes the student to errors in its own history. We train this causal student through the three successive stages illustrated in Fig. 2.

Stage I: ODE Initialization. We first initialize the causal student by regressing from intermediate states of precomputed ODE trajectories to their clean endpoints. Given an intermediate latent $\mathbf{z}_\tau^{\text{ODE}}$, its corresponding timestep τ, and the clean trajectory endpoint $\mathbf{z}_0^{\text{ODE}}$, the initialization objective is

$$\mathcal{L}_{\text{ODE}} = \mathbb{E}_{(\mathbf{z}_\tau^{\text{ODE}}, \mathbf{z}_0^{\text{ODE}}, \tau, c)} \left[\left\| G_\theta\left(\mathbf{z}_\tau^{\text{ODE}}, \tau, c\right) - \mathbf{z}_0^{\text{ODE}} \right\|_2^2 \right]. \tag{4}$$

This stage provides the student with a stable few-step causal solver before it is trained on its own autoregressive rollouts, reducing the burden on the subsequent stages.

Stage II: DMD Stabilization and Discriminator Calibration. Starting from the ODE initialization, we briefly optimize the causal student with on-policy DMD using Eq. 2 to adapt it to its own autoregressive histories. In parallel, the feature discriminator D_ψ is warmed up using real videos and decoded student rollouts. A tunable sanity gate blocks adversarial gradients from affecting the generator during this period, preventing an uninitialized discriminator from destabilizing generator training. The resulting student and calibrated discriminator are then used to initialize Stage III.

Stage III: Joint Representation Adversarial Distillation. Once the sanity gate is activated, the student is jointly optimized by on-policy DMD and representation adversarial distillation:

$$\mathcal{L}_{\text{Radian}}(t) = \lambda_{\text{DMD}} \mathcal{L}_{\text{DMD}} + g_t \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{\Phi}. \tag{5}$$

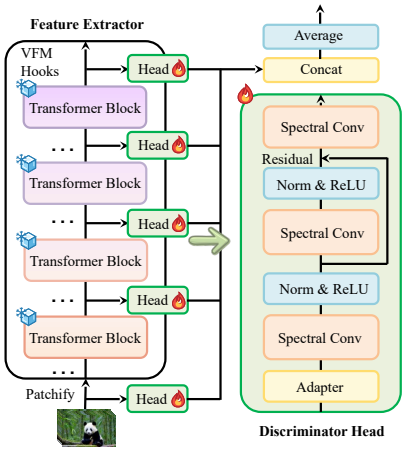


Figure 3: **Architecture of Lightweight Discriminator Head.**

$\mathcal{L}_{\text{DMD}}$ serves as an anchor that keeps the student close to the teacher's distribution, whereas $\mathcal{L}_{\text{adv}}^{\Phi}$ uses real videos to reshape the student distribution toward perceptually realistic and semantically coherent outputs. The two terms are evaluated on separate self-rollouts sampled from the same causal student distribution, with g_t as a weighting coefficient. Both objectives remain active throughout the joint stage, allowing adversarial supervision without discarding the teacher-derived generative prior.

Concretely, to compute the adversarial objective, we define the multi-level feature operator of a frozen visual encoder Φ as $\mathcal{R}_{\Phi}(\mathbf{x}) = \{\Phi_{\ell}(\mathcal{P}(\mathbf{x}))\}_{\ell \in \mathcal{S}}$, with $\mathcal{S}$ being the selected feature levels. The adversarial objective is

$$\mathcal{L}_{\text{adv}}^{\Phi} = -\mathbb{E}_{\hat{\mathbf{z}} \sim p_{\theta}^{\text{AR}}} \left[\langle H_{\psi} \left(\mathcal{R}_{\Phi} \left(\mathcal{D}_{\text{VAE}}(\hat{\mathbf{z}}) \right) \right) \rangle \right]. \tag{6}$$

Here, $\mathcal{P}$ converts decoded RGB frames into the input format required by Φ through range conversion, spatial augmentation, resizing, and normalization. H_{ψ} collects the trainable discriminator heads, and $\langle \cdot \rangle$ averages their logits over sampled frames, feature levels, and spatial tokens. $\mathcal{D}_{\text{VAE}}$ represent the VAE decoder. They remain frozen but differentiable during the training process, and are removed after training.

3.3 Multi-Level Feature Discrimination over Decoded Outputs

Eq. 6 is realized using sparse temporal sampling and multi-level feature discrimination. Given a real video $\mathbf{v} \sim p_{\text{data}}$ and a generated rollout $\hat{\mathbf{z}} \sim p_{\theta}^{\text{AR}}$, we uniformly sample M latent positions without replacement and map them to the corresponding normalized positions in the real video:

$$\begin{aligned} \mathcal{I} &= \{i_m\}_{m=1}^{M} \sim \text{UnifWOR}\left(\{0, \dots, T_z - 1\}, M\right), \\ j_m &= \text{round}\left(\frac{i_m}{T_z - 1} (T_r - 1) \right). \end{aligned} \tag{7}$$

where T_z and T_r denote the lengths of the latent rollout and real video. UnifWOR denotes uniform sampling without replacement, and j_m is the real-video position obtained by linearly mapping the sampled latent index i_m to the normalized temporal coordinate of the real video. For each sampled position, only a short latent neighborhood $\mathcal{N}_q(i_m) = i : |i - i_m| \le q$ of radius q is decoded:

$$\hat{\mathbf{x}}_m = \left[\mathcal{D}_{\text{VAE}} \left(\hat{\mathbf{z}}_{\mathcal{N}_q(i_m)} \right) \right]_{\text{ctr}}, \qquad \mathbf{x}_m = \mathbf{v}_{j_m}. \tag{8}$$

where $[\cdot]_{\text{ctr}}$ selects the center RGB frame at each sampled position. The real and generated videos are sampled independently, and the mapping only aligns their relative temporal positions. For $\mathbf{y}_m \in \mathbf{x}_m, \hat{\mathbf{x}}_m$, the frozen encoder extracts multi-level spatial features.

$$\mathbf{F}_{\ell}(\mathbf{y}_m) = \Phi_{\ell}(\mathcal{P}(\mathbf{y}_m)) \in \mathbb{R}^{N_{\ell} \times C_{\ell}}, \qquad \widetilde{\mathbf{F}}_{\ell}(\mathbf{y}_m) = \mathbf{F}_{\ell}(\mathbf{y}_m) + \mathbf{1}_{N_{\ell}} \mathbf{r}_{\ell}^{\top}, \tag{9}$$

where N_{ℓ} and C_{ℓ} denote the number and dimension of spatial tokens, and $\mathbf{r}_{\ell}$ is the global readout token. Injecting $\mathbf{r}_{\ell}$ into spatial tokens provides each local feature with global semantic context. As illustrated in Fig. 3, on top of these features, each selected level is equipped with a lightweight trainable head $h_{\psi,\ell}$ that produces dense spatial logits:

$$\mathbf{d}_{m,\ell}(\mathbf{y}_m) = h_{\psi,\ell} \left(\widetilde{\mathbf{F}}_{\ell}(\mathbf{y}_m) \right) \in \mathbb{R}^{N_{\ell}}, \qquad d_{m,\ell,n}(\mathbf{y}_m) = \left[\mathbf{d}_{m,\ell}(\mathbf{y}_m) \right]_n. \tag{10}$$

Finally, let $\mathcal{Q} = \{(m, \ell, n) \mid m \in \{1, \dots, M\}, \ell \in \mathcal{S}, n \in \{1, \dots, N_{\ell}\}\}$ collect all supervised temporal, feature-level, and spatial locations. The discriminator is trained with hinge loss:

$$\begin{aligned} \mathcal{L}_D = \mathbb{E}_{\mathbf{v} \sim p_{\text{data}}, \mathcal{I}} \left[\frac{1}{|\mathcal{Q}|} \sum_{(m,\ell,n) \in \mathcal{Q}} \left(1 - d_{m,\ell,n}(\mathbf{x}_m)\right)_+ \right] \\ + \mathbb{E}_{\hat{\mathbf{z}} \sim p_{\theta}^{\text{AR}}, \mathcal{I}} \left[\frac{1}{|\mathcal{Q}|} \sum_{(m,\ell,n) \in \mathcal{Q}} \left(1 + d_{m,\ell,n}(\hat{\mathbf{x}}_m)\right)_+ \right]. \end{aligned} \tag{11}$$

where $(u)_+ = \max(0, u)$, which penalizes real logits below 1 and generated logits above -1. Averaging the fake logits in this dense discriminator yields the generator objective in Eq. 6. The discriminator is unconditional and must distinguish real and generated samples solely from visual evidence, while text conditioning remains enforced by the conditional generator and DMD objective.

In practice, sparse decoding reduces the activation memory for the frozen VAE and VFM. Generated rollouts are decoded without gradients during discriminator updates, while generator updates retain gradients only through latent neighborhoods. Activation checkpointing and stochastic spatial augmentation reduce memory and discourage reliance on fixed local statistics. Since the VAE decoder, VFM, and discriminator heads are training-only, the supervision adds no inference cost.

3.4 Feature Space Exploration across Pretrained Backbones

Our default instantiation uses a frozen DINOv2-S/14 encoder as its adversarial feature backbone (Oquab et al., 2023). To study the effect of different adversarial feature space, we construct several comparison branches while keeping the causal student, on-policy DMD objective, autoregressive rollout, and inference architecture fixed. Their feature inputs can be summarized as

$$\mathcal{R}_{\omega,\ell} = \begin{cases} \Phi_{\omega,\ell}\left(\mathcal{P}_\omega(\hat{\mathbf{x}}_m)\right), & \text{image encoder,} \\ \mathcal{V}_{\omega,\ell}\left(\mathcal{C}_{T,s}(\hat{\mathbf{x}})\right), & \text{video encoder,} \\ \mathcal{H}_{\omega,\ell}\left(q_\tau(\hat{\mathbf{z}}; \boldsymbol{\epsilon}), \tau, \varnothing\right), & \text{DiT backbone,} \end{cases} \tag{12}$$

where ω identifies the frozen backbone, ℓ denotes an extracted feature level, $\mathcal{C}_{T,s}$ constructs a decoded clip of length T and temporal stride s, and q_τ denotes the forward corruption process at timestep τ. This common interface changes only the discriminator feature space while preserving the generator and distillation procedure as described in Sec. 3.3.

The image-level comparisons use VFM encoders DINOv2, DINOv3, and SigLIP2 (Oquab et al., 2023; Siméoni et al., 2025; Tschannen et al., 2025), with multi-level extraction and spatial discriminator heads. For video-native features, V-JEPA 2.1 and VideoMAE process decoded clips and project their tokens to discriminator width C_h for temporal discrimination (Mur-Labadia et al., 2026; Tong et al., 2022). For diffusion features, real and generated VAE latents are corrupted with a shared timestep τ and noise realization $\boldsymbol{\epsilon}$ before entering the frozen DiT teacher shared with the real-score backbone. Features from layer set $\mathcal{S}_{\mathrm{DiT}}$ are projected to width C_h and evaluated by local token-wise and pooled global heads. Sharing τ and $\boldsymbol{\epsilon}$ prevents the discriminator from exploiting corruption differences and instead focuses it on real–generated discrepancies. Unless otherwise specified, **Radian** denotes the DINOv2-S/14 configuration, with other encoders used in controlled ablations.

4 Experiments

4.1 Experimental Setup

We evaluate our method across three autoregressive T2V settings: four-step (4-NFE) chunk-wise short-video generation, its long-horizon extension to one-minute videos, and one-step (1-NFE) frame-wise short-video generation.

Baselines. For four-step chunk-wise generation, we compare our method with Self Forcing, Causal Forcing, and Salt (Huang et al., 2025a; Zhu et al., 2026; Ge et al., 2026). We additionally include a locally implemented DiT-GAN that performs adversarial distillation over internal diffusion features, enabling a controlled comparison between different-based and VFM-based adversarial feature spaces. For one-step frame-wise generation, we compare against Causal Forcing++ and One-Forcing (Zhao et al., 2026c; Feng et al., 2026a), where One-Forcing is the baseline for DiT-GAN under 1-NFE settings. For long-video generation, we compare against Rolling Forcing (Liu et al., 2026) and extend the generation results to one minute. For the baselines mentioned above, we re-evaluate the released checkpoints under our local environment.

Benchmarks. We evaluate four-step chunk-wise and one-step frame-wise short-video generation on VBench (Huang et al., 2023), reporting the normalized Total, Quality, and Semantic scores together with Dynamic Degree, Aesthetic Quality, and Imaging Quality. We additionally report VideoAlign

Table 1: Quantitative results under four-step chunk-wise (a), one-step frame-wise (b), and long-horizon autoregressive generation settings (c). #Params counts only the inference-time generator. Darker green indicates a higher rank in each column.

(a) Four-step chunk-wise autoregressive generation results.

			VBench ↑									VideoAlign ↑			
Method	#Params	NFE	**Total**	Quality	Semantic	Subj. Cons.	Dynamic	Aesthetic	Imaging	Object	Color	VQ	MQ	TA	**Total**
Self Forcing (Huang et al., 2025a)	1.3B	4	0.8393	0.8471	0.8082	0.9378	0.6972	0.6745	0.6995	0.9440	0.8680	0.0695	0.0616	0.3683	0.4995
Causal Forcing (Zhu et al., 2026)	1.3B	4	0.8400	0.8495	0.8019	0.9310	0.8694	0.6766	0.6993	0.9544	0.8311	0.0072	-0.0951	0.3542	0.2663
Salt (Ge et al., 2026)	1.3B	4	0.8434	0.8548	0.7982	0.9383	0.7926	0.6797	0.7034	0.9494	0.8291	0.0472	-0.0310	0.3774	0.3936
DiT-GAN	1.3B	4	0.8411	0.8483	0.8122	0.9492	0.5954	0.6832	0.6984	0.9563	0.8713	-0.0242	0.0855	0.3715	0.4327
Radian (CD init)	1.3B	4	0.8439	0.8550	0.7993	0.9380	0.7972	0.6792	0.7037	0.9508	0.8300	0.0521	-0.0296	0.3754	0.3979
Radian	1.3B	4	0.8444	0.8541	0.8054	0.9504	0.8056	0.6801	0.7148	0.9570	0.8722	0.2134	0.2094	0.3805	0.8033

(b) One-step frame-wise autoregressive generation results. † denotes First-Frame Enhancement, where the first latent frame uses four denoising steps and subsequent frames use one.

			VBench ↑							VideoAlign ↑	
Method	#Params	NFE	Total	Quality	Semantic	Object	App. Style	Temp. Style	Overall Cons.	TA	Total
Causal Forcing++ (Zhao et al., 2026c)	1.3B	$1^{\dagger}$	0.8383	0.8473	0.8025	0.9473	0.7248	0.6908	0.7255	0.3115	-0.1813
One Forcing (Feng et al., 2026a)	1.3B	$1^{\dagger}$	0.8415	0.8533	0.7943	0.9568	0.7433	0.6907	0.7178	0.3168	0.0448
Radian	1.3B	$1^{\dagger}$	0.8417	0.8511	0.8040	0.9663	0.7546	0.7034	0.7267	0.4105	0.2563

(c) Long-horizon generation results on one-minute videos.

			VBench-Long ↑							
Method	#Params	NFE	Total	Quality	Semantic	Subject Consist.	Temp. Flicker	Motion Smooth.	Dynamic Degree	Imaging Quality
Rolling Forcing (Liu et al., 2026)	1.3B	5	0.7805	0.8191	**0.6260**	0.9753	0.9871	0.9842	0.4532	0.7075
Radian	1.3B	4	**0.8041**	**0.8511**	0.6160	**0.9790**	**0.9882**	**0.9864**	**0.6741**	**0.7194**

(Liu et al., 2025) scores for visual quality (VQ), motion quality (MQ), text alignment (TA), and their aggregate score (Total). We evaluate the 60-second long-horizon video generation using VBench-Long (Huang et al., 2024). For each generation setting, we evaluate all methods using matched prompts, resolutions, sampling settings, and scoring protocols, detailed in Appendix A.

4.2 Main Results

4-step Chunk-wise 5-Second Video Generation. Tab. 1 (a) reports VBench and VideoAlign results for four-step chunk-wise generation. We apply **Radian** to two ODE-initialized checkpoints: Self Forcing initialized from a bidirectional teacher and Causal Forcing from an autoregressive teacher (denoted as CD init). Under the same training recipe (8 B300 GPUs, 200 iterations for Stage II, and 600 iterations for Stage III), our method achieves the best performance with a *VBench Total* of 0.8444 and a *VideoAlign Total* of 0.8033. It consistently outperforms Self Forcing, Causal Forcing and Salt across the principal metrics under the same four-step sampling budget. As further illustrated in Fig. 4 and Fig. 6, it produces sharper details, more coherent structures, better prompt alignment, and higher-quality motion while suppressing undesirable modes observed in competing methods.

1-step Frame-wise 5-Second Video Generation. Tab. 1 (b) reports VBench and VideoAlign results under the severely constrained one-step-per-frame setting ($1^{\dagger}$ NFE). **Radian** achieves a *VBench Total* of 0.8417, a *Temporal Style* score of 0.7034, and a *VideoAlign Total* of 0.2563, versus 0.0448 for One Forcing, while also outperforming Causal Forcing++ in overall performance and temporal stability. Although One Forcing, a DiT-GAN baseline using internal DiT features, attains competitive aggregate VBench scores, its clips exhibit rapid flickering, unwanted zoom-ins, and structural deformations that inflate metrics despite degrading visual quality, as shown in Fig. 4 and Fig. 7. In contrast, our method suppresses these artifacts and produces sharper frames, more stable structures, and coherent object and camera motion, indicating that decoded-output supervision in pretrained visual features provides a more reliable adversarial signal at $1^{\dagger}$ NFE.

60-Second Long-Video Generation. Tab. 1 (c) reports VBench-Long results for 957-frame videos of approximately one minute. **Radian** also establishes a new state of the art, achieving a *VBench-Long Total* of 0.8041, a *VBench-Long Quality Score* of 0.8511, a *Dynamic Degree* of 0.6741, and an *Imaging Quality* score of 0.7194. It significantly outperforms Rolling Forcing while using fewer denoising steps. As shown in Fig. 5 and Fig. 8, our method can better preserve visual details and coherent dynamics over long autoregressive rollouts of 60 seconds, whereas the baseline exhibits accumulated degradation or conservative motion. By adversarially suppressing low-quality feature modes at each autoregressive step, it reduces the local errors propagated through subsequent

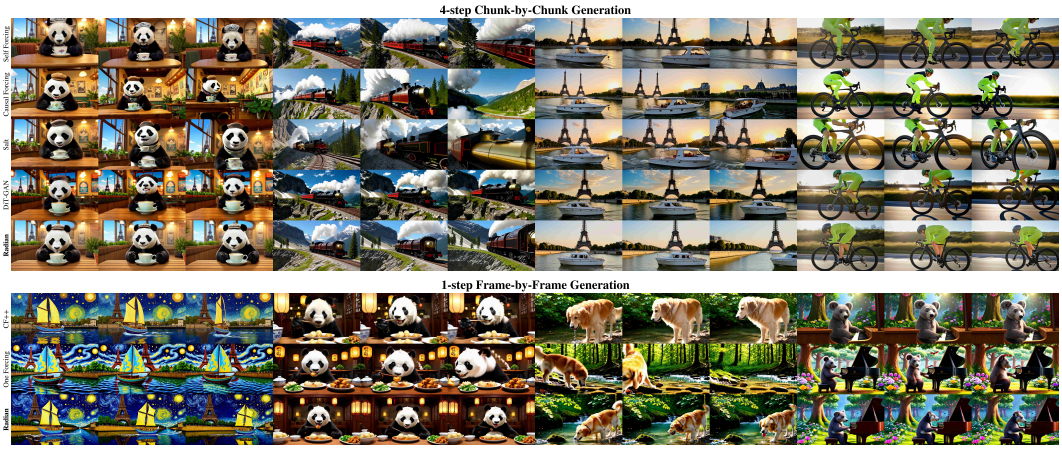


Figure 4: Selected short-video generation results using four-step chunk-wise generation and one-step frame-wise generation. More qualitative results and full prompts are provided in Appendix B.

Table 2: Ablation on different feature models for adversarial feature supervision.

Feature source	Feature model	Total ↑	Quality ↑	Semantic ↑	Multi-Obj. ↑	Color ↑	Consist. ↑	Dynamic ↑	Imaging ↑	VQ ↑	MQ ↑	TA ↑	VA Total ↑
Image VFM	DINOv2	0.8444	0.8541	0.8054	0.8343	0.8722	0.2662	0.8056	0.7148	0.2134	0.2094	0.3805	0.8033
Image VFM	DINOv3	0.8401	0.8498	0.8012	0.8528	0.8729	0.2649	0.7685	0.7104	0.1462	0.2937	0.3607	0.8006
Image VFM	SigLIP2	0.8391	0.8475	0.8052	0.8384	0.8662	0.2680	0.6343	0.6992	0.1156	0.1320	0.3751	0.6226
Video VFM	V-JEPA 2.1	0.8360	0.8430	0.8078	0.8604	0.8627	0.2676	0.5991	0.7042	0.0498	0.1275	0.4119	0.5892
Video VFM	VideoMAE	0.8372	0.8436	0.8115	0.8417	0.8787	0.2676	0.5213	0.7026	0.1000	0.1452	0.3436	0.5888
Image + Video VFM	DINOv2 + VideoMAE	0.8434	0.8525	0.8070	0.8652	0.8545	0.2679	0.6593	0.7074	0.0831	0.0575	0.3079	0.4485
Generative	DiT-GAN	0.8411	0.8483	0.8122	0.8432	0.8713	0.2674	0.5954	0.6984	-0.0242	0.0855	0.3715	0.4327

Table 3: Ablation on DINOv2 encoder size.

Size	VBench Total ↑	Quality ↑	Semantic ↑	VA Total ↑
S	0.8444	0.8541	0.8054	0.8033
B	0.8322	0.8404	0.7993	0.8910
L	0.8301	0.8366	0.8042	0.3686
G	0.8387	0.8475	0.8035	0.3257

Table 4: Ablation on DINOv2 layer number.

# Layers	VBench Total ↑	Quality ↑	Semantic ↑	VA Total ↑
1	0.8182	0.8206	0.8085	0.6747
2	0.8283	0.8389	0.7859	0.6586
4	0.8444	0.8541	0.8054	0.8033
8	0.8381	0.8469	0.8026	0.7228

conditioning contexts, thereby limiting long-horizon error accumulation and improving temporal continuity. More qualitative results and full prompts are provided in Appendix B.

4.3 Adversarial Distillation in Different Representation Spaces

Image Representation Feature Spaces. As shown in Tab. 2, the three image VFMs induce adversarial gradients with different magnitudes and directions, leading to different optimization strengths under the same objective. DINOv2-S provides a moderate and stable signal, achieving the best overall balance with a VBench *Total* of 0.8444 and a VideoAlign *Total* of 0.8033. DINOv3-S produces larger gradients and achieves the highest MQ score of 0.2937, although its aggressive updates are harder to stabilize. SigLIP2 instead provides a more semantics-oriented signal, achieving the highest *Overall Consistency* score of 0.2680 while being less sensitive to fine-grained rendering and motion artifacts. These differences arise from each encoder's feature scale, token geometry, and input Jacobian; thus, identical discriminator learning rates and adversarial weights do not imply equal supervision strength and require separate calibration.

Temporally Structured Video Features. Video VFMs encode cross-frame dependencies and tend to regularize uncertain temporal changes, improving consistency partly by reducing motion amplitude. As shown in Tab. 2, V-JEPA 2.1 achieves the highest TA score of 0.4119 while its Dynamic score drops to 0.5991, reflecting its preference for predictable and temporally persistent content. VideoMAE preserves richer local semantic and appearance cues, reaching a Semantic score of 0.8115 and a Color score of 0.8787, but at the cost of an even lower Dynamic score of 0.5213. Combining DINOv2 with VideoMAE provides a better trade-off between artifact-sensitive spatial supervision and complementary temporal regularization, raising the VBench *Total* to 0.8434, Dynamic to 0.6593, and Multi-Object to 0.8652. These results suggest that video representations are most effective as complementary temporal priors rather than replacements for DINOv2 supervision.



Figure 5: Selected long-video generation results using four-step chunk-wise generation.

Complementarity with DMD Gradients. DiT-GAN exhibits different optimization behavior from VFM-based adversarial supervision. As shown in Tab. 2, it achieves the strongest Semantic score of 0.8122, plausibly because diffusion-teacher features align with latent denoising and retain task-relevant generative semantics. Nevertheless, its VideoAlign profile remains weak: VQ, MQ, and TA are -0.0242, 0.0855, and 0.3715, yielding the worst VideoAlign *Total* of 0.4327. Our gradient-direction analysis suggests that DiT-GAN gradients are strongly coupled with DMD, reinforcing directions already present in the distillation objective. In contrast, gradients from DINO-, SigLIP-, V-JEPA-, and VideoMAE-based discriminators are approximately orthogonal to DMD. This suggests that representation-space adversarial supervision introduces complementary perceptual, semantic, or temporal directions rather than merely strengthening distillation. More qualitative comparisons are provided in Appendix B.1.

4.4 Ablation Study

Ablation on Encoder Size. We examine whether scaling the frozen DINOv2 encoder improves adversarial distillation. As shown in Tab. 3, larger encoders do not yield monotonic gains, while DINOv2-S/14 provides the best overall balance in VBench and VideoAlign. DINOv2-G/14 reaches 0.8387 and 0.3257, respectively, whereas DINOv2-B/14 achieves a higher VideoAlign *Total* of 0.8910 but a lower VBench *Total* of 0.8322. This is consistent with ADD, where DINOv2 ViT-S also outperforms ViT-L as the discriminator backbone (Sauer et al., 2024b). Larger ViTs may discard local appearance cues through increasingly invariant semantic features, while higher feature dimensionality and discriminator capacity can yield noisier gradients. Thus, a stronger encoder does not necessarily provide a better adversarial feature space.

Ablation on Feature Layer Selection. We further ablate the number and positions of DINOv2 layers used by the discriminator. In Tab. 4, the one-, two-, four-, and eight-layer settings use blocks $\{11\}$, $\{6, 11\}$, $\{6, 8, 10, 11\}$, and $\{2, 4, 6, 7, 8, 9, 10, 11\}$, respectively. Specifically, VBench *Total* increases from 0.8182 to 0.8283, peaks at 0.8444 with four layers, and drops to 0.8381 with eight, indicating diminishing returns. Too few layers provide insufficient multi-level supervision, while too many introduce redundant low-level cues. In particular, the final block overemphasizes invariant semantics and weakens local details, whereas early layers can overemphasize textures. The four layers therefore balance structural, semantic, and local appearance evidence without a high complexity. Visual results for the ablations are provided in Appendix B.1.

5 Conclusion

This paper introduced **Radian**, a representation adversarial distillation framework for few-step autoregressive video generation. While DMD transfers the diffusion teacher's generative prior in latent space, Radian complements it with real-data adversarial supervision over multi-level features from a frozen visual foundation model, directly improving decoded video quality without additional inference cost. Experiments across short, one-step, or long-horizon generation demonstrate improvements in visual fidelity, motion quality, and autoregressive stability. Our analysis further shows that different representation spaces induce distinct optimization signals: image VFMs differ in gradient strength and emphasize different perceptual or semantic cues, while video-native representations provide temporal regularization, often with reduced motion diversity. Moreover, external VFM gradients are largely complementary to DMD, whereas diffusion-internal adversarial features are more strongly coupled with the original distillation objective. These findings highlight representation-space supervision as an effective way to refine few-step autoregressive video generation quality.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript and to assist with literature retrieval and discovery. For writing, LLMs were used to refine language, improve readability and clarity, rephrase sentences, and check grammar. For literature discovery, LLMs were used to identify potentially relevant prior work and assist in locating related references. All references were subsequently reviewed and verified by the authors.

The LLMs were not involved in research ideation, methodology development, experimental design, implementation of the proposed method, data analysis, or interpretation of experimental results. All research concepts, technical contributions, experiments, analyses, and scientific conclusions were developed and conducted by the authors. The authors reviewed all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the results presented in this work. The paper provides detailed descriptions of the proposed Radian framework and its three-stage training procedure, including the representation-space adversarial objective, feature extraction and discriminator design, training schedules, model configurations, and evaluation settings.

We report the configurations used for four-step chunk-wise, one-step frame-wise, and long-horizon autoregressive generation, together with controlled ablation settings. All compared methods are evaluated using matched prompts, resolutions, sampling settings, and scoring protocols. The benchmarks used in this work, including VBench (Huang et al., 2023), VideoAlign (Liu et al., 2025), and VBench-Long (Huang et al., 2024), are publicly available.

The core code and configurations will be publicly released after the completion of the paper review process to further facilitate replication and future research.

ETHICS STATEMENT

This work studies few-step autoregressive text-to-video generation. Radian builds on the publicly released Wan2.1 (Wan et al., 2025) foundation model and uses frozen pretrained visual encoders for representation-space adversarial supervision. The resulting models may inherit biases from these pretrained components and from the video data used during post-training. More broadly, efficient and realistic video generation lowers the cost of producing synthetic media and may facilitate impersonation, misinformation, or the creation of misleading content. We therefore encourage clear disclosure of AI-generated videos, provenance tracking, and safeguards against non-consensual or deceptive use. Any models and code released from this work are intended to support research and responsible development of video generation systems.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3237–3245, 2026.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-Forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026a.

Lan Feng, Wuyang Li, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Representation distribution matching for one-step visual generation. *arXiv preprint arXiv:2607.02375*, 2026b.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling distribution matching for flow-based text-to-image distillation. *arXiv preprint arXiv:2506.00523*, 2025.

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. *arXiv preprint arXiv:2604.03118*, 2026.

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.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2025a.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live Avatar: Streaming real-time audio-driven avatar generation with infinite length, 2025b. URL https://arxiv.org/abs/2512.04677.

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, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models, 2024. URL https://arxiv.org/abs/2411.13503.

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, 2025. URL https://arxiv.org/abs/2412.03603.

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. *arXiv preprint arXiv:2606.03972*, 2026.

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. *arXiv preprint arXiv:2501.08316*, 2025a.

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. *Advances in Neural Information Processing Systems*, 38:41061–41086, 2025b.

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

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yang Luo, Shengju Qian, Xiaohang Tang, Zirui Zhu, Yong Liu, Xin Wang, and Yang You. On-policy adversarial flow distillation for autoregressive video generation. *arXiv preprint arXiv:2605.26105*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.

Team Seedance and et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder, 2026. URL https://arxiv.org/abs/2510.15301.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

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 model. *arXiv preprint*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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

Tianhang Wang, Yitong Chen, Wei Song, Zuxuan Wu, Min Li, and Jiaqi Wang. Decq: Detail-condensing queries for enhanced reconstruction and generation in representation autoencoders, 2026. URL https://arxiv.org/abs/2605.22777.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

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*, volume 2026, pp. 34858–34876, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling, 2026a. URL https://arxiv.org/abs/2607.03524.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, Fan Bao, and Jun Zhu. minWM: A full-stack open-source framework for real-time interactive video world models, 2026b. URL https://arxiv.org/abs/2605.30263.

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*, 2026c.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026a.

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*, 2026b.

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 *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026c.

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. *arXiv preprint arXiv:2602.02214*, 2026.

Appendix

Contents

Main text

1 Introduction **1**
2 Related Work **2**
3 Method **3**
3.1 Preliminaries . . . 3
3.2 Representation Adversarial Distillation for Causal Video Generation . . . 4
3.3 Multi-Level Feature Discrimination over Decoded Outputs . . . 5
3.4 Feature Space Exploration across Pretrained Backbones . . . 6
4 Experiments **6**
4.1 Experimental Setup . . . 6
4.2 Main Results . . . 7
4.3 Adversarial Distillation in Different Representation Spaces . . . 8
4.4 Ablation Study . . . 9
5 Conclusion **9**

Appendix

A Detailed Implementation and Evaluation **15**
A.1 Notation . . . 15
A.2 Evaluation Protocol . . . 15
A.3 Metric Definitions . . . 16
A.4 Models and Sampling . . . 17
A.5 Training Configuration . . . 17
A.6 Representation Configurations and Ablations . . . 19
B Additional Visualizations and Prompt Details **21**
B.1 Additional Qualitative Comparisons . . . 21
B.2 Prompts for Main-Text Visualizations . . . 22

A Detailed Implementation and Evaluation

This appendix describes notation, evaluation, sampling, training, and the image- and video-representation configurations used in Radian.

A.1 Notation

Table 5 collects the principal symbols. Diffusion time t is distinct from optimization iteration u; z denotes generator noise, whereas $\hat{z}$ denotes its generated clean latent sequence. Transformer block indices are zero-based.

A.2 Evaluation Protocol

Shared prompts and seeds. All quantitative experiments use the same randomly-chosen 15 inference seeds as the four-step five-second evaluation: 2, 4, 1805718, 3367341, 3718700, 5036436, 5279402, 5450779, 5748298, 7252721, 9274039, 9539461, 5918269, 7730833, and 2436109. This shared set applies to every compared method in the four-step chunk-wise, one-step frame-wise, and one-minute experiments, as well as all feature-model, encoder-size (including DINOv2-L), and feature-depth ablations. These are inference seeds for fixed model weights, rather than independent training runs. The set was randomly selected from a larger seed pool for further benchmarking. We additionally conducted training runs with different random seeds and observed consistent trends, indicating that the reported improvements are not specific to a particular training seed. In separate internal experiments, we also further applied Radian to a proprietary video generation model and observed consistent improvements, suggesting that the proposed approach extends beyond the Wan2.1-1.3B setting.

Table 5: Principal notation.

Symbol	Meaning
c, z	Text condition and initial Gaussian generator input.
G_θ, p_θ^{AR}	Causal student and its autoregressive output distribution.
$\hat{z}^{(b)}$, $\hat{z}^{(<b)}$	Generated latent block and its preceding context.
x_0, x_t	Clean and diffusion-corrupted video latents.
s_r, s_ϕ	Frozen teacher score and trainable fake-score model.
$\mathcal{D}_{\mathrm{VAE}}$	Frozen latent-to-RGB decoder.
x, $\hat{x}$	Real and generated RGB inputs to representation supervision.
Φ, $\mathcal{P}$, $\mathcal{R}_\Phi$	Frozen visual encoder, input preprocessing, and selected multi-level representation.
$\mathcal{S}$, F_ℓ	Selected transformer blocks and level-ℓ feature tokens.
H_ψ, $h_{\psi,\ell}$	Trainable representation discriminator and its level-specific head.
T_z, T_x, M	Latent length, decoded RGB length, and sampled image-frame count.
$\mathcal{L}_{\mathrm{DMD}}$, $\mathcal{L}_{\mathrm{adv}}^{\Phi}$, $\mathcal{L}_D$	Distribution-matching, generator adversarial, and discriminator losses.
λ_{DMD}, λ_{adv}	Generator objective weights.
$\mathrm{sg}(\cdot)$, J_f	Stop-gradient operation and Jacobian of mapping f.

Short-video evaluation uses 946 prompts at 832×480 resolution and 16 FPS. Generation uses the shared extended prompts, while scoring metadata retains their original benchmark descriptions. Within each evaluation setting, methods use the same prompts and seed set. We compute each metric with its benchmark evaluator for each seed and report the arithmetic mean over the 15 seeds. Qualitative examples are illustrative selections, not additional benchmark averages.

One-minute evaluation. Long-video comparisons use the same 946 prompts and 15 inference seeds, generating 957 RGB frames at 832×480 and 16 FPS. VBench-Long evaluates the complete generated videos.

A.3 Metric Definitions

VBench. VBench (Huang et al., 2023) comprises seven quality and nine semantic dimensions. We apply its fixed reference bounds and dimension weights to obtain Quality Q and Semantic S, and compute

$$\text{VBench Total} = \frac{4Q + S}{5}. \tag{13}$$

All 16 dimensions contribute to these aggregates, regardless of which columns appear in a displayed table. The quality group comprises subject consistency, background consistency, temporal flickering, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. The semantic group comprises object class, multiple objects, human action, color, spatial relationship, scene, appearance style, temporal style, and overall consistency. Within each group, we use the benchmark's weighted average of normalized dimension scores, not an unweighted average of the displayed columns. Dimension normalization uses $\widetilde{r}_d = \mathrm{clip}_{[0,1]}((r_d - a_d)/(b_d - a_d))$, where a_d, b_d are the benchmark reference bounds. The feature-model ablation reports native style and overall-consistency scores, whereas the one-step main table uses their benchmark-normalized counterparts; these scales should not be compared directly.

VideoAlign. VideoAlign (Liu et al., 2025) measures visual quality (VQ), motion quality (MQ), and text alignment (TA). Our evaluator applies the reward checkpoint's fixed standardization:

$$A_k = \frac{r_k - \mu_k}{\sigma_k}, \qquad \text{VA Total} = A_{\mathrm{VQ}} + A_{\mathrm{MQ}} + A_{\mathrm{TA}}. \tag{14}$$

In VQ/MQ/TA order, the means are $(3.6757, 1.1646, 2.8105)$ and standard deviations are $(2.2476, 1.3811, 2.5121)$. These constants do not depend on the methods being compared. Scores are not probabilities or bounded by $[0, 1]$: negative components and totals above one are valid. The evaluator samples at a nominal 2 FPS with at most 200,704 pixels per sampled frame.

Long-video metrics. VBench-Long (Huang et al., 2024) uses its long-video preprocessing and dimension-specific routines, including the configured static-content filter for temporal flickering.

Table 6: Sampling settings at 832×480, 16 FPS. Schedules are indices before scheduler warping.

Setting	Four-step short	One-step short	One-minute
Latent / RGB frames	21 / 81	21 / 81	240 / 957
Latent frames per block	3	1	3
Denoising schedule	$\{1000, 750, 500, 250\}$	$\{1000\}$	$\{1000, 750, 500, 250\}$
First-block exception	None	Four-step enhancement	None
Student inference CFG	Disabled	Disabled	Disabled
Evaluation seeds	15	15	15

It does not score only the first five seconds. Dynamic Degree quantifies detected movement, not necessarily plausible motion; consistency can also increase when movement decreases. We therefore interpret these measures alongside imaging quality and qualitative videos, and report short- and long-video aggregates separately.

A.4 Models and Sampling

Networks and latent geometry. The student is a causal Wan2.1-T2V-1.3B model (Wan et al., 2025). A frozen bidirectional Wan2.1-T2V-14B supplies teacher predictions, and a trainable 1.3B fake-score model estimates the student distribution. The text encoder, VAE, and visual backbone are frozen. Only the student, fake-score network, and discriminator heads/projections are optimized. At inference, the teacher, fake-score network, visual backbone, and discriminator are removed.

The causal VAE maps T_z latent frames to $T_x = 1 + 4(T_z - 1)$ RGB frames: 21 latents yield 81 frames, and 240 yield 957. The latent grid has 16 channels and spatial size 104×60. The decoder uses a temporal compression factor of four with causal context and a special first-frame boundary.

Denoising and guidance. Four-step sampling uses scheduler shift 5.0. Teacher guidance is applied during distillation, while student inference uses only the conditional branch. The implementation's guidance setting 3.0 means $\widehat{x}_r^{\mathrm{cfg}} = \widehat{x}_r(c) + 3(\widehat{x}_r(c) - \widehat{x}_r(\varnothing))$; its conditional coefficient is therefore four. Fake-score predictions are conditional and use no additional guidance term.

The frame-wise model starts from the causal ODE initialization used by Causal Forcing++ (Zhao et al., 2026c). First-Frame Enhancement uses four steps for the first latent and one for each subsequent latent. Accordingly, $1^{\dagger}$ denotes the steady-state per-frame denoising budget; a 21-latent clip uses 24 denoising evaluations in total before cache-related operations.

Rolling generation. The long-video experiment places the four-step Radian model into the rolling inference implementation of Rolling Forcing (Liu et al., 2026), without additional long-video training. A staggered window of four three-latent blocks advances through the sequence while detached key/value caches retain clean context. The full 240-latent sequence is decoded into RGB. We report NFE based on denoising evaluations, with cache-writing passes and VAE decoding treated separately.

A.5 Training Configuration

Training data. All experiments use the same 40K captioned-video training collection. Real clips are uniformly sampled to 81 frames, resized to 832×480, and represented in $[-1, 1]$ for VAE processing. Representation inputs are mapped to $[0, 1]$ before encoder-specific normalization. Real and generated videos are sampled independently, with image sampling aligned by normalized temporal positions. The DMD and GAN objectives use separate on-policy rollouts.

DMD-only continuation as reference. Self Forcing (Huang et al., 2025a) and Causal Forcing Zhu et al. (2026) serve as the DMD-based reference models for the two initialization routes. They do not use representation-space adversarial supervision, whereas Radian augments the corresponding causal generator with discriminator calibration followed by joint DMD and representation-adversarial refinement. We therefore use these baselines to assess the effect of introducing representation-space adversarial supervision.

Table 7: Default chunk-wise optimization. Learning rates refer to trainable networks; the visual encoder remains frozen.

Setting	Generator	Fake-score	Discriminator
Optimizer	AdamW	AdamW	AdamW
(β_1, β_2)	$(0, 0.999)$	$(0, 0.999)$	$(0, 0.95)$
Weight decay	0.01	0.01	0
Stage II learning rate	2×10^{-6}	4×10^{-7}	3×10^{-5}
Stage III learning rate	5×10^{-7}	4×10^{-7}	3×10^{-6}
Gradient-norm clipping	10	10	10
EMA decay	0.99	Not used	Not used

Stage I: causal ODE initialization. Stage I uses a causal ODE initialization of the pretrained foundation model. The main chunk-wise route uses the Self Forcing initialization (Huang et al., 2025a); the CD-initialized variant loads `causal_cd.pt`. These provide alternative starting weights for subsequent Radian post-training.

The inherited ODE initialization regresses clean trajectory targets from intermediate noisy states under causal context:

$$\mathcal{L}_{\mathrm{ODE}} = \mathbb{E}\left\| G_\theta(z_\tau^{\mathrm{ODE}}, \tau, c; z_0^{\mathrm{ODE},<b}) - z_0^{\mathrm{ODE},b} \right\|_2^2. \tag{15}$$

Here τ is a selected trajectory time and b is the predicted block. This objective is used during initialization. The subsequent on-policy stage conditions on the student's generated context.

Stage II: on-policy adaptation and discriminator calibration. Stage II runs for 200 outer iterations. DMD and discriminator training are active throughout. Generator-side adversarial supervision remains disabled throughout Stage II, while the discriminator is calibrated using real videos and detached student rollouts. This stage adapts the initialized student to its generated context while calibrating the discriminator for subsequent joint refinement.

Stage III: low-learning-rate joint refinement. Stage III loads the resulting generator weights and continues training with lower learning rates, using

$$\mathcal{L}_G(u) = \lambda_{\mathrm{DMD}}\mathcal{L}_{\mathrm{DMD}} + g_u\lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}}^{\Phi}, \qquad \lambda_{\mathrm{DMD}} = \lambda_{\mathrm{adv}} = 1. \tag{16}$$

Here g_u denotes the scheduled adversarial activation and equals one throughout Stage III. The default continuation uses the local-step-700 EMA model. Stage III initializes from the Stage-II generator and calibrated discriminator, and jointly optimizes DMD and representation-adversarial objectives. Table 7 summarizes the default optimization settings.

On-policy DMD. The student generates autoregressive rollouts conditioned on its own history. The frozen teacher and trainable fake-score model process the same diffusion-corrupted student latents, and their prediction difference provides the DMD update direction. The update direction is detached when optimizing the student, while the fake-score model is trained separately with a flow-denoising objective on detached student samples. The DMD and representation-adversarial objectives use separate on-policy rollouts sampled from the same student.

Batching and hardware. Training follows a five-iteration update cycle: the generator is updated once every five outer iterations, while fake-score and discriminator updates occupy the intervening iterations. The default chunk-wise run uses eight workers, one video per worker, and eight DMD accumulation microbatches, resulting in 64 videos per DMD update. The GAN branch uses one video per worker and eight sampled frames per video, resulting in eight videos per update. Losses are mean-reduced within each branch.

All final experiments run on NVIDIA B300 GPUs. The default configuration uses eight GPUs, bfloat16 mixed precision, hybrid FSDP, and activation checkpointing. Checkpointed frozen-encoder and VAE forwards retain input gradients during generator updates.

Frame-wise post-training. The one-step recipe uses 500 DMD-only iterations followed by 100 iterations of discriminator calibration before joint training. At joint activation, the generator learning

rate changes from 2×10^{-6} to 5×10^{-7}; the discriminator and fake-score learning rates are 3×10^{-6} and 4×10^{-7}, respectively. One DMD microbatch per worker gives an effective batch size of eight videos. The DMD stream uses text prompts, while the adversarial stream uses captions associated with real videos.

A.6 Representation Configurations and Ablations

Image sampling and preprocessing. The image branch samples eight of the 21 latent positions uniformly without replacement. Position i aligns to real RGB index $\mathrm{round}(i(T_r - 1)/(T_z - 1))$, or $4i$ for 81 RGB frames. Here T_r is the real clip's RGB length. Selected one-latent decoding chunks provide a sparse approximation of full-sequence decoding during training. DINOv2-S/14 (Oquab et al., 2023) uses area resizing to 518×518 and ImageNet normalization, producing a 37×37 patch grid. With probability 0.5, a square crop of side 64, 128, or 192 pixels is applied before resizing. Real and generated inputs share the same augmentation distribution.

ImageNet channel means are $(0.485, 0.456, 0.406)$ and standard deviations are $(0.229, 0.224, 0.225)$. The crop operates in RGB coordinates before encoder resizing. Real and generated videos are sampled independently, while their frames are aligned by normalized temporal position. The discriminator therefore captures distribution-level differences across real and generated samples. Cached generated samples are detached before being reused for discriminator training.

Dense discriminator. The default discriminator uses transformer blocks $\{6, 8, 10, 11\}$ together with the patch-input feature after positional embedding. At each level, the CLS readout is broadcast-added to the patch tokens. Each level has an independent head composed of spectrally normalized Conv1d layers, local normalization, LeakyReLU with slope 0.2, and a residual kernel-9 block with circular padding and scaling $1/\sqrt{2}$. A final kernel-1 layer predicts dense logits. Convolutions operate over flattened token sequences. Local normalization uses virtual groups of eight, and channel projections are trainable where required.

For real and generated logits d^{real} and d^{fake},

$$\mathcal{L}_D = \mathbb{E}\langle (1 - d^{\mathrm{real}})_+ \rangle + \mathbb{E}\langle (1 + d^{\mathrm{fake}})_+ \rangle, \tag{17}$$

$$\mathcal{L}^{\Phi}_{\mathrm{adv}} = -\mathbb{E}\langle d^{\mathrm{fake}} \rangle. \tag{18}$$

Here $(a)_+ = \max(a, 0)$ and brackets denote averaging over concatenated logits. Levels receive equal weight when their token counts match; otherwise, the reduction is token-weighted. The discriminator is unconditional.

Gradient through a frozen representation. Write $x_\theta = \mathcal{D}_{\mathrm{VAE}}(G_\theta(z, c))$ and $\mathcal{R}_\Phi = \Phi_{\mathcal{S}} \circ \mathcal{P}$. For a fixed discriminator and sampled preprocessing transform, let h_ψ denote the mean dense logit. The chain rule gives

$$\nabla_\theta \mathcal{L}^{\Phi}_{\mathrm{adv}} = -\mathbb{E}\left[J^{\mathsf{T}}_{G_\theta} J^{\mathsf{T}}_{\mathcal{D}_{\mathrm{VAE}}} J^{\mathsf{T}}_{\mathcal{P}} J^{\mathsf{T}}_{\Phi_{\mathcal{S}}} \nabla_F h_\psi(F) \right], \quad F = \mathcal{R}_\Phi(x_\theta). \tag{19}$$

Although the encoder parameters are frozen, gradients propagate through its input Jacobian. The adversarial update is therefore shaped by decoded appearance and the encoder's feature sensitivity, while the DMD update is determined by teacher and fake-score predictions. These two gradient pathways provide distinct sources of supervision to the generator. During discriminator updates, inputs are detached; during generator updates, discriminator heads are fixed while gradients remain active through the VAE, preprocessing, and frozen representation encoder.

When a head normalizes groups of inputs, F in Eq. 19 denotes the stacked group, and the corresponding Jacobian includes cross-input dependencies. For randomized crops, the gradient expression is evaluated for each sampled transform and averaged over the preprocessing distribution.

Complementary Gradients. In an auxiliary VideoMAE-based run, measurements over **127** generator updates yield a GAN–DMD gradient cosine similarity of $\mathbf{-0.0234 \pm 0.0534}$ (mean $\pm$ standard deviation), indicating approximately orthogonal directions on average in this diagnostic. The median adversarial-to-DMD gradient-norm ratio is $\mathbf{0.8381}$, showing that the adversarial signal has a comparable magnitude. These loss-weighted, pre-clipping measurements support directional complementarity in the examined run, although they do not establish universal orthogonality across representation backbones or, by themselves, demonstrate improved optimization.

Table 8: Representation designs with frozen backbones. P denotes the patch-input branch; channel arrows indicate trainable projections. Image inputs are written as frame count $\times$ spatial resolution. Video inputs use 32 frames, short side 256, and RGB stride two. The mixed variant combines DINOv2-S/14 with VideoMAE; its video head uses spatial pooling.

Backbone	Input	Feature levels	Channels	Readout / heads
DINOv2-S/14	8×518^2	$P + \{6, 8, 10, 11\}$	384	CLS + dense
DINOv3-S/16	8×512^2	$P + \{2, 5, 8, 11\}$	384	Norm. patch mean + dense
SigLIP2-B/16	8×512^2	$P + \{6, 8, 10, 11\}$	$768 \to 512$	Norm. patch mean + dense
V-JEPA 2.1-B (distilled)	Video clip	$P + \{6, 8, 10, 11\}$	$768 \to 512$	Spatiotemporal dense
VideoMAE	Video clip	$P + \{6, 8, 10, 11\}$	$768 \to 512$	Spatiotemporal dense
DINOv2 + VideoMAE	Images + clip	Image levels + video output	384; $768 \to 512$	Image dense + temporal
DiT-GAN (Wan-14B)	Noisy latents	$\{21, 28, 35, 39\}$	$5120 \to 384$	Frame-local + video-pooled

Image-encoder adaptations. DINOv3 (Siméoni et al., 2025) uses bilinear resizing and ImageNet normalization; SigLIP2 (Tschannen et al., 2025) uses bicubic resizing and channel mean/standard deviation 0.5. Both use normalized features and a final patch-mean readout in place of the default CLS readout. Selected intermediate features receive affine-free LayerNorm. SigLIP2's final output retains its pretrained post-normalization, and its 768-channel tokens are projected to 512 channels before discrimination. DINOv3 uses 384-channel heads without the default local normalization. The image feature heads preserve spatial tokens instead of reducing each frame to a single classification embedding. Table 8 summarizes the feature and head designs. Since input geometry and head adaptations vary across model families, these experiments evaluate the resulting representation-supervision configurations as complete design choices.

Video and mixed representations. V-JEPA 2.1 (Mur-Labadia et al., 2026) and VideoMAE (Tong et al., 2022) use 32-frame clips at RGB stride two, spanning 63 source positions. Resizing preserves aspect ratio with short side 256 and dimensions divisible by 16 (448×256 without cropping). A sampled spatial crop is shared across the clip, and both models use ImageNet normalization. Generated clips are decoded in overlapping four-latent pieces with overlap two.

Video-only variants use five projected spatiotemporal heads. The mixed DINOv2+VideoMAE configuration supplements the image heads with a final-layer video branch consisting of spatial pooling within each tubelet, a $768 \to 512$ projection, and temporal convolutions that produce scalar logits. Separate heads and optimizers preserve the distinct token organizations of the image and video branches.

The video-only discriminator retains dense spatiotemporal tokens, whereas the auxiliary temporal branch in the mixed configuration pools spatial locations before temporal discrimination. Image encoders process sampled frames independently, while video encoders capture cross-frame variation within the sampled clip. The two branches therefore provide complementary spatial and temporal feature structures.

Diffusion-internal representation. DiT-GAN replaces the external VFM with frozen teacher features. Real and generated latents receive a shared diffusion timestep and noise realization, with null text conditioning used for feature extraction. Projected features are passed to local and video-pooled heads. Generator updates differentiate through the frozen feature model, while discriminator updates use detached latents.

Multi-level sensitivity. A local expansion characterizes the sensitivity induced by multiple feature levels. Let $f_\ell(x)$ be a vectorized frozen feature and $a_\ell \geq 0$ fixed weights. At a differentiable input, a small perturbation δx gives

$$\sum_\ell a_\ell \| f_\ell(x + \delta x) - f_\ell(x) \|_2^2 = \delta x^\mathsf{T} M_\Phi(x) \delta x + o(\|\delta x\|_2^2), \tag{20}$$

$$M_\Phi(x) = \sum_\ell a_\ell J_{f_\ell}(x)^\mathsf{T} J_{f_\ell}(x) \succeq 0. \tag{21}$$

For positive weights, the null space of $M_\Phi(x)$ is the intersection of the null spaces of the selected feature Jacobians. Adding feature levels therefore expands the set of locally observable directions

Table 9: DINOv2-S feature-depth ablation. All variants retain a patch-input head and use the same 15 evaluation seeds.

Transformer layers	Block indices	Total heads
1	$\{11\}$	2
2	$\{6, 11\}$	3
4	$\{6, 8, 10, 11\}$	5
8	$\{2, 4, 6, 7, 8, 9, 10, 11\}$	9

whenever the corresponding sensitivities are complementary, while redundant levels contribute less additional information.

Specifically, $\delta x^\top M_\Phi \delta x = \sum_\ell a_\ell \|J_{f_\ell} \delta x\|_2^2$ vanishes exactly when every positively weighted feature has zero first-order response. Adding a feature level further intersects this set with the kernel of its Jacobian, reducing or preserving the set of locally invisible directions. The effectiveness of the resulting multi-level discriminator is then determined jointly by feature sensitivity, discriminator heads, optimization, and the generated sample distribution.

Encoder size and feature depth. The size ablation compares DINOv2-S/B/L/G, all evaluated on the same 15 seeds. The depth ablation keeps DINOv2-S fixed and varies the transformer blocks listed in Table 9. The layer count excludes the additional patch-input branch. These configurations expose different feature depths while preserving the generator inference architecture.

B Additional Visualizations and Prompt Details

B.1 Additional Qualitative Comparisons

We provide additional qualitative comparisons across generation regimes and discriminator configurations, complementing the quantitative results in the main text. The cases cover human activities, stylized scenes, animals, and object interactions, including several detailed prompts that specify both appearance and action. Each 5-second case shows three frames at 0, 2.5, and 5 seconds, while each long-video case shows six frames spanning approximately one minute. Prompts and displayed times are shared across methods within each case.

Short-video generation. Figures 6 and 7 compare Radian with Self Forcing (Huang et al., 2025a) and Causal Forcing (Zhu et al., 2026) under four-step chunk-wise generation, and with Causal Forcing++ (Zhao et al., 2026c) and One Forcing (Feng et al., 2026a) under one-step frame-wise generation. In the four-step examples, Radian maintains recognizable faces and keeps the telephone receiver or the pianist's hands and keyboard visible across the sampled frames. In the one-step examples, the Causal Forcing++ piano sample does not clearly show the requested keyboard interaction, while One Forcing exhibits larger changes in the apparent scale of the campfire and surrounding trees. Radian retains the piano-playing composition and the broader snowy scene. These observations complement the main-text results on the value of decoded-output supervision for preserving visual structure under few-step generation.

Long-video generation. Figure 8 compares Radian with Rolling Forcing (Liu et al., 2026) on four approximately 60-second videos. In the surfing and telephone examples, Rolling Forcing shows larger subject displacements and partial cropping at several sampled times. Radian keeps the panda and surfboard recognizable and retains a clearer view of the person holding the telephone throughout the displayed sequence. The cafe and European-town examples further show changes in viewpoint while preserving identifiable people, objects, and scene context. These examples are consistent with the main-text long-video results, illustrating how the benefits of representation-space supervision can persist beyond the short generation horizon.

Discriminator ablations. Figures 9–11 visualize the feature-space, encoder-size, and feature-layer configurations studied in our ablations. The feature-space examples reveal distinct appearance preferences: VideoMAE follows the black-and-white instruction more closely in the corgi scene, while DINOv2-S retains some color. The coffee, pastry, and lighthouse cases also show differences in object interaction, surface detail, and framing. In the encoder-size examples, the S configuration makes

the liquid color transition more apparent, whereas increasing encoder size does not consistently improve the depicted action. The four-layer configuration also combines recognizable structure with local detail in the corgi and astronaut examples. Together, these visualizations complement the quantitative finding that feature suitability matters more than encoder capacity alone, supporting our compact, multi-level DINOv2-S configuration.

B.2 Prompts for Main-Text Visualizations

We list the input prompts for the qualitative examples in Figures 4 and 5. Short-video prompts follow the left-to-right case order within each generation regime. Long-video prompts follow the upper-left, upper-right, lower-left, and lower-right case order within each method block. Prompt wording is reproduced from the generation records.

Four-step chunk-wise generation.

1. **Panda in a cafe.** *A panda drinking coffee in a cafe in Paris, tilt up.*
2. **Steam train.** *A steam train moving on a mountainside.*
3. **Boat on the Seine.** *A boat sailing leisurely along the Seine River with the Eiffel Tower in background, pan left.*
4. **Bicycle.** *a bicycle accelerating to gain speed.*

One-step frame-wise generation.

1. **Van Gogh-style riverboat.** *A boat sailing leisurely along the Seine River with the Eiffel Tower in background, Van Gogh style.*
2. **Panda dining.** *A cute fluffy panda eating Chinese food in a restaurant.*
3. **Dog drinking water.** *a dog drinking water.*
4. **Koala playing piano.** *A koala bear playing piano in the forest.*

Long-video generation.

1. **Oil-painted shark.** *a shark is swimming in the ocean, oil painting.*
2. **Couple in the rain.** *A couple in formal evening wear going home get caught in a heavy downpour with umbrellas, Van Gogh style.*
3. **Trumpet performance.** *A person is playing trumpet.*
4. **Cooking in a medieval castle.** *An adult person prepares vegetables in a wok, holding the handle firmly with one hand while moving a spatula through two compact circular strokes and one small controlled toss. Ingredients remain inside the pan; preserve both hands, wok, stove, vegetables, steam, and heat source. Set the scene inside a historically textured medieval castle with rough stone walls, carved oak furniture, iron fixtures, wool and linen garments, torchlight, banners, and cool daylight from arrow-slit windows. Use a stable period-film composition that keeps the complete action visible; no cuts, whip pans, or sudden costume transformations. Concentrate motion on the requested action and subtle environmental responses while keeping each person's identity, object topology, and scene layout consistent.*



Figure 6: **Additional four-step chunk-wise comparisons.** A retro-futuristic telephone conversation and piano performance, with Self Forcing, Causal Forcing, and Radian arranged from top to bottom. The full prompts appear below the corresponding cases.



Figure 7: **Additional one-step frame-wise comparisons.** A snowy campfire and piano performance, with Causal Forcing++, One Forcing, and Radian arranged from top to bottom. The sampled frames show changes in scene appearance and subject framing.

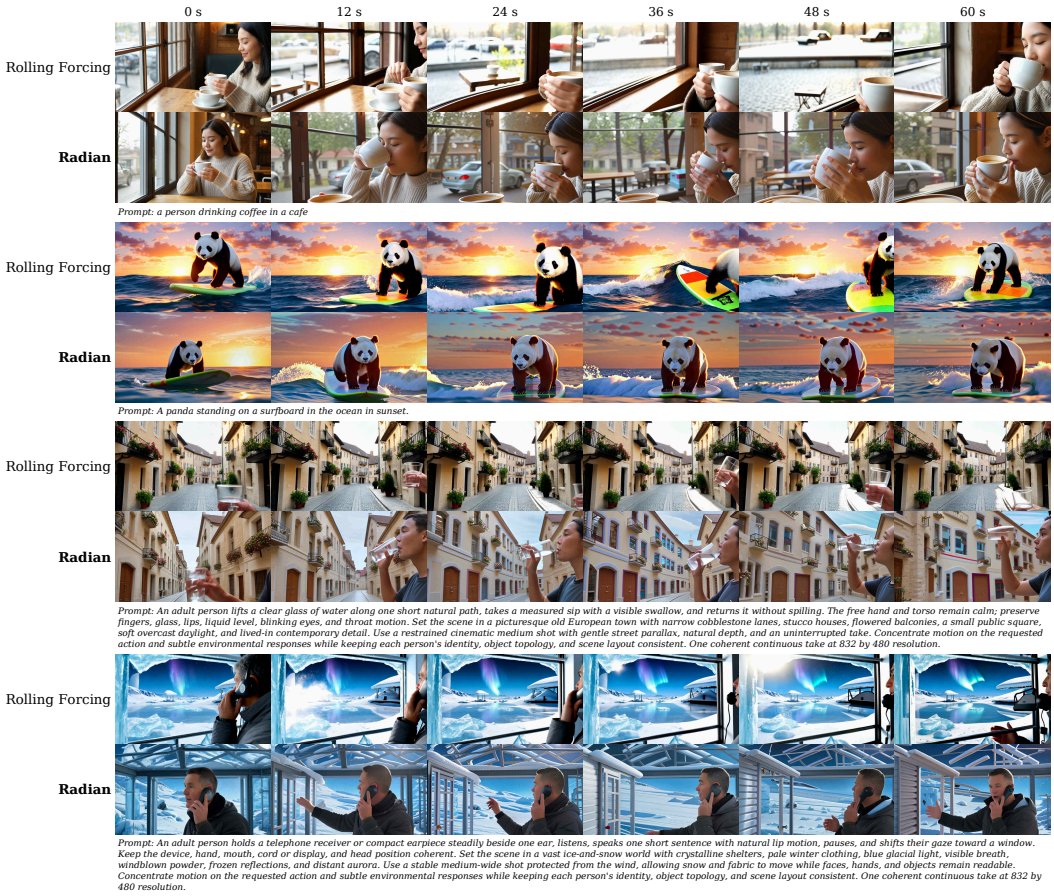


Figure 8: **Additional long-video comparisons.** From top to bottom: coffee in a cafe, panda surfing, drinking water in a European town, and a telephone conversation in the snow. Each case places Rolling Forcing above Radian and its prompt below. Six frames span each 957-frame sequence; the shared time labels are rounded to the nearest second.



Figure 9: **Feature-space comparisons.** Rows show DINOv2-S, DINOv3, SigLIP2, V-JEPA 2.1, VideoMAE, DINOv2 with VideoMAE, and DiT-GAN. The four cases depict a black-and-white corgi scene, a panda drinking coffee, glazing croissants, and a coastal lighthouse.



Figure 10: **Encoder-size comparisons.** DINOv2-S, B, and G are shown on a wolf beside a frozen stream, fireworks viewed from a bridge, a traditional waterwheel, and mixing colored liquids. Each case displays three frames with its prompt below.

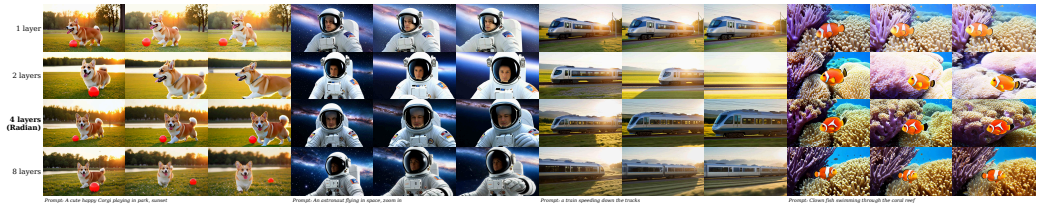


Figure 11: **Feature-layer comparisons.** Rows use one, two, four, and eight DINOv2-S feature layers, respectively. The cases show a corgi in a park, an astronaut, a train on tracks, and clownfish among coral; Radian uses the four-layer configuration.