NEXTGAN: TOWARDS SIMPLE AND SCALABLE GAN TRAINING FOR IMAGE AND VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative adversarial networks (GANs) enable one-pass sampling, but their critics are trained to separate real and generated samples rather than to provide a useful correction field. We present NEXTGAN, which directly supervises the critic's input gradient. Independent real–generated pairs are interpolated in frozen feature spaces, and the gradient is regressed onto their displacement. Random-pair regression aggregates into a distribution-dependent field that vanishes when the conditional feature distributions match; for a fixed generator, its population objective has an exact adversarial dual with a gradient-energy penalty. Without auxiliary losses, online teacher scores, or trajectories, the same objective trains pixel- and latent-space image generators and Wan2.1 video generators up to 14B parameters, remaining stable across one- and few-step sampling budgets. On ImageNet 256×256, NEXTGAN achieves FID $= 2.87$ with pixel-space JiT-B and FID $= 1.54$ with latent-space DiT-XL/1 at one evaluation, while the adversarial form reaches FID $= 1.36$. On text-to-video generation, the 1.3B model achieves the highest VBench total score among the compared 1.3B methods at one- and two-step budgets.

1 INTRODUCTION

GANs generate samples in a single forward pass (Goodfellow et al., 2014), whereas diffusion and flow models repeatedly evaluate large networks during sampling (Song et al., 2021; Lipman et al., 2023). Adapting pretrained generators through adversarial training can therefore enable one- or few-step generation from reference samples, making the critic's learning signal central.

The generator consumes the critic's input gradient, whereas conventional adversarial objectives supervise differences in critic values on real and generated samples. The resulting update direction depends on the objective, parameterization, and regularization; successful separation alone need not

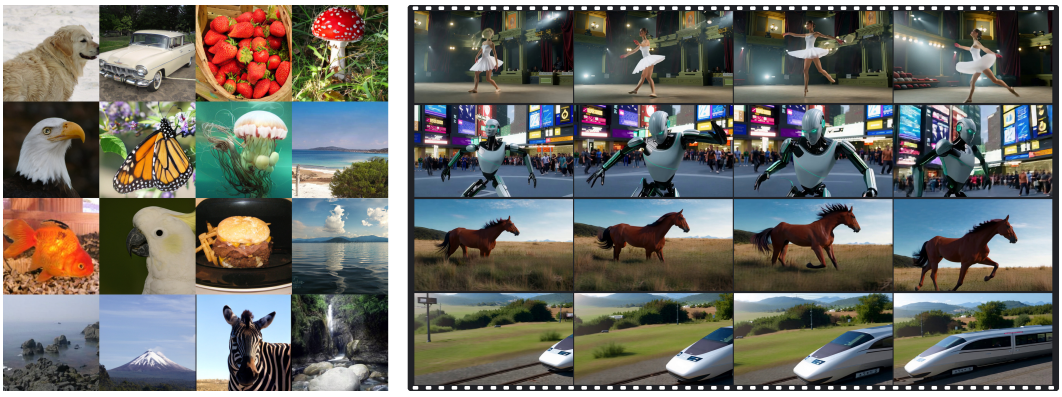


Figure 1: **Qualitative samples from NEXTGAN.** *Left:* 1-NFE class-conditional samples on ImageNet 256×256. *Right:* four frames from each of four generated videos.

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.

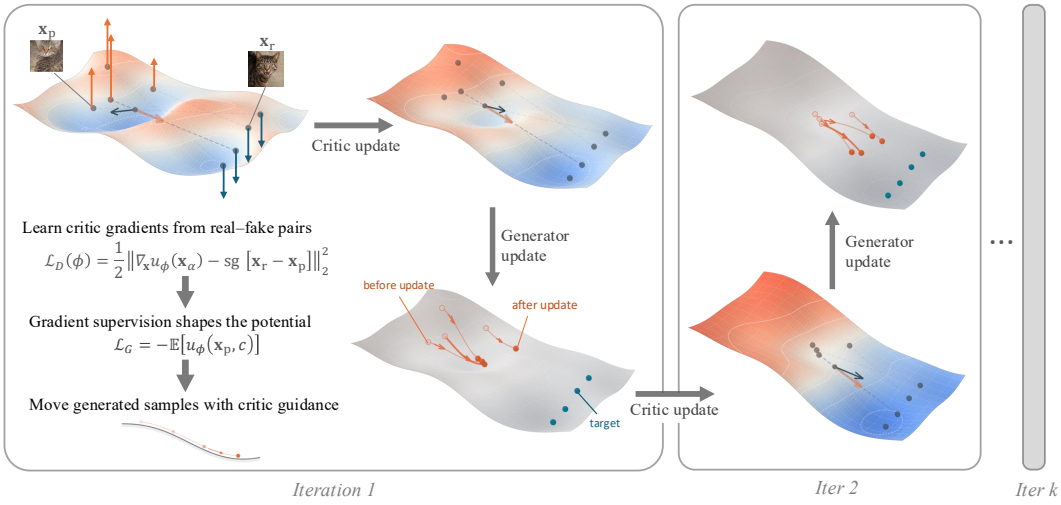


Figure 2: **How NEXTGAN learns to guide the generator.** $\mathbf{x}_p$ and $\mathbf{x}_r$ are generated and real features; $\mathbf{x}_\alpha$ lies between them. First, the critic loss $\mathcal{L}_D$ trains $\nabla_\mathbf{x} u_\phi$ to predict their displacement, with endpoints fixed (sg: stop gradient). Then, the generator loss $\mathcal{L}_G$ moves generated features uphill in the fixed potential u_ϕ. The surface depicts $-u_\phi$, so this motion appears downhill. These two updates alternate throughout training.

provide useful gradients (Arjovsky et al., 2017). Penalties on gradient magnitude constrain this signal (Gulrajani et al., 2017), but do not supply a direction target. We investigate such a target using displacements between reference samples and the current generator's outputs.

We present NEXTGAN, which trains a scalar critic by regressing its input gradient onto real–generated displacements in a frozen feature space. As illustrated in Fig. 2, independent pairs supervise the gradient along segments joining their features, and the generator then ascends the learned potential at its own features. Frozen representers provide complementary comparison spaces during training, while inference evaluates only the generator. Fig. 1 previews the resulting one-step image samples and generated video sequences.

Independent pairing makes each displacement noisy, but regression aggregates targets by interpolation location and condition. Over an idealized linear potential class, the optimal critic gradient is the conservative projection of this conditional mean displacement. When the conditional feature distributions match, exchanging the endpoints reverses the displacement without changing the interpolation law, so the mean field vanishes. Random pairs therefore supervise a distribution-dependent field without assigning samples fixed destinations; under the assumptions in Prop. 1, a sufficiently rich potential class makes the induced discrepancy zero if and only if the conditional feature distributions match.

The regression also has an exact adversarial interpretation: for a fixed generator, averaging over uniform interpolation points yields a real–generated potential gap with a gradient-energy penalty. This connects the construction to Sobolev GAN and to TBSM's conservative displacement construction, which yields the same critic objective (Mroueh et al., 2018; Sun et al., 2026). We develop this supervision for standard generator updates at generated endpoints and evaluate both regression and adversarial implementations. Their population objectives are equivalent, although their finite-minibatch updates can differ.

We evaluate NEXTGAN through controlled mode-coverage studies, image-generation experiments, and adaptation of Wan2.1 video backbones with up to 14B parameters. The controlled study shows balanced coverage of eight modes, and image ablations support combining complementary representers. On ImageNet 256×256, regression achieves 1.54 FID with a latent DiT-XL/1 and 2.87 with a pixel-space JiT-B, each in one evaluation without guidance; the adversarial form achieves 1.36 with DiT-XL/1. For video, the same 1.3B weights support one-, two-, and four-step sampling, reaching VBench total scores of 80.80 and 83.13 at one and two steps, respectively, the highest

among the compared 1.3B methods at those budgets (Tab. 2). All large-scale experiments adapt pretrained generators without online teacher scores or trajectories; the video reference samples were generated offline by Wan2.1-14B.

Our contributions are:

(a) **Displacement-supervised critic training.** We train critic gradients from independent real–generated pairs in frozen feature spaces and use the learned potentials for standard generator updates.
(b) **A distributional interpretation.** We characterize the optimal field as a projected conditional mean displacement, establish its exact population adversarial dual, and give conditions under which the discrepancy detects conditional feature-distribution mismatch.
(c) **Image and video generation at scale.** Controlled mode-coverage and representer studies accompany one-step pixel- and latent-space image generation and few-step video adaptation at 1.3B and 14B parameters, without online teacher queries.

2 Related Work

Adversarial discrepancies. GANs train a generator against a critic that separates data from generated samples (Goodfellow et al., 2014); later work changes the discrepancy or regularizes the critic with Lipschitz constraints, gradient penalties, and feature projections (Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018; Sauer et al., 2021). Sobolev GAN is closest in form because it penalizes critic-gradient energy (Mroueh et al., 2018). NextGAN differs in the supervision: the gradient itself is regressed onto real–fake displacements, and the resulting value is the corresponding negative-Sobolev discrepancy (Eq. (6)).

Flow matching and direct distribution dynamics. Flow matching regresses a velocity along a prescribed noise–data path (Lipman et al., 2023). Direct distribution-dynamics methods instead construct updates from the current model and data distributions: Drifting uses batch interactions, while Three-Body Scattering Modeling (TBSM) uses constant-size real–generated–generated interactions and tracked regression targets (Deng et al., 2026; Sun et al., 2026). TBSM's design map also contains a GAN-like displacement corner at $\rho = 1, \lambda = 0$: replacing unit bearings by real–generated displacements and constraining its tracker to be conservative yields the same path-regularized critic objective as our field regression.

We call this configuration the *TBSM GAN corner*; its generator query differs from the endpoint update used by NextGAN. NextGAN uses independent real–generated pairs, marginalizes the path time, and constrains the learned field to be conservative so that it integrates to a scalar potential.

Few-step generation from pretrained models. Distribution-matching distillation uses teacher scores or trajectories (Yin et al., 2024; Zheng et al., 2025; Gu et al., 2026), while adversarial post-training introduces a discriminator. NextGAN constructs its target from reference data and current generator samples, without querying teacher scores or trajectories during training. Throughout the paper, *teacher-free* refers specifically to the absence of such online teacher queries.

3 NextGAN

We formalize NextGAN by defining displacement targets from current real–generated pairs, deriving their adversarial dual, and specifying alternating updates (Fig. 2).

3.1 Comparison space and random-pair supervision

Let $(\bar{\mathbf{x}}_\mathrm{r}, \mathbf{c}) \sim p_\mathrm{data}(\bar{\mathbf{x}}, \mathbf{c})$ be a real sample–condition pair, draw $\mathbf{z} \sim p(\mathbf{z})$ independently of this pair, and let $\bar{\mathbf{x}}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ be a generated sample. For a frozen differentiable representation map $\boldsymbol{f}$, define the comparison vectors $\mathbf{x}_\mathrm{r} = \boldsymbol{f}(\bar{\mathbf{x}}_\mathrm{r}, \mathbf{c})$ and $\mathbf{x}_\mathrm{p} = \boldsymbol{f}(\bar{\mathbf{x}}_\mathrm{p}, \mathbf{c})$; an image representer may simply ignore $\mathbf{c}$. We write their conditional laws as $Q(\cdot \mid \mathbf{c})$ and $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$, respectively. The endpoints are independent conditional on $\mathbf{c}$. Define their displacement and the segment between them:

$$\boldsymbol{\Delta} = \mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}, \qquad \mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_\mathrm{p} + \alpha \mathbf{x}_\mathrm{r}, \qquad \alpha \sim \mathcal{U}(0,1). \tag{1}$$

All expectations use the above sampling rule with independent α. We call the law of $(\mathbf{x}_\alpha, \mathbf{c})$ the *corridor distribution* $\mu_{\boldsymbol{\theta}}$. Each displacement is a noisy observation of how the generated distribution differs from the data distribution; the pairs are not assumed to be an optimal-transport coupling.

The critic is a scalar potential $u_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c})$, with trainable parameters $\boldsymbol{\phi}$ and guidance field $\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}$. Using a potential makes the field compatible with a standard GAN generator update and permits the line-integral identity below. The representer $\boldsymbol{f}$ remains frozen.

3.2 Learning a displacement field

Conventional GAN objectives supervise critic values on real and generated samples. Here each pair supplies a vector target for the critic gradient, specifying both its direction and magnitude. With the generator fixed, the unit-weight regression objective is

$$\mathcal{L}_{\text{field}}(\boldsymbol{\phi}; \boldsymbol{\theta}) = \frac{1}{2}\,\mathbb{E}\left[\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}) - \boldsymbol{\Delta}\|_2^2\right]. \tag{2}$$

The critic observes $\mathbf{x}_\alpha$ and $\mathbf{c}$, but not α separately, so it pools supervision across interpolation times into one field for the current pair of distributions. Implementation-specific normalization and loss scaling are in App. A. Regression averages pairwise targets into the population mismatch field

$$\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{c}) = \mathbb{E}[\mathbf{X}_{\mathrm{r}} - \mathbf{X}_{\mathrm{p}} \mid \mathbf{X}_\alpha = \mathbf{x}, \mathbf{C} = \mathbf{c}], \tag{3}$$

where capital letters denote the random variables of Eq. (1). Squared-error regression recovers this conditional mean over unrestricted fields; over an idealized linear potential class, it recovers its conservative projection (Prop. 1). The field is defined $\mu_{\boldsymbol{\theta}}$-almost everywhere and changes with the generator.

Random pairing therefore has a distributional interpretation: when the conditional endpoint laws match, exchanging the endpoints and replacing α by $1-\alpha$ preserves the query and reverses the target. The conditional mean vanishes even though individual displacements remain noisy, so generated samples are not regressed directly toward randomly assigned real samples.

3.3 Adversarial dual and distribution matching

Because the field is a gradient, the fundamental theorem of calculus along Eq. (1) gives, for every pair,

$$\int_0^1 \langle \nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}), \boldsymbol{\Delta} \rangle \, \mathrm{d}\alpha = u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{r}}, \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}). \tag{4}$$

Expanding the square in Eq. (2) and applying Eq. (4) to the cross term yields

$$\begin{aligned} \mathcal{L}_{\text{field}}(\boldsymbol{\phi}; \boldsymbol{\theta}) &= \tfrac{1}{2}\mathbb{E}\|\boldsymbol{\Delta}\|_2^2 - \mathcal{V}(\boldsymbol{\phi}; \boldsymbol{\theta}), \\ \mathcal{V}(\boldsymbol{\phi}; \boldsymbol{\theta}) &= \mathbb{E}[u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{r}}, \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})] - \tfrac{1}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2. \end{aligned} \tag{5}$$

The first term does not depend on $\boldsymbol{\phi}$, so minimizing the regression loss is *exactly* maximizing the adversarial value $\mathcal{V}$: a real–generated potential gap minus the field energy on the corridor. Its supremum over potentials defines the discrepancy

$$\mathcal{D}_{\text{Next}}(P_{\boldsymbol{\theta}}, Q) = \sup_{u \in \mathcal{H}} \left\{ \mathbb{E}\big[u(\mathbf{x}_{\mathrm{r}}, \mathbf{c}) - u(\mathbf{x}_{\mathrm{p}}, \mathbf{c})\big] - \frac{1}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha, \mathbf{c})\|_2^2 \right\}. \tag{6}$$

We train with both forms: the regression loss in Eq. (2) and the adversarial critic loss $\mathcal{L}_{\mathrm{D}}^{\mathrm{adv}} = -\mathcal{V}$ (App. A.1). Their equivalence holds for a fixed generator after averaging over uniform interpolation times; their finite-minibatch updates need not coincide. The regression form supplies displacement targets, while the adversarial form exposes the distributional quantity measured at the optimum.

Proposition 1 (Population mismatch field and distribution matching)**.** *Assume finite second moments. Let $\mathcal{H}$ be a linear space of potentials measurable in $\mathbf{c}$ and C^1 in $\mathbf{x}$, with integrable endpoint values and $\nabla_{\mathbf{x}} u \in L^2(\mu_{\boldsymbol{\theta}})$. Let $\mathcal{G}$ be the $L^2(\mu_{\boldsymbol{\theta}})$-closure of $\{\nabla_{\mathbf{x}} u : u \in \mathcal{H}\}$.*

(a) Over all square-integrable vector fields, the minimizer of Eq. (2) is $\mathbf{v}_{\boldsymbol{\theta}}$; over $\mathcal{G}$, it is the orthogonal projection $\mathrm{Proj}_{\mathcal{G}}\, \mathbf{v}_{\boldsymbol{\theta}}$.

Algorithm 1 NEXTGAN training (one representer, unit weight). $\mathrm{sg}[\cdot]$ denotes stop gradient.

Require: frozen representer $\boldsymbol{f}$; critic steps per generator step n_{critic}

Initialize generator $\boldsymbol{g}_{\boldsymbol{\theta}}$ and potential $u_{\boldsymbol{\phi}}$
repeat
  Sample $(\bar{\mathbf{x}}_{\mathrm{r}}, \mathbf{c})$, $\mathbf{z} \sim p(\mathbf{z})$, and generate $\bar{\mathbf{x}}_{\mathrm{p}} \leftarrow \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$
  Compute $\mathbf{x}_{\mathrm{r}} \leftarrow \boldsymbol{f}(\bar{\mathbf{x}}_{\mathrm{r}}, \mathbf{c})$ and $\mathbf{x}_{\mathrm{p}} \leftarrow \boldsymbol{f}(\bar{\mathbf{x}}_{\mathrm{p}}, \mathbf{c})$
  for $k = 1, \ldots, n_{\mathrm{critic}}$ **do**
    Sample $\alpha \sim \mathcal{U}(0,1)$ and set $\mathbf{x}_\alpha \leftarrow \alpha \mathbf{x}_{\mathrm{r}} + (1-\alpha)\,\mathrm{sg}[\mathbf{x}_{\mathrm{p}}]$
    Update $\boldsymbol{\phi}$ by descending $\frac{1}{2}\,\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}) - \mathrm{sg}[\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}}]\|_2^2$
  end for
  Hold $\boldsymbol{\phi}$ fixed; update $\boldsymbol{\theta}$ by descending $-u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})$
until convergence

(b) $\mathcal{D}_{\mathrm{Next}}(P_{\boldsymbol{\theta}}, Q) = \frac{1}{2}\|\operatorname{Proj}_{\mathcal{G}} \mathbf{v}_{\boldsymbol{\theta}}\|^2_{L^2(\mu_{\boldsymbol{\theta}})} \geq 0$.

(c) If $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}) = Q(\cdot \mid \mathbf{c})$ *for almost every* $\mathbf{c}$*, then* $\mathbf{v}_{\boldsymbol{\theta}} = \mathbf{0}$ *and* $\mathcal{D}_{\mathrm{Next}} = 0$*. Conversely, if* $\mathcal{D}_{\mathrm{Next}} = 0$ *and* $\mathcal{H}$ *determines the joint law of* $(\mathbf{X}, \mathbf{C})$*—for example, if it contains every product* $u(\mathbf{x}, \mathbf{c}) = a(\mathbf{x})b(\mathbf{c})$ *with* a *smooth and compactly supported and* b *bounded and measurable—then the conditional feature distributions coincide for almost every* $\mathbf{c}$*.*

The proof is in App. C; finite potential networks restrict this population characterization to their parameterization.

Remark 1 (Negative-Sobolev form). Fix $\mathbf{c}$ and write q, $p_{\boldsymbol{\theta}}$, and $\omega_{\boldsymbol{\theta}}$ for the densities of $Q(\cdot \mid \mathbf{c})$, $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$, and $\mu_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$. Assume that the weighted Poisson problem admits a weak solution under suitable boundary or decay conditions. When $\mathcal{H}$ is unrestricted, the optimal potential solves $-\nabla_{\mathbf{x}} \cdot \left(\omega_{\boldsymbol{\theta}} \nabla_{\mathbf{x}} u^\star\right) = q - p_{\boldsymbol{\theta}}$ weakly. Averaging over conditions gives $\mathcal{D}_{\mathrm{Next}} = \frac{1}{2}\,\mathbb{E}_{\mathbf{c}}\|q - p_{\boldsymbol{\theta}}\|^2_{\dot{H}^{-1}(\mu_{\boldsymbol{\theta}}(\cdot|\mathbf{c}))}$. This is a squared negative-Sobolev norm weighted by the corridor measure, related to the discrepancy underlying Sobolev GAN (Mroueh et al., 2018). The regression loss equals the negative adversarial value plus the displacement-dependent constant in Eq. (5).

3.4 ALTERNATING CRITIC AND GENERATOR UPDATES

With the generator fixed, the critic learns the current displacement field. We then hold the critic fixed and update the generator by ascending the potential at generated endpoints:

$$\mathcal{L}_{\mathrm{gen}}(\boldsymbol{\theta}; \boldsymbol{\phi}) = -\mathbb{E}\, u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}), \qquad \mathbf{x}_{\mathrm{p}} = \boldsymbol{f}(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}), \mathbf{c})\,. \tag{7}$$

The chain rule pulls $\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})$ back through $\boldsymbol{f}$ and $\boldsymbol{g}_{\boldsymbol{\theta}}$. After each generator change, fresh pairs update the critic to reflect the new mismatch. Alg. 1 summarizes the loop: the generator retains the standard potential-based GAN update, while displacement supervision shapes its critic. At inference, the critic and representers are discarded.

The update queries generated endpoints, whereas regression supervises interpolated features; Prop. 1 characterizes the field only $\mu_{\boldsymbol{\theta}}$-almost everywhere. Since the corridor measure also changes with the generator, the update is not generally the full gradient of $\mathcal{D}_{\mathrm{Next}}$ (App. H.3).

3.5 TRAINING DYNAMICS AND MODE COVERAGE

To examine how this change in critic supervision affects the alternating updates in Alg. 1, we track generated samples across a mixture of eight Gaussians. NEXTGAN and a non-saturating GAN share the initialization, critic capacity, latent probes, and 30K-update budget. Fig. 3 reports samples from the same latent probes and mode imbalance, $8 - N_{\mathrm{eff}}$, where N_{eff} is the effective mode count among samples captured by the target modes. Zero indicates balanced allocation across all eight modes. Full experimental details are in App. E.

NEXTGAN reaches all eight modes and maintains a nearly balanced allocation, ending at $N_{\mathrm{eff}} = 7.995$. GAN initially spreads across the ring but later concentrates on fewer modes, ending at $N_{\mathrm{eff}} = 2.770$. The fractions of samples captured by the modes are similar (92.8% and 93.3%, respectively).

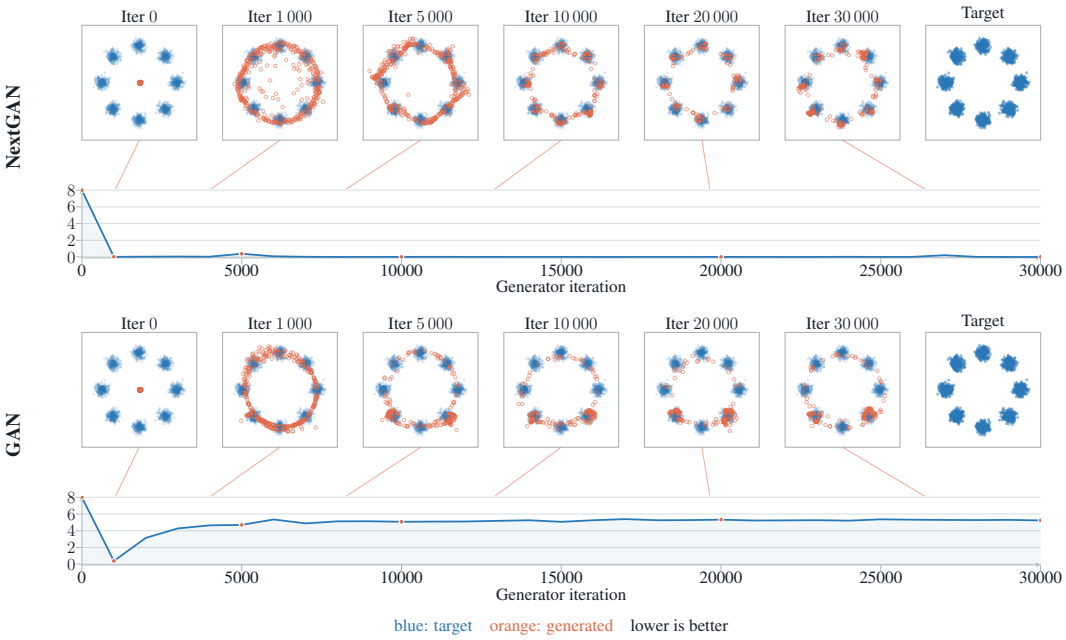


Figure 3: **Mode coverage during training.** *Rows:* NEXTGAN and GAN. *Columns:* samples from fixed latent probes at selected iterations and the target distribution. The curves below each row report mode imbalance, $8 - N_{\text{eff}}$; lower is better.

Thus, in this controlled experiment, the learned guidance supports balanced mode coverage; the main difference is how generated mass is allocated across modes, despite similar captured fractions.

3.6 PRACTICAL IMPLEMENTATION

Comparison spaces. We apply the same field-regression objective in frozen feature spaces. For images, we use frozen MAE-B/16, SigLIP2-B/16, and ResNet-18 encoders (He et al., 2022; Tschannen et al., 2025; He et al., 2016). For video, we explored VideoMAE, V-JEPA2, InternVideo-Next, RVM, and DINOv3 as pretrained representers (Tong et al., 2022; Assran et al., 2025; Wang et al., 2025; Zoran et al., 2025; Siméoni et al., 2025). In our experiments, these features tended to favor static appearance over motion, and no single representer captured all aspects of video generation. We therefore use the frozen pretrained video backbone itself as the representer. The generator uses LoRA (Hu et al., 2022); disabling the adapters provides frozen, text-conditioned features of clean video latents at selected layers.

Potential heads and optimization. Each image encoder or video layer feeds a separate four-block transformer head that outputs a conditional scalar potential. We sum the head losses and backpropagate through the frozen representers to the generator. We alternate n_{critic} critic updates with one generator update and maintain a generator EMA (Alg. 1).

Backward simulation. For video, following the backward simulation strategy of rCM (Zheng et al., 2025), training uses randomly sampled rollout lengths. Starting from pure noise, we alternate denoising to a clean estimate using the Wan velocity prediction and forward noising with fresh noise, keeping noise levels non-increasing. Here g_θ denotes the complete sampling procedure, with $\mathbf{z}$ collecting its random inputs. We backpropagate the generator loss only through the final denoising step; earlier rollout steps are detached, yielding a truncated gradient through this procedure.

Table 1: Class-conditional ImageNet-256 results. Baselines are quoted from their original papers and may use different evaluators. NEXTGAN Params denote nominal backbone sizes (implementation counts: App. A). NFE counts sequential generator evaluations, including guidance and autoregressive inner steps where applicable.

Method	Space	Backbone	Params	NFE	FID ↓	IS ↑
Consistency / flow-map distillation						
Shortcut (Frans et al., 2025)	latent	DiT-XL/2	676M	1	10.60	–
PaGoDA (Kim et al., 2024)	pixel	ADM	301M	1	1.56	259.61
Diffusion / flow matching						
MeanFlow (Geng et al., 2025)	latent	XL/2	676M	1	3.43	–
ADM (Dhariwal & Nichol, 2021)	pixel	ADM-G + ADM-U	866M	250+250	3.94	215.8
VDM++ (Kingma & Gao, 2023)	pixel	U-ViT-L/2	2B	512×2	2.12	267.7
DiT (Peebles & Xie, 2023)	latent	DiT-XL/2	675M	250×2	2.27	278.2
SiT (Ma et al., 2024)	latent	SiT-XL/2	675M	250×2	2.06	277.5
JiT (Li & He, 2025)	pixel	JiT-H/16	953M	200	1.86	303.4
Autoregressive models						
VAR (Tian et al., 2024)	discrete	VAR-d30	2B	10×2	1.92	323.1
BAR (Yu et al., 2026)	discrete	BAR-L	1.1B	2,048	0.99	296.9
MAR (Li et al., 2024)	latent	MAR-H	943M	51,200	1.55	303.7
GAN-based one-step generators						
BigGAN-deep (Brock et al., 2019)	pixel	BigGAN-deep	112M	1	6.95	171.4
GigaGAN (Kang et al., 2023)	pixel	GigaGAN	569M	1	3.45	225.5
StyleGAN-XL (Sauer et al., 2022)	pixel	StyleGAN-XL	166M	1	2.30	265.1
Direct distribution dynamics and distributional losses						
Drift (Deng et al., 2026)	pixel	DiT-L/16	464M	1	1.61	307.5
FD-loss (Yang et al., 2026)	pixel	JiT-H/16	953M	1	0.75	313.0
TBSM (Sun et al., 2026)	latent	DiT-XL/2	675M	1	1.63	243.0
NEXTGAN *(no online teacher queries)*						
NEXTGAN	pixel	JiT-B	131M	1	2.87	207.1
NEXTGAN	latent	DiT-XL/1	675M	1	1.54	243.5
NEXTGAN (adversarial form)	latent	DiT-XL/1	675M	1	1.36	241.7

4 EXPERIMENTS

Following the controlled analysis in Sec. 3.5, we evaluate NEXTGAN on pixel- and latent-space image generators and on video generators with up to 14B parameters.

Protocol. We evaluate gradient regression and its equivalent adversarial form, both with the standard GAN generator update. Training requires no online teacher scores or trajectories. App. A details the architectures, checkpoint initialization, evaluation settings, and training compute.

4.1 IMAGENET-256: ONE-STEP GENERATION

We apply the same objective to class-conditional ImageNet-256 generation (Deng et al., 2009) in both latent and pixel space. Tab. 1 reports quality and sampling cost, including evaluations used for classifier-free guidance (Ho & Salimans, 2022). Gradient regression achieves 1.54 FID with the latent DiT-XL/1 and 2.87 with the pixel-space JiT-B, each in one network evaluation without guidance. The equivalent adversarial form achieves 1.36 FID with DiT-XL/1. These results show that the training objective works with both generator types. Fig. 1 shows samples from the latent model.

App. B compares training objectives and frozen representers on JiT-B, with the training curves and feature-space ablations in Fig. 7.

4.2 TEXT-TO-VIDEO GENERATION

We fine-tune Wan2.1 text-to-video generators (Wan Team, 2025) at 1.3B and 14B parameters using LoRA (Hu et al., 2022) on the publicly released Wan_datasets from rCM (Zheng et al., 2025). We use global batch sizes of 32 and 16 for the 1.3B and 14B models, respectively. With its adapters disabled, the frozen pretrained Wan2.1 backbone serves as the representer and supplies intermediate features to the potential heads. These heads apply joint attention to text and video tokens. The field-regression objective is otherwise unchanged.

Table 2: Text-to-video generation on VBench at $832 \times 480 \times 81$. Results are grouped by model size and NFE. The 14B NEXTGAN rows select the best completed run by total score independently at each NFE; all three scores in a row come from that run. "Online teacher" indicates whether training requires online teacher scores or trajectories. Higher is better for all three scores.

Model	Online teacher	NFE	Evaluation Scores ↑		
			Quality	Semantic	Total
1.3B parameters					
Wan2.1 T2V (Wan Team, 2025)	–	50×2	82.91	80.22	82.37
rCM (Zheng et al., 2025)	yes	1	81.76	**76.34**	80.68
AnyFlow (Gu et al., 2026)	yes	1	78.74	75.63	78.12
Wan2.1 + DMD2 (Yin et al., 2024)	yes	1	80.30	71.85	78.61
NEXTGAN	no	1	**82.15**	75.41	**80.80**
rCM (Zheng et al., 2025)	yes	2	82.61	78.99	81.88
AnyFlow (Gu et al., 2026)	yes	2	83.08	**80.95**	82.66
Wan2.1 + DMD2 (Yin et al., 2024)	yes	2	**84.08**	77.56	82.78
NEXTGAN	no	2	83.98	79.73	**83.13**
rCM (Zheng et al., 2025)	yes	4	82.61	78.55	81.79
AnyFlow (Gu et al., 2026)	yes	4	84.01	**81.67**	**83.54**
Wan2.1 + DMD2 (Yin et al., 2024)	yes	4	**84.13**	76.45	82.59
NEXTGAN	no	4	83.79	78.44	82.72
14B parameters					
Wan2.1 T2V (Wan Team, 2025)	–	50×2	84.26	80.92	83.58
rCM (Zheng et al., 2025)	yes	1	80.04	74.33	78.90
AnyFlow (Gu et al., 2026)	yes	1	**81.36**	**80.71**	**81.23**
NEXTGAN	no	1	78.65	77.04	78.33
rCM (Zheng et al., 2025)	yes	2	83.72	80.58	83.09
AnyFlow (Gu et al., 2026)	yes	2	**84.12**	82.97	**83.89**
NEXTGAN	no	2	81.65	**83.02**	81.92
rCM (Zheng et al., 2025)	yes	4	83.66	81.22	83.17
AnyFlow (Gu et al., 2026)	yes	4	**84.36**	**83.07**	**84.10**
NEXTGAN	no	4	81.86	81.41	81.77

At each model scale, we evaluate one-, two-, and four-step sampling using the backward simulation procedure described in Sec. 3.6. For the 14B comparison, we report the highest completed VBench total score at each inference budget.

Tab. 2 reports self-evaluated VBench results (Huang et al., 2024); the CausVid DMD2 entries use the publicly released bidirectional checkpoint. At 1.3B parameters, a single model reaches total scores of 80.80, 83.13, and 82.72 with one, two, and four evaluations, respectively. The selected 14B results are 78.33, 81.92, and 81.77. At two steps, the 14B model attains the highest semantic score among the compared 14B methods (83.02), while AnyFlow retains the highest total score (83.89).

Comparison with GAN, WGAN, and TBSM (GAN). Fig. 4 compares four-step video generation across GAN, WGAN, TBSM (GAN), and NEXTGAN. At 1.3B parameters, NEXTGAN attains a VBench total score of 82.72, compared with 81.38 for GAN, 60.71 for WGAN, and 73.07 for TBSM (GAN). The selected qualitative examples show clearer object identity and scene structure with NEXTGAN.

Qualitative comparison. To complement the aggregate VBench scores, Fig. 5 compares single frames from rCM, AnyFlow, and NEXTGAN across six manually selected cases with matched text prompts. Each column shows a different case, with frames manually selected using the same frame index across methods within each case. In these examples, NEXTGAN produces more realistic visual appearances than rCM and AnyFlow.

Sample diversity. Fig. 6 compares four-step samples from Wan2.1+DMD2 (Yin et al., 2024) and NEXTGAN under matched prompts and per-prompt random seeds, using the same frame index for each corresponding pair of videos. Each row is one prompt and each column is a different seed; the left panel shows DMD2 and the right panel shows NEXTGAN. Across these prompts, NEXTGAN shows greater sample diversity than DMD2, with more varied layouts, colors, and subject configurations.

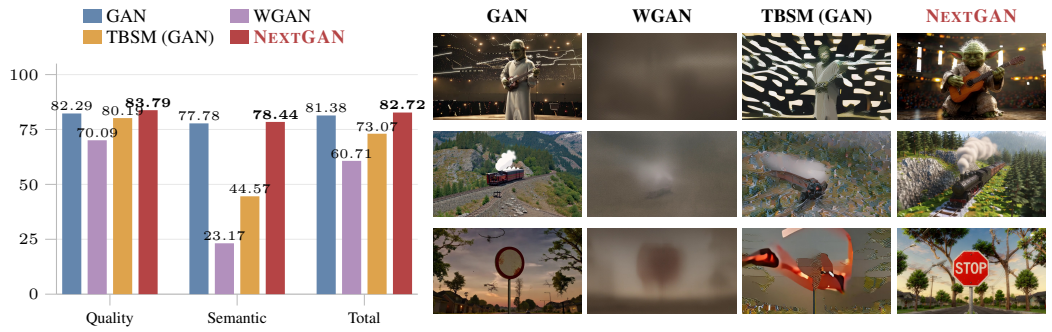


Figure 4: **Video generation with GAN, WGAN, TBSM (GAN), and NEXTGAN.** *Left:* NEXTGAN achieves the highest quality, semantic, and total VBench scores (best in bold). *Right:* NEXTGAN shows clearer subject identity, more coherent object structure, and more accurate text generation.

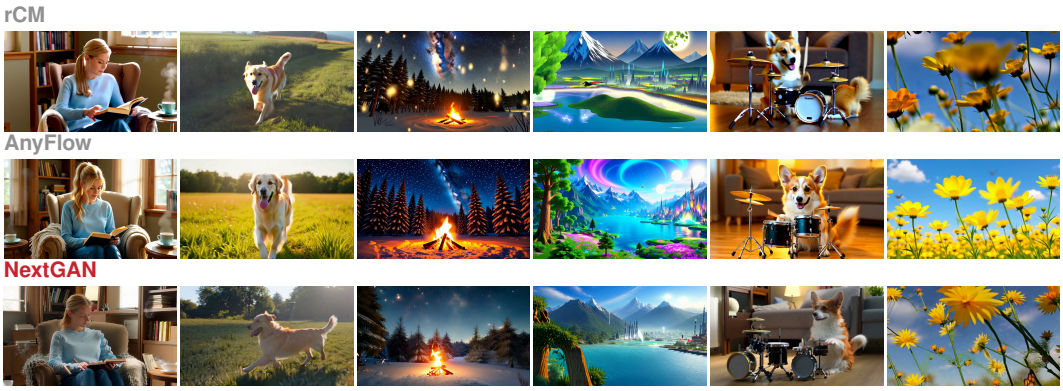


Figure 5: Six manually selected prompts (columns) with manually selected frames at matching indices across methods. Rows: rCM, AnyFlow, and NEXTGAN (all 1.3B).

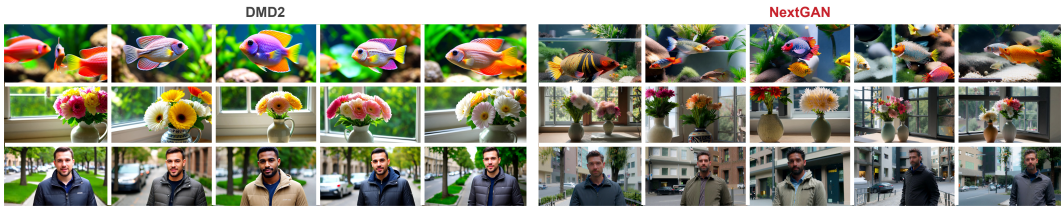


Figure 6: Intra-prompt sample diversity under four-step sampling (Wan2.1-1.3B). Left: Wan2.1+DMD2; right: NEXTGAN. Rows show fish, flowers, and man; each panel shows five seeds per prompt, with matched seed indices across methods. Corresponding videos from both methods are shown at the same frame index.

5 CONCLUSION

NEXTGAN preserves the GAN interface while directly supervising the critic gradient consumed by the generator. Regression over random real–fake pairs recovers a population mismatch field, and a conservative parameterization gives the population critic objective an exact adversarial dual over a negative-Sobolev discrepancy. The same objective fine-tunes pixel- and latent-space image generators and video generators with up to 14B parameters, supporting multiple inference budgets. The matching guarantee applies in the chosen representation spaces under the stated assumptions. Training large generators from scratch remains future work.

REPRODUCIBILITY STATEMENT

Alg. 1 and App. A specify the objective, architectures, and training protocol, and App. C contains the complete proof. We will release code, training configurations, and evaluation scripts subject to the review policy.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, code implementation, related-work literature searches, and detailed mathematical derivations. The core research ideas and the proposed method originated with the authors. The authors manually reviewed all AI-assisted content for correctness and factual accuracy and take responsibility for all claims, proofs, code, and results in this work.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026. URL https://arxiv.org/abs/2602.04770.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.12557.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. URL https://arxiv.org/abs/2505.13447.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

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

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026. URL https://arxiv.org/abs/2605.11755.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

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. *arXiv preprint arXiv:2311.17982*, 2024. URL https://arxiv.org/abs/2311.17982.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. *arXiv preprint arXiv:2303.05511*, 2023. URL https://arxiv.org/abs/2303.05511.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. PaGoDA: Progressive growing of a one-step generator from a low-resolution diffusion teacher. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.14822.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.00848.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025. URL https://arxiv.org/abs/2511.13720.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.11838.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2401.08740.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1f78iAcY7.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17480–17492. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9219adc5c42107c4911e249155320648-Abstract.html.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. doi: 10.1145/3528233.3530738. URL https://arxiv.org/abs/2202.00273.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Peng Sun, Yi Jiang, and Tao Lin. Unified continuous generative models. *arXiv preprint arXiv:2505.07447*, 2025. URL https://arxiv.org/abs/2505.07447.

Peng Sun, Zhenglin Cheng, Deyuan Liu, Jun Xie, Xinyi Shang, and Tao Lin. Three-body scattering for generative modeling. *arXiv preprint arXiv:2607.18198*, 2026. URL https://arxiv.org/abs/2607.18198.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. URL https://arxiv.org/abs/2404.02905.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

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

Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Lang Lin, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. InternVideo-Next: Towards general video foundation models without video-text supervision. *arXiv preprint arXiv:2512.01342*, 2025.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026. URL https://arxiv.org/abs/2604.28190.

Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. URL https://arxiv.org/abs/2501.01423.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. URL https://arxiv.org/abs/2405.14867.

Qihang Yu, Qihao Liu, Ju He, Xinyang Zhang, Yang Liu, Liang-Chieh Chen, and Xi Chen. Autoregressive image generation with masked bit modeling. *arXiv preprint arXiv:2602.09024*, 2026. URL https://arxiv.org/abs/2602.09024.

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. *arXiv preprint arXiv:2510.08431*, 2025. URL https://arxiv.org/abs/2510.08431.

Daniel Zoran, Nikhil Parthasarathy, Yi Yang, Drew A. Hudson, Joao Carreira, and Andrew Zisserman. Recurrent video masked autoencoders. *arXiv preprint arXiv:2512.13684*, 2025.

A IMPLEMENTATION DETAILS

Image experiments. ImageNet experiments use frozen MAE-B/16, SigLIP2-B/16, and ResNet-18 encoders (He et al., 2022; Tschannen et al., 2025; He et al., 2016) with independent condition-aware potential heads. JiT-B is initialized from the released JiT checkpoint (Li & He, 2025); DiT-XL/1 (Peebles & Xie, 2023) and the VA-VAE decoder (Yao & Wang, 2025) follow the UCGM release (Sun et al., 2025). The main DiT field-regression run uses 100K updates on eight H800 GPUs, while the controlled JiT-B ablations use 20K updates. We optimize the generator and potential heads with AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9, \beta_2 = 0.999$) at learning rates of 10^{-5} and 5×10^{-5}, respectively. We use two critic updates per generator update, a per-GPU batch size of 32 (global batch size 256 for the eight-GPU run), and a generator EMA decay of 0.999. MAE and SigLIP2 inputs are bicubically resized to 224 and 256 pixels, respectively; ResNet-18 inputs are bilinearly resized to 224 pixels. Each encoder then applies its checkpoint-specific channel normalization. Evaluation uses 50K generated images for both the main DiT run and the controlled JiT-B ablations. The 100K field-regression run consumed 158.3 H800 GPU-hours, including two checkpoint restarts (19.78 hours of eight-GPU Slurm allocation); its final segment averaged 0.669 seconds per update. Table 1 reports nominal JiT-B/16 (131M) and standard DiT-XL/2 (675M) backbone sizes (Li & He, 2025; Peebles & Xie, 2023). Our instantiated JiT-B and DiT-XL/1 generators have approximately 131.3M and 676.1M trainable parameters, respectively, excluding critics and representers.

Displacement normalization. For each representer channel, we center and standardize tokens using exponential-moving-average first and second moments. We then normalize each token to Euclidean norm $\sqrt{d}$, where d is its channel dimension. The field-regression run minimizes $\frac{\lambda}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}) - \boldsymbol{\Delta}/\lambda\|_2^2$ with $\lambda = 5$; the norm is summed over token and channel dimensions before batch averaging. Its supervised target is therefore $\boldsymbol{\Delta}/5$. Expanding the square gives the critic loss $-\mathbb{E}[u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{r}, \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{p}, \mathbf{c})] + \frac{\lambda}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2$ plus a term independent of $\boldsymbol{\phi}$. For an unrestricted potential, setting $\widetilde{u} = \lambda u$ recovers the unit-weight regression up to a positive factor; this does not assert identical optimization under the spectral normalization used for the potential heads.

Video experiments. Both model sizes use the Wan2.1-14B 480p 250K subset of the Wan_datasets released with rCM (Zheng et al., 2025). This dataset contains videos generated offline by Wan2.1-14B; these serve as the reference ("real") samples for video training. Optimization requires no online teacher scores or trajectories.

Wan2.1 is fine-tuned with LoRA (Hu et al., 2022) of rank 256 and scaling parameter $\alpha_{\mathrm{LoRA}} = 256$, applied to the attention projections and feed-forward layers. With the adapters disabled, we extract text-conditioned features from clean video latents at blocks 7, 11, and 15 (zero-based indices), using a fixed model timestep of 500 without adding noise to the latents. The features undergo 2×2 spatial average pooling, preserving the temporal dimension, and each selected layer feeds an independent text-conditioned potential head. The base-model weights remain frozen, while gradients propagate through its features to the generator. Backward simulation uses the Wan velocity parameterization.

Evaluation. We report FID (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) for ImageNet generation; sample counts for the main DiT run and controlled JiT-B ablations are specified above.

A.1 ALGORITHM AND CODE FOR THE EQUIVALENT ADVERSARIAL FORM

We use the unit-weight convention of Alg. 1 throughout this subsection. For a fixed generator, expanding the regression loss and applying Eq. (4) gives

$$\frac{1}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}) - \boldsymbol{\Delta}\|_2^2 = \underbrace{\mathbb{E}[u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{p}, \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{r}, \mathbf{c})] + \frac{1}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2}_{\mathcal{L}_\mathrm{D}^{\mathrm{adv}}} + \frac{1}{2}\mathbb{E}\|\boldsymbol{\Delta}\|_2^2. \quad (8)$$

The last term is independent of the critic parameters, and $\mathcal{L}_\mathrm{D}^{\mathrm{adv}}$ is the negative of the value in Eq. (5). The equivalence holds after averaging over uniform interpolation times; the two finite-minibatch

estimators need not coincide. The penalty is squared gradient energy, not the unit-norm penalty $(\|\nabla u\|_2 - 1)^2$ used in WGAN-GP.

Alg. 2 writes the alternating updates in adversarial form for one frozen representer and one scalar critic. Each displayed loss is averaged over the minibatch; real and generated samples are paired independently conditional on $\mathbf{c}$. The critic update detaches the endpoints from the generator, but differentiates the gradient-energy term with respect to $\boldsymbol{\phi}$. The generator update holds $\boldsymbol{\phi}$ fixed while retaining input gradients through both the critic and the frozen representer. For multiple representers, sum the corresponding losses.

Algorithm 2 NEXTGAN training in adversarial form

Require: frozen representer $\boldsymbol{f}$; critic steps per generator step n_{critic}

Initialize generator $\boldsymbol{g_\theta}$ and potential $u_{\boldsymbol{\phi}}$
repeat
  Sample $(\bar{\mathbf{x}}_{\mathrm{r}}, \mathbf{c})$, $\mathbf{z} \sim p(\mathbf{z})$, and generate $\bar{\mathbf{x}}_{\mathrm{p}} \leftarrow \boldsymbol{g_\theta}(\mathbf{z}, \mathbf{c})$
  Compute $\mathbf{x}_{\mathrm{r}} \leftarrow \boldsymbol{f}(\bar{\mathbf{x}}_{\mathrm{r}}, \mathbf{c})$ and $\mathbf{x}_{\mathrm{p}} \leftarrow \boldsymbol{f}(\bar{\mathbf{x}}_{\mathrm{p}}, \mathbf{c})$
  for $k = 1, \ldots, n_{\text{critic}}$ **do**
    Sample $\alpha \sim \mathcal{U}(0,1)$ and set $\mathbf{x}_\alpha \leftarrow \alpha \mathbf{x}_{\mathrm{r}} + (1-\alpha)\,\mathrm{sg}[\mathbf{x}_{\mathrm{p}}]$
    Update $\boldsymbol{\phi}$ by descending $u_{\boldsymbol{\phi}}(\mathrm{sg}[\mathbf{x}_{\mathrm{p}}], \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{r}}, \mathbf{c}) + \dfrac{1}{2}\,\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2$
  end for
  Hold $\boldsymbol{\phi}$ fixed; update $\boldsymbol{\theta}$ by descending $-u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})$
until convergence

The following PyTorch code implements one alternating update for one frozen representer and one scalar critic, using the endpoint-detachment convention of Alg. 1. The generator loss remains $-\mathbb{E}[u_{\boldsymbol{\phi}}(\boldsymbol{f}(\boldsymbol{g_\theta}(\mathbf{z}, \mathbf{c}), \mathbf{c}), \mathbf{c})]$. The representer includes feature preprocessing and normalization; its parameters are frozen, but input gradients are retained during the generator update. For multiple representers, sum their critic losses before the critic update and their generator losses before the generator update. Generator EMA and distributed-training plumbing are omitted.

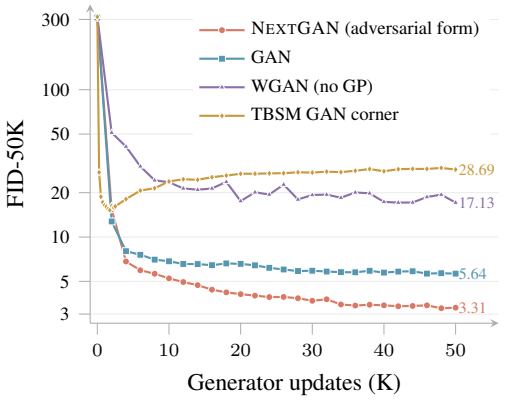
```
import torch


def nextgan_step(G, F, D, opt_g, opt_d, real, z, c,
                 n_critic=2):
    # F includes feature normalization; D returns one scalar per sample.
    # F and D must not couple different samples in a minibatch.
    F.requires_grad_(False)
    F.eval()
    D.requires_grad_(True)
    with torch.no_grad():
        xr, xp = F(real, c), F(G(z, c), c)

    for _ in range(n_critic):
        shape = (xr.shape[0],) + (1,) * (xr.ndim - 1)
        alpha = torch.rand(shape, device=xr.device, dtype=xr.dtype)
        xa = ((1 - alpha) * xp + alpha * xr).requires_grad_(True)
        grad = torch.autograd.grad(D(xa, c).sum(), xa,
                                   create_graph=True)[0]
        energy = grad.flatten(1).square().sum(1).mean()
        loss_d = D(xp, c).mean() - D(xr, c).mean() + 0.5 * energy
        opt_d.zero_grad(set_to_none=True)
        loss_d.backward()
        opt_d.step()

    D.requires_grad_(False)
    opt_g.zero_grad(set_to_none=True)
    # Keep input gradients through the frozen representer and critic.
    loss_g = -D(F(G(z, c), c), c).mean()
    loss_g.backward()
    opt_g.step()
    D.requires_grad_(True)
    return loss_d.detach(), loss_g.detach()
```

a



b

	MAE	SigLIP2	ResNet-18	FID-50K ↓
SINGLE	✓	–	–	7.08
	–	✓	–	17.78
	–	–	✓	5.55
PAIR	✓	✓	–	6.64
	✓	–	✓	4.67
	–	✓	✓	4.65
ALL	✓	✓	✓	**4.10**

Figure 7: **Training objective and feature space.** *(a)* JiT-B training curves through 50K updates (logarithmic FID-50K axis). *(b)* Frozen representer ablations on JiT-B at 20K updates. Lower is better.

B ABLATIONS: TRAINING OBJECTIVE AND FEATURE SPACE

We examine training objectives and feature spaces using JiT-B. Both the training curves in Fig. 7a and the representer ablation in panel b report FID-50K using the JiT evaluator.

Training objective. Fig. 7a shows NEXTGAN (adversarial form), logistic GAN, WGAN without gradient penalty, and the TBSM GAN corner in the JiT-B representation-space setting. At 50K updates, their FID-50K values are 3.31, 5.64, 17.13, and 28.69, respectively. The plotted TBSM curve combines a same-configuration independent run at 0–2.5K with the original run at 4–50K; at the overlapping 2K checkpoint, their FID-50K values are 16.26 and 16.01, respectively. Each curve includes the measured shared-initialization FID at step zero using 50K generated images. Lines connecting checkpoints are visual guides.

Feature space. Fig. 7b compares frozen representers after 20K updates. Individual representers yield FID-50K values from 5.55 to 17.78. Combining all three reaches 4.10, a 26% reduction from the best single representer, and also improves over every tested pair. These results support using complementary feature spaces to provide the critic with richer training signals.

C PROOF OF PROP. 1

Let $\mathbf{Y} = (\mathbf{X}_\alpha, \mathbf{C})$, let $\boldsymbol{\Delta} = \mathbf{X}_\mathrm{r} - \mathbf{X}_\mathrm{p}$, and write $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \mathbb{E}\langle \mathbf{v}_1(\mathbf{Y}), \mathbf{v}_2(\mathbf{Y}) \rangle$ for the inner product of $L^2(\mu_{\boldsymbol{\theta}})$.

Part 1. By definition, $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}[\boldsymbol{\Delta} \mid \mathbf{Y}]$. For any square-integrable field $\mathbf{w}$,

$$\frac{1}{2}\mathbb{E}\|\mathbf{w}(\mathbf{Y}) - \boldsymbol{\Delta}\|_2^2 = \frac{1}{2}\|\mathbf{w} - \mathbf{v}_{\boldsymbol{\theta}}\|^2 + \frac{1}{2}\mathbb{E}\|\boldsymbol{\Delta} - \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{Y})\|_2^2, \tag{9}$$

since the cross term $\mathbb{E}\langle \mathbf{w}(\mathbf{Y}) - \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{Y}), \mathbb{E}[\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\Delta} \mid \mathbf{Y}] \rangle$ vanishes. The second term does not depend on $\mathbf{w}$, so the unrestricted minimizer is $\mathbf{v}_{\boldsymbol{\theta}}$, and the minimizer over the closed subspace $\mathcal{G}$ is $\mathrm{Proj}_{\mathcal{G}}\, \mathbf{v}_{\boldsymbol{\theta}}$.

Part 2. For $u \in \mathcal{H}$, Eq. (4), Fubini's theorem, and the tower property give

$$\mathbb{E}\big[u(\mathbf{X}_\mathrm{r}, \mathbf{C}) - u(\mathbf{X}_\mathrm{p}, \mathbf{C})\big] = \mathbb{E}\langle \nabla_{\mathbf{x}} u(\mathbf{X}_\alpha, \mathbf{C}), \boldsymbol{\Delta} \rangle = \langle \nabla_{\mathbf{x}} u, \mathbf{v}_{\boldsymbol{\theta}} \rangle. \tag{10}$$

Hence the objective in Eq. (6) equals $\langle \mathbf{w}, \mathbf{v}_{\boldsymbol{\theta}} \rangle - \frac{1}{2}\|\mathbf{w}\|^2$ with $\mathbf{w} = \nabla_{\mathbf{x}} u$. This functional is continuous in $\mathbf{w}$, so its supremum over $\{\nabla_{\mathbf{x}} u : u \in \mathcal{H}\}$ equals its supremum over the closure $\mathcal{G}$, which is

a closed linear subspace because $\mathcal{H}$ is linear. Writing $\mathbf{w}^\star = \mathrm{Proj}_{\mathcal{G}}\, \mathbf{v}_{\boldsymbol{\theta}}$, we have $\langle \mathbf{w}, \mathbf{v}_{\boldsymbol{\theta}} \rangle = \langle \mathbf{w}, \mathbf{w}^\star \rangle$ for $\mathbf{w} \in \mathcal{G}$, and

$$\langle \mathbf{w}, \mathbf{w}^\star \rangle - \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\|\mathbf{w}^\star\|^2 - \frac{1}{2}\|\mathbf{w} - \mathbf{w}^\star\|^2, \tag{11}$$

which is maximized at $\mathbf{w} = \mathbf{w}^\star$. Therefore $\mathcal{D}_{\mathrm{Next}} = \frac{1}{2}\|\,\mathrm{Proj}_{\mathcal{G}}\, \mathbf{v}_{\boldsymbol{\theta}}\|^2 \geq 0$.

Part 3. If $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}) = Q(\cdot \mid \mathbf{c})$ for almost every $\mathbf{c}$, the joint law of $(\mathbf{X}_{\mathrm{r}}, \mathbf{X}_{\mathrm{p}}, \alpha, \mathbf{C})$ is invariant under $(\mathbf{X}_{\mathrm{r}}, \mathbf{X}_{\mathrm{p}}, \alpha) \mapsto (\mathbf{X}_{\mathrm{p}}, \mathbf{X}_{\mathrm{r}}, 1 - \alpha)$. This map leaves $\mathbf{X}_\alpha$ unchanged and negates $\boldsymbol{\Delta}$. Hence the conditional law of $\boldsymbol{\Delta}$ given $(\mathbf{X}_\alpha, \mathbf{C})$ is symmetric about zero, so $\mathbf{v}_{\boldsymbol{\theta}} = \mathbf{0}$ almost surely and $\mathcal{D}_{\mathrm{Next}} = 0$ by Part 2.

Conversely, suppose $\mathcal{D}_{\mathrm{Next}} = 0$ and fix $u \in \mathcal{H}$. Since $\mathcal{H}$ is linear, $tu \in \mathcal{H}$ for every $t \in \mathbb{R}$, and

$$0 \;\geq\; t\,\langle \nabla_{\mathbf{x}} u, \mathbf{v}_{\boldsymbol{\theta}} \rangle - \frac{t^2}{2}\|\nabla_{\mathbf{x}} u\|^2 \qquad \text{for all } t \in \mathbb{R}. \tag{12}$$

Letting $t \to 0^{\pm}$ gives $\langle \nabla_{\mathbf{x}} u, \mathbf{v}_{\boldsymbol{\theta}} \rangle = 0$, so by Eq. (10), $\mathbb{E}[u(\mathbf{X}_{\mathrm{r}}, \mathbf{C})] = \mathbb{E}[u(\mathbf{X}_{\mathrm{p}}, \mathbf{C})]$ for every $u \in \mathcal{H}$. Because $\mathcal{H}$ determines the joint law, these equalities imply equality of the real and generated joint laws of $(\mathbf{X}, \mathbf{C})$. The product test functions in the proposition are one sufficient choice of such a class. Since both share the marginal of $\mathbf{C}$, their conditional laws agree for almost every $\mathbf{c}$. The argument uses independent real–generated sampling and requires no optimal-transport coupling.

D Negative-Sobolev form (Rem. 1)

Fix $\mathbf{c}$ and write $\mu = \mu_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$. Let q and $p_{\boldsymbol{\theta}}$ be the densities of $Q(\cdot \mid \mathbf{c})$ and $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$, respectively, and let ω be the density of μ. For unrestricted u, the objective $J(u) = \int u(q - p_{\boldsymbol{\theta}})\,\mathrm{d}\mathbf{x} - \frac{1}{2}\int \|\nabla_{\mathbf{x}} u\|^2 \mathrm{d}\mu$ is concave. Its first variation in direction φ is $\int \varphi(q - p_{\boldsymbol{\theta}})\,\mathrm{d}\mathbf{x} - \int \langle \nabla_{\mathbf{x}} u, \nabla_{\mathbf{x}} \varphi \rangle\,\mathrm{d}\mu$, so a maximizer satisfies $-\nabla_{\mathbf{x}} \cdot (\omega \nabla_{\mathbf{x}} u^\star) = q - p_{\boldsymbol{\theta}}$ weakly. Testing this equation with $u^\star$ gives $\int u^\star(q - p_{\boldsymbol{\theta}})\,\mathrm{d}\mathbf{x} = \int \|\nabla_{\mathbf{x}} u^\star\|^2 \mathrm{d}\mu$, hence $J(u^\star) = \frac{1}{2}\int \|\nabla_{\mathbf{x}} u^\star\|^2 \mathrm{d}\mu = \frac{1}{2}\|q - p_{\boldsymbol{\theta}}\|^2_{\dot{H}^{-1}(\mu)}$, using the dual definition $\|q - p_{\boldsymbol{\theta}}\|_{\dot{H}^{-1}(\mu)} = \sup\{\int u(q - p_{\boldsymbol{\theta}})\,\mathrm{d}\mathbf{x} : \int \|\nabla_{\mathbf{x}} u\|^2 \mathrm{d}\mu \leq 1\}$.

E Protocol for the controlled 2D study

The released script `experiments/toy8_evolution_figure.py` uses a mixture of eight Gaussians on a circle (radius 2, standard deviation 0.16), three-hidden-layer SiLU MLP potential and generator networks (width 128), batch size 512, Adam with learning rate 2×10^{-4} for both networks, and two critic updates per generator update. The paired visualization in Fig. 3 uses seed 0 and 30K generator updates for each method. In that rerun, NextGAN covers all eight modes with effective mode count 7.995 and captured fraction 0.928; the matched GAN baseline has effective mode count 2.770 and captured fraction 0.933. The raw independent-pair cosine is not a field-recovery metric because the population target is a conditional average rather than an individual displacement. If reporting direction accuracy, we first compute the numerical conditional field on a 2D grid.

F Regularity of the field identity

For fixed conditioning variables and a differentiable potential, Eq. (4) is the fundamental theorem of calculus applied to the straight path in Eq. (1). The identity concerns the population field-regression objective and requires the same conditioning variables along each path. It does not require a pointwise optimal-transport coupling; the real–fake pairing is the one sampled by the training procedure.

G Additional Discussion

Throughout this appendix, we fix a condition $\mathbf{c}$ and suppress it in the notation; $P = P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$ and $Q = Q(\cdot \mid \mathbf{c})$. We use α for interpolation time and t for distributional evolution. This appendix separates derivation from verification. We first develop the geometric construction from quadratic

transport to the practical NEXTGAN objective and clarify its scope. App. H then states the formal assumptions and proves the conditional-regression, variational, differentiation, and minibatch results.

G.1 WASSERSTEIN GRADIENT FLOWS OF ENERGY FUNCTIONALS

We first place the NextGAN construction in the general Wasserstein-gradient-flow (WGF) framework (Han et al., 2026). Let P_t be an evolving probability measure on the comparison space and let Q be a fixed target measure. For an energy functional $\mathcal{F}(P;Q)$ with first variation $\delta\mathcal{F}/\delta P$, its formal WGF is

$$\partial_t P_t = \nabla_{\mathbf{x}} \cdot \left(P_t \nabla_{\mathbf{x}} \frac{\delta \mathcal{F}}{\delta P}(P_t; Q) \right), \tag{13}$$

or, equivalently, the continuity equation

$$\partial_t P_t + \nabla_{\mathbf{x}} \cdot (P_t \mathbf{v}_t) = 0, \qquad \mathbf{v}_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \frac{\delta \mathcal{F}}{\delta P}(P_t; Q)(\mathbf{x}). \tag{14}$$

When the differentiations are justified and there is no boundary flux, the energy decreases along the flow:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(P_t;Q) = -\int \left\| \nabla_{\mathbf{x}} \frac{\delta \mathcal{F}}{\delta P}(P_t;Q)(\mathbf{x}) \right\|^2 P_t(\mathrm{d}\mathbf{x}) \le 0. \tag{15}$$

Thus an energy specifies a distributional velocity field, while a numerical algorithm only approximates the corresponding flow.

G.1.1 THE NEXTGAN ENERGY AS A DISTRIBUTIONAL FUNCTIONAL

For a fixed nonnegative path-occupation measure μ, define the population NextGAN energy of a generated feature distribution P relative to a real feature distribution Q by

$$\mathcal{F}^{\mu}_{\mathrm{cor}}(P;Q) = \sup_{u \in \mathcal{H}_\mu} \left\{ \int u \,\mathrm{d}Q - \int u \,\mathrm{d}P - \frac{1}{2} \int \|\nabla_{\mathbf{x}} u(\mathbf{x})\|^2 \,\mu(\mathrm{d}\mathbf{x}) \right\}, \tag{16}$$

where $\mathcal{H}_\mu$ is the finite-energy space modulo the kernel of the seminorm $u \mapsto (\int \|\nabla_{\mathbf{x}} u\|^2 \mathrm{d}\mu)^{1/2}$. Under the assumptions in App. H, let $u^\star_P$ be an optimizer of Eq. (16). The envelope theorem gives the first variation with respect to the generated measure:

$$\frac{\delta \mathcal{F}^{\mu}_{\mathrm{cor}}}{\delta P}(P;Q)(\mathbf{x}) = -u^\star_P(\mathbf{x}) + \mathrm{const}, \tag{17}$$

where the additive constant is immaterial because probability-preserving perturbations have zero total mass. Substituting Eq. (17) into the WGF velocity in Eq. (14) yields

$$\mathbf{v}_t(\mathbf{x}) = \nabla_{\mathbf{x}} u^\star_{P_t}(\mathbf{x}). \tag{18}$$

Therefore, at the population level and for fixed μ, the NextGAN discriminator potential is the scalar potential whose gradient gives the WGF velocity of the generated comparison-space distribution. The generator loss $-\mathbb{E}_{\mathbf{x}\sim P_t}[u^\star_{P_t}(\mathbf{x})]$ moves generated particles in the same direction, since gradient descent on the particle energy $-u^\star_{P_t}$ produces $\dot{\mathbf{x}} = \nabla_{\mathbf{x}} u^\star_{P_t}(\mathbf{x})$.

For fixed μ, Eq. (15) specializes to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}^{\mu}_{\mathrm{cor}}(P_t;Q) = -\int \|\nabla_{\mathbf{x}} u^\star_{P_t}(\mathbf{x})\|^2 P_t(\mathrm{d}\mathbf{x}). \tag{19}$$

The energy term in Eq. (16) is instead integrated against the path measure μ. These two measures coincide only under an additional identification such as $\mu = P_t$; if μ changes with t, the geometry-variation term derived in App. H must also be included. The WGF interpretation is exact for the population NEXTGAN functional under the stated assumptions. It does not extend automatically to alternative logistic or WGAN objectives, or to finite-width alternating neural optimization.

G.2 Quadratic Wasserstein Geometry

G.2.1 Static and dynamic formulations

The Kantorovich definition is

$$W_2^2(P,Q) = \inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(\mathbf{x}_\mathrm{p},\mathbf{x}_\mathrm{r})\sim\pi}\|\mathbf{x}_\mathrm{r}-\mathbf{x}_\mathrm{p}\|^2. \tag{20}$$

Under the standard assumptions of the Benamou–Brenier formulation, this is equivalently

$$\frac{1}{2}W_2^2(P,Q) = \inf_{\rho_t,\mathbf{v}_t}\left\{\int_0^1\int \frac{1}{2}\rho_t(\mathbf{x})\|\mathbf{v}_t(\mathbf{x})\|^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \;\middle|\; \begin{array}{ll}\partial_t\rho_t+\nabla_\mathbf{x}\cdot(\rho_t\mathbf{v}_t)=0, & \\ \rho_{t=0}=P, & \rho_{t=1}=Q\end{array}\right\}. \tag{21}$$

The objective is the kinetic energy of the transporting velocity, and the constraints enforce mass conservation and the endpoint marginals (Benamou & Brenier, 2000). Introducing a multiplier $\Phi(t,\mathbf{x})$ for the continuity equation and integrating both the time-derivative and divergence terms by parts, the Lagrangian (up to fixed-endpoint boundary terms) becomes

$$\mathcal{L}(\rho,\mathbf{v},\Phi) = \int_0^1\int \rho_t(\mathbf{x})\left[\frac{1}{2}\|\mathbf{v}_t(\mathbf{x})\|^2 - \partial_t\Phi(t,\mathbf{x}) - \nabla_\mathbf{x}\Phi(t,\mathbf{x})^\top\mathbf{v}_t(\mathbf{x})\right]\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{22}$$

Stationarity with respect to $\mathbf{v}_t$ gives $\rho_t(\mathbf{v}_t-\nabla_\mathbf{x}\Phi)=0$, hence

$$\mathbf{v}_t(\mathbf{x}) = \nabla_\mathbf{x}\Phi(t,\mathbf{x}) \qquad \text{on the support of } \rho_t. \tag{23}$$

Substituting this relation into Eq. (22) and taking the first variation with respect to ρ_t gives the Hamilton–Jacobi equation

$$\partial_t\Phi(t,\mathbf{x}) + \frac{1}{2}\|\nabla_\mathbf{x}\Phi(t,\mathbf{x})\|^2 = 0 \qquad \text{where } \rho_t(\mathbf{x})>0. \tag{24}$$

If Φ is twice continuously differentiable, equality of mixed partials makes the corresponding velocity locally irrotational in any dimension:

$$\partial_{x_i}(\mathbf{v}_t)_j - \partial_{x_j}(\mathbf{v}_t)_i = 0, \qquad i,j=1,\dots,d. \tag{25}$$

Thus, the conservative velocity and the Hamilton–Jacobi potential follow from the dynamic OT first-order conditions. The potential is generally time-dependent; the curl-free statement is specific to the OT optimum and is not imposed on an arbitrary flow-matching field.

G.2.2 Kantorovich dual and the OT boundary

For completeness, the Kantorovich dual of the quadratic cost is

$$\frac{1}{2}W_2^2(P,Q) = \sup_{(\varphi,\psi)}\left\{\int\varphi\,\mathrm{d}P + \int\psi\,\mathrm{d}Q:\ \varphi(\mathbf{x}_\mathrm{p})+\psi(\mathbf{x}_\mathrm{r}) \le \frac{1}{2}\|\mathbf{x}_\mathrm{r}-\mathbf{x}_\mathrm{p}\|^2\right\}. \tag{26}$$

The constraint in Eq. (26) is imposed for every endpoint pair and the coupling is optimized implicitly by the duality theorem. In contrast, Eq. (33) uses one static potential, samples the fixed product coupling, and penalizes its gradient only on the resulting straight-path occupation measure. Consequently, replacing the OT coupling by $P\times Q$ changes the objective before any neural approximation is introduced.

G.3 Flow Matching for a Prescribed Coupling

G.3.1 Displacement paths

For any fixed coupling $\pi\in\Pi(P,Q)$, draw $(\mathbf{x}_\mathrm{p},\mathbf{x}_\mathrm{r})\sim\pi$ and define the straight path

$$\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_\mathrm{p} + \alpha\mathbf{x}_\mathrm{r}. \tag{27}$$

The induced law is $\rho_\alpha = \mathrm{Law}(\mathbf{X}_\alpha)$. If $\pi=\pi^\star$ is an optimal coupling, this law is the constant-speed W_2 geodesic and its Eulerian velocity satisfies

$$\mathbf{v}_\alpha^\star(\mathbf{x}) = \mathbb{E}_{\pi^\star}[\mathbf{X}_\mathrm{r}-\mathbf{X}_\mathrm{p}\mid\mathbf{X}_\alpha=\mathbf{x},\alpha], \qquad \int_0^1\mathbb{E}_{\mathbf{x}\sim\rho_\alpha}\|\mathbf{v}_\alpha^\star(\mathbf{x})\|^2\mathrm{d}\alpha = W_2^2(P,Q). \tag{28}$$

For an arbitrary coupling, optimality of W_2 gives

$$W_2^2(P,Q) \le \mathbb{E}_\pi \|\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}\|^2, \tag{29}$$

with equality only when the chosen coupling is optimal (up to non-uniqueness). This inequality records the distinction between optimizing the OT coupling and prescribing a coupling for flow matching. Under the product coupling used by NextGAN, the straight-path law serves as a tractable reference path distribution and need not coincide with the Wasserstein geodesic.

G.3.2 Conditional flow-matching regression

For a general prescribed coupling, define the conditional velocity $\mathbf{v}_\alpha^\pi(\mathbf{x}) = \mathbb{E}_\pi[\mathbf{X}_\mathrm{r} - \mathbf{X}_\mathrm{p} \mid \mathbf{X}_\alpha = \mathbf{x}, \alpha]$. For any measurable field $\mathbf{w}_\alpha$,

$$\mathbb{E}\|\mathbf{w}_\alpha(\mathbf{x}_\alpha) - (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p})\|^2 = \mathbb{E}\|\mathbf{w}_\alpha(\mathbf{x}_\alpha) - \mathbf{v}_\alpha^\pi(\mathbf{x}_\alpha)\|^2 + \mathbb{E}\|\mathbf{v}_\alpha^\pi(\mathbf{x}_\alpha) - (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p})\|^2, \tag{30}$$

Therefore, the flow-matching regression is minimized by $\mathbf{w}_\alpha = \mathbf{v}_\alpha^\pi$ almost everywhere. This statement is valid for any chosen coupling and does not require that the coupling be optimal.

G.4 From Flow Matching to the NextGAN Objective

At the fixed condition $\mathbf{c}$, $\mathcal{L}_{\mathrm{field}}(u)$ denotes the conditional version of Eq. (2) for a generic potential u. Under the product coupling $\pi = P \times Q$ and the conservative potential field described in App. G.1, the NextGAN flow-matching regression is

$$2\mathcal{L}_{\mathrm{field}}(u) = \mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha) - (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p})\|^2. \tag{31}$$

G.4.1 Algebraic reduction

Using $u(\mathbf{x}_\mathrm{r}) - u(\mathbf{x}_\mathrm{p}) = \int_0^1 \nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)^\top (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p})\, \mathrm{d}\alpha$ and uniform α, expansion of Eq. (31) gives

$$2\mathcal{L}_{\mathrm{field}}(u) = \mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)\|^2 - 2\mathbb{E}[u(\mathbf{x}_\mathrm{r}) - u(\mathbf{x}_\mathrm{p})] + C, \qquad C = \mathbb{E}\|\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}\|^2. \tag{32}$$

The constant C is independent of u. Therefore, minimizing Eq. (32) over the conservative field is equivalent to maximizing

$$\mathcal{V}(u) = \mathbb{E}[u(\mathbf{x}_\mathrm{r}) - u(\mathbf{x}_\mathrm{p})] - \frac{1}{2}\mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)\|^2. \tag{33}$$

Eq. (33) is the NextGAN objective. It is a distinct adversarial functional motivated by the quadratic-transport geometry and instantiated through flow matching, rather than an exact evaluation of W_2.

G.5 Scope and Relation to Optimal Transport

Exact quadratic OT optimizes the endpoint coupling and a density path under the continuity equation, with a time-dependent cost-conjugate potential (Villani, 2009; Benamou & Brenier, 2000). NextGAN instead fixes $P \times Q$, prescribes straight paths, and uses a static potential. These restrictions explain why Eq. (33) should be interpreted as a W_2-motivated adversarial objective rather than as an exact Wasserstein distance.

The formal variational characterization, well-posedness conditions, and implementation-level gradient rules are stated and proved in App. H.

H Formal Statements and Proofs

This section gives the formal results used by the construction in App. G. The results concern population objectives and their Monte Carlo estimators; they do not constitute a convergence theorem for finite-width neural-network alternating optimization.

H.1 Assumptions and endpoint conventions

Assumption 1 (NextGAN regularity and endpoint convention). *The comparison space is $\mathbb{R}^d$, and $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$, the set of probability measures with finite second moment. We use $\mathbf{x}_{\mathrm{p}} \sim P$ for a generated comparison vector and $\mathbf{x}_{\mathrm{r}} \sim Q$ for a real comparison vector. The product coupling $P \times Q$ is used in the practical objective, while $\pi^\star$ denotes an optimizer of the quadratic transport problem when it exists. Expectations involving a path also integrate an independent $\alpha \sim \mathcal{U}(0,1)$. The functions and vector fields appearing below are measurable and square-integrable under the stated path laws, and differentiation under an expectation is used only when the corresponding integrability and regularity conditions hold.*

The regularity conditions above are sufficient for the displayed formal calculations; the neural-network implementation is a finite-dimensional approximation and does not satisfy them automatically.

Proposition 2 (Conditional flow-matching decomposition). *Let $(\mathbf{x}_{\mathrm{p}}, \mathbf{x}_{\mathrm{r}}) \sim \pi$ have finite second moment, let $\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_{\mathrm{p}} + \alpha \mathbf{x}_{\mathrm{r}}$, and define $\mathbf{v}^\pi_\alpha(\mathbf{x}) = \mathbb{E}[\mathbf{X}_{\mathrm{r}} - \mathbf{X}_{\mathrm{p}} \mid \mathbf{X}_\alpha = \mathbf{x}, \alpha]$. For every square-integrable measurable field $\mathbf{w}_\alpha$,*

$$\mathbb{E}\|\mathbf{w}_\alpha(\mathbf{x}_\alpha) - (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}})\|^2 = \mathbb{E}\|\mathbf{w}_\alpha(\mathbf{x}_\alpha) - \mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha)\|^2 + \mathbb{E}\|\mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha) - (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}})\|^2. \tag{34}$$

Thus $\mathbf{v}^\pi_\alpha$ is the unique minimizer up to ρ_α-null sets for almost every α, where $\rho_\alpha = \mathrm{Law}(\mathbf{X}_\alpha)$.

Proof. Write $\mathbf{a} = \mathbf{w}_\alpha(\mathbf{x}_\alpha) - \mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha)$ and $\mathbf{b} = \mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha) - (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}})$. Expanding $\|\mathbf{a} + \mathbf{b}\|^2$ leaves the cross term $2\mathbb{E}[\mathbf{a}^\top \mathbf{b}]$. Conditioning on $(\mathbf{x}_\alpha, \alpha)$ gives

$$\mathbb{E}[\mathbf{a}^\top \mathbf{b} \mid \mathbf{x}_\alpha, \alpha] = \mathbf{a}^\top \mathbb{E}[\mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha) - (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}}) \mid \mathbf{x}_\alpha, \alpha] = 0,$$

by the definition of the conditional expectation. Taking expectations proves Eq. (34); the first term is nonnegative and vanishes exactly when $\mathbf{w}_\alpha(\mathbf{x}_\alpha) = \mathbf{v}^\pi_\alpha(\mathbf{x}_\alpha)$ almost everywhere. □

Proposition 3 (Algebraic reduction to the NextGAN value). *Assume $u \in C^1$ and that the terms below are integrable. Under any coupling with straight paths,*

$$\mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha) - (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}})\|^2 = \mathbb{E}\|\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)\|^2 - 2\mathbb{E}[u(\mathbf{x}_{\mathrm{r}}) - u(\mathbf{x}_{\mathrm{p}})] + \mathbb{E}\|\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}}\|^2. \tag{35}$$

Therefore minimizing the flow-matching regression over the conservative potential class used by the method is equivalent to maximizing Eq. (33).

Proof. Expand the square and use the fundamental theorem of calculus along each line segment:

$$u(\mathbf{x}_{\mathrm{r}}) - u(\mathbf{x}_{\mathrm{p}}) = \int_0^1 \nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)^\top (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}}) \, \mathrm{d}\alpha.$$

Since α is uniform, $\mathbb{E}[\nabla_{\mathbf{x}} u(\mathbf{x}_\alpha)^\top (\mathbf{x}_{\mathrm{r}} - \mathbf{x}_{\mathrm{p}})] = \mathbb{E}[u(\mathbf{x}_{\mathrm{r}}) - u(\mathbf{x}_{\mathrm{p}})]$. Substitution into the expanded square gives Eq. (35); the final displacement term is independent of u. □

H.2 Population variational characterization

For a fixed condition, let p, q be the generated and real feature densities and let μ be the occupation measure of the sampled paths, with density ω. Assume that $r = q - p$ defines a bounded linear functional on the finite-energy space and annihilates the kernel of its energy seminorm. The population objective is

$$\mathcal{V}_\mu(p, u) = \int u(\mathbf{x})(q(\mathbf{x}) - p(\mathbf{x})) \, \mathrm{d}\mathbf{x} - \frac{1}{2} \int \omega(\mathbf{x}) \|\nabla_{\mathbf{x}} u(\mathbf{x})\|^2 \, \mathrm{d}\mathbf{x}. \tag{36}$$

For a perturbation $u + \varepsilon v$, its first variation is

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{V}_\mu(p, u + \varepsilon v)\Big|_{\varepsilon=0} = \int v(\mathbf{x})(q(\mathbf{x}) - p(\mathbf{x})) \, \mathrm{d}\mathbf{x} - \int \omega(\mathbf{x}) \nabla_{\mathbf{x}} u(\mathbf{x}) \cdot \nabla_{\mathbf{x}} v(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{37}$$

Stationarity for every admissible test function v gives the weak Euler–Lagrange equation

$$-\nabla_{\mathbf{x}} \cdot (\omega \nabla_{\mathbf{x}} u^\star) = q - p. \tag{38}$$

To state the associated dual value, define the energy form

$$a_\mu(u_1, u_2) = \int \omega(\mathbf{x}) \nabla_{\mathbf{x}} u_1(\mathbf{x})^\top \nabla_{\mathbf{x}} u_2(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{39}$$

and assume that it is coercive after quotienting out its kernel (or after fixing a gauge such as zero mean). Define the dual seminorm

$$\|r\|_{\dot{H}^{-1}(\mu)} = \sup_{a_\mu(v,v)>0} \frac{\int r(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x}}{a_\mu(v,v)^{1/2}}. \tag{40}$$

Theorem 1 (Weighted Poisson characterization of the NextGAN value)**.** *Under the assumptions above, there is a unique equivalence class $u^\star$ satisfying*

$$a_\mu(u^\star, v) = \int r(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \qquad \textit{for every admissible } v, \tag{41}$$

which is the weak form of $-\nabla_{\mathbf{x}} \cdot (\omega \nabla_{\mathbf{x}} u^\star) = r$. Moreover,

$$\sup_u \left\{ \int u r \, \mathrm{d}\mathbf{x} - \frac{1}{2} a_\mu(u,u) \right\} = \frac{1}{2} a_\mu(u^\star, u^\star) = \frac{1}{2} \|r\|^2_{\dot{H}^{-1}(\mu)}. \tag{42}$$

Proof. Coercivity and boundedness allow the Riesz representation of the functional $v \mapsto \int r v \, \mathrm{d}\mathbf{x}$ in the energy inner product, yielding Eq. (41). For any admissible u, write $u = u^\star + w$. Using Eq. (41),

$$\int u r \, \mathrm{d}\mathbf{x} - \frac{1}{2} a_\mu(u,u) = \frac{1}{2} a_\mu(u^\star, u^\star) - \frac{1}{2} a_\mu(w,w) \le \frac{1}{2} a_\mu(u^\star, u^\star).$$

Equality holds exactly for w in the kernel, proving the first equality in Eq. (42). Finally, the Riesz representer identity gives

$$\|r\|^2_{\dot{H}^{-1}(\mu)} = a_\mu(u^\star, u^\star),$$

which proves the second equality. □

The theorem is a population statement for a fixed path geometry. It does not imply that alternating stochastic gradient updates find $u^\star$, nor that a finite neural discriminator realizes the whole energy space. It also does not identify the value with $W_2^2(P,Q)$, because the product coupling, straight paths, static potential, and occupation measure μ are fixed by the NextGAN construction.

H.3 Variation of the path geometry

Let s index endpoint densities p_s, q_s and occupation measures μ_s with densities ω_s. Write $r_s = q_s - p_s$, and suppose that u_s is an exact stationary maximizer of

$$\mathcal{V}(s,u) = \int u(\mathbf{x}) r_s(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \frac{1}{2} \int \omega_s(\mathbf{x}) \|\nabla_{\mathbf{x}} u(\mathbf{x})\|^2 \, \mathrm{d}\mathbf{x}. \tag{43}$$

Under differentiability and an envelope-theorem interchange of derivative and integral,

$$\frac{\mathrm{d}}{\mathrm{d}s} \mathcal{V}(s, u_s) = \int u_s(\mathbf{x}) \, \dot{r}_s(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \frac{1}{2} \int \dot{\omega}_s(\mathbf{x}) \|\nabla_{\mathbf{x}} u_s(\mathbf{x})\|^2 \, \mathrm{d}\mathbf{x}. \tag{44}$$

The terms involving $\dot{u}_s$ vanish because u_s is stationary in the discriminator variable. The second term is the geometry-variation contribution; its sign is determined by $\dot{\omega}_s$ and therefore is not fixed in general. Eq. (44) is why the fixed-geometry interpretation cannot be transferred unchanged to an adaptive path distribution.

H.4 Differentiation through a frozen representer

For a raw generated sample $\bar{\mathbf{x}}_{\mathrm{p}} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ and its comparison-space representation $\mathbf{x}_{\mathrm{p}} = \boldsymbol{f}(\bar{\mathbf{x}}_{\mathrm{p}}, \mathbf{c})$, the generator-side chain rule is

$$\nabla_{\boldsymbol{\theta}} u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}) = J_{\boldsymbol{g}_{\boldsymbol{\theta}}}(\mathbf{z}, \mathbf{c})^\top J_{\boldsymbol{f}}(\bar{\mathbf{x}}_{\mathrm{p}}, \mathbf{c})^\top \nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}). \tag{45}$$

Freezing $\boldsymbol{f}$ means that its parameters receive no update; it does not remove the input Jacobian $J_{\boldsymbol{f}}$ needed to propagate the discriminator signal to $\boldsymbol{g}_{\boldsymbol{\theta}}$. With multiple independent representation fields and scalar weights w_ℓ (unit weights for the summed head losses), linearity gives

$$\nabla_{\boldsymbol{\theta}} \sum_{\ell=1}^{L} w_\ell u_{\boldsymbol{\phi}_\ell}(\boldsymbol{f}_\ell(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{c}),\mathbf{c}),\mathbf{c}) = \sum_{\ell=1}^{L} w_\ell J_{\boldsymbol{g}_{\boldsymbol{\theta}}}^\top J_{\boldsymbol{f}_\ell}^\top \nabla_{\mathbf{x}_\ell} u_{\boldsymbol{\phi}_\ell}. \tag{46}$$

During the discriminator update, generated and real endpoint vectors are detached from $\boldsymbol{g}_{\boldsymbol{\theta}}$ and the representation-map parameters; during the generator update, discriminator parameters are held fixed while the chain rule above is retained.

H.5 Finite-sample alternating objectives

For a minibatch of size B, independently draw $\mathbf{c}_i \sim p_{\mathrm{data}}(\mathbf{c})$ and endpoint pairs $(\mathbf{x}_{\mathrm{p}}^i, \mathbf{x}_{\mathrm{r}}^i) \sim P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}_i) \times Q(\cdot \mid \mathbf{c}_i)$. Independently draw $\alpha_i \sim \mathcal{U}(0,1)$ and set $\mathbf{x}_{\alpha_i}^i = (1-\alpha_i)\mathbf{x}_{\mathrm{p}}^i + \alpha_i \mathbf{x}_{\mathrm{r}}^i$. The Monte Carlo discriminator loss corresponding to Eq. (33) is

$$\widehat{\mathcal{L}}_{\mathrm{D}}^{\mathrm{adv}} = \frac{1}{B}\sum_{i=1}^{B}\left[u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}^i,\mathbf{c}_i) - u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{r}}^i,\mathbf{c}_i) + \frac{1}{2}\left\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_{\alpha_i}^i,\mathbf{c}_i)\right\|^2\right]. \tag{47}$$

For this update, the endpoint vectors are treated as constants with respect to $\boldsymbol{g}_{\boldsymbol{\theta}}$ and the representation-map parameters, while $\mathbf{x}_{\alpha_i}^i$ remains a differentiable input of $u_{\boldsymbol{\phi}}$ so that the field energy can be evaluated. With discriminator parameters held fixed at $\boldsymbol{\phi}$, the generator loss is

$$\widehat{\mathcal{L}}_{\mathrm{gen}} = -\frac{1}{B}\sum_{i=1}^{B} u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}^i,\mathbf{c}_i). \tag{48}$$

For parameters fixed independently of the sampled minibatch, Eqs. (47)–(48) are unbiased estimators of their population counterparts before the nonlinear optimization update. The gradient of Eq. (48) is taken through $\mathbf{x}_{\mathrm{p}}^i = \boldsymbol{f}(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_i,\mathbf{c}_i),\mathbf{c}_i)$, whereas no gradient is propagated into $\boldsymbol{\phi}$ or the frozen representation-map parameters.

Proposition 4 (Unbiasedness of the minibatch objectives)**.** *Under Assump. 1, fix $\boldsymbol{\theta}, \boldsymbol{\phi}$ independently of the minibatch. If the endpoint pairs and interpolation times in Eq. (47) are independent and identically distributed, then*

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{\mathrm{D}}^{\mathrm{adv}}\right] = \mathcal{L}_{\mathrm{D}}^{\mathrm{adv}}, \qquad \mathbb{E}\left[\widehat{\mathcal{L}}_{\mathrm{gen}}\right] = \mathcal{L}_{\mathrm{gen}}. \tag{49}$$

Proof. Each summand in Eqs. (47) and (48) has the same expectation as the corresponding population integrand by the sampling rule and Assump. 1. Linearity of expectation and the factor $1/B$ then give Eq. (49). The statement concerns the objective estimator at fixed parameters; it does not assert unbiasedness of a finite-step optimizer after parameter updates. □

I Convergence of the NextGAN Particle Dynamics

The preceding sections identify the population NextGAN energy and its WGF velocity. We now analyze the comparison-space particle dynamics induced by that velocity, following the particle-to-continuum distinction used in distributional WGF analyses. This result is an idealized dynamics statement: the critic is represented by a velocity oracle up to an explicit approximation error, and the conclusion is not a convergence theorem for finite-width neural-network alternating optimization.

I.1 NextGAN velocity and empirical dynamics

Fix a nonnegative occupation measure μ and a real comparison-space distribution Q. For a generated comparison-space distribution P, let $u_{P,Q}^{\mu}$ be a maximizer of Eq. (16) and define

$$\mathbf{v}_{P,Q}^{\mu}(\mathbf{x}) := \nabla_{\mathbf{x}} u_{P,Q}^{\mu}(\mathbf{x}). \tag{50}$$

By Eq. (17), this is the formal WGF velocity of the NextGAN energy with respect to its generated marginal when μ is fixed. The corresponding continuum dynamics are

$$\partial_t P_t + \nabla_{\mathbf{x}} \cdot \Big(P_t \mathbf{v}^{\mu}_{P_t,Q} \Big) = 0, \qquad P_{t=0} = P_0. \tag{51}$$

Given generated comparison-space particles $\{\mathbf{x}_i^{(0)}\}_{i=1}^N$, a real empirical measure $\widehat{Q}_M = M^{-1}\sum_{j=1}^M \delta_{\mathbf{y}_j}$, and step size $\eta > 0$, define

$$\widehat{P}_k = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i^{(k)}}, \qquad \mathbf{x}_i^{(k+1)} = \mathbf{x}_i^{(k)} + \eta \widetilde{\mathbf{v}}_k(\mathbf{x}_i^{(k)}), \tag{52}$$

where $\widetilde{\mathbf{v}}_k$ is the velocity produced by the learned discriminator. The velocity residual appearing below is evaluated directly at the current particles, without introducing an additional error variable. Let $\widehat{P}_t^{\eta}$ denote the empirical measure obtained by linearly interpolating the particles between $t_k = k\eta$ and t_{k+1}.

I.2 Regularity assumption

Assumption 2 (NextGAN velocity regularity). *Write $m_2(R) := (\int \|\mathbf{x}\|_2^2 \, \mathrm{d}R(\mathbf{x}))^{1/2}$. There exist finite constants L_0, L_x, L_P, L_Q such that, for all comparison-space measures $P, \widetilde{P}, Q, \widetilde{Q}$ in the relevant bounded-moment class:*

$$\|\mathbf{v}^{\mu}_{P,Q}\|_{L^2(P)} \le L_0\big(1 + m_2(P) + m_2(Q)\big), \tag{53}$$

$$\|\mathbf{v}^{\mu}_{P,Q}(\mathbf{x}) - \mathbf{v}^{\mu}_{P,Q}(\mathbf{x}')\| \le L_x \|\mathbf{x} - \mathbf{x}'\|, \tag{54}$$

$$\|\mathbf{v}^{\mu}_{P,Q} - \mathbf{v}^{\mu}_{\widetilde{P},Q}\|_{L^2(\widetilde{P})} \le L_P W_2(P, \widetilde{P}), \tag{55}$$

$$\|\mathbf{v}^{\mu}_{P,Q} - \mathbf{v}^{\mu}_{P,\widetilde{Q}}\|_{L^2(P)} \le L_Q W_2(Q, \widetilde{Q}). \tag{56}$$

The initial comparison-space particles are i.i.d. from P_0, the real particles are i.i.d. from Q, and the velocity residuals evaluated at the current particles are square-integrable. The fixed measure μ and the admissible comparison-space domain are shared by the continuum and empirical dynamics.

Assump. 2 is the direct analogue of the velocity regularity conditions used for particle-dynamics convergence. For NextGAN, it is a condition to be checked for the selected representation geometry and discriminator class; spectral normalization and bounded comparison-space domains can support such a check, but do not by themselves imply all four inequalities.

I.3 Continuum and particle convergence

Theorem 2 (Convergence of NextGAN-induced particle dynamics). *Assume Assump. 2 and let $T > 0$. Set $L = L_x + L_P$ (with $(e^{LT} - 1)/L$ interpreted as T when $L = 0$) and $K_T = \lceil T/\eta \rceil$. Then Eq. (51) has a unique weak solution $P \in C([0,T]; \mathcal{P}_2(\mathbb{R}^d))$. There is a constant $C_T < \infty$, depending only on T, the constants in Assump. 2, and the shared moment bound, such that*

$$\begin{aligned} \sup_{t\in[0,T]} W_2\Big(\widehat{P}_t^{\eta}, P_t\Big) \le{}& e^{LT} W_2(\widehat{P}_0, P_0) + \frac{L_Q}{L}\big(e^{LT} - 1\big) W_2(\widehat{Q}_M, Q) \\ &+ C_T \eta + e^{LT} \eta \sum_{k=0}^{K_T - 1} \left(\frac{1}{N}\sum_{i=1}^N \Big\| \widetilde{\mathbf{v}}_k(\mathbf{x}_i^{(k)}) - \mathbf{v}^{\mu}_{\widehat{P}_k, \widehat{Q}_M}(\mathbf{x}_i^{(k)}) \Big\|^2 \right)^{1/2}. \end{aligned} \tag{57}$$

Consequently, if $N, M \to \infty$, $\eta \to 0$, and the explicitly displayed weighted residual sum converges to zero, then

$$\sup_{t\in[0,T]} W_2\Big(\widehat{P}_t^{\eta}, P_t\Big) \longrightarrow 0 \qquad \textit{almost surely}. \tag{58}$$

Proof. We separate the proof into the continuum stability, the empirical particle system, the Euler error, and the learned-velocity error.

Continuum stability. Let P_t and $\widetilde{P}_t$ solve Eq. (51) with initial measures $P_0, \widetilde{P}_0$ and target measures $Q, \widetilde{Q}$. Couple their characteristic curves using an optimal initial coupling. The spatial and source-measure bounds in Assump. 2, followed by Cauchy–Schwarz, give for the induced distance $D(t)$

$$\frac{\mathrm{d}}{\mathrm{d}t} D(t) \le L D(t) + L_Q W_2(Q, \widetilde{Q}).$$

Gronwall's inequality yields

$$W_2(P_t, \widetilde{P}_t) \le e^{Lt} W_2(P_0, \widetilde{P}_0) + \frac{L_Q}{L}\big(e^{Lt} - 1\big) W_2(Q, \widetilde{Q}). \tag{59}$$

Taking $\widetilde{P}_t$ to be the continuous-time empirical particle solution driven by $\mathbf{v}^{\mu}_{\widehat{P}_t, \widehat{Q}_M}$ gives the first two terms of Eq. (57).

Finite-dimensional well-posedness. For $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, equip $(\mathbb{R}^d)^N$ with $\|\mathbf{X}\|_{2,N}^2 = N^{-1} \sum_i \|\mathbf{x}_i\|^2$. Define

$$\mathbf{v}_N(\mathbf{X}) = \left(\mathbf{v}^{\mu}_{\widehat{P}_{\mathbf{X}}, \widehat{Q}_M}(\mathbf{x}_1), \dots, \mathbf{v}^{\mu}_{\widehat{P}_{\mathbf{X}}, \widehat{Q}_M}(\mathbf{x}_N) \right),$$

where $\widehat{P}_{\mathbf{X}} = N^{-1} \sum_i \delta_{\mathbf{x}_i}$. Matching $\mathbf{x}_i$ with $\mathbf{x}'_i$ gives $W_2(\widehat{P}_{\mathbf{X}}, \widehat{P}_{\mathbf{X}'}) \le \|\mathbf{X} - \mathbf{X}'\|_{2,N}$. Hence

$$\|\mathbf{v}_N(\mathbf{X}) - \mathbf{v}_N(\mathbf{X}')\|_{2,N} \le (L_x + L_P)\|\mathbf{X} - \mathbf{X}'\|_{2,N} = L\|\mathbf{X} - \mathbf{X}'\|_{2,N}.$$

The growth bound gives linear growth, so Picard–Lindelöf yields a unique global ODE solution for the continuous empirical particles. Its empirical measure is a weak solution of the corresponding empirical continuity equation.

Euler discretization. Let $\bar{\mathbf{X}}(t)$ denote the continuous empirical solution driven by the exact empirical velocity and let $\bar{\mathbf{X}}^{\eta}(t)$ be its piecewise-linear Euler interpolation. The global Lipschitz bound and the growth estimate imply the standard Euler estimate

$$\sup_{t \in [0,T]} \|\bar{\mathbf{X}}^{\eta}(t) - \bar{\mathbf{X}}(t)\|_{2,N} \le C_T \eta. \tag{60}$$

The coupling that matches particles with the same index gives the same bound for the associated empirical measures in W_2.

Learned-velocity error. Let $\mathbf{X}^{\eta}$ be the interpolation generated by Eq. (52) and compare it with $\bar{\mathbf{X}}^{\eta}$ one step at a time. Directly expanding the update and applying the spatial and source-measure Lipschitz bounds gives

$$\begin{aligned}\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\|_{2,N} &\le (1 + L\eta)\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\|_{2,N} \\ &\quad + \eta \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \widetilde{\mathbf{v}}_k(\mathbf{x}_i^{(k)}) - \mathbf{v}^{\mu}_{\widehat{P}_k, \widehat{Q}_M}(\mathbf{x}_i^{(k)}) \right\|^2 \right)^{1/2}.\end{aligned}$$

Discrete Gronwall therefore gives

$$\sup_{t \in [0,T]} \|\mathbf{X}^{\eta}(t) - \bar{\mathbf{X}}^{\eta}(t)\|_{2,N} \le e^{LT} \eta \sum_{k=0}^{K_T - 1} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \widetilde{\mathbf{v}}_k(\mathbf{x}_i^{(k)}) - \mathbf{v}^{\mu}_{\widehat{P}_k, \widehat{Q}_M}(\mathbf{x}_i^{(k)}) \right\|^2 \right)^{1/2}. \tag{61}$$

Combining Eqs. (59)–(61) with the triangle inequality proves Eq. (57). Finally, empirical measures formed from i.i.d. samples converge almost surely in W_2 under finite second moments, which gives Eq. (58). □

I.4 Interpretation for neural NextGAN

Thm. 2 applies to comparison-space particles driven by a NextGAN velocity oracle. A neural generator introduces an additional map-realization error. If its generated distribution P_k^{gen} satisfies

$$W_2\Big(P_{k+1}^{\mathrm{gen}}, (\mathrm{Id} + \eta \widetilde{\mathbf{v}}_k)_{\#} P_k^{\mathrm{gen}}\Big) \le \gamma_k, \tag{62}$$

then the same discrete Gronwall argument adds a term bounded by $C_T \sup_k \gamma_k / \eta$ (or by the corresponding weighted sum when the errors are nonuniform) to Eq. (57). Establishing a bound on γ_k requires a separate approximation theorem for the generator parameterization and its optimization procedure. Thus the particle result supplies a principled convergence target and an error decomposition for NextGAN, while full neural-training convergence remains outside the present theorem.

J LIMITATIONS

Representation and theoretical scope. The distribution-matching characterization in Prop. 1 concerns the chosen feature spaces and requires a sufficiently rich potential class. Matching these feature distributions does not by itself establish equality of the underlying image or video distributions when the representers discard information. The feature-space ablations in App. B also show that representer choice affects generation quality. Moreover, the population characterization and particle convergence result do not establish convergence of finite neural critics and generators under alternating optimization; controlling the generator realization error remains an open requirement.

Dependence on pretraining and reference data. Our large-scale image and video experiments fine-tune pretrained generators, and the video experiments use LoRA. They therefore do not establish that NEXTGAN can train comparably sized generators from scratch. The video reference samples are generated offline by Wan2.1-14B, as described in App. A. Although training requires no online teacher scores or trajectories, this setup still relies on a pretrained model to construct the reference distribution. Its effectiveness with directly collected real-video training data remains to be evaluated.

Video scaling and evaluation. The video results do not show uniform improvements across model sizes and sampling budgets. In Tab. 2, the 14B model remains below the strongest compared baseline in total VBench score at each tested budget. In addition, the reported 14B rows select the best completed run independently at each budget, so they do not establish the same performance from a single shared checkpoint. Further controlled experiments are needed to assess the consistency of gains at this scale.