FLOW-COUPLED DRIFTING FOR ONE-STEP GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generation produces samples with a single network evaluation, but training high-quality one-step generators remains challenging. Drifting Models train such generators with attraction and repulsion fields, yet estimating these fields requires many real and generated samples per condition and therefore large batches. With only one real sample, the positive, and one other generated sample, the negative, the normalized field reduces to the positive–negative difference, which ignores where the current output lies relative to the positive and the negative. We propose Flow-Coupled Drifting (FCD) to make such single-pair training effective. First, we scale the attraction and repulsion displacements separately according to their distances from the current output. The resulting force depends on the output, and its expectation is proportional to the formal 2-Wasserstein gradient-flow velocity of a smoothed energy distance in pretrained feature spaces. Second, to pair each generator input with its own positive, we form two inputs by interpolating the same positive with independent Gaussian noises at a shared time, and the generator maps them to the output and the negative. We also train at the pure-noise endpoint used for one-step inference. On class-conditional ImageNet 256×256, FCD adapts pretrained SiT models into one-step generators, achieves an FID of 1.38, and its performance improves consistently with model size. Ablations show that flow-coupled training accelerates convergence and remains effective with a batch size of 16. The code will be made publicly available.

1 INTRODUCTION

Generative models have shown strong performance in image and video synthesis (Ho et al., 2020; Song et al., 2020; Lipman et al., 2022; Liu et al., 2022). Both diffusion and flow-based models rely on iterative sampling for generation. Recent research has devoted substantial attention to few-step generation, especially one-step generation. Distillation methods (Salimans & Ho, 2022; Yin et al., 2024b; Zhou et al., 2024) rely on a pretrained teacher, while consistency and MeanFlow models (Song et al., 2023; Geng et al., 2026a) can suffer from unstable training or require Jacobian-vector products (JVPs). Drifting Models (Deng et al., 2026) take a different approach by training a one-step generator directly from an attraction-repulsion field between real and generated samples.

The drifting field is estimated from a batch of real samples, the positives, and generated samples, the negatives, through kernel-weighted expectations. The attraction term pulls each generated sample toward the positives, and the repulsion term pushes it away from the negatives. However, a reliable estimate of the field at a generated sample requires many positives and negatives of the same condition. For example, the main models of Deng et al. (2026) use 64 generated samples per class label and an effective batch size of 8192. This requirement limits the scalability of Drifting Models and their applicability to existing pretrained models. It also cannot be met when only one real sample is available for a condition. A naive approach to remove this requirement is to retain one positive and one negative per generated sample. However, each kernel factor is then divided out by its own normalizer. The field reduces to the positive–negative difference, whose direction no longer depends on the generated sample or on the kernel distance.

In this work, we propose *Flow-Coupled Drifting* (FCD), which has two designs. First, to address this degeneracy, we replace the normalized single-pair field with separately scaled attraction and repulsion displacements. The direction of the resulting force depends on the generated sample. Its

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



Figure 1: One-step samples generated by FCD-XL/2 on class-conditional ImageNet 256×256.

expectation is the formal 2-Wasserstein gradient-flow velocity of one half of a smoothed energy distance in pretrained feature spaces. Second, in pure-noise training, the positive is sampled independently of the generator input, so each attraction term points toward a randomly drawn real sample. To pair each generator input with its own positive, we also train the generator on linear interpolations between the positive and Gaussian noise, as in Flow Matching (Lipman et al., 2022; Liu et al., 2022). A time-conditioned generator maps two such interpolations of the same positive, with independent noises and a shared time, to the generated sample and the negative. Experiments show that this design accelerates convergence and improves generation quality. In expectation, the semi-gradient of the resulting loss with detached targets equals the gradient of an energy. This energy combines the smoothed energy distance at the pure-noise endpoint with conditional energies at earlier times. Sampling remains one-step from pure noise.

We evaluate FCD on ImageNet (Deng et al., 2009) 256×256. Initialized from a pretrained SiT (Ma et al., 2024), FCD achieves an FID of 1.38 for one-step generation and an FDr^6 of 6.16 under the metric proposed by FD-loss (Yang et al., 2026). Its performance improves consistently from B/2 to XL/2. Reducing our force to the positive–negative difference fails to produce reasonable results. FCD also remains effective with a batch size of 16. Our contributions are as follows:

- We show that, with one positive and one negative, the Drifting field degenerates to the positive–negative difference, whose direction does not depend on the generated sample. We address this degeneracy by separately scaling the attraction and repulsion, which makes the expected force a formal gradient-flow velocity of a smoothed energy distance.
- We propose Flow-Coupled Drifting, which pairs each generator input with its own positive by training on linear interpolations between the positive and Gaussian noise. The negative is generated from another interpolation of the same positive at the same time.
- With one positive and one negative per generated sample, Flow-Coupled Drifting adapts a pretrained SiT into a one-step generator with an FID of 1.38 on ImageNet 256×256. Its performance improves consistently with model size.

2 Related Work

Diffusion and flow-based models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Karras et al., 2022) define a forward noising process and learn the reverse-time denoising dynamics. Flow Matching and Rectified Flow (Lipman et al., 2022; Liu et al., 2022) instead learn velocity fields that transport a reference distribution toward the data distribution along predefined interpolation paths between noise and data. Latent diffusion (Rombach et al., 2022) and representation autoencoders (Zheng et al., 2026) operate in the latent space of a pretrained autoencoder, often with transformer backbones (Peebles & Xie, 2023; Ma et al., 2024). Despite their strong generation quality, these models typically require many sampling steps.

Flow-based few-step generation and distillation. One approach distills a pretrained multi-step diffusion model into a few-step model through progressive distillation (Salimans & Ho, 2022) or score distillation (Yin et al., 2024b;a; Zhou et al., 2024), often with auxiliary score networks, adversarial discriminators, or multi-stage training. Flow-map methods (Boffi et al., 2024) instead learn mappings between two times along the probability-flow trajectory, including consistency models (Song et al., 2023; Song & Dhariwal, 2024; Lu & Song, 2025; Kim et al., 2024; Peng et al., 2026) and average-velocity or shortcut models (Frans et al., 2025; Geng et al., 2026a;b). Some of these methods suffer from unstable training or require Jacobian-vector products (JVPs) (Lu & Song, 2025; Geng et al., 2026a).

One-step generator mappings. Generative Adversarial Networks (GANs) train one-step generators with adversarial gradients, but minimax optimization is often unstable (Goodfellow et al., 2014; Arjovsky et al., 2017). Drifting Models (Deng et al., 2026) require no discriminator and evolve the pushforward distribution during training rather than at inference. They displace generated samples by a kernel-induced attraction and repulsion field in a pretrained feature space. Under Gaussian kernels, recent analyses show that the drifting field equals the score difference between smoothed distributions (Lai et al., 2026; Turan et al., 2026). Other work connects drifting-style dynamics to Sinkhorn divergences and Wasserstein gradient flows (He et al., 2026; Cao et al., 2026; Han et al., 2026). Teacher-Feature Drifting (Zhang et al., 2026) uses the field to distill pretrained diffusion models and measures it in the teacher features. These works estimate the field from multiple real and generated samples. In contrast, we show that the Drifting field degenerates with one positive and one negative per generated sample, and we propose Flow-Coupled Drifting to address this degeneracy.

3 Method

3.1 Preliminaries

Flow Matching. Let p denote the data distribution, $\boldsymbol{y} \sim p$ be a clean latent, and $\boldsymbol{\epsilon} \sim p_\epsilon = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ be Gaussian noise. Flow Matching (Lipman et al., 2022; Liu et al., 2022) defines a probability path between data and noise. For the linear path used by Rectified Flow, the interpolant is:

$$\boldsymbol{z}_t = (1-t)\boldsymbol{y} + t\boldsymbol{\epsilon}, \qquad t \in [0,1], \tag{1}$$

where $t = 0$ corresponds to data and $t = 1$ to pure noise. The conditional velocity is the path derivative $\dot{\boldsymbol{z}}_t = \boldsymbol{\epsilon} - \boldsymbol{y}$, and the marginal velocity field is:

$$\boldsymbol{v}^\star(\boldsymbol{z}_t, t) = \mathbb{E}[\boldsymbol{\epsilon} - \boldsymbol{y} \mid \boldsymbol{z}_t, t]. \tag{2}$$

Flow Matching trains a time-conditioned velocity network to predict this field by regressing to the conditional velocity:

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{y},\boldsymbol{\epsilon}} \|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - (\boldsymbol{\epsilon} - \boldsymbol{y})\|_2^2. \tag{3}$$

Drifting Models. A Drifting Model (Deng et al., 2026) trains a one-step generator by moving its generated distribution toward the data distribution during training. Let $f_{\boldsymbol{\theta}}$ map noise $\boldsymbol{\epsilon} \sim p_\epsilon$ to a generated sample $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})$, and let $q_{\boldsymbol{\theta}} = (f_{\boldsymbol{\theta}})_{\#} p_\epsilon$ be the pushforward distribution. For a point $\boldsymbol{x}$, a real sample $\boldsymbol{y}^+ \sim p$, and a generated sample $\boldsymbol{y}^- \sim q_{\boldsymbol{\theta}}$, Drifting defines attraction and repulsion fields:

$$\boldsymbol{V}_p^+(\boldsymbol{x}) = \frac{\mathbb{E}_{\boldsymbol{y}^+\sim p}[k(\boldsymbol{x},\boldsymbol{y}^+)(\boldsymbol{y}^+ - \boldsymbol{x})]}{\mathbb{E}_{\boldsymbol{y}^+\sim p}[k(\boldsymbol{x},\boldsymbol{y}^+)]}, \qquad \boldsymbol{V}_{q_{\boldsymbol{\theta}}}^-(\boldsymbol{x}) = \frac{\mathbb{E}_{\boldsymbol{y}^-\sim q_{\boldsymbol{\theta}}}[k(\boldsymbol{x},\boldsymbol{y}^-)(\boldsymbol{y}^- - \boldsymbol{x})]}{\mathbb{E}_{\boldsymbol{y}^-\sim q_{\boldsymbol{\theta}}}[k(\boldsymbol{x},\boldsymbol{y}^-)]}. \tag{4}$$

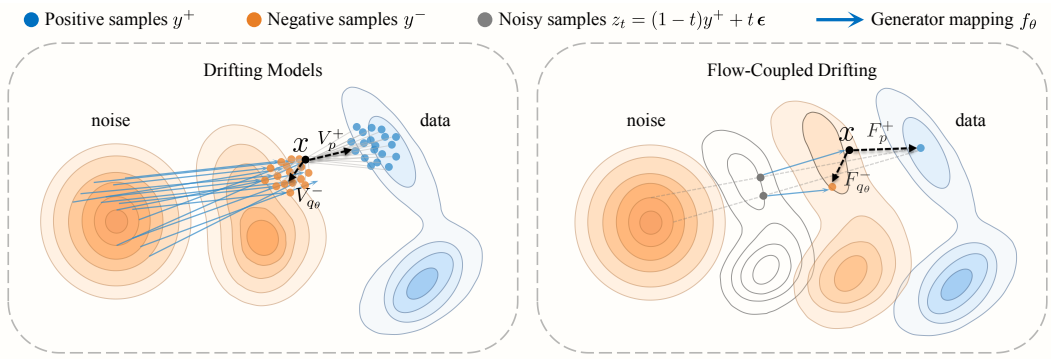


Figure 2: Overview of Flow-Coupled Drifting. **Left:** Drifting Models move each generated sample $\boldsymbol{x}$ by the attraction $\boldsymbol{V}_p^+$ toward many positives and the repulsion $\boldsymbol{V}_{q_\theta}^-$ from many negatives, both estimated with kernel weights as in Equation 4. **Right:** Flow-Coupled Drifting uses one positive $\boldsymbol{y}^+$ and one negative $\boldsymbol{y}^-$ per generated sample. The generator maps two noisy samples of $\boldsymbol{y}^+$ with independent noises and a shared time t to $\boldsymbol{x}$ and $\boldsymbol{y}^-$. The force on $\boldsymbol{x}$ is the attraction $\boldsymbol{F}_p^+$ minus the repulsion $\boldsymbol{F}_{q_\theta}^-$ in Equation 8. At $t = 1$, the noisy samples are pure noise, as at inference.

The drifting field is their difference, $\boldsymbol{V}_{p,q_\theta}(\boldsymbol{x}) = \boldsymbol{V}_p^+(\boldsymbol{x}) - \boldsymbol{V}_{q_\theta}^-(\boldsymbol{x})$. In practice, the kernel is typically instantiated as an exponential distance kernel:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2}{r}\right), \tag{5}$$

with radius $r > 0$, and the expectations are approximated by kernel-weighted averages over finite real and generated samples. The generator is trained by regressing its output to a detached drifted target:

$$\mathcal{L}_{\text{drift}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon}} \left\| f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}) - \text{sg}[f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}) + \boldsymbol{V}_{p,q_{\boldsymbol{\theta}}}(f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}))] \right\|_2^2. \tag{6}$$

Deng et al. (2026) show that the field vanishes when $q_{\boldsymbol{\theta}} = p$ and provide an identifiability heuristic for the converse. In practice, however, estimating the field at each generated sample requires many positive and negative samples of the same condition, and hence a large batch.

3.2 Single-pair degeneracy in Drifting

To remove this requirement, consider one positive $\boldsymbol{y}^+ \sim p$ and one negative $\boldsymbol{y}^- \sim q_{\boldsymbol{\theta}}$ for each generated sample $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})$. In this pure-noise baseline, the generated sample, the positive, and the negative are sampled independently, and the negative is a separate generated sample rather than $\boldsymbol{x}$ itself. With only one positive and one negative, Equation 4 reduces to:

$$\begin{aligned} \widehat{\boldsymbol{V}}(\boldsymbol{x}) &= \frac{k(\boldsymbol{x}, \boldsymbol{y}^+)(\boldsymbol{y}^+ - \boldsymbol{x})}{k(\boldsymbol{x}, \boldsymbol{y}^+)} - \frac{k(\boldsymbol{x}, \boldsymbol{y}^-)(\boldsymbol{y}^- - \boldsymbol{x})}{k(\boldsymbol{x}, \boldsymbol{y}^-)} \\ &= \boldsymbol{y}^+ - \boldsymbol{y}^-. \end{aligned} \tag{7}$$

Each kernel factor is divided out by its own normalizer, and the two $\boldsymbol{x}$ terms sum to zero. The implementation of Deng et al. (2026) instead normalizes the kernel jointly over positives and negatives, and with one positive and one negative for a generated sample it gives a positive multiple of $\boldsymbol{y}^+ - \boldsymbol{y}^-$ (Appendix B.1). In both cases, the direction of the field is that of $\boldsymbol{y}^+ - \boldsymbol{y}^-$. For Equation 7, the per-sample loss becomes $\ell_{\text{single}} = \|\boldsymbol{x} - \text{sg}[\boldsymbol{x} + \boldsymbol{y}^+ - \boldsymbol{y}^-]\|_2^2$, and its autodiff semi-gradient, computed with the target treated as a constant, is $-\nabla_{\boldsymbol{x}}^{\text{sg}} \ell_{\text{single}} = 2(\boldsymbol{y}^+ - \boldsymbol{y}^-)$. Thus, for a fixed positive–negative pair, the update direction is determined only by $\boldsymbol{y}^+ - \boldsymbol{y}^-$ and no longer depends on $\boldsymbol{x}$ or on the kernel distance. Even with the flow-coupled training of Section 3.4, training with this update fails to produce reasonable results, as shown by the positive–negative difference in Table 1(b).

3.3 SINGLE-PAIR FORCE AND LOSS

To restore the dependence of the update direction on $\boldsymbol{x}$, we replace the normalized single-pair Drifting field with separately scaled attraction and repulsion displacements:

$$\boldsymbol{F}_p^+(\boldsymbol{x};\boldsymbol{y}^+) = \frac{\boldsymbol{y}^+ - \boldsymbol{x}}{(\|\boldsymbol{y}^+ - \boldsymbol{x}\|_2^2 + \tau^2)^{\gamma/2}}, \qquad \boldsymbol{F}_{q_{\boldsymbol{\theta}}}^-(\boldsymbol{x};\boldsymbol{y}^-) = \frac{\boldsymbol{y}^- - \boldsymbol{x}}{(\|\boldsymbol{y}^- - \boldsymbol{x}\|_2^2 + \tau^2)^{\gamma/2}}, \tag{8}$$

where $0 < \gamma < 2$ and $\tau > 0$ is a smoothing parameter. The force is their difference, $\boldsymbol{F}(\boldsymbol{x};\boldsymbol{y}^+,\boldsymbol{y}^-) = \boldsymbol{F}_p^+(\boldsymbol{x};\boldsymbol{y}^+) - \boldsymbol{F}_{q_{\boldsymbol{\theta}}}^-(\boldsymbol{x};\boldsymbol{y}^-)$. The coefficients $(\|\boldsymbol{y}^\pm - \boldsymbol{x}\|_2^2 + \tau^2)^{-\gamma/2}$ depend on the distances from $\boldsymbol{x}$ to the positive and the negative, and they differ unless the two distances are equal. The two $\boldsymbol{x}$ terms therefore no longer sum to zero in general, and, unlike Equation 7, the direction of $\boldsymbol{F}$ depends on $\boldsymbol{x}$. The exponent γ controls how the magnitude of each term scales with distance. The boundary case $\gamma = 0$ recovers $\boldsymbol{y}^+ - \boldsymbol{y}^-$ in Equation 7. For $\gamma = 1$ and $\tau \to 0$, each term at a nonzero displacement is a unit vector.

For $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})$ and $\boldsymbol{y}^- = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}^-)$ with independent $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^- \sim p_\epsilon$ and $\boldsymbol{y}^+ \sim p$, we regress each generated sample to a new target:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}^-,\boldsymbol{y}^+} \left\| \boldsymbol{x} - \mathrm{sg}[\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x};\boldsymbol{y}^+,\boldsymbol{y}^-)] \right\|_2^2. \tag{9}$$

We next relate the force to a distributional objective. Let q denote the distribution of generated samples $\boldsymbol{x}$, and assume that q and p have finite second moments, so that they lie in the 2-Wasserstein space (Ambrosio et al., 2005) and all expectations below are finite. We define:

$$\mathcal{E}(q,p) = 2\mathbb{E}_{\boldsymbol{x}\sim q,\boldsymbol{y}\sim p}\psi(\|\boldsymbol{x}-\boldsymbol{y}\|_2) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\psi(\|\boldsymbol{x}-\boldsymbol{x}'\|_2) - \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim p}\psi(\|\boldsymbol{y}-\boldsymbol{y}'\|_2), \tag{10}$$

where $\boldsymbol{x}, \boldsymbol{x}' \sim q$ and $\boldsymbol{y}, \boldsymbol{y}' \sim p$ are independent, and:

$$\psi(d) = \frac{(d^2+\tau^2)^{1-\gamma/2} - \tau^{2-\gamma}}{2-\gamma}. \tag{11}$$

As $\tau \to 0$, $\psi(d) \to d^{2-\gamma}/(2-\gamma)$, so up to the factor $1/(2-\gamma)$, $\mathcal{E}$ reduces to the energy distance with exponent $2-\gamma$ (Székely & Rizzo, 2013), and $\gamma = 1$ gives the standard energy distance. We therefore refer to $\mathcal{E}$ as a smoothed energy distance. Moreover, $\mathcal{E}(q,p) \geq 0$ with equality if and only if $q = p$. For a fixed $\boldsymbol{x}$, with $\boldsymbol{y}^+ \sim p$ and $\boldsymbol{y}^- \sim q$ sampled independently, the expected force satisfies:

$$\mathbb{E}[\boldsymbol{F}(\boldsymbol{x};\boldsymbol{y}^+,\boldsymbol{y}^-) \mid \boldsymbol{x}] = -\nabla_{\boldsymbol{x}} \frac{\delta(\frac{1}{2}\mathcal{E})(q,p)}{\delta q}(\boldsymbol{x}), \tag{12}$$

where $\delta/\delta q$ denotes the first variation with respect to q. Formally, the velocity of the 2-Wasserstein gradient flow of a functional of q is the negative spatial gradient of its first variation (Ambrosio et al., 2005; Han et al., 2026). The 2-Wasserstein gradient flow of the squared maximum mean discrepancy has a velocity of the same form (Arbel et al., 2019). By Equation 12, $\boldsymbol{F}$ is therefore an unbiased estimator of this formal velocity for $\frac{1}{2}\mathcal{E}(q,p)$ viewed as a functional of q. In parameter space, the semi-gradient of Equation 9 equals $\nabla_{\boldsymbol{\theta}}\mathcal{E}(q_{\boldsymbol{\theta}},p)$ (Proposition 5). Appendices B.2 and B.3 provide the proofs.

3.4 FLOW-COUPLED TRAINING

The modified force addresses the single-pair degeneracy, but pure-noise training still requires the generator to learn the entire noise-to-data map from inputs at $t = 1$ only. At $t = 1$, the positive is sampled independently of the input, so each attraction term points toward a randomly drawn real sample. We therefore also train on inputs at times $t < 1$ on the path in Equation 1, which are constructed from their positives. The attraction term then relates each input to its own positive, which accelerates convergence, as shown in Figure 3. Only the input sampling changes, and the force and loss remain the same.

Let $f_{\boldsymbol{\theta}}(\boldsymbol{z},t)$ be the time-conditioned extension of the generator, with $f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon},1) = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})$, so its pure-noise output distribution remains $q_{\boldsymbol{\theta}}$. For each generated sample, with probability $p_{\mathrm{flow}} \in [0,1]$ we sample $t \sim \mathcal{U}[0,1)$, and otherwise we set $t = 1$. For a shared positive $\boldsymbol{y}^+ \sim p$ and independent

Algorithm 1 Training step of Flow-Coupled Drifting

Require: Generator $f_{\boldsymbol{\theta}}$, batch size B
Require: $p_{\text{flow}} \in [0,1]$, $\gamma \in (0,2)$, $\tau > 0$
1: **for** $i = 1, \dots, B$ **do**
2: Sample a positive $\boldsymbol{y}_i^+ \sim p$ and $\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i^- \overset{\text{i.i.d.}}{\sim} p_\epsilon$
3: With probability p_{flow}, sample $t_i \sim \mathcal{U}[0,1)$, otherwise set $t_i \leftarrow 1$ ▷ $t = 1$: pure noise
4: $\boldsymbol{z}_i \leftarrow (1-t_i)\boldsymbol{y}_i^+ + t_i\boldsymbol{\epsilon}_i, \quad \boldsymbol{z}_i^- \leftarrow (1-t_i)\boldsymbol{y}_i^+ + t_i\boldsymbol{\epsilon}_i^-$
5: $\boldsymbol{x}_i \leftarrow f_{\boldsymbol{\theta}}(\boldsymbol{z}_i, t_i), \quad \boldsymbol{y}_i^- \leftarrow \text{sg}[f_{\boldsymbol{\theta}}(\boldsymbol{z}_i^-, t_i)]$
6: $\boldsymbol{F}_i \leftarrow \boldsymbol{F}(\boldsymbol{x}_i; \boldsymbol{y}_i^+, \boldsymbol{y}_i^-)$ ▷ Equation 8
7: **end for**
8: $\mathcal{L}_{\text{FCD}} \leftarrow \frac{1}{B}\sum_{i=1}^{B} \|\boldsymbol{x}_i - \text{sg}[\boldsymbol{x}_i + \boldsymbol{F}_i]\|_2^2$
9: Update $\boldsymbol{\theta}$ using $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{FCD}}$

$\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^- \sim p_\epsilon$, let $\boldsymbol{z}_t$ and $\boldsymbol{z}_t^-$ be the interpolants in Equation 1 with $\boldsymbol{y} = \boldsymbol{y}^+$ and noises $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}^-$, respectively. The generated sample and the negative are:

$$\boldsymbol{x}_t = f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t), \qquad \boldsymbol{y}_t^- = f_{\boldsymbol{\theta}}(\boldsymbol{z}_t^-, t). \tag{13}$$

Figure 2 illustrates this construction. Substituting $\boldsymbol{x}_t$ and $\boldsymbol{y}_t^-$ into Equation 9 gives the training loss, which we use instead of the Flow Matching loss in Equation 3:

$$\mathcal{L}_{\text{FCD}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{y}^+,\boldsymbol{\epsilon},\boldsymbol{\epsilon}^-} \left\| \boldsymbol{x}_t - \text{sg}[\boldsymbol{x}_t + \boldsymbol{F}(\boldsymbol{x}_t; \boldsymbol{y}^+, \boldsymbol{y}_t^-)] \right\|_2^2. \tag{14}$$

For $p_{\text{flow}} = 0$, training reduces to the pure-noise training of Section 3.3. For $p_{\text{flow}} = 1$, the inference endpoint receives no direct supervision.

The same construction also admits an energy interpretation. For $t < 1$, the generated sample depends on its positive, so Equation 12 does not apply with the data distribution p. Given the positive $\boldsymbol{y}^+$ and the time t, however, the generated sample and the negative are independent draws from the same conditional distribution $q_{\boldsymbol{\theta},t}^{\boldsymbol{y}^+}$, which is the distribution of $f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ over the noise. If this conditional distribution also has finite second moments, Equation 12 applies with q replaced by it and p replaced by the Dirac distribution $\delta_{\boldsymbol{y}^+}$ at the positive. This leads to the energy:

$$\mathcal{J}(\boldsymbol{\theta}) = (1 - p_{\text{flow}})\, \mathcal{E}(q_{\boldsymbol{\theta}}, p) + p_{\text{flow}}\, \mathbb{E}_{t\sim\mathcal{U}[0,1)} \mathcal{A}_t(\boldsymbol{\theta}), \qquad \mathcal{A}_t(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}^+\sim p}\, \mathcal{E}\big(q_{\boldsymbol{\theta},t}^{\boldsymbol{y}^+}, \delta_{\boldsymbol{y}^+}\big), \tag{15}$$

where $\mathcal{E}(q, \delta_{\boldsymbol{y}^+})$ is Equation 10 with p replaced by $\delta_{\boldsymbol{y}^+}$. At the pure-noise endpoint, the conditional distribution equals $q_{\boldsymbol{\theta}}$, and Appendix B.4 shows that $\mathcal{A}_1$ differs from $\mathcal{E}(q_{\boldsymbol{\theta}}, p)$ only by a constant independent of $\boldsymbol{\theta}$. Hence, up to a constant, $\mathcal{J}$ is the conditional energy $\mathcal{A}_t$ averaged over the time distribution above. Under the standard conditions allowing differentiation under the expectation (Mohamed et al., 2020), Appendix B.4 proves:

$$\nabla_{\boldsymbol{\theta}}^{\text{sg}} \mathcal{L}_{\text{FCD}} = \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}), \tag{16}$$

where $\nabla_{\boldsymbol{\theta}}^{\text{sg}}$ is the autodiff semi-gradient with detached targets. Thus, the minibatch semi-gradient of Equation 14 is an unbiased estimate of $\nabla_{\boldsymbol{\theta}}\mathcal{J}$. The first term of Equation 15 depends only on the output at $t = 1$, and the second only on the outputs at $t < 1$. Hence, for $p_{\text{flow}} < 1$, every minimizer of $\mathcal{J}$ over all measurable time-conditioned generators has a pure-noise output distribution equal to p (Appendix B.5).

Feature spaces and class conditioning. Following Deng et al. (2026), we compute the force and loss on the embedded samples $\phi(\boldsymbol{x})$, $\phi(\boldsymbol{y}^+)$, and $\phi(\boldsymbol{y}^-)$ of a pretrained feature encoder ϕ, while gradients are backpropagated through ϕ to update the generator. With multiple encoders, the force and loss are computed separately in each feature space, and the losses are summed (Appendix A). The results above then hold for each encoder, with every sample replaced by its embedding and $f_{\boldsymbol{\theta}}$ replaced by $\phi \circ f_{\boldsymbol{\theta}}$. In particular, $\mathcal{E}$ and its gradient-flow velocity are defined in the feature space, and $\mathcal{E}(\phi_\# q_{\boldsymbol{\theta}}, \phi_\# p) = 0$ implies $\phi_\# q_{\boldsymbol{\theta}} = \phi_\# p$. Accordingly, the minimizers above satisfy $\phi_\# q_{\boldsymbol{\theta}} = \phi_\# p$. For class-conditional generation, the generator is conditioned on the class label of the positive for both inputs, all distributions are conditioned on this label, and the results above hold for each class.

Algorithm 1 summarizes training, and inference uses a single evaluation $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, 1)$.

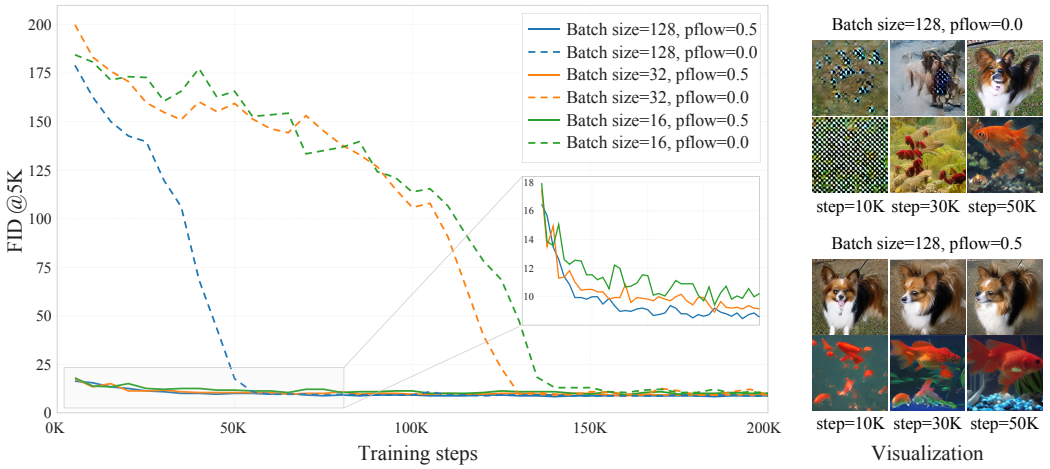


Figure 3: Effect of flow-coupled training. **Left:** FID evaluated on 5K generated samples over training steps for different batch sizes with $p_{\text{flow}} = 0.5$ and $p_{\text{flow}} = 0.0$. The inset enlarges the curves with $p_{\text{flow}} = 0.5$. **Right:** One-step visualizations after 10K, 30K, and 50K training steps with batch size 128 for $p_{\text{flow}} = 0.0$ and $p_{\text{flow}} = 0.5$.

4 Experiments

Experimental setting. We conduct our experiments on class-conditional ImageNet (Deng et al., 2009) generation at 256×256 resolution. We evaluate Fréchet Inception Distance (FID) (Heusel et al., 2017) on 50K generated images and study one-step generation by default. We also report Inception Score (IS) (Salimans et al., 2016) and FDr^6 (Yang et al., 2026). Following recent one-step diffusion and flow acceleration methods (Peng et al., 2026; Wang et al., 2026), we use a pretrained VAE tokenizer (Rombach et al., 2022) to represent images as latents. We initialize each model from a pretrained SiT (Ma et al., 2024) of the same size and train it into a one-step generator with Flow-Coupled Drifting for 100 epochs.

In the ablation study, each variant is initialized from SiT-XL/2 and trained for 20 epochs. By default, we use $p_{\text{flow}} = 0.5$, $\gamma = 1.0$, batch size 128, and both the MAE-B (He et al., 2022) and SigLIP-B (Tschannen et al., 2025) encoders.

4.1 Ablation Study

Flow-coupled training. Our method reduces to pure-noise training when $p_{\text{flow}} = 0$, where each attraction term points toward a randomly drawn real sample. Table 1(a) compares different values of p_{flow}. Training only on inputs at $t < 1$ already outperforms pure-noise training, although the generator never sees the pure-noise inputs used at inference. Mixing the two kinds of inputs further improves the results. Inputs at $t < 1$ relate each input to its own positive, while inputs at $t = 1$ match those used at inference, and training benefits from both. The best FID is achieved at $p_{\text{flow}} = 0.5$, and $p_{\text{flow}} = 0.2$ performs almost as well. Figure 3 further shows that flow coupling substantially accelerates convergence, and with $p_{\text{flow}} = 0.5$ the generator already produces recognizable samples after 10K training steps.

Exponent γ. The exponent γ determines how each term of Equation 8 scales with the distance to $\boldsymbol{x}$, and $\gamma = 0$ reduces the force to the degenerate positive–negative difference of Section 3.2. Table 1(b) compares different values of γ. With $\gamma = 0$, training fails to produce reasonable results even with flow coupling. This is consistent with the corresponding energy, which reduces to the squared distance between the means of the generated and data distributions. All positive values of γ yield meaningful one-step results, consistent with separately scaling the attraction and repulsion, which restores the dependence of the update direction on $\boldsymbol{x}$. The value $\gamma = 1.0$ achieves the best result, and $\gamma = 1.5$ performs almost as well.

Table 1: Ablations of flow-coupled training and the exponent γ. (a) p_{flow} is varied at $\gamma = 1.0$, and the endpoint probability $1 - p_{\text{flow}}$ is the probability of $t = 1$. (b) γ is varied at $p_{\text{flow}} = 0.5$, where d is the distance from $\boldsymbol{x}$ to the positive or the negative, and the rates hold when $d \gg \tau$. All rows use one positive and one negative per generated sample and one-step sampling at $t = 1$.

(a) Flow-coupled training

p_{flow}	endpoint probability	FID ↓
1.0	0.00	2.54
0.5	0.50	2.10
0.2	0.80	2.12
0.0	1.00	2.89

(b) Exponent γ

γ	force	FID ↓
0.0	positive–negative difference	179.08
0.5	magnitude grows as $d^{1/2}$	2.37
1.0	smoothed unit displacements	2.10
1.5	magnitude decays as $d^{-1/2}$	2.13

Table 2: Encoder ablation at $p_{\text{flow}} = 0.5$ and $\gamma = 1.0$.

encoder	FID ↓
MAE-B	6.87
SigLIP-B	7.42
MAE-B + SigLIP-B	2.10

Table 3: Positive-only objective at $\gamma = 1.0$, with the repulsion term removed from the force.

p_{flow}	inputs	related work	FID ↓
1.0	$t < 1$ only	PFM	34.87
0.5	mixed	–	41.49
0.0	$t = 1$ only	perceptual loss	54.15

Feature encoder. Following FD-loss (Yang et al., 2026), we compare forces computed in single and multiple feature spaces. Table 2 reports the results. Each single encoder yields meaningful one-step results, while combining MAE-B and SigLIP-B is substantially more effective than using either one alone. This is consistent with the summed energy, which vanishes only when the embedding distributions match under both encoders.

Is the negative necessary? We remove the repulsion term $\boldsymbol{F}^{-}_{q_\theta}$ from Equation 8 at $\gamma = 1.0$ and keep the rest of Equation 14 unchanged. This positive-only objective relates to two existing works. (i) With inputs at $t = 1$ only, it resembles a perceptual loss (Zhang et al., 2018), except that the real sample is drawn independently of the noise. (ii) With inputs at $t < 1$ only, it pulls each output toward the positive from which its input is constructed, as in Perceptual Flow Matching (PFM) (Zhao et al., 2026). As shown in Table 3, all settings yield FIDs above 34, compared with 2.10 to 2.89 with the negative in Table 1(a). Without the repulsion, the expected force on a generated sample $\boldsymbol{x}$ at $t = 1$ is $-\nabla_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}\sim p} \psi(\|\boldsymbol{x} - \boldsymbol{y}\|_2)$. For $\gamma = 1$, this function is strictly convex in $\boldsymbol{x}$, so the attraction alone moves every generated sample toward the same minimizer rather than toward p. With the repulsion, the full energy in Equation 10 vanishes only at $q = p$, so the negative is necessary.

Training batch size. The force itself always uses one pair per generated sample, independent of the training batch size. Table 4 varies the batch size, and Figure 3 shows the corresponding training curves with and without flow coupling. With flow coupling, all batch sizes reach a low FID early in training, and smaller batches only slightly increase the final FID. Thus, Flow-Coupled Drifting remains effective at small batch sizes.

Table 4: Batch size ablation.

batch size	FID ↓
128	2.10
32	2.23
16	2.41

Scalability. Table 5 reports our method at three model sizes, B/2, L/2, and XL/2, each initialized from a pretrained SiT of the same size and trained for 100 epochs. FID, IS, and FDr^6 all improve consistently as the model size increases. Consistent with the behavior of Transformer-based diffusion and flow models (Peebles & Xie, 2023; Ma et al., 2024), our method exhibits promising scalability for one-step generation.

4.2 Comparison with Prior Work

ImageNet 256×256 comparisons. Table 5 compares our method with multi-step diffusion and flow models, GANs, one-step diffusion and flow models, and Drifting Models. Unlike Drifting Models, which train the generator from scratch, FCD is initialized from a pretrained SiT. In one-

Table 5: Class-conditional generation on ImageNet 256×256. "$\times 2$" indicates that CFG doubles the NFE per sampling step. Bold indicates the best result among one-step methods.

Method	Params	NFE	FID ↓	IS ↑	FDr^6 ↓
Multi-step Diffusion/Flows					
ADM-G (Dhariwal & Nichol, 2021)	554M	250×2	4.59	186.7	–
DiT-XL/2 (Peebles & Xie, 2023)	675M	250×2	2.27	278.2	–
SiT-XL/2 (Ma et al., 2024)	675M	250×2	2.12	256.7	8.44
LightningDiT-XL (Yao et al., 2025)	675M	250×2	1.42	294.3	4.57
RAE-XL (Zheng et al., 2026)	839M	50×2	1.16	261.0	3.26
JiT-H/16 (Li & He, 2026)	953M	100×2	1.97	296.0	7.66
PixelDiT-XL (Yu et al., 2026)	797M	100×2	1.61	292.7	–
GANs					
BigGAN (Brock et al., 2018)	112M	1	6.95	152.8	–
GigaGAN (Kang et al., 2023)	569M	1	3.45	225.5	–
StyleGAN-XL (Sauer et al., 2022)	166M	1	2.30	265.1	–
Single-step Diffusion/Flows					
IMM-XL/2 (Zhou et al., 2025)	675M	1×2	7.77	–	–
iCT-XL/2 (Song & Dhariwal, 2024)	675M	1	34.24	–	–
Shortcut-XL/2 (Frans et al., 2025)	675M	1	10.60	–	–
MeanFlow-XL/2 (Geng et al., 2026a)	676M	1	3.43	–	–
TiM-XL/2 (Wang et al., 2026)	664M	1	3.26	210.3	–
iMF-XL/2 (Geng et al., 2026b)	610M	1	1.82	278.9	8.39
pMF-H (Lu et al., 2026)	935M	1	2.29	267.2	6.87
FACM (Peng et al., 2026)	675M	1	1.70	295.0	–
Drifting Models					
Drift-L/2 (latent) (Deng et al., 2026)	463M	1	1.53	257.2	10.92
Drift-L/16 (pixel) (Deng et al., 2026)	465M	1	1.43	**305.8**	10.51
FCD-B/2	130M	1	2.45	255.8	9.76
FCD-L/2	458M	1	1.72	264.7	7.41
FCD-XL/2	675M	1	**1.38**	277.5	**6.16**

step generation, FCD-XL/2 achieves an FID of 1.38 and an FDr^6 of 6.16. This FID is the lowest among the one-step methods in Table 5, and is also lower than that of its initialization, the multi-step SiT-XL/2 (FID 2.12), which has an NFE of 250×2. Figure 1 shows one-step samples generated by FCD-XL/2.

5 Conclusion

We present Flow-Coupled Drifting, a framework that trains one-step generators with one positive and one negative per generated sample. Its expected force is proportional to the formal 2-Wasserstein gradient-flow velocity of a smoothed energy distance, and flow coupling pairs each generator input with its own positive. This removes the need for many samples of each condition and a large batch, and allows existing pretrained multi-step models to be adapted into one-step generators. We hope that our work will extend Drifting Models to settings with only one real sample per condition, such as text-to-image generation.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, including drafting and revising parts of the manuscript to improve clarity and readability. We did not use generative AI tools to propose research ideas, develop the method, design or run experiments, generate data, or create figures. All AI-assisted text was reviewed and revised by the authors, and all results reported in this paper come from the authors' own experiments. We take responsibility for the final content of this work, including all text and claims.

ETHICS STATEMENT

This work studies generative modeling. Generative models can be misused to produce deceptive content, and we encourage responsible deployment of the methods presented here. Our experiments use publicly available datasets and do not involve human subjects.

REPRODUCIBILITY STATEMENT

Algorithm 1 summarizes the training procedure of Flow-Coupled Drifting, and Sections 3.3 and 3.4 define the force, the time sampling, and the training loss. Section 4 describes the dataset, the evaluation metrics, and the ablation settings, and Appendix A describes the latent representation, the initialization, the feature encoders, and the loss normalization across encoders. Table 6 lists all hyperparameters for the ablations and for each model size. For the theoretical results, the assumptions are stated in each proposition, and the complete proofs are given in Appendices B.2 to B.5. We will release the source code and trained models upon publication.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 32, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. pmlr, 2017.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via wasserstein gradient flows of kde-approximated divergences. *arXiv preprint arXiv:2603.10592*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026b.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 2002.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10134. IEEE, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of score-based and drifting models. *arXiv preprint arXiv:2603.07514*, 2026.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

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*, pp. 23–40. Springer, 2024.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. Facm: Flow-anchored consistency models. In *International Conference on Learning Representations*, volume 2026, pp. 7520–7543, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

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

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29178–29189, 2026.

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

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

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

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

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yuan Zhang, Chenyi Li, Haodong Yu, Guoqing Ma, Jiajun Zha, Yuanming Yang, Bo Wang, Wei Tang, Wenbo Li, Haoyang Huang, et al. Teacher-feature drifting: One-step diffusion distillation with pretrained diffusion representations. *arXiv preprint arXiv:2605.07327*, 2026.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling. *arXiv preprint arXiv:2607.03524*, 2026.

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

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. *arXiv preprint arXiv:2503.07565*, 2025.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

APPENDIX

A IMPLEMENTATION DETAILS

Table 6: Configurations on ImageNet 256×256. The ablation column is used in Section 4.1, where a table reference indicates that the entry is varied in that table and otherwise follows the default in Section 4. The other columns correspond to Table 5.

configs	ablation	B/2	L/2	XL/2
Generator architecture				
initialization	SiT-XL/2	SiT-B/2	SiT-L/2	SiT-XL/2
params (M)	675	130	458	675
depth	28	12	24	28
hidden dim	1152	768	1024	1152
heads	16	12	16	16
patch size	2×2			
input size	$32 \times 32 \times 4$			
Feature encoders				
encoders	Table 2	MAE-B + SigLIP-B		
Optimizer				
optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)			
learning rate	1e-5			
lr schedule	constant			
weight decay	0.0			
EMA decay	0.999			
batch size	Table 4	128	128	128
training epochs	20	100	100	100
Flow-Coupled Drifting				
time t with prob. p_{flow}	$\mathcal{U}[0, 1)$			
smoothing τ	1e-4			
flow probability p_{flow}	Table 1(a)	0.5	0.5	0.5
exponent γ	Table 1(b)	1.0	1.0	1.0

Table 6 lists the configurations. We use a pretrained VAE tokenizer (Rombach et al., 2022) to encode each 256×256 image into a $32 \times 32 \times 4$ latent, which is the input and output space of the generator. Each generator is initialized from the pretrained SiT of the same size (Ma et al., 2024). The generated latents are decoded into images by the VAE decoder and then passed to the feature encoders, and gradients are backpropagated through the decoder and the encoders to the generator.

Multiple encoders. The scale of the loss can differ across feature spaces. Following FD-loss (Yang et al., 2026), we normalize the loss of each encoder ϕ_i by its detached value and sum the normalized losses with equal weights:

$$\mathcal{L} = \sum_i \frac{\mathcal{L}_{\phi_i}}{\text{sg}[\mathcal{L}_{\phi_i}] + c}, \tag{17}$$

where $\mathcal{L}_{\phi_i}$ is the minibatch loss of Equation 14 computed in the feature space of ϕ_i, and $c = 0.01$ is a small constant for numerical stability. This keeps each term close to unit scale regardless of the feature space. The analysis in Section 3 applies to the losses without this normalization. The normalization multiplies the semi-gradient of each $\mathcal{L}_{\phi_i}$ by a positive scalar that depends on the minibatch and the encoder. It therefore preserves the direction of each sampled force but changes the relative weights across minibatches and encoders.

B Additional Derivations

Throughout this appendix, $\beta = 1 - \gamma/2 \in (0, 1)$, so ψ in Equation 11 reads $\psi(d) = ((d^2 + \tau^2)^\beta - \tau^{2\beta})/(2\beta)$.

B.1 Single-pair field in the implementation of Drifting

The implementation of Deng et al. (2026) normalizes the kernel jointly over the positives and negatives, and over the generated samples, which gives a positive weight to each positive and each negative of a generated sample $\boldsymbol{x}$. It then multiplies the weight of each positive by the total weight of the negatives, multiplies the weight of each negative by the total weight of the positives, and subtracts the weighted sum of the negatives from that of the positives. With one positive and one negative of weights $a^+, a^- > 0$, both products equal a^+a^-, and the field is:

$$\widehat{\boldsymbol{V}}(\boldsymbol{x}) = a^+a^-\boldsymbol{y}^+ - a^+a^-\boldsymbol{y}^- = a^+a^-(\boldsymbol{y}^+ - \boldsymbol{y}^-). \tag{18}$$

The weights depend on $\boldsymbol{x}$ through the kernel, but the direction of the field is that of $\boldsymbol{y}^+ - \boldsymbol{y}^-$.

B.2 Distributional objective for the single-pair force

Preliminaries. Direct differentiation of Equation 11 gives:

$$\psi'(d) = \frac{d}{(d^2 + \tau^2)^{\gamma/2}}. \tag{19}$$

Since $\tau > 0$, the map $\boldsymbol{v} \mapsto \psi(\|\boldsymbol{v}\|_2)$ is continuously differentiable on $\mathbb{R}^D$, including $\boldsymbol{v} = \boldsymbol{0}$. We denote its gradient by:

$$\boldsymbol{h}(\boldsymbol{v}) = \nabla_{\boldsymbol{v}}\psi(\|\boldsymbol{v}\|_2) = \frac{\boldsymbol{v}}{(\|\boldsymbol{v}\|_2^2 + \tau^2)^{\gamma/2}}. \tag{20}$$

The function $\boldsymbol{h}$ is odd, and since $\gamma > 0$, it satisfies the linear bound:

$$\|\boldsymbol{h}(\boldsymbol{v})\|_2 \le \tau^{-\gamma}\|\boldsymbol{v}\|_2. \tag{21}$$

In this notation, $\boldsymbol{F}_p^+(\boldsymbol{x}; \boldsymbol{y}^+) = \boldsymbol{h}(\boldsymbol{y}^+ - \boldsymbol{x})$ and $\boldsymbol{F}_{q_{\boldsymbol{\theta}}}^-(\boldsymbol{x}; \boldsymbol{y}^-) = \boldsymbol{h}(\boldsymbol{y}^- - \boldsymbol{x})$ in Equation 8, so the force reads:

$$\boldsymbol{F}(\boldsymbol{x}; \boldsymbol{y}^+, \boldsymbol{y}^-) = \boldsymbol{h}(\boldsymbol{y}^+ - \boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{y}^- - \boldsymbol{x}). \tag{22}$$

The map $a \mapsto a^\beta$ is concave on $[0, \infty)$ and vanishes at zero, so it is subadditive, and $(d^2 + \tau^2)^\beta \le d^{2\beta} + \tau^{2\beta}$. Hence:

$$0 \le \psi(d) \le \frac{d^{2-\gamma}}{2-\gamma} \le \frac{1 + d^2}{2 - \gamma}. \tag{23}$$

Proposition 1 (Energy gradient)**.** *Let p and q be distributions on $\mathbb{R}^D$ with finite second moments. Then all expectations in Equation 10 are finite, and the first variation of $\mathcal{E}$ with respect to q, up to an additive constant independent of $\boldsymbol{x}$, is:*

$$\frac{\delta\mathcal{E}(q, p)}{\delta q}(\boldsymbol{x}) = 2\mathbb{E}_{\boldsymbol{y}\sim p}\psi(\|\boldsymbol{x} - \boldsymbol{y}\|_2) - 2\mathbb{E}_{\boldsymbol{x}'\sim q}\psi(\|\boldsymbol{x} - \boldsymbol{x}'\|_2). \tag{24}$$

Moreover, for $\boldsymbol{y}^+ \sim p$ and $\boldsymbol{y}^- \sim q$ sampled independently, Equation 12 holds.

Proof. Finiteness follows from Equation 23 and the finite second moments. In Equation 10, q enters the first term linearly and the second term quadratically, and the third term does not depend on q. Varying either copy of q in the second term gives the same contribution, which yields the factor two in Equation 24.

By Equation 21 and the finite moments of p and q, dominated convergence allows differentiation under the expectations in Equation 24. Using $\nabla_{\boldsymbol{x}}\psi(\|\boldsymbol{x} - \boldsymbol{y}\|_2) = \boldsymbol{h}(\boldsymbol{x} - \boldsymbol{y}) = -\boldsymbol{h}(\boldsymbol{y} - \boldsymbol{x})$, we obtain:

$$-\frac{1}{2}\nabla_{\boldsymbol{x}}\frac{\delta\mathcal{E}(q, p)}{\delta q}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}^+\sim p}\boldsymbol{h}(\boldsymbol{y}^+ - \boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}^-\sim q}\boldsymbol{h}(\boldsymbol{y}^- - \boldsymbol{x}). \tag{25}$$

For a fixed $\boldsymbol{x}$ and independent $\boldsymbol{y}^+ \sim p$ and $\boldsymbol{y}^- \sim q$, the right-hand side equals $\mathbb{E}[\boldsymbol{F}(\boldsymbol{x}; \boldsymbol{y}^+, \boldsymbol{y}^-) \mid \boldsymbol{x}]$ by Equation 22. Since $\delta(\frac{1}{2}\mathcal{E})/\delta q = \frac{1}{2}\delta\mathcal{E}/\delta q$, this proves Equation 12. □

B.3 DISTRIBUTION IDENTIFICATION

Proposition 2 (Identification). *Let p and q be distributions on $\mathbb{R}^D$ with finite second moments. Then:*

$$\mathcal{E}(q,p) \geq 0, \qquad \mathcal{E}(q,p) = 0 \iff q = p. \tag{26}$$

Proof. Let Γ denote the gamma function. For $a \geq 0$, integration by parts gives:

$$\int_0^\infty \left(1 - e^{-sa}\right) s^{-1-\beta}\, \mathrm{d}s = \frac{\Gamma(1-\beta)}{\beta}\, a^\beta. \tag{27}$$

Applying Equation 27 to $a = d^2 + \tau^2$ and $a = \tau^2$ and subtracting yields:

$$\psi(d) = \frac{1}{2\Gamma(1-\beta)} \int_0^\infty e^{-\tau^2 s} \left(1 - e^{-sd^2}\right) s^{-1-\beta}\, \mathrm{d}s. \tag{28}$$

Let $k_s(\boldsymbol{x}, \boldsymbol{y}) = e^{-s\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}$ denote the Gaussian kernel with parameter $s > 0$. The corresponding population squared maximum mean discrepancy (Gretton et al., 2012) is:

$$\mathrm{MMD}^2_{k_s}(q,p) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q} k_s(\boldsymbol{x},\boldsymbol{x}') + \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim p} k_s(\boldsymbol{y},\boldsymbol{y}') - 2\mathbb{E}_{\boldsymbol{x}\sim q,\boldsymbol{y}\sim p} k_s(\boldsymbol{x},\boldsymbol{y}). \tag{29}$$

The integrand in Equation 28 is nonnegative, so by Tonelli's theorem each of the three expectations in Equation 10 can be exchanged with the integral over s. Each expectation is finite by Proposition 1. Substituting Equation 28 into Equation 10, the constant terms cancel, and we obtain:

$$\mathcal{E}(q,p) = \frac{1}{2\Gamma(1-\beta)} \int_0^\infty e^{-\tau^2 s} s^{-1-\beta}\, \mathrm{MMD}^2_{k_s}(q,p)\, \mathrm{d}s. \tag{30}$$

For every $s > 0$, the Gaussian kernel k_s is characteristic on $\mathbb{R}^D$ (Sriperumbudur et al., 2010), so $\mathrm{MMD}^2_{k_s}(q,p) \geq 0$, with equality if and only if $q = p$. The weight $e^{-\tau^2 s} s^{-1-\beta}$ is strictly positive, so $\mathcal{E}(q,p) \geq 0$. If $q = p$, the integrand vanishes and $\mathcal{E}(q,p) = 0$. Conversely, if $\mathcal{E}(q,p) = 0$, the nonnegative integrand vanishes for almost every $s > 0$. Hence $\mathrm{MMD}^2_{k_s}(q,p) = 0$ for some $s > 0$, and $q = p$. □

Relation to MMD gradient flows. By Equation 30, $\mathcal{E}$ is a weighted integral of squared MMDs with Gaussian kernels. The 2-Wasserstein gradient flow of the squared MMD was studied by Arbel et al. (2019), and drifting-style dynamics have been related to Wasserstein gradient flows by Cao et al. (2026) and Han et al. (2026). Our contribution lies in estimating this velocity without bias from one positive and one negative per generated sample through the force in Equation 8, and in extending this estimate to the conditional energies of flow-coupled training in Equation 15.

B.4 ENERGY INTERPRETATION OF FLOW-COUPLED TRAINING

We use the notation of Section 3.4. As in the feature-space paragraph of Section 3.4, the generated sample, the negative and the positive below denote their embeddings in one feature space. We assume that p and every $q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}$ have finite second moments.

Proposition 3 (Endpoint energy). *Let $C_p = \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'\sim p}\psi(\|\boldsymbol{y}-\boldsymbol{y}'\|_2)$. Then:*

$$\mathcal{A}_1(\boldsymbol{\theta}) = \mathcal{E}(q_{\boldsymbol{\theta}}, p) + C_p, \qquad \mathbb{E}_t \mathcal{A}_t(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta}) + (1 - p_{\mathrm{flow}}) C_p, \tag{31}$$

where $\mathbb{E}_t$ is taken over the time distribution in Section 3.4.

Proof. Since $\psi(0) = 0$, replacing p by $\delta_{\boldsymbol{y}^+}$ in Equation 10 gives, for every q with a finite second moment:

$$\mathcal{E}(q, \delta_{\boldsymbol{y}^+}) = 2\mathbb{E}_{\boldsymbol{x}\sim q}\psi(\|\boldsymbol{x}-\boldsymbol{y}^+\|_2) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\psi(\|\boldsymbol{x}-\boldsymbol{x}'\|_2). \tag{32}$$

At $t = 1$, $\boldsymbol{z}_1 = \boldsymbol{\epsilon}$, so $q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},1} = (f_{\boldsymbol{\theta}}(\cdot, 1))_{\#} p_\epsilon = q_{\boldsymbol{\theta}}$ for every $\boldsymbol{y}^+$. Taking the expectation of Equation 32 over $\boldsymbol{y}^+ \sim p$ gives:

$$\mathcal{A}_1(\boldsymbol{\theta}) = 2\mathbb{E}_{\boldsymbol{x}\sim q_{\boldsymbol{\theta}},\boldsymbol{y}\sim p}\psi(\|\boldsymbol{x}-\boldsymbol{y}\|_2) - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q_{\boldsymbol{\theta}}}\psi(\|\boldsymbol{x}-\boldsymbol{x}'\|_2). \tag{33}$$

Comparing Equation 33 with Equation 10 gives the first identity. Since $t = 1$ with probability $1 - p_{\text{flow}}$ and $t \sim \mathcal{U}[0, 1)$ otherwise:

$$\mathbb{E}_t \mathcal{A}_t = (1 - p_{\text{flow}})\mathcal{A}_1 + p_{\text{flow}} \mathbb{E}_{t \sim \mathcal{U}[0,1)} \mathcal{A}_t. \tag{34}$$

Substituting the first identity into Equation 34 and comparing with Equation 15 gives the second identity. □

Proposition 4 (Conditional force)**.** *Fix $(\boldsymbol{y}^+, t)$ and write $q = q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}$. Then:*

$$\mathbb{E}\big[\boldsymbol{F}(\boldsymbol{x}_t; \boldsymbol{y}^+, \boldsymbol{y}_t^-) \mid \boldsymbol{\epsilon}, \boldsymbol{y}^+, t\big] = -\frac{1}{2}\, \boldsymbol{g}_t(\boldsymbol{x}_t), \qquad \boldsymbol{g}_t(\boldsymbol{x}) = 2\boldsymbol{h}(\boldsymbol{x} - \boldsymbol{y}^+) - 2\mathbb{E}_{\boldsymbol{x}' \sim q}\boldsymbol{h}(\boldsymbol{x} - \boldsymbol{x}'), \tag{35}$$

and $\boldsymbol{g}_t(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\, \delta\mathcal{E}(q, \delta_{\boldsymbol{y}^+})/\delta q(\boldsymbol{x})$. Hence Equation 12 holds with q replaced by $q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}$ and p replaced by $\delta_{\boldsymbol{y}^+}$.

Proof. Given $(\boldsymbol{y}^+, t)$, $\boldsymbol{x}_t$ is a function of $\boldsymbol{\epsilon}$ only and $\boldsymbol{y}_t^-$ is a function of $\boldsymbol{\epsilon}^-$ only. Since $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}^-$ are independent, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t^-$ are independent draws from q. Taking the expectation of Equation 22 over $\boldsymbol{\epsilon}^-$ and using that $\boldsymbol{h}$ is odd gives Equation 35. The distribution $\delta_{\boldsymbol{y}^+}$ has a finite second moment, so Proposition 1 applies with p replaced by $\delta_{\boldsymbol{y}^+}$. Taking the spatial gradient of Equation 24 then gives the expression for $\boldsymbol{g}_t$. □

As t decreases, both inputs move toward the positive and toward each other, since their distances to the positive and to each other scale linearly with t. At $t = 0$, both inputs equal the positive, so the generated sample and the negative coincide. The repulsion term then vanishes, and the loss only attracts the generated sample toward the positive.

Proposition 5 (Gradient identity)**.** *Assume that the expectations below are finite and that differentiation with respect to $\boldsymbol{\theta}$ under the expectation is permitted, which are the standard conditions for pathwise gradient estimators (Mohamed et al., 2020). Then Equation 16 holds. For $p_{\text{flow}} = 0$, it reduces to $\nabla^{\text{sg}}_{\boldsymbol{\theta}} \mathcal{L}_{\text{FCD}} = \nabla_{\boldsymbol{\theta}} \mathcal{E}(q_{\boldsymbol{\theta}}, p)$.*

Proof. Let $\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t$ denote the Jacobian of $f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ with respect to $\boldsymbol{\theta}$. The target in Equation 14 is detached, and $\boldsymbol{y}_t^-$ enters the loss only through the target. Differentiating Equation 14 therefore gives:

$$\nabla^{\text{sg}}_{\boldsymbol{\theta}} \mathcal{L}_{\text{FCD}} = 2\mathbb{E}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t)^\top \big(\boldsymbol{x}_t - \text{sg}[\boldsymbol{x}_t + \boldsymbol{F}]\big)\big] = -2\mathbb{E}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t)^\top \boldsymbol{F}(\boldsymbol{x}_t; \boldsymbol{y}^+, \boldsymbol{y}_t^-)\big]. \tag{36}$$

Given $(\boldsymbol{\epsilon}, \boldsymbol{y}^+, t)$, the Jacobian $\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t$ is fixed. Conditioning on $(\boldsymbol{\epsilon}, \boldsymbol{y}^+, t)$ and applying Proposition 4 gives:

$$\nabla^{\text{sg}}_{\boldsymbol{\theta}} \mathcal{L}_{\text{FCD}} = \mathbb{E}_{t, \boldsymbol{y}^+, \boldsymbol{\epsilon}}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t)^\top \boldsymbol{g}_t(\boldsymbol{x}_t)\big]. \tag{37}$$

We now compute $\nabla_{\boldsymbol{\theta}} \mathbb{E}_t \mathcal{A}_t$. Fix $(\boldsymbol{y}^+, t)$, write $\boldsymbol{x} = \boldsymbol{x}_t$, and let $\boldsymbol{x}' = f_{\boldsymbol{\theta}}((1 - t)\boldsymbol{y}^+ + t\boldsymbol{\epsilon}', t)$, where $\boldsymbol{\epsilon}'$ is an independent copy of $\boldsymbol{\epsilon}$. By Equation 32:

$$\mathcal{E}\big(q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}, \delta_{\boldsymbol{y}^+}\big) = 2\mathbb{E}_{\boldsymbol{\epsilon}} \psi(\|\boldsymbol{x} - \boldsymbol{y}^+\|_2) - \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'} \psi(\|\boldsymbol{x} - \boldsymbol{x}'\|_2). \tag{38}$$

Differentiating Equation 38 under the expectation and applying the chain rule gives:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathcal{E}\big(q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}, \delta_{\boldsymbol{y}^+}\big) = {} & 2\mathbb{E}_{\boldsymbol{\epsilon}}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x})^\top \boldsymbol{h}(\boldsymbol{x} - \boldsymbol{y}^+)\big] \\ & - \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x})^\top \boldsymbol{h}(\boldsymbol{x} - \boldsymbol{x}') + (\partial_{\boldsymbol{\theta}} \boldsymbol{x}')^\top \boldsymbol{h}(\boldsymbol{x}' - \boldsymbol{x})\big]. \end{aligned} \tag{39}$$

Since $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}'$ are identically distributed and independent, exchanging them shows that the last two expectations are equal. Hence, by the definition of $\boldsymbol{g}_t$ in Equation 35:

$$\nabla_{\boldsymbol{\theta}} \mathcal{E}\big(q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}, \delta_{\boldsymbol{y}^+}\big) = \mathbb{E}_{\boldsymbol{\epsilon}}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t)^\top \boldsymbol{g}_t(\boldsymbol{x}_t)\big]. \tag{40}$$

The distributions of $\boldsymbol{y}^+$ and t do not depend on $\boldsymbol{\theta}$. Taking the expectation of Equation 40 over $\boldsymbol{y}^+ \sim p$ and t therefore gives:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_t \mathcal{A}_t(\boldsymbol{\theta}) = \mathbb{E}_{t, \boldsymbol{y}^+, \boldsymbol{\epsilon}}\big[(\partial_{\boldsymbol{\theta}} \boldsymbol{x}_t)^\top \boldsymbol{g}_t(\boldsymbol{x}_t)\big]. \tag{41}$$

By Proposition 3, $\mathbb{E}_t \mathcal{A}_t$ and $\mathcal{J}$ differ by the constant $(1 - p_{\text{flow}})C_p$, which does not depend on $\boldsymbol{\theta}$. Comparing Equation 37 with Equation 41 proves Equation 16. For $p_{\text{flow}} = 0$, Equation 15 gives $\mathcal{J} = \mathcal{E}(q_{\boldsymbol{\theta}}, p)$. □

B.5 MINIMIZERS OF THE FLOW-COUPLED ENERGY

In this subsection, we write the generator as a measurable map $f : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$ instead of $f_{\boldsymbol{\theta}}$, and we write $q_f = (f(\cdot,1))_{\#}p_\epsilon$ and $q^{\boldsymbol{y}^+}_{f,t}$ for the distribution of $f(\boldsymbol{z}_t, t)$ over the noise. For a feature encoder ϕ, let $\mathcal{J}(f)$ denote Equation 15 computed in the feature space of ϕ, that is, with $q_{\boldsymbol{\theta}}$, p, $q^{\boldsymbol{y}^+}_{\boldsymbol{\theta},t}$, and $\delta_{\boldsymbol{y}^+}$ replaced by $\phi_{\#}q_f$, $\phi_{\#}p$, $\phi_{\#}q^{\boldsymbol{y}^+}_{f,t}$, and $\delta_{\phi(\boldsymbol{y}^+)}$. As in Section 3.3, $\phi_{\#}p$ has a finite second moment. Let $\mathcal{G}$ be the set of all measurable f for which $\phi_{\#}q_f$ and every $\phi_{\#}q^{\boldsymbol{y}^+}_{f,t}$ have finite second moments and $\mathbb{E}_{t\sim\mathcal{U}[0,1)}\mathcal{A}_t(f) < \infty$.

Proposition 6 (Endpoint minimizer)**.** *Let $p_{\text{flow}} \in [0,1)$. If $f^\star \in \mathcal{G}$ satisfies $\mathcal{J}(f^\star) \le \mathcal{J}(f)$ for every $f \in \mathcal{G}$, then $\phi_{\#}q_{f^\star} = \phi_{\#}p$.*

Proof. By Equation 15, for every $f \in \mathcal{G}$:

$$\mathcal{J}(f) = (1 - p_{\text{flow}})\,\mathcal{E}(\phi_{\#}q_f, \phi_{\#}p) + p_{\text{flow}}\,\mathbb{E}_{t\sim\mathcal{U}[0,1)}\mathcal{A}_t(f). \tag{42}$$

Both terms are finite, the first by Proposition 1 and the second by the definition of $\mathcal{G}$. The first term depends on f only through $f(\cdot,1)$. Since $t \sim \mathcal{U}[0,1)$ satisfies $t < 1$, the second term depends on f only through its restriction to $\mathbb{R}^D \times [0,1)$.

The space $\mathbb{R}^D$ is Polish and hence a Borel space, so there exists a measurable $g : [0,1] \to \mathbb{R}^D$ such that $g(\vartheta) \sim p$ for $\vartheta \sim \mathcal{U}(0,1)$ (Kallenberg, 2002, Lemma 3.22). Let Φ be the standard normal cumulative distribution function and ϵ_1 the first coordinate of $\boldsymbol{\epsilon} \sim p_\epsilon$. Then $\Phi(\epsilon_1) \sim \mathcal{U}(0,1)$, so $T(\boldsymbol{\epsilon}) = g(\Phi(\epsilon_1))$ is measurable and satisfies $T_{\#}p_\epsilon = p$. Define $\tilde{f}$ by $\tilde{f}(\cdot,1) = T$ and $\tilde{f}(\cdot,t) = f^\star(\cdot,t)$ for $t < 1$. Then $\tilde{f}$ is measurable, $q_{\tilde{f}} = p$, and $q^{\boldsymbol{y}^+}_{\tilde{f},t} = q^{\boldsymbol{y}^+}_{f^\star,t}$ for every $t < 1$. Hence $\tilde{f} \in \mathcal{G}$, $\mathbb{E}_{t\sim\mathcal{U}[0,1)}\mathcal{A}_t(\tilde{f}) = \mathbb{E}_{t\sim\mathcal{U}[0,1)}\mathcal{A}_t(f^\star)$, and $\mathcal{E}(\phi_{\#}q_{\tilde{f}}, \phi_{\#}p) = 0$ by Proposition 2. By Equation 42 and the minimality of $f^\star$:

$$0 \le \mathcal{J}(\tilde{f}) - \mathcal{J}(f^\star) = -(1 - p_{\text{flow}})\,\mathcal{E}(\phi_{\#}q_{f^\star}, \phi_{\#}p). \tag{43}$$

Since $1 - p_{\text{flow}} > 0$, this gives $\mathcal{E}(\phi_{\#}q_{f^\star}, \phi_{\#}p) \le 0$. By Proposition 2, $\mathcal{E}(\phi_{\#}q_{f^\star}, \phi_{\#}p) = 0$ and $\phi_{\#}q_{f^\star} = \phi_{\#}p$. □

With multiple encoders, $\mathcal{J}$ is the sum of the energies of the encoders, and $\mathcal{G}$ requires the conditions above for every encoder. The same $\tilde{f}$ makes every endpoint energy vanish, so the argument above shows that every endpoint energy vanishes at $f^\star$, and the conclusion holds for each encoder. For class-conditional generation, the argument applies to each class.

C LIMITATIONS

Like Drifting Models, FCD still computes the force and loss in the feature spaces of pretrained encoders. Exploring representations that do not require a pretrained encoder is an important direction for future research. Moreover, although FCD performs well on class-conditional ImageNet, we have not yet applied it to text-to-image generation, where each condition typically has only one real sample. Since FCD requires only one real sample per generated sample, extending it to this setting is a promising next step.

D ADDITIONAL VISUALIZATIONS

Figure 4 shows additional class-conditional one-step samples generated by FCD-XL/2 on ImageNet 256×256.

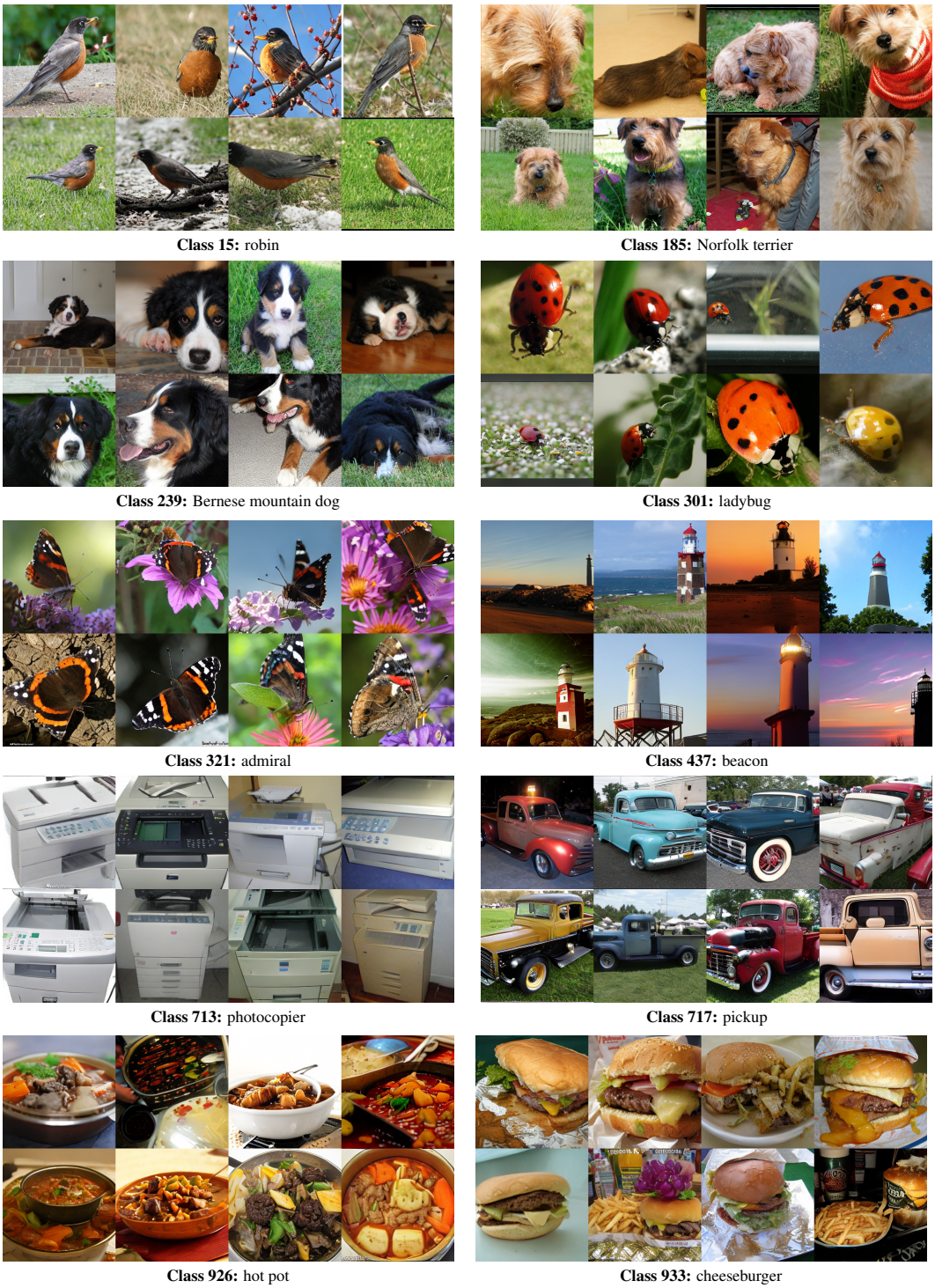


Figure 4: Additional one-step samples generated by FCD-XL/2 (FID 1.38) on class-conditional ImageNet 256×256.