Generative Drifting is Secretly Score Matching: a Spectral and Variational Perspective

Anonymous authors
Paper under double-blind review

Abstract

Generative Modeling via Drifting (Deng et al., 2026) has recently achieved state-of-the-art one-step image generation through a kernel-based drift operator, yet the success is largely empirical and its theoretical foundations remain poorly understood. In this paper, we make the following observation: *under a Gaussian kernel, the drift operator is exactly a score difference on smoothed distributions*. This insight allows us to answer all three key questions, which were left open in the original work: (1) whether a vanishing drift guarantees equality of distributions ($V_{p,q} = 0 \Rightarrow p = q$), (2) how to choose between kernels, and (3) why the stop-gradient operator is indispensable for stable training. Our observations position drifting within the well-studied score-matching family and enable a rich theoretical perspective for subsequent analysis. By linearizing the McKean-Vlasov dynamics resulting from our formulation and probing these dynamics in Fourier space, we reveal frequency-dependent convergence timescales comparable to *Landau damping* in plasma kinetic theory: the Gaussian kernel suffers an exponential high-frequency bottleneck, potentially explaining the empirical preference for the Laplacian kernel. Our analysis also suggests a fix: an exponential bandwidth annealing schedule $\sigma(t) = \sigma_0 e^{-rt}$ that reduces convergence time from $\exp(O(K_{\max}^2))$ to $O(\log K_{\max})$. Finally, by formalizing drifting as a Wasserstein gradient flow of the smoothed KL divergence, we prove that the stop-gradient operator is not a heuristic but is derived directly from the frozen-field discretization mandated by the Jordan, Kinderlehrer and Otto (JKO) scheme, and removing it severs training from any gradient-flow guarantee. This variational perspective further provides a general template for constructing novel drift operators, which we demonstrate with a Sinkhorn divergence drift. We validate our analysis on toy datasets and scale it up to ImageNet.

1 Introduction

The dominant paradigm in continuous generative modeling relies on learning the *score function* of data distributions and then using the learned score to guide the generation process. Energy-Based Models exploit the score to bypass the intractable partition function Grathwohl et al. (2019); Du & Mordatch (2019); Song & Kingma (2021); Score-Based Generative Models and Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) are unified under this framework via Denoising Score Matching (Vincent, 2011) which established that training a denoiser is theoretically equivalent to score matching; furthermore, Flow Matching (Lipman et al., 2023) has recently been shown to fit within the same family (Gao et al., 2024). Bridging these paradigms has allowed researchers to develop robust mathematical machinery and principled training practices across all of these models.

A recent approach, *Generative Modeling via Drifting* (Deng et al., 2026), appears to depart from this formalism. Instead of learning a score or velocity field, drifting prescribes a *kernel-based drift operator* $V_{p,q}$ that pulls generated samples toward data while pushing them away from one another. The generator is trained to match its own drifted outputs until the drift vanishes, achieving impressive one-step image generation without distillation, teacher models, or adversarial training.

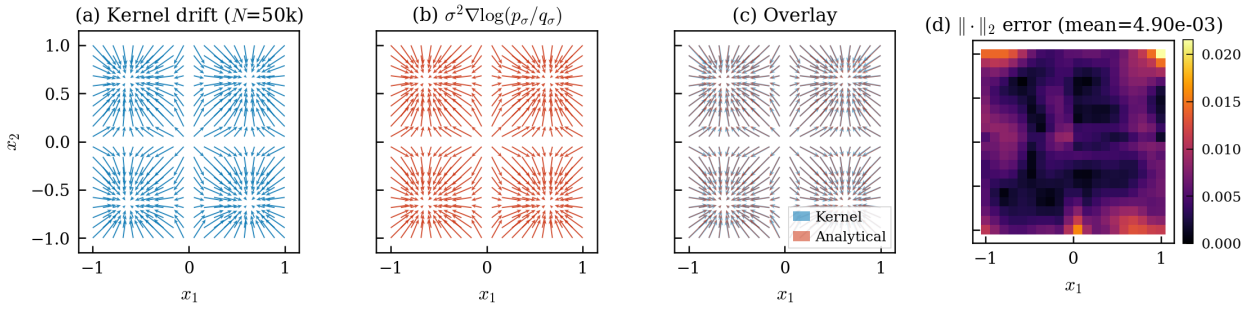

Figure 1: Numerical confirmation of Theorem 4.1 on a 4-mode Gaussian mixture. **(a)** Empirical kernel mean-shift drift ($N = 50$k samples). **(b)** Analytical score-difference form $\sigma^2 \nabla \log(p_\sigma / q_\sigma)$. **(c)** Overlay: the two fields are visually indistinguishable. **(d)** Pointwise ℓ_2 error heatmap (mean 4.9×10^{-3}). Details are provided in Section F.1.

Despite its impressive empirical results, the mathematical structure of drifting remains largely underexplored. Specifically, Deng et al. (2026) left three foundational open questions:

1. *Identifiability.* Does $V_{p,q} = 0$ guarantee $p = q$?
2. *Kernel selection.* The drift operator depends on the choice of the kernel; how should they be defined and selected?
3. *Algorithmic stability.* Is the stop-gradient operator essential, and, if so, what are its theoretical justifications?

All three questions share a common root: we do not know what the drift operator *actually computes*.

In this paper we show that the drift operator, under the Gaussian kernel, *is* a score difference on smoothed distributions:

$$V_{p,q}^{(\sigma)}(x) \;=\; \sigma^2 \, \nabla_x \log \frac{p_\sigma(x)}{q_\sigma(x)}. \tag{1}$$

This identity, which we derive from first principles by direct kernel substitution, immediately positions drifting within the score-matching family (see Figure 1 for direct numerical confirmation). More importantly, it provides the theoretical framework required to resolve all three open questions. Specifically, we prove identifiability via the injectivity of Gaussian convolution (Sec. 4), kernel selection via a Fourier-space stability analysis of the linearized McKean-Vlasov equation (Sec. 5.2), and stop-gradient necessity via the JKO discretization of a Wasserstein gradient flow (Sec. 5.3).

To summarize, our key contributions are:

1. **Score-matching identity and identifiability** (§4). Direct kernel substitution gives $V_{p,q}^{(\sigma)} = \sigma^2 \nabla \log(p_\sigma / q_\sigma)$, placing drifting in the score-matching family and reducing identifiability to Fourier injectivity.
2. **Landau damping and kernel diagnosis** (§5.2). Linearizing the McKean–Vlasov dynamics around equilibrium yields mode-resolved convergence timescales: exponential slowdown for the Gaussian kernel, polynomial for the Laplacian, a first principled explanation of the empirical kernel preference.
3. **Wasserstein gradient flow and stop-gradient necessity** (§5.3). Drifting is the JKO-discretized Wasserstein gradient flow of the smoothed KL divergence, and the stop-gradient operator is precisely the frozen-field structure this discretization requires; removing it severs training from any descent guarantee.
4. **Algorithmic improvements** (§6). An exponential bandwidth schedule $\sigma(t) = \sigma_0 e^{-rt}$ provably reduces convergence time from $\exp(\mathcal{O}(K_{\max}^2))$ to $\mathcal{O}(\log K_{\max})$ while retaining Gaussian identifiability, and the variational template $V = -\nabla(\delta \mathcal{F} / \delta q)$ yields new operators such as a Sinkhorn-divergence drift.

2 Background

We briefly recall the basic principles behind score-matching and discuss how Diffusion and Flow Models are unified under this lens.

Score matching and Energy-Based Models. The score function $\nabla_x \log p_\theta(x)$ bypasses the intractable normalization constant in energy-based models and enables Langevin sampling: $X_{t+1} = X_t + \frac{\eta^2}{2}\nabla_x \log p_\theta(X_t) + \eta z$, where $z \sim \mathcal{N}(0, I)$ and η is the step size. Because the raw data score is unavailable, Vincent (2011) introduced denoising score matching: by working with the Gaussian-perturbed distribution $q(\tilde{x}) = \int \mathcal{N}(\tilde{x}; x, \sigma^2 I) p_D(x)\, dx$, the objective is replaced by matching the score of the transition kernel, which simplifies to $-(\tilde{x} - x)/\sigma^2$. This perturbed score will reappear at the heart of our analysis.

Diffusion and Flow Models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) connect the noise-predicting network ϵ_θ to the score via Tweedie's formula, $\epsilon_\theta(x_t) \approx -\sigma_t \nabla_{x_t} \log p_t(x_t)$. This connection is most clear when viewing the forward perturbation as a Stochastic Differential Equation (SDE), $dx = f(x,t)\, dt + g(t)\, dw$, where $f(x,t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and w is standard Brownian motion. The generative process is then given by the exact reverse-time SDE:

$$dx = [f(x,t) - g(t)^2 \nabla_x \log p_t(x)]\, dt + g(t)\, d\bar{w}, \tag{2}$$

where $\bar{w}$ is standard Brownian motion flowing backward in time, and $p_t(x)$ is the marginal distribution of the perturbed data at time t.

Flow models (Lipman et al., 2023) learn a velocity field v_θ via $\dot{x}_t = v_\theta(x_t)$, sampling by ODE integration

$$x_1 = x_0 + \int_0^1 v_\theta(x_\tau)\, d\tau \tag{3}$$

Recent work (Gao et al., 2024) has shown that these two families are equivalent, completing the unification under score matching. In both cases, *the dynamics operate at inference time*: the generator is a learned field that is integrated to produce samples.

3 Drifting Models

Drifting (Deng et al., 2026) departs from the above paradigm by shifting all dynamics from inference to training. Let p be a data distribution on $\mathbb{R}^d$ and $f_\theta : \mathbb{R}^k \to \mathbb{R}^d$ a generator whose pushforward $q = (f_\theta)_\# \mathcal{N}(0, I)$ should approximate p. The *drift operator* $V_{p,q} = V_p^+ - V_q^-$ is defined as:

$$V_p^+(x) = \frac{\mathbb{E}_{y\sim p}[k(x,y)(y-x)]}{\mathbb{E}_{y\sim p}[k(x,y)]}, \qquad V_q^-(x) = \frac{\mathbb{E}_{y\sim q}[k(x,y)(y-x)]}{\mathbb{E}_{y\sim q}[k(x,y)]}, \tag{4}$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{>0}$ is a positive kernel. The attractive term V_p^+ pulls generated samples toward nearby data; the repulsive term V_q^- pushes them away from one another to prevent mode collapse.

During training, noise samples $\epsilon \in \mathbb{R}^k$ are mapped to $x = f_\theta(\epsilon)$ and drifted to target states $\tilde{x} = x + V_{p,q}(x)$. Note the target data distribution p only enters the optimization through the drift operator $V_{p,q}$. The generator is trained via a *stop-gradient loss*:

$$\mathcal{L}(\theta) = \mathbb{E}_\epsilon[\| f_\theta(\epsilon) - \mathrm{sg}[\tilde{x}] \|^2]. \tag{5}$$

Deng et al. (2026) established $q = p \Rightarrow V_{p,q} = 0$ but left the converse, *identifiability*, open. Moreover, as mentioned above, their reliance on a Laplacian kernel with grid-searched bandwidths lacks theoretical justification, and the necessity of $\mathrm{sg}[\cdot]$ remains a heuristic. We resolve all three in the sections that follow.

4 Drifting is Score Matching

4.1 The Core Identity

The following result is the foundation for all subsequent analysis.

Theorem 4.1 (Gaussian drift as score difference)**.** *Under the Gaussian kernel φ_σ, the drift operator admits the closed form expression:*

$$V_{p,q}^{(\sigma)}(x) \;=\; \sigma^2\, \nabla_x \log \frac{p_\sigma(x)}{q_\sigma(x)}, \tag{6}$$

where $p_\sigma := p * \varphi_\sigma$ and $q_\sigma := q * \varphi_\sigma$.

Proof. See A.1. □

This identity, which we derive by direct kernel substitution within the drifting architecture, inserts drifting squarely into the score-matching family under the Gaussian kernel. It was independently established by Weber (2024) from the opposite direction, starting from a KL-minimizing probability flow; the two derivations share the identity and little else. Ours is the starting point for the Fourier stability analysis, the annealing schedule, and the gradient-flow formalism that follow.

4.2 Connection to Denoising Score Matching

Theorem 4.1 places drifting in direct correspondence with the score-matching framework (Vincent, 2011; Song et al., 2021). Classical Denoising Score Matching trains a network ψ_θ to *approximate* the score $\nabla \log p_\sigma$ of a smoothed data distribution by regressing against the perturbation-kernel score $-(x-y)/\sigma^2$ (Vincent, 2011); at inference, ψ_θ is then evaluated repeatedly to integrate an SDE or ODE. Drifting (Deng et al., 2026) instead parameterizes a *sampler* f_θ whose pushforward defines $q = (f_\theta)_\# \mathcal{N}(0, I)$, and reads off the score difference $\sigma^2\big(\nabla \log p_\sigma - \nabla \log q_\sigma\big)$ exactly and non-parametrically from Parzen estimates of p_σ and q_σ on the current minibatch. The neural network thus never represents a vector field: the score is computed from particles on the fly, used only at training time to push f_θ toward p, and *vanishes from the pipeline at inference*, where sampling reduces to a single forward pass of f_θ.

4.3 Continuous-Time Limit and McKean–Vlasov Structure

The discrete training update $x_{n+1} = x_n + V_{p,q_n}(x_n)$ naturally admits a continuous-time limit in which each sample evolves as

$$\frac{dx_t}{dt} = V_{p,\, q(t)}(x_t), \tag{7}$$

with the generator distribution satisfying the continuity equation

$$\partial_t q(t,x) + \nabla_x \cdot \big(q(t,x)\, V_{p,\, q(t)}(x)\big) = 0. \tag{8}$$

Derivation details are in B.1. Under the Gaussian kernel, Eq. equation 8 is the *McKean–Vlasov equation*, well-studied in kinetic theory (Villani, 2002). Contrary to Diffusion and Flow models, because the velocity field is prescribed (not learned), the dynamics are amenable to the stability analysis in §5.

5 Theoretical Consequences of the Identity

Theorem 4.1 allow us to resolve all three foundational questions. We treat each in turn.

5.1 Identifiability

Proposition 5.1 (Identifiability)**.** *If $V_{p,q}^{(\sigma)}(x) = 0$ for all $x \in \mathbb{R}^d$ and $\sigma > 0$, then $p = q$.*

Proof. See A.2. □

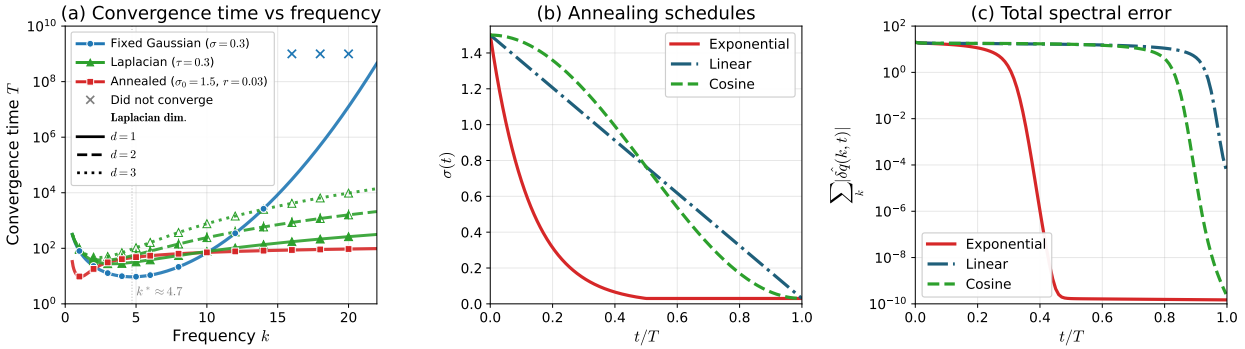

Figure 2: Spectral validation. **(a)** Convergence time $T(k)$ vs. frequency: lines are analytical predictions from Theorem 5.2, markers are measured decay times. The fixed-bandwidth Gaussian exhibits exponential slowdown (Landau damping); the Laplacian kernel yields polynomial scaling; the annealed Gaussian eliminates the bottleneck entirely. **(b)** Annealing schedules $\sigma(t)$. **(c)** Total spectral error under different schedules. Details in F.2.

5.2 Why the Laplacian Kernel? Landau Damping in Generative Models

We can exploit the analytical expression of the drift operator in Equation 6 and linearize the PDE in Eq. equation 8 around equilibrium, writing $q(t,x) = p(x) + \delta(t,x)$ with δ small. Under the local-homogeneity approximation (B.2), the linearized dynamics for the smoothed perturbation $\delta_\sigma = \varphi_\sigma * \delta$ reduce to

$$\partial_t \delta(x,t) = \sigma^2 \Delta\big(\delta_\sigma(x,t)\big). \tag{9}$$

Theorem 5.2 (Mode-resolved convergence timescales)**.** *Let $\kappa \in L^1(\mathbb{R}^d)$ be an even kernel. Under the linearized dynamics $\partial_t \delta = c_\kappa \Delta(\kappa * \delta)$, each Fourier mode evolves as*

$$\partial_t \hat{\delta}(\xi,t) = -\lambda_\kappa(\xi)\,\hat{\delta}(\xi,t), \qquad \lambda_\kappa(\xi) = c_\kappa |\xi|^2 \hat{\kappa}(\xi).$$

The convergence timescale of mode ξ is $\tau_\kappa(\xi) = 1/\lambda_\kappa(\xi) = (c_\kappa|\xi|^2\hat{\kappa}(\xi))^{-1}$.

Proof. See B.2 □

For the Gaussian kernel, the effective linearized convolution kernel is the Gaussian kernel itself. For the exponential kernel, the same linearization yields an equation of the form in Theorem 5.2, but with the companion kernel $h_\tau(r) \propto \tau(r+\tau)e^{-r/\tau}$.

Remark 5.3 (Landau damping analogy). In plasma physics, perturbations in a collisionless plasma decay via *Landau damping*: frequency modes are damped at rates controlled by the medium's spectral properties (Villani, 2002; Mouhot & Villani, 2011). In generative drifting, the *kernel* plays the role of the medium. The dispersion relation $\lambda_\kappa(\xi) \propto |\xi|^2\hat{\kappa}(\xi)$ is the precise analogue. To our knowledge, this identification has not appeared in prior work on generative models.

Applying Theorem 5.2 to the two kernels of interest makes the bottleneck explicit.

Corollary 5.4 (Gaussian vs. Laplacian convergence times)**.** *To reduce all Fourier modes up to $|k| = K_{\max}$ by a factor $1/\epsilon$:*

$$T_{\text{Gauss}} = \frac{\log(1/\epsilon)}{\sigma^2 K_{\max}^2} \exp\Big(\frac{\sigma^2 K_{\max}^2}{2}\Big), \qquad \textit{(exponential in } K_{\max}^2\textit{)}, \tag{10}$$

$$T_{\text{exp}} \propto \log(1/\epsilon)\ \tau^{d+1}\, K_{\max}^{d+1}, \qquad \textit{(polynomial in } K_{\max}\textit{)}. \tag{11}$$

Proof. See B.4 □

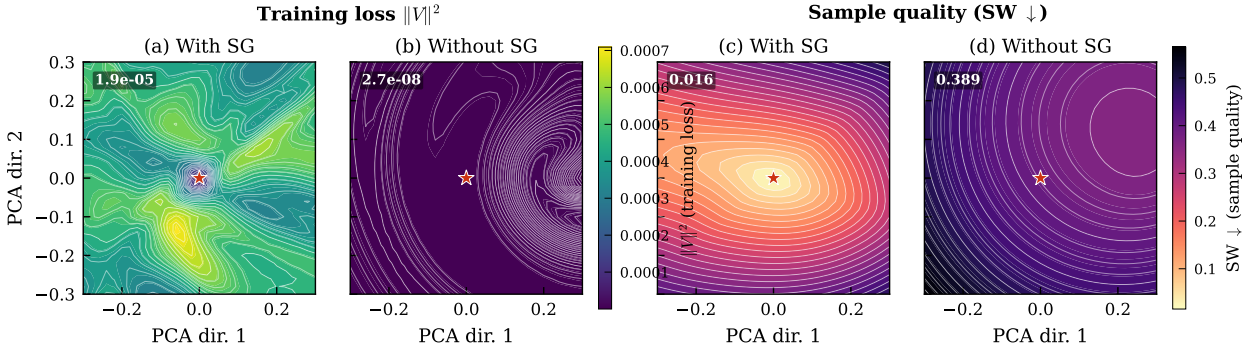

Figure 3: Loss landscapes with and without stop-gradient, projected onto the top two principal gradient-variation directions. **(a,b)** Training loss $\|V\|^2$: without SG the minimum is $\sim$100$\times$ deeper. **(c,d)** Sliced Wasserstein distance: with SG the loss minimum coincides with low distributional error; without SG the deep minimum corresponds to poor sample quality. Red star: trained solution. Details in F.3.

The Gaussian kernel at bandwidth σ acts as a low-pass filter with cutoff $|k| \sim 1/\sigma$: modes above the cutoff are exponentially suppressed. This bottleneck is intrinsic to the *kernelized particle dynamics*, not the neural network; it persists even with a perfect function approximator, and is mechanistically distinct from the spectral bias of neural networks (Rahaman et al., 2019; Tancik et al., 2020). Corollary 5.4 gives a first principled explanation of why Deng et al. (2026) may have favor the Laplacian kernel. It may also suggest their use of a multi-scale drift operator of the form $V^{\text{multi}}_{p,q} = \sum_\sigma V^{(\sigma)}_{p,q}$. In Figure 2(a) we validate the analytical predictions numerically, details are presented in F.2.

5.3 Stop-Gradient as a Structural Necessity

A distinctive feature of drifting is that its dynamics unfold during *training* rather than inference: the sequence of generators $\{f_{\theta_i}\}$ induces distributions $q_i = (f_{\theta_i})_\# \nu$ that define a training-time probability flow. We now show this flow is the Jordan–Kinderlehrer–Otto (JKO) discretization of a Wasserstein gradient flow (Santambrogio, 2015; Ambrosio & Savaré, 2007) of a smoothed KL functional, and that the stop-gradient operator is not a stabilization trick but the frozen-field discretization this variational scheme requires, earning the iterates $\{q_i\}$ monotone-descent guarantees toward $q = p$.

The smoothed KL energy. Fix $p \in \mathcal{P}(\Omega)$ and $\sigma > 0$. Define

$$F_\sigma[q] := \sigma^2 \, D_{\mathrm{KL}}(q_\sigma \| p_\sigma) = \sigma^2 \int_{\mathbb{R}^d} q_\sigma(x) \log \frac{q_\sigma(x)}{p_\sigma(x)} \, dx. \tag{12}$$

Proposition 5.5 (First variation and drift recovery)**.**

(i) *F_σ is lower-semi-continuous.*

(ii) *$\delta F_\sigma / \delta q(x) = \sigma^2 (\varphi_\sigma * \log(q_\sigma / p_\sigma))(x)$ (up to an additive constant), and is C^∞ on $\mathbb{R}^d$.*

(iii) *The Wasserstein gradient field $v_\sigma[q] := -\nabla_x (\delta F_\sigma / \delta q)$ satisfies $v_\sigma[q](x) \approx -V^{(\sigma)}_{p,q}(x)$ when $\log(q_\sigma / p_\sigma)$ varies slowly at scale σ; the error is $O(\sigma^2 \|\nabla^2 \log(q_\sigma / p_\sigma)\|_\infty)$ (D.1.3).*

Proof. See D.1.2. □

The drifting continuity equation 8 is thus (up to this quantified approximation) the Wasserstein gradient flow of F_σ.

The JKO scheme and the frozen-field step. The Jordan–Kinderlehrer–Otto scheme (Jordan et al., 1998) discretizes the gradient flow as

$$q_{n+1}^\tau \in \operatorname*{arg\,min}_{q\in\mathcal{P}(\Omega)} \Big\{ F_\sigma[q] + \frac{1}{2\tau} W_2^2(q, q_n^\tau) \Big\}. \tag{13}$$

We show that, given the functional 12, this scheme is well-posed (unique minimizer, monotone energy descent; D.2.1) and converges as $\tau \to 0$ (D.2.4). The Euler–Lagrange condition reveals an implicit velocity:

$$\frac{T_n(x) - x}{\tau} = v_\sigma[\hat{q}](x) = -\sigma^2 \nabla(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma))(x), \quad \hat{q}\text{-a.e.}, \tag{14}$$

where T_n is the optimal map from $\hat{q} := q_{n+1}^\tau$ to q_n^τ. The velocity has the same form as the drift, *but is evaluated at the unknown minimizer* $\hat{q}$, making the JKO step implicit and intractable.

The practical drifting algorithm resolves this by replacing q_{n+1}^τ with q_n^τ in the velocity field, the *frozen-field* (explicit Euler) approximation:

$$S_n(x) := x + \tau v_\sigma[q_n^\tau](x), \qquad \tilde{q}_{n+1}^\tau := (S_n)_\# q_n^\tau. \tag{15}$$

Where $(S_n)_\# q_n^\tau$ represents the pushforward of the distribution q_n^τ by map S_n. We show in detail in D.3, that given our assumptions, this approximation introduces an error of only $W_2(\tilde{q}_{n+1}^\tau, q_{n+1}^\tau) = O(\tau^{3/2})$. The chain of approximations is:

$$\underbrace{q_{n+1}^\tau = \arg\min\{F_\sigma + \tfrac{1}{2\tau} W_2^2(\cdot, q_n^\tau)\}}_{\text{JKO (implicit, intractable)}} \xrightarrow{\text{freeze velocity}} \underbrace{\tilde{q}_{n+1}^\tau = (S_n)_\# q_n^\tau}_{\text{Explicit Euler}} \xrightarrow{\text{parametric fit}} \underbrace{\min_\theta \mathcal{L}(\theta)}_{\text{Stop-gradient loss}}$$

Parametric implementation and necessity proof. Fitting the frozen-field target with a generator G_θ yields the loss

$$\mathcal{L}(\theta) = \mathbb{E}_{\varepsilon\sim\nu}\Big[\big\| G_\theta(\varepsilon) - \mathrm{sg}\big[G_{\theta_n}(\varepsilon) + \tau v_\sigma[q_{\theta_n}](G_{\theta_n}(\varepsilon)) \big] \big\|^2 \Big], \tag{16}$$

which is precisely the drifting loss of Deng et al. (2026) with $\eta = \tau$.

Theorem 5.6 (Stop-gradient preserves Wasserstein discretization)**.**

(i) (Structural correspondence.) *A global minimizer of equation 16 satisfies* $q_{\theta_{n+1}} = (S_n)_\# q_{\theta_n}$*, the frozen-field explicit Euler step of the Wasserstein gradient flow of* F_σ*.*

(ii) (Stop-gradient necessity.) *Removing* $\mathrm{sg}[\cdot]$ *yields* $\mathcal{L}_{\text{coupled}}(\theta) = \tau^2 \|v_\sigma[q_\theta]\|^2_{L^2(q_\theta)}$*, whose gradient includes distribution-feedback terms* $(D_q v_\sigma)\nabla_\theta q_\theta$*. Stationarity can now be achieved by* drift collapse*—reducing the velocity norm without transporting mass toward* p*—rather than by distributional convergence.*

Proof. See D.4. □

Figures 3 and 4 provide direct empirical confirmation: with stop-gradient the loss minimum aligns with low distributional error; without it, drift collapse produces a spuriously deep minimum with no distributional improvement.

6 From Theory to Practice

This extensive analysis is not only descriptive but also prescriptive, indeed the score-matching identity, spectral analysis, and gradient-flow formalism each yield a concrete practical contribution. We treat them in turn.

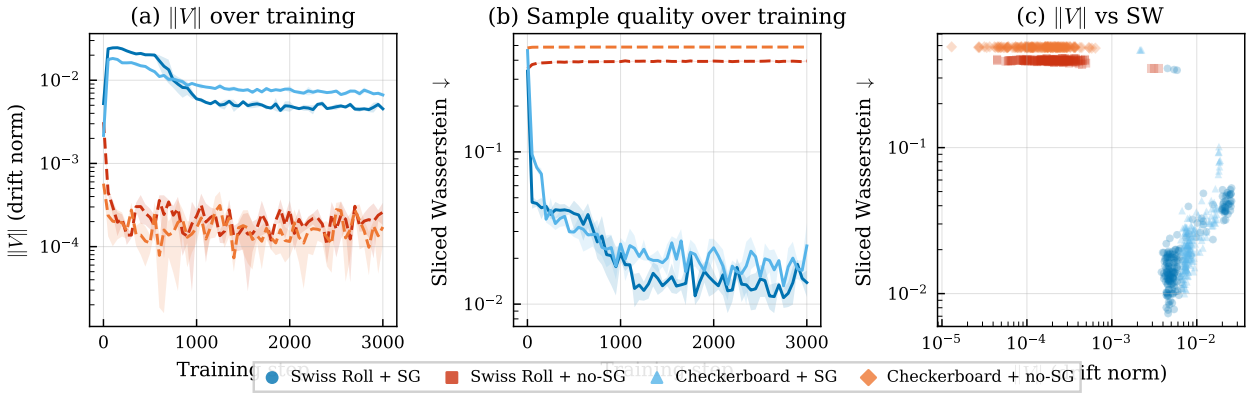

Figure 4: Drift norm vs. distributional distance during training on synthetic 2D targets. **(a)** Mean drift norm $\|V\|$. **(b)** Sliced Wasserstein distance. **(c)** Log–log scatter across training steps and seeds. With stop-gradient (solid), the two quantities are strongly correlated ($r > 0.95$) and jointly decay. Without stop-gradient (dashed), the drift norm collapses to $\sim 10^{-8}$ while the Wasserstein distance remains at 0.389—a direct demonstration of drift collapse (Theorem 5.6(ii)). Details in F.3.

6.1 Kernel Comparison: Why the Laplacian Works

Applying Theorem 5.2 to the Gaussian kernel gives the mode-resolved timescale

$$\tau(k, \sigma) = \frac{1}{\sigma^2 |k|^2} \exp\Big(\frac{\sigma^2 |k|^2}{2}\Big). \tag{17}$$

For $\sigma|k| \gg 1$, this grows exponentially: modes above the cutoff $|k| \sim 1/\sigma$ are frozen out. For the Laplacian kernel, $\hat{k}_\tau(\xi) \propto (1+\tau^2|\xi|^2)^{-(d+1)/2}$, giving the high-frequency rate $\lambda_{\exp}(k) \asymp \tau^{-(d+1)}|k|^{-(d-1)}$, only polynomial slowdown. Corollary 5.4 quantifies the difference, providing the first principled justification for the empirical kernel preference in Deng et al. (2026).

6.2 Exponential Bandwidth Annealing

The spectral analysis reveals a potential issue: the Gaussian kernel provides identifiability and a clean score-matching form, but suffers an exponential high-frequency bottleneck; the Laplacian kernel avoids this bottleneck but lacks these analytical properties. We resolve this problem with the following bandwidth annealing schedule:

$$\sigma(t) = \sigma_0 \, e^{-rt}, \tag{18}$$

held constant at $\sigma_{\min}$ once reached. The exponential form ensures that the optimal-rate window ($\sigma^2|k|^2 = 2$) sweeps continuously across increasing frequencies, activating each mode at its maximal convergence rate.

Theorem 6.1 (Convergence time under exponential annealing)**.** *Let $\sigma(t)$ follow Eq.equation 18 until $\sigma_{\min}$, then remain constant. To reduce all modes $|k| \le K_{\max}$ by factor $1/\epsilon$,*

$$T_{\epsilon,\text{anneal}} \frac{1}{r} \log\Big(\frac{\sigma_0}{\sigma_{\min}}\Big) + \frac{1}{\lambda_{\min}(K_{\max})} \log(1/\epsilon), \quad \lambda_{\min} = \sigma_{\min}^2 K_{\max}^2 \exp\Big(-\tfrac{1}{2}\sigma_{\min}^2 K_{\max}^2\Big). \tag{19}$$

Proof. Full proof in B.5. □

The total time depends only *logarithmically* on $K_{\max}$, compared to exponentially for the fixed-bandwidth Gaussian. Figure 2(b,c) confirms that the exponential schedule achieves the fastest spectral error reduction across all modes.

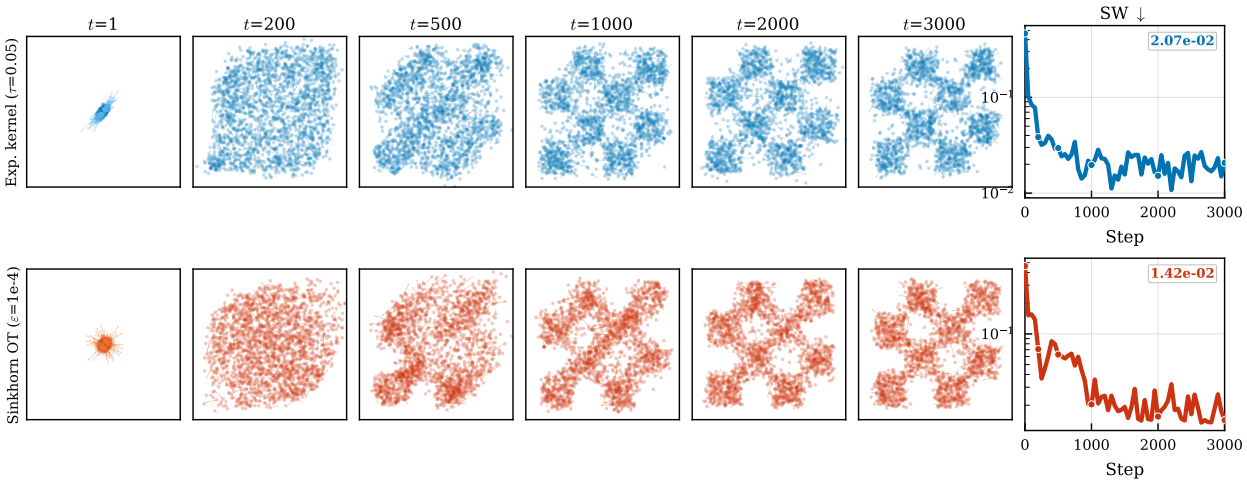

Figure 5: Sinkhorn-derived drift (bottom row) vs. Laplacian-kernel drift (top row) on the checkerboard distribution. Training snapshots with drift vectors overlaid; rightmost panels show sliced Wasserstein distance over training. Both converge successfully (final SW 1.42×10^{-2} and 2.07×10^{-2} respectively), demonstrating that the gradient-flow template of §6.3 yields practical operators beyond the original kernel family. Details are provided in F.4.

6.3 Principled Drift Operator Construction

The original operator introduced in Deng et al. (2026) was motivated through anti-symmetry properties of the kernel mean-shift, and was presented as one specific instantiation of a drifting model. Our gradient-flow formalism reveals the kernel drift as a particular instance of a universal variational template: given any sufficiently regular discrepancy functional F on $\mathcal{P}(\Omega)$, define

$$V(x) = -\nabla_x \frac{\delta F}{\delta q}(x). \tag{20}$$

The entire JKO–frozen-field–stop-gradient chain of §5.3 applies verbatim to any F satisfying: **(a)** lower semi-continuity in W_2; **(b)** existence and smoothness of $\delta F/\delta q$; and **(c)** $F[q] \geq 0$ with $F[q] = 0 \Leftrightarrow q = p$. This supersedes the anti-symmetry requirement of Deng et al. (2026): condition (c) alone is what guarantees that the gradient flow drives q to p, and (a)–(b) ensure that the JKO scheme is well-posed and that the velocity field is well-defined.

The Sinkhorn divergence as an energy functional. A natural candidate that satisfies (a)–(c) is the *Sinkhorn divergence* of Feydy et al. (2018), built from entropy-regularized optimal transport. For two probability measures $\mu, \nu \in \mathcal{P}(\Omega)$ and a cost $c(x, y) = \|x - y\|^r$ ($r \in \{1, 2\}$), the entropic OT problem with regularization $\varepsilon > 0$ is

$$\mathrm{OT}_\varepsilon(\mu, \nu) \;=\; \min_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y) \;+\; \varepsilon\, D_{\mathrm{KL}}\big(\pi \,\big\|\, \mu \otimes \nu\big), \tag{21}$$

where $\Pi(\mu, \nu)$ denotes couplings with marginals μ, ν. The unique minimizer has the Gibbs form $\pi^\star(x, y) \propto \exp\big((f(x) + g(y) - c(x, y))/\varepsilon\big)$, with dual potentials (f, g) obtained by Sinkhorn iterations (Peyré & Cuturi, 2020); we use the standard log-domain implementation, fully detailed in G.4. The Sinkhorn divergence is the *debiased* version

$$S_\varepsilon(q, p) \;=\; \mathrm{OT}_\varepsilon(q, p) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(q, q) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(p, p), \tag{22}$$

which interpolates between the MMD ($\varepsilon \to \infty$) and unregularized OT ($\varepsilon \to 0$), and removes the entropic bias of OT_ε alone. By Theorem 1 of Feydy et al. (2018), S_ε is symmetric, convex in each argument, positive-definite, and metrizes weak convergence; in particular $S_\varepsilon(q, p) = 0 \Leftrightarrow q = p$, so condition (c) holds. Lower semi-continuity and smoothness of the first variation follow from Proposition 2 there, securing (a)–(b). Plugging

$F[q] := S_\varepsilon(q, p)$ into the template equation 20 therefore yields a drift with full JKO–frozen-field–stop-gradient guarantees.

The Sinkhorn drift operator. We can compute the variational gradient explicitly at the particle level. For an empirical generator distribution $q = \frac{1}{N}\sum_i \delta_{x_i}$, let $\pi^\star_{q\to p}$ and $\pi^\star_{q\to q}$ be the entropic couplings between (q, p) and (q, q), and define the *barycentric projections*

$$T_{q\to p}(x_i) \;=\; \frac{\sum_j [\pi^\star_{q\to p}]_{ij}\, y_j}{\sum_j [\pi^\star_{q\to p}]_{ij}}, \qquad T_{q\to q}(x_i) \;=\; \frac{\sum_j [\pi^\star_{q\to q}]_{ij}\, x_j}{\sum_j [\pi^\star_{q\to q}]_{ij}}. \tag{23}$$

The map $T_{q\to p}$ sends each generated particle to its entropic OT image in the data distribution, and $T_{q\to q}$ does the same against the generator's own particles.

Proposition 6.2 (Sinkhorn drift)**.** *Let $F[q] = S_\varepsilon(q, p)$ with S_ε as in equation 22. Then the variational drift operator equation 20 admits the closed form*

$$V_{\mathrm{SK}}(x_i) \;\propto\; T_{q\to p}(x_i) \;-\; T_{q\to q}(x_i), \tag{24}$$

where $T_{q\to p}, T_{q\to q}$ are the barycentric projections of equation 23. The operator V_{SK} inherits all JKO–frozen-field–stop-gradient guarantees of §5.3.

Proof. Full derivation in F.4. □

Equation equation 24 is a striking parallel to the kernel drift $V_{p,q} = V^+_p - V^-_q$: an *attractive* term pulling particles toward p, minus a *repulsive* term pushing them away from one another. But here the structure does not stem from a hand-chosen kernel and anti-symmetry argument; it arises mechanically from differentiating a divergence satisfying (a)–(c), with the attractive and repulsive halves coming from the two OT_ε terms in the debiasing. The geometry has changed from kernel mean-shift to entropic optimal transport, but the gradient-flow scaffolding is unchanged. We plug V_{SK} into the standard stop-gradient loss $\mathcal{L}(\theta) = \mathbb{E}_\epsilon[\|f_\theta(\epsilon) - \mathrm{sg}[\tilde{x}]\|^2]$ with $\tilde{x} = x + V_{\mathrm{SK}}(x)$; full algorithmic details, including log-domain Sinkhorn iterations and a diagonal-masking heuristic that prevents $T_{q\to q}$ from collapsing to the identity as $q \to p$, are deferred to G.4.

Figure 5 shows that V_{SK} converges comparably to the Laplacian kernel drift on the checkerboard target, demonstrating that the gradient-flow template yields practical operators well beyond the original kernel family.

6.4 Image Generation Experiments

We now put our theoretical insights to the test on the realistic ImageNet generation setting, placing ourselves in the same ablation experimental setup as Deng et al. (2026) and testing three variants suggested by the theory: the Gaussian-kernel drift (which has the clean score-matching identity of Theorem 4.1), the exponentially annealed Gaussian (which targets the spectral bottleneck of §5.2), and the Sinkhorn-divergence drift of Proposition 6.2 (which exercises the gradient-flow template).

Experimental setup. We train class-conditional generators on ImageNet-256 (1000 classes) in the latent space of the Stable Diffusion VAE; FID is computed after VAE-decoding back to pixels. The generator is `DitGen-B`, a one-step DiT backbone. The drift loss is computed on multi-scale activations of a frozen MAE-pretrained ResNet-18, summed across streams. All variants share the same architecture, SSL representation, and AdamW optimizer and are trained for 30,000 total steps on 16 H100s, *only the drift operator differs* across rows of Table 1. We report one-step (NFE = 1) FID and IS, selecting the best CFG scale per checkpoint. Full hyperparameters, kernel and Sinkhorn algorithms, qualitative samples, and per-step wall-clock comparisons are provided in G.

Discussion. Three observations stand out. *First*, the fixed-bandwidth Gaussian kernel is competitive with the Laplacian baseline (FID 8.69 vs. 8.55, IS 157.0 vs. 148.0), validating that the kernel admitting a clean score-matching identity (Theorem 4.1) and full identifiability guarantees (Proposition 5.1) is also

Table 1: Ablation on ImageNet 256×256 class-conditional generation. All variants share the same B/2 architecture, latent-MAE SSL representation, and training schedule of Deng et al. (2026); only the drift operator differs. We report FID (↓) and IS (↑) under one-step (NFE = 1) generation. Best in **bold**, second-best underlined.

Drift operator	**Bandwidth**	**NFE**	**FID↓**	**IS↑**
Laplacian (Deng et al., 2026)	τ fixed	1	<u>8.55</u>	148.0
Gaussian (ours)	σ fixed	1	8.69	**157.0**
Gaussian (ours)	$\sigma(t) = \sigma_0 e^{-rt}$	1	**8.36**	<u>154.2</u>
Sinkhorn (ours)	ε fixed	1	8.81	135.48

a strong empirical performer; the theoretical convenience of the Gaussian does not come at an empirical cost. *Second*, the annealed schedule $\sigma(t) = \sigma_0 e^{-rt}$ improves FID over both the fixed Gaussian and the Laplacian baseline (8.36 vs. 8.69 and 8.55), confirming that the spectral analysis of §5.2 ports from the linearized toy regime to a high-dimensional generative setting. We note that in this regime the high-frequency bottleneck is not the only relevant axis, since the polynomial rate of the Laplacian kernel itself depends on dimension (Corollary 5.4); that the annealing prescription nonetheless yields a measurable improvement suggests the spectral diagnosis remains informative beyond its strict linearization assumptions. *Third*, the Sinkhorn-divergence drift converges to a regime comparable with the kernel-based operators (FID 8.81) despite arising from optimal-transport geometry rather than kernel mean-shift; this is direct evidence that the gradient-flow template $V = -\nabla(\delta\mathcal{F}/\delta q)$ of §5.3 is constructive, and that the anti-symmetry property invoked in Deng et al. (2026) is not fundamental: descent only requires conditions (a)–(c) of §6.3. Taken together, these results show that each of the three theoretical contributions (the score-matching identity, the spectral analysis, and the variational template) translates into a concrete experimental signal at ImageNet scale, and not merely on the linearized toy problems of §F.

7 Related Work

Generative modeling and the three-objective tradeoff. Modern generative modeling balances sample quality, mode coverage, and sampling speed. GANs (Goodfellow et al., 2014; Brock et al., 2019) optimize quality and speed but are prone to mode collapse; diffusion and flow models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Lipman et al., 2023) attain state-of-the-art quality and coverage at the cost of tens to hundreds of network evaluations per sample, with distillation (Zhou et al., 2024; Song et al., 2023) compressing them via expensive two-stage pipelines. Drifting (Deng et al., 2026) obtains one-step generation directly; the closest contemporary entries are Inductive Moment Matching (Zhou et al., 2025), using MMD with Gaussian kernels, and Li & Zhu (2026), who reinterpret drifting through long-horizon flow maps.

Unification under score matching. A parallel line of work has progressively unified continuous generative modeling under the score-matching umbrella: Vincent (2011) identified denoiser training with score estimation, Song et al. (2021) reframed diffusion's noise prediction via Tweedie's formula, and Gao et al. (2024) placed flow matching in the same family. This unification has been productive, percolating denoising-score-matching theory and convergence analyses across paradigms. Theorem 4.1 extends the trajectory to drifting and reduces identifiability to Fourier injectivity. The score-difference identity was independently obtained by Weber (2024) and Lai et al. (2026); neither develops the Fourier stability analysis, Landau-damping diagnosis, annealing schedule, or gradient-flow formalism it enables.

Spectral methods. Frequency analyses have been productive across kernel learning and optimization, with a recurring finding that learning is biased toward low frequencies (Rahaman et al., 2019; Tancik et al., 2020). Our analysis is mechanistically distinct from this network-level bias: the timescales we identify are intrinsic to the *kernelized particle dynamics* and persist with a perfect approximator, originating in the dispersion relation $\lambda_\kappa(\xi) = c_\kappa|\xi|^2\hat{\kappa}(\xi)$. The Landau-damping interpretation (Villani, 2002; Mouhot & Villani, 2011) makes this sharp and, to our knowledge, is new in generative modeling. Carrillo et al. (2024) prove

exponential KL convergence for continuous SVGD (Liu & Wang, 2019; Liu, 2017) via a Stein–log-Sobolev inequality but yield only a single global rate; we instead recover mode-resolved timescales.

Optimal transport. OT plays increasingly central roles in modern generative modeling: stabilizing GAN losses (Salimans et al., 2018), straightening flow-matching couplings (Tong et al., 2024), and underpinning Schrödinger-bridge methods (Bortoli et al., 2023). We use it in a different register as a tool of *analysis*. The Jordan–Kinderlehrer–Otto framework (Jordan et al., 1998; Santambrogio, 2015; Ambrosio & Savaré, 2007) reveals drifting as the explicit-Euler discretization of a Wasserstein gradient flow of the smoothed KL energy, with stop-gradient as the required frozen-field structure. This grounds drifting in a well-understood variational theory and, as §5.3 shows, opens a constructive route to new operators; our Sinkhorn-divergence drift, drawing on entropic OT (Peyré & Cuturi, 2020; Feydy et al., 2018), is one such instance.

8 Conclusion

We began with a simple question, what does the drifting kernel actually compute? We demonstrate that under a Gaussian kernel, it is a score difference on smoothed distributions. This single identity serves as a foundation that helps resolve all three open theoretical questions simultaneously, each through a distinct analytical lens.

Identifiability follows from Fourier injectivity. Landau damping offers an explanation for kernel selection: the Gaussian kernel exponentially suppresses high-frequency modes, justifying the empirical preference for the Laplacian and motivating the exponential annealing schedule $\sigma(t) = \sigma_0 e^{-rt}$ that reduces convergence time from $\exp(O(K_{\max}^2))$ to $O(\log K_{\max})$. The stop-gradient operator, far from being a heuristic, is the frozen-field discretization mandated by the JKO scheme; removing it leads to drift collapse, a spurious minimization that reduces the loss without transporting mass toward the data distribution.

Beyond resolving these open problems, the gradient-flow formalism yields a modular template $V = -\nabla(\delta F/\delta q)$ that generalizes drift construction beyond the original kernel family, demonstrated here with a Sinkhorn divergence drift.

9 Limitations and Future Work

Scope of the spectral analysis. The Landau-damping diagnosis of §5.2 is local: it linearizes around equilibrium and treats the background density as homogeneous, so its predicted timescales hold strictly only in the small-perturbation regime. This linearization around fixed points yields necessary conditions and points toward concrete fixes, whose worth is then settled by experiment. The annealing schedule of §6.2 is such a fix, and its gains on ImageNet (Table 1) suggest the diagnosis remains informative well beyond its strict assumptions. A fully nonlinear treatment, perhaps via the machinery of Mouhot & Villani (2011), would extend the guarantees into the regime where most training occurs.

Toward optimal kernels. Our analysis exposes a tension between analytical structure (the Gaussian kernel yields the score-matching identity and full identifiability) and high-frequency resolution (the Laplacian kernel's polynomial decay resolves fine modes faster). Annealing reconciles the two empirically, but the broader point is that kernel choice jointly controls the available theory and the convergence dynamics. A natural open direction is to design kernels with prescribed spectral profiles, tailored to the target distribution's frequency content, while retaining the Gaussian's analytical scaffolding.

Beyond the JKO discretization. JKO is one discretization of the Wasserstein gradient flow, and the stop-gradient loss inherits its first-order explicit-Euler character. Other minimizing-movement schemes, higher-order discretizations, splitting methods, or Nesterov-style acceleration on Wasserstein space may yield faster or more stable training. Our score-matching identity and gradient-flow formalism make this principled rather than speculative. A systematic study of these discretizations is a natural next step.

References

Luigi Ambrosio and Giuseppe Savaré. Chapter 1 - gradient flows of probability measures. volume 3 of *Handbook of Differential Equations: Evolutionary Equations*, pp. 1–136. North-Holland, 2007. doi: https://doi.org/10.1016/S1874-5717(07)80004-1. URL https://www.sciencedirect.com/science/article/pii/S1874571707800041.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling, 2023. URL https://arxiv.org/abs/2106.01357.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis, 2019. URL https://arxiv.org/abs/1809.11096.

José A. Carrillo, Jakub Skrzeczkowski, and Jethro Warnett. The stein-log-sobolev inequality and the exponential rate of convergence for the continuous stein variational gradient descent method, 2024. URL https://arxiv.org/abs/2412.10295.

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

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018. URL https://arxiv.org/abs/1810.08278.

Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. URL https://diffusionflow.github.io/.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of drifting and score-based models, 2026. URL https://arxiv.org/abs/2603.07514.

Zhiqi Li and Bo Zhu. A long-short flow-map perspective for drifting models, 2026. URL https://arxiv.org/abs/2602.20463.

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

Qiang Liu. Stein variational gradient descent as gradient flow, 2017. URL https://arxiv.org/abs/1704.07520.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2019. URL https://arxiv.org/abs/1608.04471.

Clément Mouhot and Cédric Villani. On landau damping. *Acta Mathematica*, 207(1):29–201, 2011. ISSN 0001-5962. doi: 10.1007/s11511-011-0068-9. URL http://dx.doi.org/10.1007/s11511-011-0068-9.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019. URL https://arxiv.org/abs/1806.08734.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport, 2018. URL https://arxiv.org/abs/1803.05573.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015. ISBN 9783319208282. doi: 10.1007/978-3-319-20828-2. URL http://dx.doi.org/10.1007/978-3-319-20828-2.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020. URL https://arxiv.org/abs/2006.10739.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2024. URL https://arxiv.org/abs/2302.00482.

Cédric Villani. Chapter 2 - a review of mathematical topics in collisional kinetic theory. volume 1 of *Handbook of Mathematical Fluid Dynamics*, pp. 71–74. North-Holland, 2002. doi: https://doi.org/10.1016/S1874-5792(02)80004-0. URL https://www.sciencedirect.com/science/article/pii/S1874579202800040.

Cédric Villani. *Optimal transport : old and new / Cédric Villani*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 7 2011. ISSN 1530-888X. doi: 10.1162/neco_a_00142. URL http://dx.doi.org/10.1162/neco_a_00142.

Romann M. Weber. The score-difference flow for implicit generative modeling, 2024. URL https://arxiv.org/abs/2304.12906.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching, 2025. URL https://arxiv.org/abs/2503.07565.

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, 2024. URL https://arxiv.org/abs/2404.04057.

Appendix Contents

To ease navigation, here is the table of contents of the appendix.

A Detailed Proofs for Section 4

A.1 Proof of Proposition 6 (Gaussian Drift as Score Difference)

Proof. Recall the Gaussian kernel

$$\varphi_\sigma(z) := (2\pi\sigma^2)^{-d/2} \exp\Big(-\frac{\|z\|^2}{2\sigma^2}\Big), \qquad p_\sigma := p * \varphi_\sigma.$$

We start from the attractive drift definition

$$V_p^+(x) = \frac{\int \varphi_\sigma(x-y)(y-x)p(y)\,dy}{\int \varphi_\sigma(x-y)p(y)\,dy}.$$

A direct computation gives the gradient identity (differentiate w.r.t. x)

$$\nabla_x \varphi_\sigma(x-y) = -\frac{x-y}{\sigma^2}\,\varphi_\sigma(x-y),$$

hence

$$(y-x)\,\varphi_\sigma(x-y) = \sigma^2 \nabla_x \varphi_\sigma(x-y).$$

Substituting into the numerator,

$$\begin{aligned}\int \varphi_\sigma(x-y)(y-x)p(y)\,dy &= \sigma^2 \int \nabla_x \varphi_\sigma(x-y)\,p(y)\,dy \\ &= \sigma^2 \nabla_x \int \varphi_\sigma(x-y)\,p(y)\,dy \\ &= \sigma^2 \nabla_x p_\sigma(x).\end{aligned}$$

The interchange of ∇_x and $\int dy$ is justified by dominated convergence: both $\varphi_\sigma(\cdot)$ and $\nabla\varphi_\sigma(\cdot)$ are bounded by Gaussian envelopes integrable against $p(y)\,dy$.

The denominator is exactly $p_\sigma(x)$, so

$$V_p^+(x) = \sigma^2 \frac{\nabla p_\sigma(x)}{p_\sigma(x)} = \sigma^2 \nabla \log p_\sigma(x).$$

The same calculation applied to the repulsive drift V_q^- (with q in place of p) yields

$$V_q^-(x) = \sigma^2 \nabla \log q_\sigma(x), \qquad q_\sigma := q * \varphi_\sigma.$$

Therefore,

$$V_{p,q}^{(\sigma)}(x) = V_p^+(x) - V_q^-(x) = \sigma^2 \nabla \log \frac{p_\sigma(x)}{q_\sigma(x)}.$$

□

A.2 Proof of Proposition 5.1 (Identifiability)

Proof. Assume that for some fixed $\sigma > 0$,

$$V_{p,q}^{(\sigma)}(x) = 0 \quad \text{for all } x.$$

By Equation (6),

$$\sigma^2 \nabla \log \frac{p_\sigma(x)}{q_\sigma(x)} = 0 \quad \Rightarrow \quad \nabla \log \frac{p_\sigma(x)}{q_\sigma(x)} = 0,$$

so $\log(p_\sigma/q_\sigma)$ is constant on $\mathbb{R}^d$. Hence there exists $C > 0$ such that

$$p_\sigma(x) = C\, q_\sigma(x) \qquad \text{for all } x.$$

Integrating both sides and using that p_σ, q_σ are probability densities (indeed $\int p_\sigma = \int p = 1$ and likewise for q), we obtain $C = 1$, hence $p_\sigma = q_\sigma$.

Now take Fourier transforms. Convolution with a Gaussian multiplies Fourier transforms:

$$\widehat{p_\sigma}(\xi) = \hat{p}(\xi)\,\widehat{\varphi_\sigma}(\xi), \qquad \widehat{\varphi_\sigma}(\xi) = e^{-\sigma^2\|\xi\|^2/2},$$

and similarly for q. Since $p_\sigma = q_\sigma$,

$$\hat{p}(\xi)\, e^{-\sigma^2\|\xi\|^2/2} = \hat{q}(\xi)\, e^{-\sigma^2\|\xi\|^2/2}.$$

Because $e^{-\sigma^2\|\xi\|^2/2} > 0$ for all ξ, we conclude $\hat{p}(\xi) = \hat{q}(\xi)$ for all ξ. The Fourier transform uniquely determines a probability measure (equivalently, characteristic functions are injective), hence $p = q$. □

B Detailed Proofs for Section 4.3

B.1 Continuous-Time Limit

We sketch the (standard) passage from the discrete particle update to the McKean–Vlasov continuity equation, emphasizing what is formal and what can be made rigorous.

Consider the particle-level update

$$x_{n+1} = x_n + \varepsilon\, V_{p,q_n}(x_n), \qquad \varepsilon > 0,$$

and define the piecewise-constant interpolation

$$x^\varepsilon(t) := x_n, \qquad t \in [n\varepsilon, (n+1)\varepsilon).$$

Formally,

$$\frac{x^\varepsilon(t+\varepsilon) - x^\varepsilon(t)}{\varepsilon} = V_{p,q_n}(x_n).$$

If, as $\varepsilon \to 0$: (i) $q_n \to q(t)$ in W_2 at $t = n\varepsilon$, (ii) for each fixed q the map $x \mapsto V_{p,q}(x)$ is locally Lipschitz with at most linear growth, and (iii) the dependence $q \mapsto V_{p,q}$ is continuous in W_2 in a suitable sense, then standard stability results for ODE schemes imply that $x^\varepsilon(\cdot)$ converges (along subsequences) to a limit $x(\cdot)$ satisfying the McKean–Vlasov ODE

$$\dot{x}(t) = V_{p,q(t)}(x(t)).$$

Let $q(t)$ denote the law of $x(t)$. Assume $q(t)$ admits a density and $v(t,x) := V_{p,q(t)}(x)$ is sufficiently regular so that the chain rule holds. For any test function $\phi \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\phi(x(t)) = \nabla\phi(x(t)) \cdot v(t, x(t)).$$

Taking expectation gives

$$\frac{d}{dt}\int \phi(x)\, dq(t)(x) = \int \nabla\phi(x) \cdot v(t,x)\, dq(t)(x).$$

Integrating by parts (justified since ϕ is compactly supported) yields the weak form

$$\frac{d}{dt}\int \phi\, dq(t) = -\int \phi(x)\, \nabla \cdot (q(t,x)v(t,x))\, dx,$$

which is exactly the continuity equation equation 8 in distributional form:

$$\partial_t q + \nabla \cdot (q\, V_{p,q}) = 0.$$

B.2 Linearization and Dispersion Relation

We derive the dispersion relation in a way that separates (a) exact identities from (b) simplifying assumptions used to obtain a closed-form Fourier rate.

Proof. Write $q(t,x) = p(x) + \delta(t,x)$ with $|\delta| \ll p$. Denote Gaussian smoothing by $f_\sigma = \varphi_\sigma * f$, so $q_\sigma = p_\sigma + \delta_\sigma$. For the Gaussian kernel the drift admits the score form

$$V_{p,q}^{(\sigma)}(x) = \sigma^2 \nabla \log \frac{p_\sigma(x)}{q_\sigma(x)}.$$

Substituting $q = p + \delta$ gives

$$V_{p,p+\delta}^{(\sigma)} = \sigma^2 \nabla \log \frac{p_\sigma}{p_\sigma + \delta_\sigma} = -\sigma^2 \nabla \log\left(1 + \frac{\delta_\sigma}{p_\sigma}\right).$$

Linearizing for small δ yields

$$V_{p,p+\delta}^{(\sigma)} = -\sigma^2 \nabla\left(\frac{\delta_\sigma}{p_\sigma}\right) + O(\delta^2).$$

The continuity equation

$$\partial_t q + \nabla \cdot (q V_{p,q}) = 0$$

then implies, keeping only first-order terms,

$$\partial_t \delta = -\nabla \cdot (p V_{p,p+\delta}^{(\sigma)}) = \sigma^2 \nabla \cdot \left(p \, \nabla\left(\frac{\delta_\sigma}{p_\sigma}\right)\right) + O(\delta^2).$$

Under the local-homogeneity approximation $p_\sigma \approx \text{const}$, this reduces to

$$\partial_t \delta(x,t) = \sigma^2 \Delta(\delta_\sigma(x,t)).$$

□

B.3 Proof of Theorem 5.2

Proof Take the Fourier transform in space of the linearized PDE

$$\partial_t \delta = c_\kappa \Delta(\kappa * \delta).$$

Using that convolution becomes multiplication and that

$$\widehat{\Delta f}(\xi) = -|\xi|^2 \hat{f}(\xi),$$

we obtain

$$\partial_t \hat{\delta}(\xi,t) = c_\kappa \widehat{\Delta(\kappa * \delta)}(\xi,t) = -c_\kappa |\xi|^2 \widehat{\kappa * \delta}(\xi,t) = -c_\kappa |\xi|^2 \hat{\kappa}(\xi) \hat{\delta}(\xi,t).$$

Hence each Fourier mode evolves independently according to

$$\partial_t \hat{\delta}(\xi,t) = -\lambda_\kappa(\xi) \hat{\delta}(\xi,t), \qquad \lambda_\kappa(\xi) = c_\kappa |\xi|^2 \hat{\kappa}(\xi).$$

This is a scalar ODE, with solution

$$\hat{\delta}(\xi,t) = e^{-\lambda_\kappa(\xi) t} \hat{\delta}(\xi,0).$$

Therefore the mode ξ decays exponentially at rate $\lambda_\kappa(\xi)$, so its characteristic convergence timescale is

$$\tau_\kappa(\xi) = \lambda_\kappa(\xi)^{-1} = \frac{1}{c_\kappa |\xi|^2 \hat{\kappa}(\xi)}.$$

This proves the claim. □

Exponential kernel. We now show that the exponential-kernel drift also linearizes to the form required by Theorem 5.2, but with an effective convolution kernel different from the original exponential kernel.

Let

$$k_\tau(r) = e^{-r/\tau}.$$

Define the companion kernel

$$h_\tau(r) := \tau(r+\tau)e^{-r/\tau}.$$

Since

$$h'_\tau(r) = -re^{-r/\tau},$$

we have, for $r = \|x-y\|$,

$$\nabla_x h_\tau(\|x-y\|) = h'_\tau(r)\frac{x-y}{r} = (y-x)e^{-\|x-y\|/\tau}.$$

Therefore

$$V_p^+(x) = \frac{\int k_\tau(\|x-y\|)(y-x)p(y)\,dy}{\int k_\tau(\|x-y\|)p(y)\,dy} = \frac{\nabla(h_\tau * p)(x)}{(k_\tau * p)(x)}.$$

Similarly,

$$V_q^-(x) = \frac{\nabla(h_\tau * q)(x)}{(k_\tau * q)(x)}.$$

Hence

$$V_{p,q}^{\exp}(x) = \frac{\nabla(h_\tau * p)(x)}{(k_\tau * p)(x)} - \frac{\nabla(h_\tau * q)(x)}{(k_\tau * q)(x)}.$$

Now write $q = p + \delta$, with $|\delta| \ll p$. Under the same local-homogeneity approximation used above,

$$p(x) \approx \rho_0, \qquad (k_\tau * p)(x) \approx \rho_0 Z_\tau, \qquad \nabla(h_\tau * p)(x) \approx 0,$$

where

$$Z_\tau := \int_{\mathbb{R}^d} k_\tau(z)\,dz.$$

Keeping only first-order terms gives

$$V_{p,p+\delta}^{\exp}(x) \approx -\frac{1}{\rho_0 Z_\tau}\nabla(h_\tau * \delta)(x).$$

Substituting into the continuity equation,

$$\partial_t q + \nabla \cdot (qV_{p,q}) = 0,$$

and retaining only first-order terms yields

$$\partial_t \delta = -\nabla \cdot (pV_{p,p+\delta}^{\exp}) \approx \frac{1}{Z_\tau}\Delta(h_\tau * \delta).$$

Equivalently,

$$\partial_t \delta = \Delta(\kappa_\tau * \delta), \qquad \kappa_\tau := \frac{h_\tau}{Z_\tau}.$$

Thus the exponential-kernel drift falls under Theorem 5.2 with $c_\kappa = 1$ and $\kappa = \kappa_\tau$.

B.4 Proof of Corollary 5.4

Proof. From Theorem 5.2, each Fourier mode evolves as

$$\hat{\delta}(k,t) = e^{-\lambda_\kappa(k)t}\hat{\delta}(k,0), \qquad \lambda_\kappa(k) = c_\kappa|k|^2\hat{\kappa}(k).$$

To reduce the amplitude of mode k by a factor $1/\epsilon$, we require

$$e^{-\lambda_\kappa(k)T} \le \epsilon \quad \Longrightarrow \quad T \ge \frac{\log(1/\epsilon)}{\lambda_\kappa(k)}.$$

To damp all modes with $|k| \le K_{\max}$, the required time is governed by the slowest-decaying mode in that range:

$$T = \max_{|k|\le K_{\max}} \frac{\log(1/\epsilon)}{\lambda_\kappa(k)}.$$

Gaussian kernel. For the Gaussian kernel $\kappa = \varphi_\sigma$,

$$\hat{\kappa}(k) = e^{-\sigma^2|k|^2/2}, \qquad c_\kappa = \sigma^2,$$

so

$$\lambda_{\text{Gauss}}(k) = \sigma^2|k|^2 e^{-\sigma^2|k|^2/2}.$$

For large $K_{\max}$, the smallest rate on $[0, K_{\max}]$ occurs at $k = K_{\max}$, giving

$$T_{\text{Gauss}} = \frac{\log(1/\epsilon)}{\sigma^2 K_{\max}^2} \exp\Big(\frac{\sigma^2 K_{\max}^2}{2}\Big).$$

Exponential (Laplacian) kernel. For the exponential kernel, the effective kernel in Theorem 5.2 is

$$\kappa_\tau = \frac{h_\tau}{Z_\tau}, \qquad h_\tau(r) = \tau(r+\tau)e^{-r/\tau}, \qquad Z_\tau = \int_{\mathbb{R}^d} e^{-\|z\|/\tau}\, dz.$$

Using $Z_\tau \asymp \tau^d$ and $\widehat{h}_\tau(k) \asymp \tau^{d+2}(1+\tau^2|k|^2)^{-(d+3)/2}$,

$$\widehat{\kappa_\tau}(k) \asymp \tau^2(1+\tau^2|k|^2)^{-(d+3)/2}.$$

Hence

$$\lambda_{\exp}(k) \asymp |k|^2\tau^2(1+\tau^2|k|^2)^{-(d+3)/2}.$$

For large $|k|$,

$$\lambda_{\exp}(k) \asymp \tau^{-(d+1)}|k|^{-(d+1)}.$$

Evaluating at the slowest mode $k = K_{\max}$ gives

$$T_{\exp} \asymp \log(1/\epsilon)\, \tau^{d+1} K_{\max}^{d+1}.$$

□

B.5 Proof of Theorem 6.1 (Annealing Convergence Time)

Proof. Fix any mode $|k| \le K_{\max}$ and define the cumulative decay

$$\Lambda(k,t) := \int_0^t \sigma(s)^2|k|^2 e^{-\sigma(s)^2|k|^2/2}\, ds,$$

so that $\hat{\delta}(k,t) = \hat{\delta}(k,0)e^{-\Lambda(k,t)}$. Reducing $|\hat{\delta}(k,t)|$ by a factor $1/\epsilon$ is equivalent to $\Lambda(k,t) \ge \log(1/\epsilon)$. We analyse the two phases separately.

Annealing phase ($t \in [0, T_{\text{ann}}]$). Substituting $u = \sigma(s)^2|k|^2$, so $du = -2ru\,ds$, gives

$$\Lambda(k, T_{\text{ann}}) = \frac{1}{2r} \int_{\sigma_{\min}^2|k|^2}^{\sigma_0^2|k|^2} e^{-u/2}\,du = \frac{1}{r}\Big[e^{-\sigma_{\min}^2|k|^2/2} - e^{-\sigma_0^2|k|^2/2}\Big] \geq 0,$$

since the integrand is strictly positive. Modes with $\sigma_{\min}^2|k|^2 < 2$ pass through their peak rate $f(2) = 2/e$ during this phase and accumulate decay of order $1/r$; they are not the bottleneck.

Constant phase ($t > T_{\text{ann}}$). Once σ freezes at $\sigma_{\min}$, mode k decays at the fixed rate $\lambda(k, \sigma_{\min}) = \sigma_{\min}^2|k|^2 e^{-\sigma_{\min}^2|k|^2/2}$. Under the assumption $\sigma_{\min}^2 K_{\max}^2 \geq 2$, the function $f(u) = ue^{-u/2}$ is strictly decreasing for $u \geq 2$, so the rate is minimised at $k = K_{\max}$:

$$\lambda(k, \sigma_{\min}) \geq \lambda_{\min} := \sigma_{\min}^2 K_{\max}^2 \exp\Big(-\tfrac{1}{2}\sigma_{\min}^2 K_{\max}^2\Big) > 0, \qquad \forall\, |k| \leq K_{\max}.$$

Conclusion. Combining both phases and using $\Lambda(k, T_{\text{ann}}) \geq 0$,

$$\Lambda(k,\, T_{\text{ann}} + \Delta t) \geq \lambda_{\min}\,\Delta t.$$

Setting $\Delta t = \frac{1}{\lambda_{\min}} \log(1/\epsilon)$ ensures $\Lambda \geq \log(1/\epsilon)$ for all $|k| \leq K_{\max}$, yielding

$$T_{\epsilon,\text{anneal}} \leq \frac{1}{r} \log\Big(\frac{\sigma_0}{\sigma_{\min}}\Big) + \frac{1}{\lambda_{\min}(K_{\max})} \log(1/\epsilon).$$

Choosing $\sigma_{\min} = \sqrt{2}/K_{\max}$ gives $\lambda_{\min} = 2e^{-1}$ independently of $K_{\max}$, so the second term is $\frac{e}{2}\log(1/\epsilon)$ and the first is $\frac{1}{r}\log(\sigma_0 K_{\max}/\sqrt{2})$, establishing the $O\big(\frac{1}{r}\log K_{\max}\big)$ headline bound. □

C A Hitchhiker's Guide to Optimal Transport and Wasserstein Gradient Flows

This appendix summarizes the minimal elements of optimal transport and Wasserstein gradient flows needed in Sections 4–5. We state definitions and structural results without full proofs; detailed treatments can be found in Santambrogio (2015), Villani (2009) and Ambrosio & Savaré (2007).

Throughout, $\Omega \subset \mathbb{R}^d$ is compact and convex, and $\mathcal{P}_2(\Omega)$ denotes probability measures with finite second moment.

C.1 Optimal Transport and the 2-Wasserstein Distance

Monge formulation. Given $\mu, \nu \in \mathcal{P}_2(\Omega)$, the quadratic Monge problem seeks

$$\inf_{T_{\#}\mu = \nu} \int_\Omega \|x - T(x)\|^2\, d\mu(x), \tag{25}$$

where T pushes μ onto ν. This formulation may fail to admit a solution because maps are restrictive.

Kantorovich relaxation. The relaxed problem considers transport plans:

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x - y\|^2\, d\pi(x, y), \tag{26}$$

where $\Pi(\mu, \nu)$ is the set of couplings with marginals μ and ν. This problem always admits a minimizer.

The induced distance W_2 satisfies non-negativity, symmetry, and the triangle inequality; thus $(\mathcal{P}_2(\Omega), W_2)$ is a metric space.

Existence of transport maps. For quadratic cost, if μ is absolutely continuous, then the optimal plan is induced by a map

$$T = \nabla\varphi$$

for a convex potential φ (Brenier's theorem). In the smooth density setting of Section 5, Wasserstein updates are therefore realized by transport maps.

C.2 Functionals and First Variations

A *functional* is a map

$$F : \mathcal{P}_2(\Omega) \to \mathbb{R}.$$

Examples used in this paper include:

- the smoothed KL divergence $F_\sigma[q] = \sigma^2 \mathrm{KL}(q_\sigma \| p_\sigma)$,
- entropic optimal transport divergences,
- entropy $\int q \log q$.

First variation. Let $q \in \mathcal{P}_2(\Omega)$ with density and consider perturbations

$$q_s = (Id + s\xi)_\# q,$$

for smooth compactly supported vector fields ξ. The first variation of F at q is defined via

$$\frac{d}{ds} F[q_s]\Big|_{s=0} = \int_\Omega \nabla \frac{\delta F}{\delta q}(x) \cdot \xi(x)\, q(x)\, dx. \tag{27}$$

The function $\frac{\delta F}{\delta q}$ is the *functional derivative* of F.

For example,

$$\frac{\delta}{\delta q} \mathrm{KL}(q\|p) = \log \frac{q}{p} + 1.$$

C.3 Continuity Equation and Velocity Fields

If particles evolve according to

$$\dot{x}(t) = v(t, x),$$

then their density satisfies the continuity equation

$$\partial_t q + \nabla \cdot (qv) = 0. \tag{28}$$

Thus probability evolution is fully determined by a velocity field.

C.4 Wasserstein Gradient Flows

The space $\mathcal{P}_2(\Omega)$ admits a formal Riemannian structure in which:

- tangent vectors at q are velocity fields v,
- the metric is

$$\|v\|_{T_q}^2 = \int_\Omega \|v(x)\|^2 q(x)\, dx.$$

Under this geometry, the steepest descent flow of a functional F satisfies

$$\partial_t q = \nabla \cdot \left(q \nabla \frac{\delta F}{\delta q} \right). \tag{29}$$

Equivalently,

$$\partial_t q + \nabla \cdot (qv) = 0, \qquad v = -\nabla \frac{\delta F}{\delta q}.$$

Thus the velocity field is the negative spatial gradient of the functional derivative. This principle underlies the drift structure derived in Section 5.

C.5 The JKO Scheme (Minimizing Movements)

The Wasserstein gradient flow of F can be discretized via the Jordan–Kinderlehrer–Otto (JKO) scheme:

$$q_{n+1} = \arg\min_{q \in \mathcal{P}_2(\Omega)} \left\{ F[q] + \frac{1}{2\tau} W_2^2(q, q_n) \right\}. \tag{30}$$

This is the implicit Euler scheme in Wasserstein space.

First-order optimality. Let T_{n+1} be the optimal map from q_{n+1} to q_n. Then the Euler–Lagrange condition yields

$$\frac{Id - T_{n+1}}{\tau} = -\nabla \frac{\delta F}{\delta q}(q_{n+1}). \tag{31}$$

Thus

$$T_{n+1}(x) = x + \tau v_{n+1}(x), \qquad v_{n+1} = -\nabla \frac{\delta F}{\delta q}(q_{n+1}).$$

As $\tau \to 0$, the JKO interpolants converge to the solution of the gradient-flow PDE.

C.6 Frozen-Field Approximation

The JKO update is implicit because the velocity depends on q_{n+1}. A tractable explicit approximation freezes the velocity at q_n:

$$q_{n+1} \approx (Id + \tau v[q_n])_{\#} q_n.$$

This is the explicit Euler discretization of the Wasserstein gradient flow. Section 5 shows that the stop-gradient training objective implements precisely this frozen-field step.

Connection to this work. In Section 5:

- The smoothed KL energy defines a functional F_σ.
- Its functional derivative yields the drift velocity.
- The drifting PDE is its Wasserstein gradient flow.
- The stop-gradient loss corresponds to the frozen-field explicit Euler approximation of the JKO scheme.

This geometric viewpoint applies to any sufficiently regular divergence functional.

D Complete Proofs for Section 5.3

Standing assumptions (densities). Throughout this appendix we work on a compact convex domain $\Omega \subset \mathbb{R}^d$ with nonempty interior. We assume that all measures $q \in \mathcal{P}(\Omega)$ considered in the JKO scheme admit densities (still denoted q) with respect to Lebesgue measure on Ω, and we extend them by 0 outside Ω. Moreover, for each fixed $\sigma > 0$ we assume the iterates satisfy

$$0 < m_\sigma \;\le\; q_\sigma(x) \;\le\; M_\sigma < \infty \qquad \text{for all } x \in \Omega, \tag{32}$$

for constants m_σ, M_σ that may depend on σ and Ω but not on the iterate index. (On a compact Ω, a uniform lower bound holds automatically for q_σ when q is a probability measure, since φ_σ is strictly positive and Ω is bounded; a uniform upper bound is mild and holds, e.g., if $q \in L^\infty(\Omega)$.)

We fix $\sigma > 0$ and a data distribution $p \in \mathcal{P}(\Omega)$ with density p. We define

$$q_\sigma := q * \varphi_\sigma, \qquad p_\sigma := p * \varphi_\sigma,$$

where convolution is taken in $\mathbb{R}^d$ after extending q, p by 0 outside Ω. The energy functional is

$$F_\sigma[q] := \sigma^2 \, D_{\mathrm{KL}}(q_\sigma \| p_\sigma) = \sigma^2 \int_{\mathbb{R}^d} q_\sigma(x) \log \frac{q_\sigma(x)}{p_\sigma(x)} \, dx.$$

Because φ_σ is smooth and strictly positive and Ω is bounded, $q_\sigma, p_\sigma \in C^\infty(\mathbb{R}^d)$ and $q_\sigma, p_\sigma > 0$ everywhere.

D.1 Properties of the Smoothed KL Energy (Proposition 5.5)

We collect the analytic properties of F_σ that underpin the variational theory: lower semicontinuity (needed for existence in the JKO scheme), the first variation (which recovers the drift velocity), and the convolution approximation error (which quantifies the gap between $v_\sigma[q]$ and $-\sigma^2 \nabla \log(q_\sigma/p_\sigma)$).

D.1.1 Lower Semicontinuity of F_σ (Proposition 5.5(i))

Proof. Let $q^{(j)} \to q$ narrowly in $\mathcal{P}(\Omega)$. Since φ_σ is bounded and continuous,

$$q_\sigma^{(j)}(x) = \int_\Omega \varphi_\sigma(x-y) \, q^{(j)}(dy) \;\longrightarrow\; \int_\Omega \varphi_\sigma(x-y) \, q(dy) = q_\sigma(x)$$

pointwise for every $x \in \mathbb{R}^d$. Moreover, on a compact Ω the smoothed densities are uniformly bounded: $q_\sigma^{(j)}, q_\sigma \in [m_\sigma, M_\sigma]$ for all j.

The integrand $u \mapsto u \log(u/p_\sigma(x))$ is continuous and convex on $(0, \infty)$. On the bounded interval $[m_\sigma, M_\sigma]$ it is uniformly bounded, so dominated convergence gives

$$\int_{\mathbb{R}^d} q_\sigma^{(j)}(x) \log \frac{q_\sigma^{(j)}(x)}{p_\sigma(x)} \, dx \;\longrightarrow\; \int_{\mathbb{R}^d} q_\sigma(x) \log \frac{q_\sigma(x)}{p_\sigma(x)} \, dx.$$

In particular, $F_\sigma[q^{(j)}] \to F_\sigma[q]$, which is stronger than lower semicontinuity. (In fact, under the standing bounds equation 32, F_σ is continuous with respect to narrow convergence on $\mathcal{P}(\Omega)$.) □

D.1.2 First Variation of F_σ

Proof of Proposition 5.5(ii, iii). We compute the directional derivative of F_σ along transport perturbations.

Let $q \in \mathcal{P}(\Omega)$ (with density) and let $\xi \in C_c^\infty(\Omega; \mathbb{R}^d)$. For $|s|$ small, define $T_s := \mathrm{Id} + s\xi$ and $q_s := (T_s)_\# q$. Then q_s satisfies the continuity equation in the sense of distributions:

$$\partial_s q_s \big|_{s=0} = -\nabla \cdot (q \, \xi).$$

Convolving with φ_σ and using commutation of convolution with derivatives,

$$\partial_s q_{s,\sigma} \big|_{s=0} = (\partial_s q_s|_{s=0}) * \varphi_\sigma = -(\nabla \cdot (q\xi)) * \varphi_\sigma = -\nabla \cdot \big((q\xi) * \varphi_\sigma\big).$$

Differentiate $F_\sigma[q_s]$ at $s = 0$:

$$\begin{aligned} \frac{d}{ds} F_\sigma[q_s]\Big|_{s=0} &= \sigma^2 \int_{\mathbb{R}^d} \Big(1 + \log \frac{q_\sigma}{p_\sigma}\Big)(x) \, \partial_s q_{s,\sigma}\big|_{s=0}(x) \, dx \\ &= -\sigma^2 \int_{\mathbb{R}^d} \Big(1 + \log \frac{q_\sigma}{p_\sigma}\Big)(x) \, \nabla \cdot \big((q\xi) * \varphi_\sigma\big)(x) \, dx. \end{aligned}$$

We now integrate by parts on $\mathbb{R}^d$. This is justified because $(q\xi) * \varphi_\sigma$ decays at least Gaussianly as $\|x\| \to \infty$ (since $q\xi$ is compactly supported in Ω), while $\nabla \log(q_\sigma/p_\sigma)$ has at most polynomial growth; hence boundary terms vanish. Thus,

$$\frac{d}{ds} F_\sigma[q_s]\Big|_{s=0} = \sigma^2 \int_{\mathbb{R}^d} \nabla \log \frac{q_\sigma}{p_\sigma}(x) \cdot \big((q\xi) * \varphi_\sigma\big)(x) \, dx.$$

Write $(q\xi) * \varphi_\sigma(x) = \int_\Omega \varphi_\sigma(x-y)\,\xi(y)\,q(y)\,dy$ and apply Fubini:

$$\begin{aligned}\frac{d}{ds}F_\sigma[q_s]\Big|_{s=0} &= \sigma^2 \int_\Omega \left(\int_{\mathbb{R}^d} \nabla \log \frac{q_\sigma}{p_\sigma}(x)\,\varphi_\sigma(x-y)\,dx \right) \cdot \xi(y)\,q(y)\,dy \\ &= \int_\Omega \sigma^2 \big(\varphi_\sigma * \nabla \log(q_\sigma/p_\sigma)\big)(y) \cdot \xi(y)\,q(y)\,dy.\end{aligned}$$

By the definition of first variation in Wasserstein calculus, this identifies

$$\nabla \frac{\delta F_\sigma}{\delta q}(y) = \sigma^2 \big(\varphi_\sigma * \nabla \log(q_\sigma/p_\sigma)\big)(y),$$

and hence (up to an additive constant)

$$\frac{\delta F_\sigma}{\delta q}(y) = \sigma^2 \big(\varphi_\sigma * \log(q_\sigma/p_\sigma)\big)(y) + C.$$

Regularity follows since $\log(q_\sigma/p_\sigma) \in C^\infty(\mathbb{R}^d)$ and convolution with φ_σ preserves smoothness. □

D.1.3 The Convolution Approximation Error

The velocity field $v_\sigma[q] = -\nabla(\delta F_\sigma/\delta q)$ differs from $-\sigma^2 \nabla \log(q_\sigma/p_\sigma)$ by a convolution smoothing. The following quantifies this gap.

Proposition D.1. *Let $g \in C^2(\mathbb{R}^d)$. Then for all $x \in \mathbb{R}^d$:*

$$\big|(\varphi_\sigma * g)(x) - g(x)\big| \;\le\; \frac{\sigma^2 d}{2}\,\|\nabla^2 g\|_\infty.$$

In particular, the velocity field satisfies

$$\big\|v_\sigma[q] + \sigma^2 \nabla \log(q_\sigma/p_\sigma)\big\|_\infty \;\le\; \frac{\sigma^4 d}{2}\,\big\|\nabla^2 \log(q_\sigma/p_\sigma)\big\|_\infty.$$

Proof. Taylor-expand g around x:

$$g(x+z) = g(x) + \nabla g(x)\cdot z + \frac{1}{2} z^\top \nabla^2 g(\xi) z$$

for some ξ on the segment between x and $x+z$. Integrate against $\varphi_\sigma(z)\,dz$. The linear term vanishes by symmetry. The remainder is bounded by $\frac{1}{2}\|\nabla^2 g\|_\infty\,\mathbb{E}[\|Z\|^2] = \frac{\sigma^2 d}{2}\|\nabla^2 g\|_\infty$ for $Z \sim \mathcal{N}(0, \sigma^2 I)$. Apply this with $g = \log(q_\sigma/p_\sigma)$ and take a gradient, producing the extra σ^2 factor. □

D.1.4 Lipschitz Continuity of the Velocity Field

The following Lipschitz estimate is used in the convergence proof (Theorem D.5) and the consistency bound $W_2(\tilde{q}^\tau_{n+1}, q^\tau_{n+1}) = O(\tau^{3/2})$.

Lemma D.2. *For fixed $\sigma > 0$, under the standing bounds equation 32 the map $\mu \mapsto v_\sigma[\mu]$ is Lipschitz from $(\mathcal{P}(\Omega), W_2)$ to $(C(\Omega;\mathbb{R}^d), \|\cdot\|_\infty)$.*

Proof. Let $\mu, \nu \in \mathcal{P}(\Omega)$. By definition,

$$v_\sigma[\mu] - v_\sigma[\nu] = -\sigma^2 \nabla\Big(\varphi_\sigma * \big[\log(\mu_\sigma/p_\sigma) - \log(\nu_\sigma/p_\sigma)\big]\Big) = -\sigma^2 \nabla\Big(\varphi_\sigma * \log(\mu_\sigma/\nu_\sigma)\Big).$$

Thus,

$$\|v_\sigma[\mu] - v_\sigma[\nu]\|_\infty \le \sigma^2 \|\nabla \varphi_\sigma\|_{L^1(\mathbb{R}^d)}\,\|\log(\mu_\sigma/\nu_\sigma)\|_{L^\infty(\Omega)}.$$

Using the lower bound m_σ in equation 32 and the mean value theorem for log,

$$\|\log(\mu_\sigma/\nu_\sigma)\|_\infty \le \frac{1}{m_\sigma}\|\mu_\sigma - \nu_\sigma\|_\infty.$$

Finally, for any $\mu, \nu \in \mathcal{P}(\Omega)$ and $x \in \Omega$, the map $y \mapsto \varphi_\sigma(x-y)$ is Lipschitz with constant $\|\nabla\varphi_\sigma\|_\infty$, hence

$$\|\mu_\sigma - \nu_\sigma\|_{L^\infty(\Omega)} \le \|\nabla\varphi_\sigma\|_\infty W_1(\mu,\nu) \le \|\nabla\varphi_\sigma\|_\infty W_2(\mu,\nu), \tag{33}$$

where we used $W_1 \le W_2$ on a bounded metric space. Combining the inequalities yields

$$\|v_\sigma[\mu] - v_\sigma[\nu]\|_\infty \le L_\sigma W_2(\mu,\nu),$$

for a constant L_σ depending on σ, Ω, and m_σ. □

D.2 The JKO Scheme for F_σ

We now establish that the JKO minimizing-movement scheme applied to F_σ is well-posed, enjoys monotone energy descent, and converges as $\tau \to 0$ to a solution of the gradient-flow PDE $\partial_t q + \nabla\cdot(q\, v_\sigma[q]) = 0$. Our treatment follows closely the exposition in Chapter 8 in (Santambrogio, 2015).

D.2.1 Well-Posedness: Existence, Uniqueness, and Energy Descent

Proposition D.3 (JKO well-posedness)**.** *For each $n \ge 0$ and $\tau > 0$, the JKO functional $\mathcal{J}(q) := F_\sigma[q] + \frac{1}{2\tau}W_2^2(q, q_n^\tau)$ admits a unique minimizer $q_{n+1}^\tau \in \mathcal{P}(\Omega)$, and the energy decreases monotonically: $F_\sigma[q_{n+1}^\tau] \le F_\sigma[q_n^\tau]$.*

Proof. Define the JKO functional

$$\mathcal{J}(q) := F_\sigma[q] + \frac{1}{2\tau}W_2^2(q, q_n^\tau), \qquad q \in \mathcal{P}(\Omega).$$

(i) Existence. Let $(q^{(j)})_{j\ge1}$ be a minimizing sequence. Since Ω is compact, $\mathcal{P}(\Omega)$ is compact for narrow convergence (equivalently for W_2). Hence, up to a subsequence, $q^{(j)} \to \hat{q}$ narrowly. By Proposition 5.5(i), F_σ is narrowly lower semicontinuous. On a compact domain, $W_2(\cdot, q_n^\tau)$ is continuous under narrow convergence. Therefore $\mathcal{J}(\hat{q}) \le \liminf_j \mathcal{J}(q^{(j)}) = \inf \mathcal{J}$ and $\hat{q}$ is a minimizer.

(ii) Uniqueness. We show strict convexity of $\mathcal{J}$ along *mixtures*. The map $q \mapsto q_\sigma$ is linear, and the functional $\rho \mapsto D_{\mathrm{KL}}(\rho\|p_\sigma) = \int \rho\log(\rho/p_\sigma)$ is strictly convex in ρ on strictly positive densities. Gaussian convolution is injective on probability measures, so $q_1 \ne q_2$ implies $q_{1,\sigma} \ne q_{2,\sigma}$. Hence for $\lambda \in (0,1)$,

$$F_\sigma[\lambda q_1 + (1-\lambda)q_2] = \sigma^2 D_{\mathrm{KL}}(\lambda q_{1,\sigma} + (1-\lambda)q_{2,\sigma}\|p_\sigma) < \lambda F_\sigma[q_1] + (1-\lambda)F_\sigma[q_2].$$

Moreover, $q \mapsto W_2^2(q, q_n^\tau)$ is convex along mixtures: if π_i is optimal (or ε-optimal) between q_i and q_n^τ, then $\lambda\pi_1 + (1-\lambda)\pi_2$ is a coupling between $\lambda q_1 + (1-\lambda)q_2$ and q_n^τ, giving

$$W_2^2(\lambda q_1 + (1-\lambda)q_2, q_n^\tau) \le \lambda W_2^2(q_1, q_n^\tau) + (1-\lambda)W_2^2(q_2, q_n^\tau).$$

Thus $\mathcal{J}$ is strictly convex along mixtures, hence its minimizer is unique.

(iii) Energy descent. By optimality of q_{n+1}^τ and choosing the competitor $q = q_n^\tau$,

$$F_\sigma[q_{n+1}^\tau] + \frac{1}{2\tau}W_2^2(q_{n+1}^\tau, q_n^\tau) \le F_\sigma[q_n^\tau],$$

which implies $F_\sigma[q_{n+1}^\tau] \le F_\sigma[q_n^\tau]$. □

D.2.2 Euler–Lagrange Condition (Equation 14)

Proof. Let $\hat{q} := q_{n+1}^\tau$ denote the unique minimizer of the JKO step

$$\hat{q} \in \operatorname*{arg\,min}_{q \in \mathcal{P}(\Omega)} \left\{ F_\sigma[q] + \frac{1}{2\tau} W_2^2(q, q_n^\tau) \right\}.$$

Under the standing density assumptions, $\hat{q}$ is absolutely continuous on Ω. Hence Brenier's theorem applies: there exists an optimal transport map $T_n : \Omega \to \Omega$ pushing $\hat{q}$ to q_n^τ, and a Kantorovich potential $\bar{\varphi}$ (unique up to additive constants) such that

$$T_n(x) = x - \nabla\bar{\varphi}(x) \qquad \hat{q}\text{-a.e.}$$

Step 1: First-order optimality of $\hat{q}$. Consider perturbations $q_\varepsilon = (\mathrm{Id} + \varepsilon\xi)_\# \hat{q}$ with $\xi \in C_c^\infty(\Omega; \mathbb{R}^d)$. Optimality of $\hat{q}$ implies

$$\frac{d}{d\varepsilon} \left(F_\sigma[q_\varepsilon] + \frac{1}{2\tau} W_2^2(q_\varepsilon, q_n^\tau) \right) \Big|_{\varepsilon=0} = 0.$$

For the F_σ term, Proposition 5.5(ii) yields

$$\frac{d}{d\varepsilon} F_\sigma[q_\varepsilon]\Big|_{\varepsilon=0} = \int_\Omega \nabla \left(\frac{\delta F_\sigma}{\delta q} \right)(x) \cdot \xi(x)\, \hat{q}(x)\, dx, \qquad \frac{\delta F_\sigma}{\delta q}(x) = \sigma^2(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma))(x) + C.$$

For the Wasserstein term, standard OT first-variation calculus (e.g., (Santambrogio, 2015), Prop. 7.17) gives

$$\frac{d}{d\varepsilon} \frac{1}{2\tau} W_2^2(q_\varepsilon, q_n^\tau)\Big|_{\varepsilon=0} = -\int_\Omega \frac{1}{\tau} \nabla\bar{\varphi}(x) \cdot \xi(x)\, \hat{q}(x)\, dx.$$

Summing and using arbitrariness of ξ gives, $\hat{q}$-a.e.,

$$\nabla \left(\sigma^2(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma)) \right)(x) - \frac{1}{\tau} \nabla\bar{\varphi}(x) = 0,$$

hence (integrating in x) there exists a constant C such that

$$\sigma^2(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma))(x) + \frac{1}{\tau}\bar{\varphi}(x) = C \qquad \hat{q}\text{-a.e.},$$

which is equation 14.

Step 2: Velocity identification. Differentiating the previous equation:

$$\frac{1}{\tau} \nabla\bar{\varphi}(x) = -\nabla \Big(\sigma^2(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma)) \Big)(x).$$

Using $T_n(x) = x - \nabla\bar{\varphi}(x)$ yields

$$\frac{T_n(x) - x}{\tau} = -\frac{1}{\tau} \nabla\bar{\varphi}(x) = -\nabla \Big(\sigma^2(\varphi_\sigma * \log(\hat{q}_\sigma / p_\sigma)) \Big)(x) = v_\sigma[\hat{q}](x), \qquad \hat{q}\text{-a.e.},$$

which is equation 14. □

D.2.3 A Priori Estimates

Proposition D.4 (A priori estimates)**.** *Let $(q_k^\tau)_{k \geq 0}$ be the JKO iterates and let $\tilde{q}^\tau$ denote the W_2-geodesic interpolation between successive iterates.*

(i) (Energy summability.) $\displaystyle\sum_{k=0}^{N-1} \frac{W_2^2(q_{k+1}^\tau, q_k^\tau)}{\tau} \leq 2(F_\sigma[q_0] - \inf F_\sigma) =: C_0.$

(ii) (Hölder regularity.) $W_2(\tilde{q}^\tau(t), \tilde{q}^\tau(s)) \le C_0^{1/2}\,|t-s|^{1/2}$ *for all* $0 \le s < t \le T$.

Proof. **(i) Energy summability.** By optimality of q^τ_{k+1} and competitor q^τ_k,

$$F_\sigma[q^\tau_{k+1}] + \frac{1}{2\tau} W_2^2(q^\tau_{k+1}, q^\tau_k) \le F_\sigma[q^\tau_k]. \tag{34}$$

Rearrange and sum over $k = 0, \dots, N-1$:

$$\sum_{k=0}^{N-1} \frac{W_2^2(q^\tau_{k+1}, q^\tau_k)}{\tau} \le 2 \sum_{k=0}^{N-1} (F_\sigma[q^\tau_k] - F_\sigma[q^\tau_{k+1}]) = 2(F_\sigma[q_0] - F_\sigma[q^\tau_N]) \le 2(F_\sigma[q_0] - \inf F_\sigma) =: C_0.$$

(ii) Hölder regularity. Let $\tilde{q}^\tau$ be the W_2-geodesic interpolation between successive iterates. Its metric derivative satisfies for $t \in (k\tau, (k+1)\tau)$:

$$|(\tilde{q}^\tau)'|(t) = \frac{W_2(q^\tau_{k+1}, q^\tau_k)}{\tau}.$$

Thus for $0 \le s < t \le T$, by Cauchy–Schwarz,

$$\begin{aligned} W_2(\tilde{q}^\tau(t), \tilde{q}^\tau(s)) &\le \int_s^t |(\tilde{q}^\tau)'|(r)\, dr \le |t-s|^{1/2} \left(\int_s^t |(\tilde{q}^\tau)'|^2(r)\, dr \right)^{1/2} \\ &\le |t-s|^{1/2} \left(\int_0^T |(\tilde{q}^\tau)'|^2(r)\, dr \right)^{1/2} = |t-s|^{1/2} \left(\sum_k \frac{W_2^2(q^\tau_{k+1}, q^\tau_k)}{\tau} \right)^{1/2} \le C_0^{1/2} |t-s|^{1/2}. \end{aligned}$$

□

D.2.4 Convergence to Gradient Flow

Theorem D.5 (Convergence to gradient flow)**.** *As* $\tau \to 0$*, the JKO interpolants converge (up to subsequences) uniformly in* W_2 *to a limit* $q \in C^{0,1/2}([0,T]; \mathcal{P}(\Omega))$ *satisfying* $\partial_t q + \nabla \cdot (q\, v_\sigma[q]) = 0$ *in distributions, with* $F_\sigma[q(t)] \le F_\sigma[q(s)]$ *for* $0 \le s < t \le T$ *and* $q(0) = q_0$.

Proof. **Step 1: Compactness and extraction.** By Proposition D.4(ii), $\{\tilde{q}^\tau\}_{\tau>0}$ is equicontinuous in $C([0,T]; \mathcal{P}(\Omega))$ with the W_2 metric. Since $\mathcal{P}(\Omega)$ is compact (compact Ω), Arzelà–Ascoli gives $\tau_j \to 0$ and $q \in C^{0,1/2}([0,T]; \mathcal{P}(\Omega))$ such that $\tilde{q}^{\tau_j} \to q$ uniformly in W_2 on $[0,T]$. The piecewise-constant interpolation $q^\tau(t)$ converges to the same limit.

Step 2: Continuity equation for the geodesic interpolant and momentum bounds. Fix $k \ge 0$ and let $T_{k+1} : \Omega \to \Omega$ be the optimal transport map pushing q^τ_{k+1} to q^τ_k. By Euler-Lagrange condition 14,

$$T_{k+1}(x) = x + \tau\, v_\sigma[q^\tau_{k+1}](x) \qquad q^\tau_{k+1}\text{-a.e.} \tag{35}$$

Define the displacement interpolation on $(k\tau, (k+1)\tau]$ by

$$s(t) := \frac{t - k\tau}{\tau} \in (0,1], \qquad X_{k+1,s} := (1-s)\, + s\, T_{k+1}, \qquad \tilde{q}^\tau(t) := (X_{k+1,s(t)})_\# q^\tau_{k+1}.$$

(Thus $\tilde{q}^\tau(k\tau^+) = q^\tau_{k+1}$ and $\tilde{q}^\tau((k+1)\tau) = q^\tau_k$.)

On this interval the curve $\tilde{q}^\tau$ solves the continuity equation

$$\partial_t \tilde{q}^\tau + \nabla \cdot (\tilde{q}^\tau\, w^\tau) = 0 \quad \text{in } \mathcal{D}'((0,T) \times \Omega),$$

with the (a.e.-defined) velocity field

$$w^\tau(t, \cdot) := \frac{T_{k+1} - }{\tau} \circ (X_{k+1,s(t)})^{-1}.$$

Moreover, the Benamou–Brenier action along each geodesic segment yields

$$\int_{k\tau}^{(k+1)\tau}\int_\Omega \|w^\tau(t,x)\|^2\, d\tilde{q}^\tau(t)(x)\, dt = \frac{1}{\tau}\, W_2^2(q_{k+1}^\tau, q_k^\tau). \tag{36}$$

Summing over k and using Proposition D.4(i) gives the uniform bound

$$\int_0^T \|w^\tau(t)\|^2_{L^2(\tilde{q}^\tau(t))}\, dt = \sum_k \frac{W_2^2(q_{k+1}^\tau, q_k^\tau)}{\tau} \le C_0.$$

Define the flux measures $E^\tau := \tilde{q}^\tau\, w^\tau$ on $[0,T]\times\Omega$. The above bound implies $\{E^\tau\}$ is uniformly bounded as vector-valued Radon measures, so along the subsequence τ_j we have $E^{\tau_j} \overset{*}{\rightharpoonup} E$ and, passing to the limit in the continuity equation,

$$\partial_t q + \nabla\cdot E = 0 \quad \text{in } \mathcal{D}'((0,T)\times\Omega).$$

Step 3: Identification $E = q\, v_\sigma[q]$. We first relate w^τ to the JKO velocity at the right endpoint. By equation 35 we have

$$\frac{T_{k+1}-}{\tau} = v_\sigma[q_{k+1}^\tau] \qquad q_{k+1}^\tau\text{-a.e.}$$

Hence for $t\in(k\tau,(k+1)\tau]$,

$$w^\tau(t,\cdot) = v_\sigma[q_{k+1}^\tau]\circ (X_{k+1,s(t)})^{-1} \qquad \tilde{q}^\tau(t)\text{-a.e.}$$

Next, observe that for $t\in(k\tau,(k+1)\tau]$,

$$W_2\big(\tilde{q}^\tau(t),\, q_{k+1}^\tau\big) \le W_2(q_{k+1}^\tau, q_k^\tau),$$

since displacement interpolants are constant-speed geodesics. Therefore, using Lemma D.2 and Proposition D.4(i),

$$\int_0^T \big\|v_\sigma[\tilde{q}^\tau(t)] - v_\sigma[q_{k(t)+1}^\tau]\big\|_\infty\, dt \;\le\; L_\sigma \int_0^T W_2(\tilde{q}^\tau(t), q_{k(t)+1}^\tau)\, dt \;\le\; L_\sigma \sum_k \tau\, W_2(q_{k+1}^\tau, q_k^\tau) \;\xrightarrow[\tau\to0]{}\; 0,$$

where $k(t)$ denotes the unique index with $t\in(k\tau,(k+1)\tau]$ and we used Cauchy–Schwarz together with $\sum_k \frac{W_2^2(q_{k+1}^\tau,q_k^\tau)}{\tau}\le C_0$.

Now test against any $f\in C([0,T]\times\Omega;\mathbb{R}^d)$. Using the representation of E^τ and the previous estimate,

$$\begin{aligned}\int_0^T\!\!\int_\Omega f\cdot dE^\tau &= \int_0^T\!\!\int_\Omega f(t,x)\cdot w^\tau(t,x)\, d\tilde{q}^\tau(t)(x)\, dt \\ &= \int_0^T\!\!\int_\Omega f(t,x)\cdot v_\sigma[\tilde{q}^\tau(t)](x)\, d\tilde{q}^\tau(t)(x)\, dt \;+\; o(1).\end{aligned}$$

Passing to the limit $\tau_j\to0$ using $\tilde{q}^{\tau_j}\to q$ uniformly in W_2 and the Lipschitz continuity of $v_\sigma[\cdot]$ (Lemma D.2) yields

$$\int_0^T\!\!\int_\Omega f\cdot dE = \int_0^T\!\!\int_\Omega f(t,x)\cdot v_\sigma[q(t)](x)\, dq(t)(x)\, dt.$$

Hence $E = q\, v_\sigma[q]$, and therefore

$$\partial_t q + \nabla\cdot\big(q\, v_\sigma[q]\big) = 0$$

in distributions on $(0,T)\times\Omega$.

Step 4: Energy monotonicity. From equation 34, for each τ the map $t\mapsto F_\sigma[q^\tau(t)]$ is nonincreasing. Fix $0\le s<t\le T$. Then $F_\sigma[q^\tau(t)]\le F_\sigma[q^\tau(s)]$. Along $\tau_j\to0$, we have $q^{\tau_j}(r)\to q(r)$ in W_2, hence by equation 33 $q_\sigma^{\tau_j}(r)\to q_\sigma(r)$ uniformly on Ω. Under the standing bounds equation 32, the integrand $u\mapsto u\log(u/p_\sigma)$ is continuous and dominated on $[m_\sigma, M_\sigma]$, so dominated convergence yields $F_\sigma[q^{\tau_j}(r)]\to F_\sigma[q(r)]$ for $r=s,t$. Passing to the limit gives $F_\sigma[q(t)]\le F_\sigma[q(s)]$.

Step 5: Initial datum. By construction $q^\tau(0)=q_0$ and the 1/2-Hölder bound gives $q(0)=q_0$. □

D.3 Proof of Consistency: ie, Equation 5.3 (Implicit–Explicit Consistency)

Proof. Let $q^* := q^\tau_{n+1}$ be the JKO minimizer and let

$$\tilde{q} := (S_n)_\# q^\tau_n, \qquad S_n(x) = x + \tau v_\sigma[q^\tau_n](x),$$

denote the frozen-field (explicit Euler) update.

Let T_n be the optimal transport map pushing q^* to q^τ_n. By Euler-Langrange Optimality condition 14,

$$T_n(x) = x + \tau v_\sigma[q^*](x), \qquad q^\tau_n = (T_n)_\# q^*.$$

Define

$$R(x) := x + \tau v_\sigma[q^*](x).$$

Then $R = T_n$ and $q^\tau_n = R_\# q^*$.

Step 1: Reduce to pushforwards of the same base measure.

Define

$$\bar{q} := (S_n)_\# q^\tau_n.$$

Using $q^\tau_n = R_\# q^*$,

$$\bar{q} = (S_n)_\#(R_\# q^*) = (S_n \circ R)_\# q^*.$$

Thus $\tilde{q} = \bar{q}$, and the explicit update corresponds to transporting q^* with the map $S_n \circ R$.

To compare with q^*, it is convenient to rewrite both measures as pushforwards of q^τ_n.

Since $q^\tau_n = R_\# q^*$, we have

$$q^* = (R^{-1})_\# q^\tau_n.$$

Hence

$$\tilde{q} = (S_n)_\# q^\tau_n, \qquad q^* = (R^{-1})_\# q^\tau_n.$$

Step 2: Wasserstein stability of pushforwards.

For any maps $A, B : \Omega \to \Omega$ and probability measure μ, the coupling $(A, B)_\# \mu$ yields

$$W_2(A_\# \mu, B_\# \mu) \le \|A - B\|_{L^2(\mu)}.$$

Applying this with $\mu = q^\tau_n$, $A = S_n$, and $B = R^{-1}$ gives

$$W_2(\tilde{q}, q^*) = W_2((S_n)_\# q^\tau_n, (R^{-1})_\# q^\tau_n) \le \|S_n - R^{-1}\|_{L^2(q^\tau_n)}.$$

Step 3: Approximation of the inverse map.

Since

$$R(x) = x + \tau v_\sigma[q^*](x),$$

a first-order Taylor expansion shows that

$$R^{-1}(x) = x - \tau v_\sigma[q^*](x) + O(\tau^2),$$

uniformly on Ω, provided $v_\sigma[q^*]$ is Lipschitz in x.

Therefore

$$S_n(x) - R^{-1}(x) = \tau\big(v_\sigma[q^\tau_n](x) - v_\sigma[q^*](x)\big) + O(\tau^2).$$

Taking the $L^2(q_n^\tau)$ norm yields

$$\|S_n - R^{-1}\|_{L^2(q_n^\tau)} \le \tau \|v_\sigma[q_n^\tau] - v_\sigma[q^*]\|_\infty + O(\tau^2).$$

Step 4: Controlling the velocity difference.

By Lemma D.2,

$$\|v_\sigma[q_n^\tau] - v_\sigma[q^*]\|_\infty \le L_\sigma W_2(q_n^\tau, q^*).$$

From the JKO optimality inequality equation 34,

$$W_2(q_n^\tau, q^*)^2 \le 2\tau\big(F_\sigma[q_n^\tau] - F_\sigma[q^*]\big) \le 2\tau F_\sigma[q_n^\tau].$$

Since $F_\sigma[q_n^\tau]$ is bounded along the scheme,

$$W_2(q_n^\tau, q^*) = O(\sqrt{\tau}).$$

Therefore

$$\|v_\sigma[q_n^\tau] - v_\sigma[q^*]\|_\infty = O(\sqrt{\tau}).$$

Step 5: Final estimate.

Combining the previous bounds,

$$W_2(\tilde{q}, q^*) \le \tau\, O(\sqrt{\tau}) + O(\tau^2) = O(\tau^{3/2}).$$

This proves the consistency bound equation 5.3. □

D.4 Proof of Theorem 5.6 (Stop-Gradient Preserves Wasserstein Discretization)

Proof. **(i) Structural correspondence.** Write the stop-gradient loss as

$$\mathcal{L}(\theta) = \mathbb{E}_{\varepsilon\sim\nu}\big[\|G_\theta(\varepsilon) - t(\varepsilon)\|^2\big], \quad t(\varepsilon) := G_{\theta_n}(\varepsilon) + \tau v_\sigma[q_{\theta_n}]\big(G_{\theta_n}(\varepsilon)\big),$$

where t is treated as a fixed target (no gradient flows through it). If the model class is realizable for t (i.e., there exists θ_{n+1} with $G_{\theta_{n+1}}(\varepsilon) = t(\varepsilon)$ ν-a.e.), then any global minimizer achieves $\mathcal{L}(\theta_{n+1}) = 0$ and satisfies $G_{\theta_{n+1}} = t$ ν-a.e., i.e.

$$G_{\theta_{n+1}} = S_n \circ G_{\theta_n} \qquad \nu\text{-a.e.}, \quad S_n(x) := x + \tau v_\sigma[q_{\theta_n}](x).$$

Pushing forward ν yields $q_{\theta_{n+1}} = (S_n)_\# q_{\theta_n}$, which is exactly the frozen-field explicit Euler step.

(ii) Necessity (effect of removing stop-gradient). Without stop-gradient, the coupled loss becomes

$$\mathcal{L}_{\text{coupled}}(\theta) = \mathbb{E}_\varepsilon\Big[\big\|G_\theta(\varepsilon) - \big(G_\theta(\varepsilon) + \tau v_\sigma[q_\theta](G_\theta(\varepsilon))\big)\big\|^2\Big] = \tau^2\, \mathbb{E}_\varepsilon\big[\|v_\sigma[q_\theta](G_\theta(\varepsilon))\|^2\big] = \tau^2 \|v_\sigma[q_\theta]\|^2_{L^2(q_\theta)}.$$

Its gradient includes both the pathwise term and the distribution-feedback term:

$$\nabla_\theta \mathcal{L}_{\text{coupled}} = 2\tau^2\, \mathbb{E}_\varepsilon\Big[v \cdot \big((D_x v)\nabla_\theta G_\theta + (D_q v)\nabla_\theta q_\theta\big)\Big], \quad v := v_\sigma[q_\theta](G_\theta(\varepsilon)).$$

Unlike the stop-gradient objective, this is *not* the regression of an explicit Euler target for a fixed velocity field: as θ changes, the target velocity itself changes via q_θ (the $(D_q v)\nabla_\theta q_\theta$ term). Consequently, minimizing $\mathcal{L}_{\text{coupled}}$ does not implement the frozen-field discretization of the Wasserstein gradient flow; it can admit spurious stationary points/minima where the loss decreases by reducing the velocity norm on the current support of q_θ ("drift collapse") without transporting mass toward p. □

E Schedule Ablation

We compare exponential ($\sigma(t) = \sigma_0 e^{-rt}$), linear ($\sigma(t) = \sigma_0(1 - t/T)$), and cosine ($\sigma(t) = \sigma_0 \cos(\pi t/2T)$) schedules, all sweeping from $\sigma_0 = 1.5$ to $\sigma_{\min} = 0.03$. The exponential schedule converges fastest for all k. This is consistent with the activation-time analysis: the exponential schedule reaches $\sigma^2|k|^2 = 2$ earliest for each k.

F Toy Experiments

Our experiments validate the theoretical predictions of §3–§5.3 on synthetic benchmarks. All models use 3-layer MLPs ($d_{\text{hidden}} = 256$, ReLU, Adam, lr 10^{-3}, batch size 2048).

F.1 Score-matching verification (1).

Let p be a 4-mode Gaussian mixture and $q = \mathcal{N}(0, \sigma_q^2 I)$. For each bandwidth σ, we evaluate the empirical kernel mean-shift drift $\hat{V}_{p,q}^{(\sigma)}(x) = \frac{\sum_i \varphi_\sigma(x - x_i^p)\,(x_i^p - x)}{\sum_i \varphi_\sigma(x - x_i^p)} - \frac{\sum_j \varphi_\sigma(x - x_j^q)\,(x_j^q - x)}{\sum_j \varphi_\sigma(x - x_j^q)}$ at $N = 50$k samples and compare it against the analytical form $V^{(\sigma)}(x) = \sigma^2\big(\nabla \log p_\sigma(x) - \nabla \log q_\sigma(x)\big)$, where the smoothed densities $p_\sigma = p * \mathcal{N}(0, \sigma^2 I)$ and $q_\sigma = q * \mathcal{N}(0, \sigma^2 I)$ admit closed-form scores. Writing $s^2 = \sigma_p^2 + \sigma^2$ for the inflated component variance and $\phi_k(x) = \mathcal{N}(x;\, \mu_k,\, s^2 I)$:

$$\nabla \log p_\sigma(x) \;=\; \frac{\sum_k w_k\, \phi_k(x)\,(\mu_k - x)}{s^2 \sum_k w_k\, \phi_k(x)}, \qquad \nabla \log q_\sigma(x) \;=\; -\,\frac{x - \mu_q}{\sigma_q^2 + \sigma^2}\,. \tag{37}$$

The pointwise ℓ_2 error $e(x) = \|\hat{V}_{p,q}^{(\sigma)}(x) - V^{(\sigma)}(x)\|_2$ has mean 4.9×10^{-3} at $\sigma = 0.3$ confirming Equation 6.

F.2 Spectral convergence times: Figure (2).

We initialize a small perturbation $\delta q(x, 0) = A \sum_{k \in \mathcal{K}} \cos(kx)$ with amplitude $A = 10^{-6}$ and tracked modes $\mathcal{K} = \{1, 2, \ldots, 20\}$, then evolve the linearized Fourier dynamics $\hat{\delta q}(k, t{+}\mathrm{d}t) = \hat{\delta q}(k, t)\, \exp\big(-\lambda(k, \sigma(t))\,\mathrm{d}t\big)$ with per-mode decay rates $\lambda_{\mathrm{G}}(k, \sigma) = \sigma^2 k^2 e^{-\sigma^2 k^2/2}$ (Gaussian kernel) and $\lambda_{\mathrm{E}}(k, \tau) = 2\tau^3 k^2/(1 + \tau^2 k^2)$ (exponential kernel), and define the convergence time as the first T at which $|\hat{\delta q}(k, T)| < \varepsilon\, |\hat{\delta q}(k, 0)|$ with $\varepsilon = 10^{-3}$. For fixed kernels, $T(k) = \log(1/\varepsilon)\,/\,\lambda(k)$; for the annealed Gaussian schedule $\sigma(t) = \sigma_0 e^{-rt}$, $T(k)$ solves the integral equation $\int_0^T \lambda_{\mathrm{G}}(k, \sigma(t))\,\mathrm{d}t = \log(1/\varepsilon)$. Panel (a) shows that numerical threshold-crossing times (markers) match the analytical curves (lines) across modes $k = 1, \ldots, 20$: the fixed Gaussian ($\sigma = 0.3$) exhibits exponential slowdown past $k^* = \sqrt{2}/\sigma \approx 4.7$, the exponential kernel ($\tau = 0.3$) yields polynomial scaling, and the annealed schedule eliminates the bottleneck entirely. Panel (c) plots the total spectral error $E(t) = \sum_k |\hat{\delta q}(k, t)|$ under three annealing schedules (exponential, linear, cosine); the exponential schedule achieves the fastest decay, with k=20 converging $\sim$6$\times$ faster than linear.

F.3 Stop-gradient necessity: Figures 3 and 4

We train generators G_θ on 2D targets (Swiss roll, checkerboard) and track two quantities over training: the mean drift norm $\bar{V} = \frac{1}{M}\sum_{i=1}^{M} \|V_{\mathrm{sg}}(G_\theta(z_i),\, p)\|_2$ (where V_{sg} denotes the exponential-kernel drift with τ=0.05 evaluated at M=2048 generated points with the target detached), and the sliced Wasserstein distance $\mathrm{SW}(G_\theta, p)$ between 5k generated and 5k target samples (200 projections). With stop-gradient, $\bar{V}$ and SW are strongly correlated (log-log $r > 0.95$, 3 seeds) and jointly decay to near zero (final SW = 0.016). Without stop-gradient,where the loss $\|G_\theta(z) - (G_\theta(z) + V(G_\theta(z), p))\|^2$ backpropagates through the drift,$\bar{V}$ drops to $\sim 10^{-8}$ while SW remains at 0.389, demonstrating drift collapse (5.6(ii)). The loss landscape (3) provides geometric intuition: we project the parameter space onto the top-2 PCA directions of training gradients, $\theta(\alpha, \beta) = \theta^* + \alpha\, d_1 + \beta\, d_2$, and evaluate both the training loss $\|V\|^2$ and the sample quality SW on a 31×31 grid. With stop-gradient, the loss minimum aligns with a basin of low SW; without it, the coupled objective admits a $\sim$100$\times$ deeper minimum that corresponds to poor sample quality.

F.4 Sinkhorn drift feasibility: Figure (5).

Energy functional and validity conditions. We use the Sinkhorn divergence of Feydy et al. (2018) as energy functional $F[q] := S_\varepsilon(q, p)$, defined in equation 22. By Theorem 1 of that reference, S_ε is symmetric positive definite, convex in each argument, and metrizes the convergence in law; in particular $S_\varepsilon(q, p) = 0 \Leftrightarrow q = p$. Lower semicontinuity and smoothness of $\delta F/\delta q$ follow from the same reference (Proposition 2). All three conditions (a)-(c) of §6.3 are therefore satisfied, and the template $V = -\nabla_x(\delta F/\delta q)$ yields a valid drift with full JKO-frozen-field-stop-gradient guarantees.

Particle-level gradient. For empirical $q = \frac{1}{N}\sum_i \delta_{x_i}$, the gradient of $S_\varepsilon(q, p)$ with respect to particle x_i follows from Proposition 2 of Feydy et al. (2018) (equations 26-27):

$$\nabla_{x_i} S_\varepsilon(q, p) \;=\; \nabla_{x_i} \mathrm{OT}_\varepsilon(q, p) \;-\; \nabla_{x_i} \mathrm{OT}_\varepsilon(q, q), \tag{38}$$

Concretely, denoting by π_{qp} and π_{qq} the optimal entropic couplings, the barycentric projections

$$T_{q\to p}(x_i) = \frac{\sum_j [\pi_{qp}]_{ij}\, y_j}{\sum_j [\pi_{qp}]_{ij}}, \qquad T_{q\to q}(x_i) = \frac{\sum_j [\pi_{qq}]_{ij}\, x_j}{\sum_j [\pi_{qq}]_{ij}},$$

give the drift

$$V_{\mathrm{SK}}(x_i) \;\propto\; T_{q\to p}(x_i) - T_{q\to q}(x_i), \tag{39}$$

We plug this drift-operator in the stop-gradient training $\mathcal{L}(\theta) = \mathbb{E}_\epsilon[||f_\theta(\epsilon) - sg[\tilde{x}]||^2]$ with $\tilde{x} = x + V_{\mathrm{SK}}(x)$.

Training We train on the checkerboard distribution using the exponential-kernel drift (τ=0.05) and the Sinkhorn-derived drift from the debiased entropic cost $S_\varepsilon(q, p)$ with ε=10^{-4}. Both drifts successfully transport the generator to the target, achieving final SW of 2.07×10^{-2} and 1.42×10^{-2} respectively. The Sinkhorn drift converges to the prescribed distribution, confirming that the gradient-flow formulation of §6.3 yields practical drift operators beyond the original kernel family.

G Image Generation Experimental Details

G.1 Dataset, representation space, and architecture

We train class-conditional generators on ImageNet-256 (1000 classes) in the latent space of the Stable Diffusion VAE, which maps a 256×256×3 image to a 32×32×4 latent. FID is computed after VAE-decoding back to pixels.

The generator is `DitGen-B`, a one-step DiT backbone (hidden size 768, depth 12, 12 heads, patch size 2) with RoPE, RMSNorm, QK-norm, SwiGLU MLPs, and AdaLN class+CFG conditioning.

The feature extractor is a frozen MAE-pretrained ResNet-18, from which we extract multi-scale activations at `conv1`, the four stages `layer1`–`layer4`, and every second residual block within each stage. The drift loss is computed independently on each stream and summed.

G.2 Training protocol

All training details are reported in Table 2

G.3 Drift-operator implementations

G.3.1 Gaussian kernel drift

The Gaussian-kernel drift is a drop-in replacement for the Laplacian baseline. Following the convention of Deng et al. (2026) we compute the kernel as a doubly-stochastic softmax (rather than the raw $\exp(-\|x-y\|^2/2\sigma^2)$ form), with quadratic logits in place of the linear Laplacian logits. Concretely, with bandwidth σ,

$$\mathrm{logits}(x, y) = -\frac{\|x - y\|^2}{2\sigma^2}, \qquad \tilde{k}(x, y) = \sqrt{\mathrm{softmax}_y(\mathrm{logits}) \cdot \mathrm{softmax}_x(\mathrm{logits})}. \tag{40}$$

Table 2: Training hyperparameters for the image-generation ablations (`latent_ablation` configuration).

Optimization	
Optimizer	AdamW (β_1=0.9, β_2=0.95, ϵ=10^{-8})
Weight decay	0.01
Peak learning rate	2×10^{-4}
LR schedule	linear warmup (5,000 steps), then constant
Gradient clipping	$\|g\|_2 \leq 2.0$
Total steps	30,000
Batching	
Labels per step	1024
Generated samples per label	64
GPUs	16 H100
Memory bank	
Positive bank size (per class)	128
Negative bank size (shared)	1000
Pushes per step	64
Positives / negatives drawn per label	64 / 16
CFG & EMA	
CFG sampling range	$[1.0, 4.0]$
CFG power-law exponent	3
Unconditional-only fraction	0
EMA decay	0.999

The doubly-stochastic normalization preserves the antisymmetry $V_{p,q} = -V_{q,p}$ and is required for the empirical-mean Monte-Carlo interpretation of Deng et al. (2026) (Eq. 17 there). All other steps: masking self-interactions, splitting positive/negative streams, computing the attractive and repulsive coefficients—are unchanged from the Laplacian baseline; only the logit transformation differs. Algorithm 1 summarizes the procedure with the bandwidth $\sigma(t)$ written generically: setting $\sigma(t) \equiv \sigma$ recovers the fixed-bandwidth Gaussian drift, while the exponential schedule $\sigma(t) = \max(\sigma_0 e^{-rt}, \sigma_{\min})$ implements the annealed variant of §6.2.

G.4 Sinkhorn-Drift operator implementations

We now describe the practical implementation of the Sinkhorn-divergence drift introduced in §6.3. The key ingredients are: entropic regularization of the optimal transport problem, a numerically stable log-space Sinkhorn solver, and a diagonal-masking heuristic that is necessary in the regime $q \approx p$ encountered late in training.

Entropic optimal transport. Given two empirical distributions $\hat{q} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ and $\hat{p} = \frac{1}{M}\sum_{j=1}^{M} \delta_{y_j}$ on $\mathbb{R}^d$, with cost matrix $C_{ij} = \|x_i - y_j\|^r$ ($r \in \{1, 2\}$), the entropic optimal transport problem reads

$$\mathrm{OT}_\varepsilon(\hat{q}, \hat{p}) \;=\; \min_{\pi \in \Pi(\hat{q}, \hat{p})} \sum_{i,j} \pi_{ij} C_{ij} \;+\; \varepsilon \sum_{i,j} \pi_{ij} \big(\log \pi_{ij} - 1\big), \tag{41}$$

where $\Pi(\hat{q}, \hat{p})$ is the set of couplings with marginals $\frac{1}{N}\mathbf{1}_N$ and $\frac{1}{M}\mathbf{1}_M$. The unique minimizer admits the Gibbs form $\pi^\star_{ij} = \exp\big((f_i + g_j - C_{ij})/\varepsilon\big)$, where the dual potentials (f, g) are obtained by alternately enforcing the two marginal constraints, the Sinkhorn iterations (Peyré & Cuturi, 2020). To stabilize entropic regularization across data scales we set $\varepsilon = \alpha\, \mathbb{E}\|x - y\|^r$ with α a dimensionless temperature; this keeps the kernel exponent $-C_{ij}/\varepsilon$ at unit mean across batches and across cost powers.

Log-space Sinkhorn iterations. At small ε, the kernel $K_{ij} = \exp(-C_{ij}/\varepsilon)$ underflows to zero, and the standard matrix-scaling form of Sinkhorn becomes numerically unusable. We instead iterate on the dual

Algorithm 1 Gaussian-kernel drift step (with optional bandwidth schedule)

Require: generated batch $X = \{x_i\}_{i=1}^{N}$, data batch $Y = \{y_j\}_{j=1}^{N_{\text{pos}}}$, negatives $Z = \{z_k\}_{k=1}^{N_{\text{neg}}}$ (typically $Z = X$), training step t, bandwidth schedule $\sigma(\cdot)$ (constant σ or $\sigma(t) = \max(\sigma_0 e^{-rt}, \sigma_{\min})$)

// 1. Bandwidth selection (fixed or annealed)
$\sigma \leftarrow \sigma(t)$
// 2. Pairwise distances and self-masking
$d_{ij}^{\text{pos}} \leftarrow \|x_i - y_j\|, \quad d_{ik}^{\text{neg}} \leftarrow \|x_i - z_k\|$
$d_{ii}^{\text{neg}} \leftarrow +\infty$ (mask self-interactions when $Z = X$)
// 3. Quadratic Gaussian logits
$\ell_{ij}^{\text{pos}} \leftarrow -(d_{ij}^{\text{pos}})^2/(2\sigma^2), \quad \ell_{ik}^{\text{neg}} \leftarrow -(d_{ik}^{\text{neg}})^2/(2\sigma^2)$
// 4. Doubly-stochastic softmax (preserves antisymmetry)
$\ell \leftarrow [\ell^{\text{pos}} \,\|\, \ell^{\text{neg}}]$ (concatenate along the column axis)
$A^{\text{row}} \leftarrow \text{softmax}_{\text{row}}(\ell), \quad A^{\text{col}} \leftarrow \text{softmax}_{\text{col}}(\ell)$
$A \leftarrow \sqrt{A^{\text{row}} \odot A^{\text{col}}}$ (entrywise; doubly-stochastic kernel)
$A^{\text{pos}}, A^{\text{neg}} \leftarrow$ split A along columns into $[N_{\text{pos}}, N_{\text{neg}}]$
// 5. Drift via attractive / repulsive coefficients
$W_{ij}^{\text{pos}} \leftarrow A_{ij}^{\text{pos}} \cdot \sum_k A_{ik}^{\text{neg}}$
$W_{ik}^{\text{neg}} \leftarrow A_{ik}^{\text{neg}} \cdot \sum_j A_{ij}^{\text{pos}}$
$V^+(x_i) \leftarrow \sum_j W_{ij}^{\text{pos}} y_j, \quad V^-(x_i) \leftarrow \sum_k W_{ik}^{\text{neg}} z_k$
$V(x_i) \leftarrow V^+(x_i) - V^-(x_i)$
// 6. Stop-gradient drift loss
$\mathcal{L} \leftarrow \frac{1}{N} \sum_i \left\| x_i - \text{sg}[x_i + V(x_i)] \right\|^2$
return $\mathcal{L}$

potentials (f, g) directly, using the log-sum-exp trick (Peyré & Cuturi, 2020; Feydy et al., 2018). With uniform marginals $\log a_i = -\log N$ and $\log b_j = -\log M$, the updates read

$$g_j \leftarrow \log b_j - \text{logsumexp}_i\big(f_i - C_{ij}/\varepsilon\big), \tag{42}$$

$$f_i \leftarrow \log a_i - \text{logsumexp}_j\big(g_j - C_{ij}/\varepsilon\big). \tag{43}$$

We iterate equation 42–equation 43 for at most T steps, with adaptive stopping when $\|f^{(t+1)} - f^{(t)}\|_\infty < \text{tol} \cdot \varepsilon$. The coupling is reconstructed as $\log \pi_{ij} = f_i + g_j - C_{ij}/\varepsilon$ and exponentiated in a single pass.

Diagonal masking when $q \approx p$. The Sinkhorn drift is the difference of two barycentric maps, $V(x_i) \propto T_{q\to p}(x_i) - T_{q\to q}(x_i)$, and its validity rests on a non-trivial self-coupling π_{qq}. As training progresses and $q \to p$, however, the cross cost C_{qp} and the self cost C_{qq} become comparable in magnitude, but the latter has a free zero on its diagonal: at any finite ε the entropic minimizer puts increasing mass on π_{qq}^{ii}, and in the limit $\pi_{qq} \to \frac{1}{N}\text{diag}(\mathbf{1}_N)$, so $T_{q\to q}(x_i) \to x_i$ and the drift collapses to the *biased* form $V \approx T_{q\to p}(x) - x$, which is no longer a Wasserstein gradient of the Sinkhorn divergence and reintroduces the bias that the debiasing in equation 22 was designed to remove. We employ a simple and effective remedy: mask the diagonal of C_{qq} by setting $C_{qq}^{ii} = +\infty$ before the Sinkhorn solve, which forces $\pi_{qq}^{ii} = 0$ and turns $T_{q\to q}(x_i)$ into a leave-one-out entropic centroid of the remaining particles. The drift then retains genuine self-interaction signal at all ε. We note that this heuristic introduces a small shift in the fixed point of the flow (since π_{qp} is unmasked while π_{qq} is masked, V no longer vanishes *exactly* at $q = p$); empirically we find this trade-off to be strongly favourable, and study it further in §H.4.

Drift operator and loss. With the entropic couplings π_{qp}, π_{qq} in hand, the barycentric maps are computed by a single batched matrix product,

$$T_{q\to p}(x_i) \;=\; N \sum_j [\pi_{qp}]_{ij}\, y_j, \qquad T_{q\to q}(x_i) \;=\; N \sum_j [\pi_{qq}]_{ij}\, x_j, \tag{44}$$

Algorithm 2 Sinkhorn-divergence drift step

Require: generated batch $X = \{x_i\}_{i=1}^N$, data batch $Y = \{y_j\}_{j=1}^M$, temperature α, max iterations T, tolerance tol, cost power $r \in \{1, 2\}$, drift scale η

// 1. Cost matrices
$C_{qp} \leftarrow \|x_i - y_j\|^r, \quad C_{qq} \leftarrow \frac{1}{2}\big(\|x_i - x_j\|^r + \|x_j - x_i\|^r\big)$ (symmetrize for fp robustness)
$\varepsilon \leftarrow \alpha \cdot \mathrm{mean}(C_{qp})$
// 2. Diagonal masking on the self-cost
$C_{qq}^{ii} \leftarrow +\infty \quad$ for $i = 1, \ldots, N$
// 3. Log-space Sinkhorn for both couplings
$\pi_{qp} \leftarrow \mathrm{LogSinkhorn}(C_{qp}, \varepsilon, T, \mathrm{tol})$ (Eqs. equation 42–equation 43)
$\pi_{qq} \leftarrow \mathrm{LogSinkhorn}(C_{qq}, \varepsilon, T, \mathrm{tol})$
// 4. Drift via barycentric maps
$T_{q\to p}(x_i) \leftarrow N \sum_j [\pi_{qp}]_{ij}\, y_j$
$T_{q\to q}(x_i) \leftarrow N \sum_j [\pi_{qq}]_{ij}\, x_j$
$V_{\mathrm{SK}}(x_i) \leftarrow \eta\,\big(T_{q\to p}(x_i) - T_{q\to q}(x_i)\big)$
// 5. Stop-gradient drift loss
$\mathcal{L} \leftarrow \frac{1}{N} \sum_i \big\| x_i - \mathrm{sg}[x_i + V_{\mathrm{SK}}(x_i)] \big\|^2$
return $\mathcal{L}$

(the prefactor N corrects for the uniform source marginal $1/N$), and the Sinkhorn drift is

$$V_{\mathrm{SK}}(x_i) \;=\; \eta\,\big(T_{q\to p}(x_i) - T_{q\to q}(x_i)\big), \tag{45}$$

where η is a step-size hyperparameter (analogous to τ in the JKO derivation). The loss is the standard stop-gradient drift loss of Eq. equation 5,

$$\mathcal{L}_{\mathrm{SK}}(\theta) \;=\; \mathbb{E}_\epsilon\big[\, \big\| f_\theta(\epsilon) - \mathrm{sg}\big[f_\theta(\epsilon) + V_{\mathrm{SK}}(f_\theta(\epsilon))\big] \big\|^2 \,\big]. \tag{46}$$

The full procedure for a single training step is summarized in Algorithm 2.

All operations are batched over the leading dimension and run end-to-end on GPU; the per-step overhead relative to the kernel baseline is dominated by the T Sinkhorn iterations on two $N \times M$ and $N \times N$ cost matrices, as quantified in Table 3.

G.5 Qualitative samples

Figures 6, 7, 8, and 9 show 48 uncurated 256×256 samples from each of our four models, exponential baseline (Deng et al., 2026), Gaussian-drift, Gaussian-drift with annealing, and Sinkhorn-drift after 30,000 training steps. Each grid mixes samples from 10 ImageNet classes (macaw, hummingbird, flamingo, golden retriever, tiger, monarch butterfly, African elephant, sports car, cheeseburger, volcano).

G.6 Compute, memory, and implementation notes

All ablations run on 16 NVIDIA H100 GPUs. A full 30,000-step ablation takes roughly 10–11 wall-clock hours depending on the loss (the two-distribution Sinkhorn variant solves three Sinkhorn problems per step when CFG is active). Checkpoints are evaluated at all CFG scales; we report the best.

All drift variants share the same $\mathcal{O}(B^2 d)$ pairwise-distance cost; the Sinkhorn drift additionally runs S Sinkhorn iterations on the entropic couplings π_{qp} and π_{qq}. Table 3 reports per-step wall-clock on a single H100, averaged over 100 runs after 10 warm-up iterations, for representative (B, d) shapes.

Analysis. Three observations stand out. (i) The three kernel-based variants (softmax, Gaussian, annealed Gaussian) are within 2–4% of one another at every shape: the bandwidth schedule and kernel choice contribute negligible overhead on top of the shared $B^2 d$ pairwise-distance computation. (ii) The Sinkhorn drift carries a fixed iteration overhead from the S Sinkhorn sweeps on the two coupling matrices, which dominates at small B:

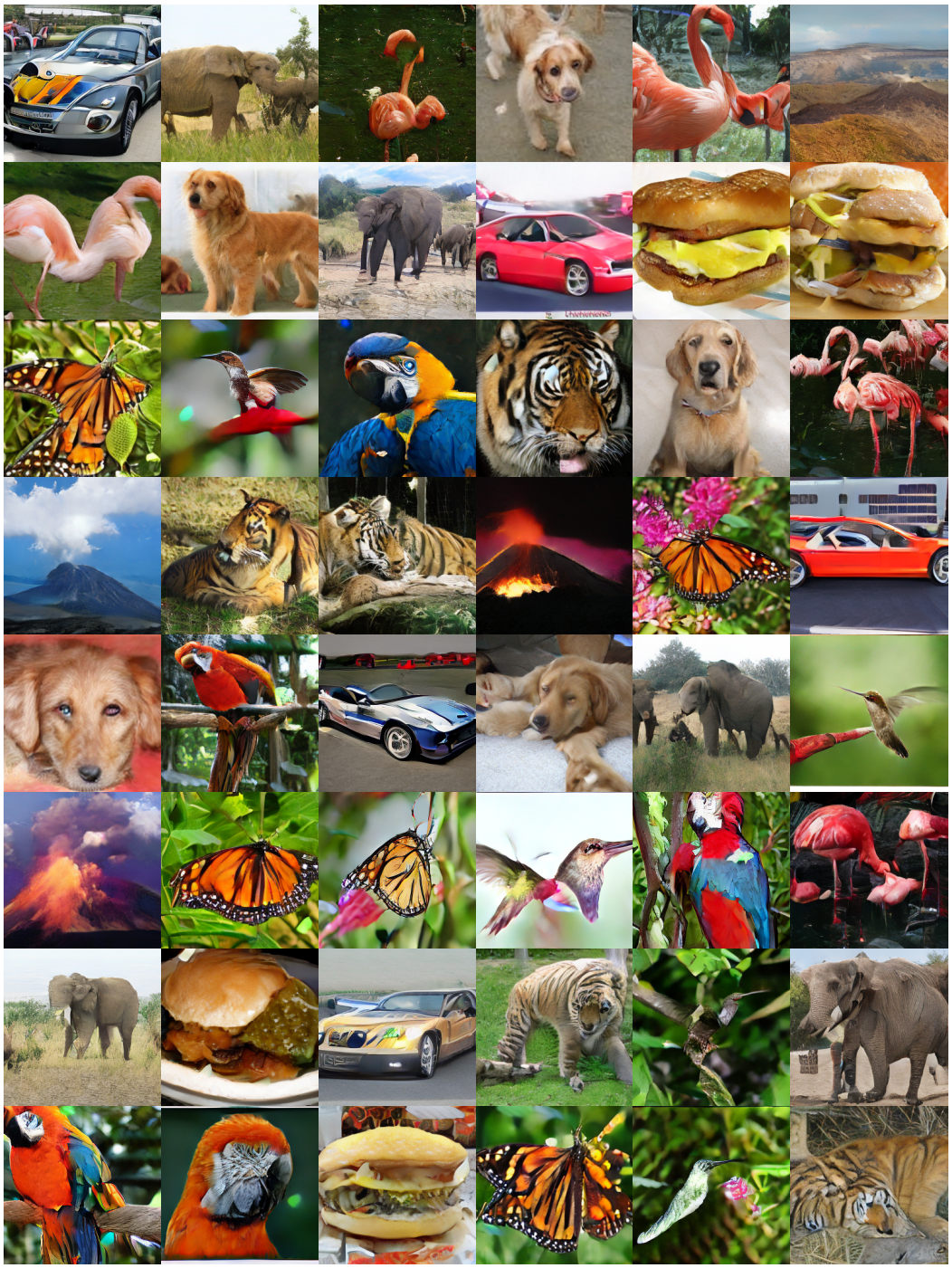

Figure 6: Uncurated samples from the exponential baseline (Deng et al., 2026) at 30,000 steps, CFG w=2.2. Each cell is an independently generated 256×256 image; samples from 10 ImageNet classes (macaw, hummingbird, flamingo, golden retriever, tiger, monarch butterfly, African elephant, sports car, cheeseburger, volcano) are shuffled across the grid.

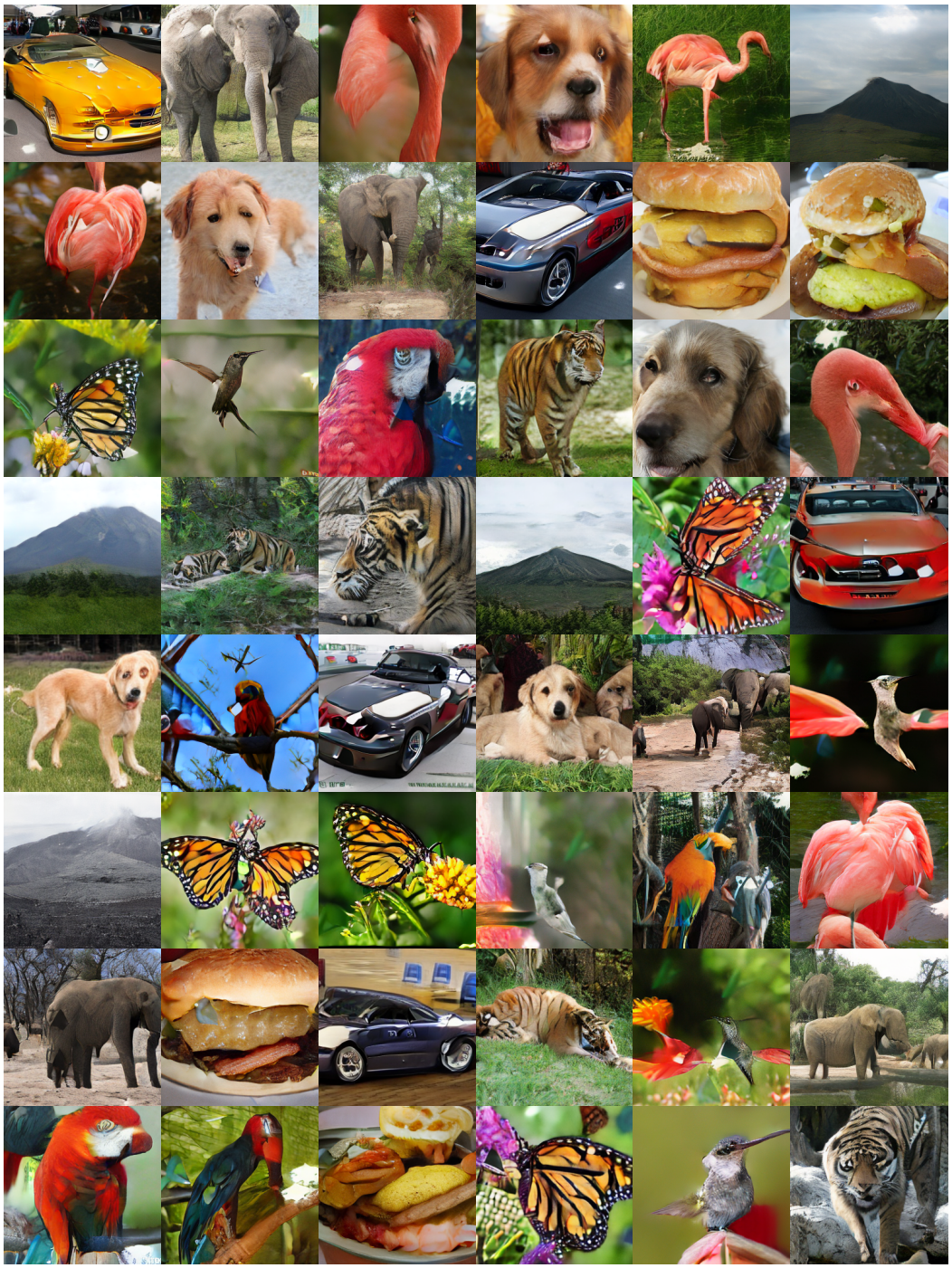

Figure 7: Uncurated samples from the Gaussian-drift model at 30,000 steps, CFG w=1.833. Each cell is an independently generated 256×256 image; samples from 10 ImageNet classes (macaw, hummingbird, flamingo, golden retriever, tiger, monarch butterfly, African elephant, sports car, cheeseburger, volcano) are shuffled across the grid.

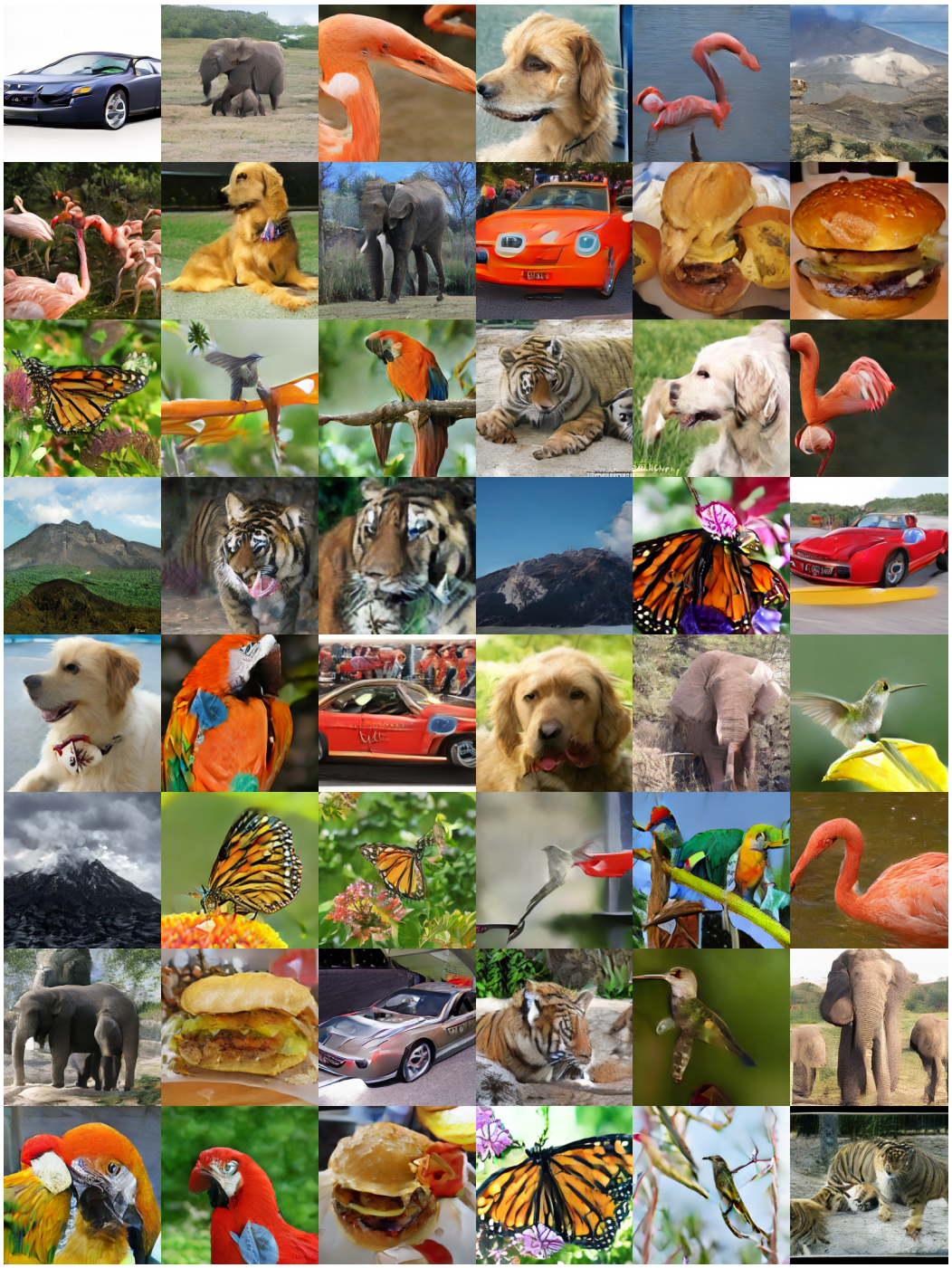

Figure 8: Uncurated samples from the Gaussian-drift model with annealing at 30,000 steps, CFG w=1.944. Each cell is an independently generated 256×256 image; samples from 10 ImageNet classes (macaw, hummingbird, flamingo, golden retriever, tiger, monarch butterfly, African elephant, sports car, cheeseburger, volcano) are shuffled across the grid.

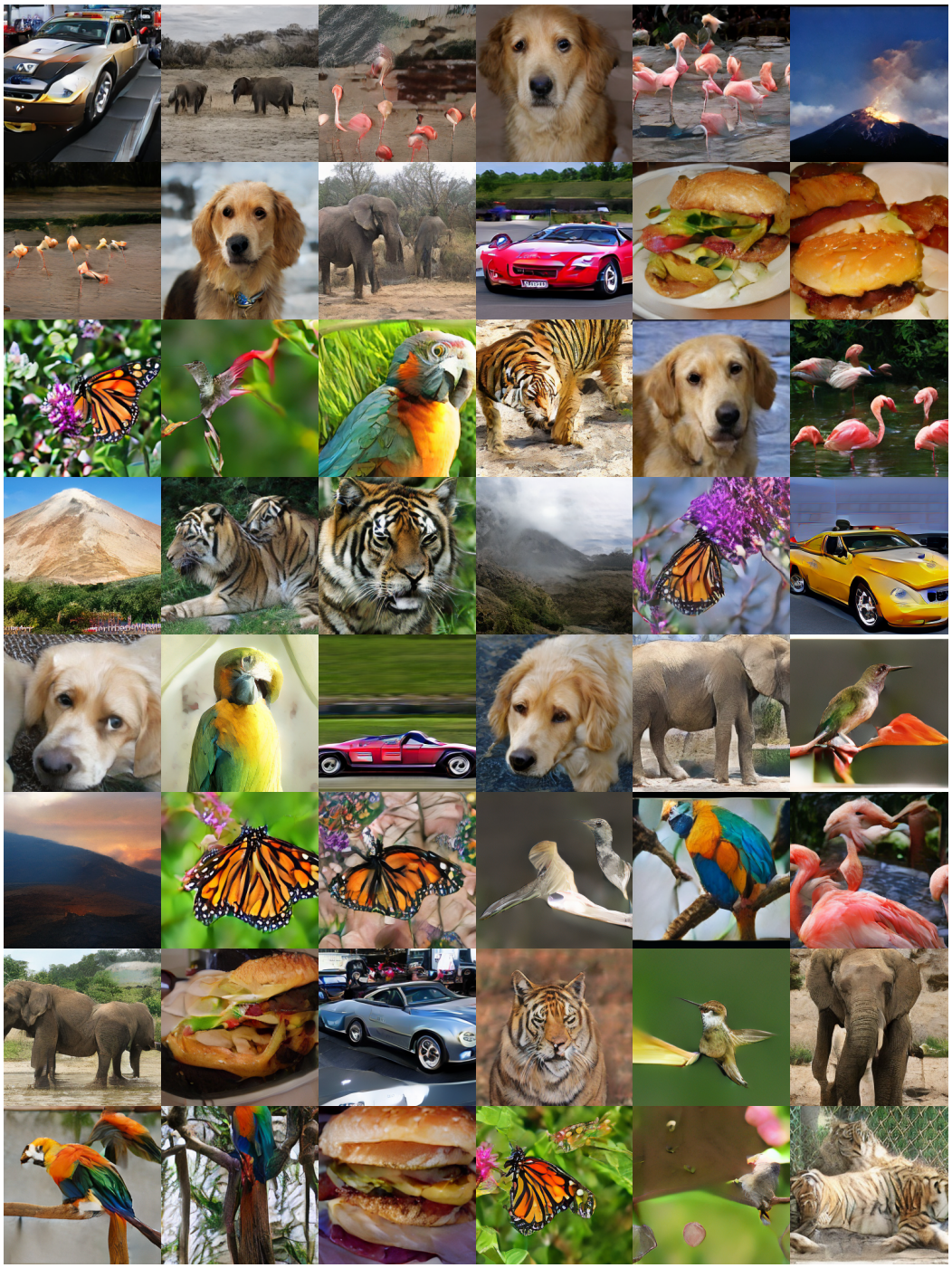

Figure 9: Uncurated samples from the Sinkhorn-drift model at 30,000 steps, CFG w=1.167. Each cell is an independently generated 256×256 image; samples from 10 ImageNet classes (macaw, hummingbird, flamingo, golden retriever, tiger, monarch butterfly, African elephant, sports car, cheeseburger, volcano) are shuffled across the grid.

Table 3: Per-step drift wall-clock on a single H100, in milliseconds (mean ± std over 100 iterations after 10 warm-ups). The "softmax" column is the Laplacian-kernel softmax baseline; "annealed (3R)" is the Gaussian drift with the $3R$ exponential bandwidth schedule of Section 6.2. **Bold** marks the fastest variant in each row.

Shape		**Wall-clock per step (ms)**			
B	d	softmax	Gaussian	annealed (3R)	Sinkhorn
64	256	$\mathbf{0.761 \pm 0.020}$	0.775 ± 0.012	0.780 ± 0.014	1.004 ± 0.105
256	256	1.214 ± 0.004	1.265 ± 0.005	1.261 ± 0.004	$\mathbf{1.042 \pm 0.011}$
1024	256	4.211 ± 0.005	4.343 ± 0.005	4.341 ± 0.005	$\mathbf{2.552 \pm 0.011}$
64	768	$\mathbf{0.899 \pm 0.005}$	0.930 ± 0.005	0.935 ± 0.006	1.040 ± 0.088

at (B, d)=$(64, 256)$ it is $\sim$32% slower than the softmax baseline. (iii) This overhead is rapidly amortized as B grows. At $(256, 256)$ Sinkhorn is already on par with the kernel variants (1.04 vs. 1.21 ms), and at $(1024, 256)$ it is in fact $\sim 1.65\times$ *faster* (2.55 vs. 4.21 ms): the B^2 pairwise-distance and softmax-normalization passes that the kernel drifts execute eagerly are absorbed into the fused Sinkhorn iterations, whose constant factor is smaller. The picture is qualitatively unchanged at d=768: Sinkhorn pays a fixed-cost premium at B=64, and the gap is dominated by the iteration count rather than the embedding dimension. In summary, Sinkhorn is not a meaningful bottleneck at the batch sizes used in our image-generation experiments ($B \geq 256$), and is strictly cheaper than the kernel baselines once B is large enough for the $B^2 d$ cost to dominate.