PT-FLOW: ONE-STEP SCHRÖDINGER-BRIDGE GENERATION BY PROXIMAL TILTING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generative models usually compress a many-step integration of a velocity field into a single step, using distillation, consistency conditions, trajectory straightening, or averaged fields. This compression tends to give up the optimality of the learned map or an explicit transport objective. PT-Flow takes a different route. We choose the governing law of the field (a viscous potential flow, i.e. the Schrödinger bridge) so that its time evolution has an exact closed-form solution through the Cole–Hopf transform. The single step is then not a shortened trajectory but the closed-form solution of the flow. The one hard part is a single Gaussian expectation whose naive estimator has variance that grows sharply as the flow becomes sharp. We remove this obstruction with a prox-tilted importance estimator: we center the proposal at the proximal point and scale it by the local curvature, which cancels the leading variation and makes the remaining variance shrink as the viscosity decreases. The estimator therefore becomes more accurate exactly where generation is sharp, and the same network that builds the proposal is the one-step generator itself, so sampling is a single network evaluation with no teacher, no adversarial training, and no ODE solver at inference. The learned scalar potential also defines a normalized Schrödinger-bridge marginal. On ImageNet 256×256, PT-Flow reaches an FID of 1.54 with one function evaluation. It combines one-step sampling with teacher-free, adversarial-free training and optimal-transport consistency at convergence.

1 INTRODUCTION

Diffusion and flow-based generative models generate high-quality samples by transporting noise to the data distribution (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023). Sampling, however, typically requires integrating a time-dependent vector field over many steps. Recent work reduces this cost through distillation, consistency training, trajectory straightening, and averaged velocity fields (Song et al., 2023; Liu et al., 2023; Frans et al., 2025; Geng et al., 2025a). These methods have made one-step generation practical, but generally learn a shortcut to a multi-step process. They focus on fast sampling, while an explicit transport objective usually remains outside their design.

This paper explores a different route: can we choose transport dynamics whose full evolution has a closed-form solution? We consider pressureless viscous potential flow, which is equivalent to a Schrödinger bridge with a Brownian reference process (Schrödinger, 1931; Léonard, 2014). Its velocity potential satisfies the viscous Hamilton–Jacobi equation, which becomes the linear heat equation under the Cole–Hopf transform (Hopf, 1950; Cole, 1951). The evolution over the entire time interval can therefore be expressed as a Gaussian expectation, without integrating an ODE or SDE. In the small-viscosity regime, the resulting deterministic endpoint predictor approaches the proximal operator of the learned potential (Moreau, 1965), directly connecting Schrödinger-bridge transport, optimal transport, and one-step generation (Benamou & Brenier, 2000; Peyré & Cuturi, 2019).

The main obstacle is computational. Naive Monte Carlo estimation has relative variance that grows exponentially with the transport magnitude divided by viscosity. This is most severe for high-dimensional generation with small viscosity, and helps explain why existing simulation-free Schrödinger-bridge methods typically rely on restrictive parametric potentials (Korotin et al., 2024) or iterative sampling procedures (Mokrov et al., 2024).

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

We address this problem with **PT-Flow**, a one-step model based on proximal tilting. We construct an importance proposal centered at the potential's proximal point and scaled by its local curvature. This matches the dominant local geometry of the target integral and cancels its leading variation. Under the conditions in Theorem 3.9, the remaining log-weight variance is $O(\varepsilon)$: unlike naive Monte Carlo, it shrinks with the viscosity ε.

The proximal proposal also provides the generator. We train a network to predict the proximal point from its first-order optimality residual and use the same prediction in the importance estimator. At inference, this network maps noise to a sample in one evaluation. The learned scalar potential also defines a normalized Schrödinger-bridge marginal. PT-Flow thus combines one-step sampling and optimal-transport consistency without a pretrained teacher or adversarial optimization.

Our contributions are:

- We derive a one-step formulation from the closed-form evolution of a Schrödinger bridge and show that its endpoint predictor approaches a proximal map at small viscosity.
- We identify the exponential variance of naive Monte Carlo and introduce a proximally tilted estimator with $O(\varepsilon)$ log-weight variance under explicit conditions.
- We develop a teacher-free framework in which the tilting network is also the one-step generator, while the potential approaches optimal transport as viscosity vanishes

2 PRELIMINARIES

Notation. We first introduce the setting of PT-Flow. We work on $\mathbb{R}^d$. The noise distribution is the standard Gaussian $\rho_0 = \mathcal{N}(0, I_d)$, and the data distribution is $\rho_1 = \rho_{\mathrm{data}}$. Two time variables run in opposite directions. Generative time t goes from $t = 0$ at the noise side to $t = 1$ at the data side. Smoothing time $s = 1 - t$ runs the other way, from $s = 0$ at the data side to $s = 1$ at the noise side. Two further symbols recur throughout. We write $G_{\sigma^2}(z) = (2\pi\sigma^2)^{-d/2} e^{-\|z\|^2/2\sigma^2}$ for the Gaussian kernel, and $\varepsilon > 0$ for the viscosity, which equals half the variance rate of the Brownian reference process. The network learns a single scalar function, the log-potential $u_\theta : \mathbb{R}^d \to \mathbb{R}$; everything else is built from it. The potential in physical units is $\varphi_\theta = 2\varepsilon\, u_\theta$, and the data-side Schrödinger potential is $\psi_1^\theta = e^{-u_\theta} = e^{-\varphi_\theta/2\varepsilon}$. We mark how these objects evolve in smoothing time with a superscript (s). Two values matter most: $\psi_0 := \psi^{(1)}$ and $\varphi_0 := -2\varepsilon \log \psi_0$, both living on the noise side. The complete notation table and the standing assumptions (A0)-(A6) are collected in Appendix B. Two are worth keeping in mind from the outset: coercivity, $\varphi_\theta(y) + \frac{1-\delta}{2}\|y\|^2 \to +\infty$ (A1), and weak convexity, $\nabla^2 \varphi_\theta \succeq -\lambda I$ with $\lambda < 1$ (A2).

Entropic optimal transport and the Schrödinger bridge. The Monge problem seeks a map T that minimizes $\mathbb{E}_{x\sim\rho_0}\|x - T(x)\|^2$ subject to $T_\# \rho_0 = \rho_1$. Its Kantorovich relaxation, taken over couplings $\pi \in \Pi(\rho_0, \rho_1)$, attains the value $\frac{1}{2}W_2^2(\rho_0, \rho_1)$, and Brenier's theorem shows the optimal coupling is a deterministic gradient map (Villani, 2009). The dynamic form of the same problem (Benamou & Brenier, 2000) has an appealing reading: optimal mass travels along straight lines at constant speed, driven by a velocity $v = -\nabla\varphi$ whose potential φ solves a Hamilton–Jacobi equation. Now add an entropic penalty of strength 2ε. This smooths the deterministic map into a stochastic kernel whose conditional width is $O(\sqrt{\varepsilon})$, and as $\varepsilon \to 0$ it Γ-converges back to optimal transport (Peyré & Cuturi, 2019; Cuturi, 2013). The same object has a second description as the Schrödinger bridge (Schrödinger, 1931; Léonard, 2014). Among all Brownian path laws with variance rate 2ε pinned to the marginals ρ_0 and ρ_1, it is the one closest in KL divergence to free Brownian motion. Its static coupling factorizes as

$$\pi^*(x_0, x_1) \;=\; \hat{\psi}_0(x_0)\, G_{2\varepsilon}(x_1 - x_0)\, \psi_1(x_1), \tag{1}$$

where the two potentials propagate by the heat semigroup, $\psi_0 = G_{2\varepsilon} * \psi_1$ and $\hat{\psi}_1 = G_{2\varepsilon} * \hat{\psi}_0$, subject to the marginal constraints $\hat{\psi}_0\psi_0 = \rho_0$ and $\hat{\psi}_1\psi_1 = \rho_1$. Under finite second moments and mild integrability (A0), the solution exists and is unique up to a scalar gauge (Léonard, 2014). The fact we build on is this: the entire bridge, its coupling, marginals, and drift is determined by the single scalar function ψ_1. Every other object follows from ψ_1 by heat-semigroup propagation and division by known Gaussians. Learning ψ_1 therefore means learning the whole transport.

3 METHOD

We develop PT-Flow in six steps. Each step produces exactly the object the next one needs. §3.1 chooses a governing law whose time evolution has an exact solution operator, so that "one step" is analytic rather than a compressed integral. This yields the smoothed potential φ_0. §3.2 turns φ_0 into a deterministic one-step sampler and identifies that sampler with a proximal operator. §3.3 shows that a single scalar loss trains the potential, and that this loss comes with normalization, convexity, and an exact semi-dual reading. §3.4 isolates the one place the program breaks. The culprit is an estimator whose sample cost is exponential in the transport-to-viscosity ratio, which we state as a variance barrier. §3.5 removes this obstruction with a proximally-tilted estimator whose variance improves as ε cools. §3.6 then observes that the tilting network is the generator, which collapses training and sampling into a single architecture. §3.7 assembles the full algorithm and guidance. All proofs are deferred to the appendix.

3.1 A GOVERNING LAW WITH A CLOSED-FORM SOLUTION OPERATOR

Our goal is a map $T : \mathbb{R}^d \to \mathbb{R}^d$ with $T_{\#}\rho_0 \approx \rho_{\text{data}}$ whose evaluation costs a single network pass. The dominant strategy learns a time-dependent velocity field and then compresses its flow by distillation, consistency conditions, trajectory straightening, or averaged fields. We take the opposite route: we choose the field's governing law so that its time evolution admits an exact solution operator. One step is then not a compressed journey but the journey's analytic solution.

Concretely, we impose pressureless viscous potential flow. This law fixes three things at once. The velocity is a gradient by construction, $v = -\nabla\varphi$, so the field is curl-free architecturally rather than as a training outcome. The mass evolves by the Fokker–Planck equation $\partial_t\rho + \nabla\cdot(\rho v) = \varepsilon\Delta\rho$. And the potential obeys the viscous Hamilton–Jacobi equation in smoothing time,

$$\partial_s\varphi + \tfrac{1}{2}\|\nabla\varphi\|^2 = \varepsilon\,\Delta\varphi, \qquad \varphi(\cdot, s{=}0) = \varphi_\theta, \tag{2}$$

with boundary conditions $\rho(\cdot, 0) = \rho_0$ and $\rho(\cdot, 1) = \rho_{\text{data}}$ derived in Appendix C. This particular law earns its place twice over. At $\varepsilon = 0$ it is exactly the Benamou–Brenier optimality system: mass moves along W_2 geodesics-straight lines at constant speed. The straightness that rectified flow buys through repeated retraining (Liu et al., 2023) is here simply a property of the law. At $\varepsilon > 0$ it becomes the Schrödinger bridge with Brownian reference of variance rate 2ε, which supplies the boundary machinery of §3.3. This viscosity that makes it well-posed is precisely what makes it solvable.

Lemma 3.1 (Cole–Hopf linearization)**.** *Let $\psi = e^{-\varphi/2\varepsilon}$ with φ twice differentiable and $\psi > 0$. Then φ solves equation 2 if and only if ψ solves the linear heat equation $\partial_s\psi = \varepsilon\Delta\psi$.*

(Proof in Appendix D.1, following Hopf, 1950; Cole, 1951.)

The exponential change of variables absorbs the nonlinearity exactly, not just approximately. Linear diffusion then integrates in closed form:

Corollary 3.2 (Closed-form propagation)**.** *The heat semigroup with generator $\varepsilon\Delta$ has Gaussian kernel $G_{2\varepsilon s}$, so $\psi^{(s)} = G_{2\varepsilon s} * \psi_1^\theta$. At $s = 1$,*

$$\psi_0(x) \;=\; (G_{2\varepsilon} * \psi_1^\theta)(x) \;=\; \mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[e^{-u_\theta(x+\sqrt{2\varepsilon}\,z)}\big]. \tag{3}$$

Equation 3 is the structural source of one-step generation: a single Gaussian expectation propagates the potential across the entire time interval. There is no ODE solver, no discretization, and no time input to the network. There was never an integral to compress in the first place. For small ε, equation 3 takes the form of a Laplace integral. Defining $F_x(y) := \varphi_\theta(y) + \frac{1}{2}\|x - y\|^2$, one obtains (Appendix D.2, under (A1)-(A2))

$$\varphi_0(x) \;=\; \underbrace{\min_y F_x(y)}_{\text{Moreau envelope } M_{\varphi_\theta}(x)} + \varepsilon \log\det\nabla^2 F_x(y^*) \;+\; O(\varepsilon^2 \text{ per dim}). \tag{4}$$

The correction is the term $\varepsilon\log\det\nabla^2 F_x(y^*) = \varepsilon\sum_{i=1}^d \log\lambda_i$, a sum of d order-one contributions and hence $O(\varepsilon d)$. The Hopf–Lax regime therefore requires $\varepsilon d \to 0$ rather than merely $\varepsilon \to 0$.

We flag this now because the same window $\varepsilon \lesssim 1/d$ reappears in §3.5 from a completely independent variance argument, which serves as a coherence check on the theory. The minimizer in equation 4 is by definition the proximal operator, $y^*(x) = \mathrm{prox}_{\varphi_\theta}(x)$. The Moreau identity then gives $\nabla M_{\varphi_\theta}(x) = x - \mathrm{prox}_{\varphi_\theta}(x)$. Both objects are about to become the sampler.

3.2 From the smoothed potential to a one-step sampler

Corollary 3.2 delivers ψ_0. A one-step generator must turn the stochastic bridge kernel $p(x_1 \mid x_0)$ into a single deterministic output. The natural choice is its mean, $\mathbb{E}[X_1 \mid X_0 = x_0]$, which is the minimum-mean-squared-error predictor of the data endpoint. A Tweedie-type identity makes this mean computable in closed form from ψ_0, with no extra network and no integration over x_1.

Theorem 3.3 (Tweedie identity; uses (A1)). $\mathbb{E}[X_1 \mid X_0 = x_0] = x_0 + 2\varepsilon \nabla \log \psi_0(x_0)$. *Equivalently, in potential form,*

$$T_\varepsilon(x_0) \;=\; x_0 - \nabla \varphi_0(x_0). \tag{5}$$

(Proof: Appendix E.1.)

Combining equation 5 with the Moreau identity below equation 4 yields the structural fact the rest of the paper exploits:

$$T(x_0) \;\approx\; x_0 - \big(x_0 - \mathrm{prox}_{\varphi_\theta}(x_0)\big) \;=\; \mathrm{prox}_{\varphi_\theta}(x_0), \tag{6}$$

exact up to the $O(\varepsilon)$ gradient of the Laplace correction in equation 4. In words, the one-step generator approximates the proximal operator of the learned potential, exact only as $\varepsilon \to 0$. As a sanity check, take $\rho_1 = \delta_a$: then $\varphi_0 = \frac{1}{2}\|x - a\|^2 + \mathrm{const}$, and $T(x_0) = a$ for every x_0.

One honesty clause must be stated before we exploit equation 6. The deterministic map T replaces the stochastic bridge kernel $p(x_1|x_0)$ by its mean. As a result, two distributions that we must not conflate come apart: the pushforward $T_\# \rho_0$, and the model's bridge marginal $\hat{\rho}_1 := \hat{\psi}_1 \psi_1^\theta$. The latter is the distribution the learning objective fits in §3.3. The gap between them is quantitative and small in the operating regime.

Proposition 3.4 (Sampler–marginal gap; uses (A2)). $W_2\big(T_\# \rho_0, \hat{\rho}_1\big) \;\le\; \big(\mathbb{E}_{x_0}[\mathrm{tr}\,\mathrm{Cov}(X_1|X_0 = x_0)]\big)^{1/2} \;\approx\; \sqrt{2\varepsilon d/(1+\bar{\lambda})}$, *where $\bar{\lambda}$ is the mean curvature of φ_θ at the prox points. (Proof: Appendix E.2.)*

At the operating point this is a per-coordinate deficit $\sqrt{2\varepsilon/(1+\bar{\lambda})} \approx 0.14$, small relative to the unit data scale, and auditable rather than rhetorical.

3.3 Learning the data-side potential

Sections 3.1–3.2 solved the dynamics for a given φ_θ. It remains to choose φ_θ so that the transported noise lands on the data. Here the Schrödinger structure of §2 pays off: we parameterize only the data-side potential, $\psi_1^\theta = e^{-u_\theta}$, whose positivity is automatic. Every other bridge object is then closed-form. The noise-side potential ψ_0 is the Gaussian expectation equation 3. The conjugate $\hat{\psi}_0 = \rho_0/\psi_0$ is exact, since the noise marginal ρ_0 is an analytic Gaussian; this is why the noise-side constraint holds by construction. The remaining conjugate $\hat{\psi}_1 = G_{2\varepsilon} * \hat{\psi}_0$ is a second Gaussian expectation. Only the data-side constraint $\hat{\psi}_1 \psi_1 = \rho_1$ is left unenforced, and it becomes the entire training signal.

First, the model density is normalized for free.

Theorem 3.5 (Automatic normalization; uses (A1)). *For every θ, $\int \hat{\psi}_1 \psi_1^\theta \, \mathrm{d}x = 1$. Consequently the model marginal $\hat{\rho}_1 = \hat{\psi}_1 \psi_1^\theta$ is exactly normalized. (Proof: one line of heat-semigroup self-adjointness; Appendix F.1.)*

Unlike an energy-based model, the global normalization holds identically in θ by an algebraic property of the heat kernel. The per-input potential $\psi_0(x)$ still requires the Gaussian integral equation 3, which is the cost addressed in §3.4

Second, the remaining boundary condition is the stationarity condition of a plain scalar loss:

$$\boxed{\mathcal{L}(\theta) \;=\; \mathbb{E}_{x_0\sim\rho_0}\big[\log\psi_0(x_0)\big] \;+\; \mathbb{E}_{x_1\sim\rho_1}\big[u_\theta(x_1)\big]} \tag{7}$$

This is one expectation over noise and one over data, at a single Monte-Carlo level. The second term needs no estimator at all.

Theorem 3.6 (The objective solves the bridge; uses (A0)–(A1))**.** *(i)* ***Stationarity:*** *the first variation of equation 7 in ψ_1 vanishes precisely when $\hat{\psi}_1\psi_1 = \rho_1$. Minimizing equation 7 therefore solves the Schrödinger system rather than a surrogate for it. (ii)* ***Convexity:*** *$\mathcal{L}$ is convex in the log-potential $\log\psi_1$, with the harmless scalar gauge of (A0) as its only flat direction. (iii)* ***Semi-dual identity:*** *minimizing equation 7 is exactly maximizing the entropic-OT Kantorovich semi-dual with cost $\frac{1}{2}\|x-y\|^2$ and regularization 2ε, the noise-side (c,ε)-transform having been performed analytically. (Proofs: Appendix F.)*

Part (iii) locates PT-Flow precisely in the literature. It is the same objective family that LightSB (Korotin et al., 2024) optimizes over Gaussian-mixture potentials, now carried by a free-form network. It also reveals the alternating scheme of §3.7 as a function-space half-Sinkhorn: the noise side is transformed exactly, and the data side is fit by gradient steps. At the optimum both boundary conditions hold, so by uniqueness (A0) the learned coupling is the entropic-OT plan. As $\varepsilon \to 0$ it converges to the Brenier map, so optimal, straight transport holds at convergence.

3.4 The obstruction: a variance barrier

Everything so far is exact except the Laplace expansion of equation 4, and runs through the single Gaussian expectation equation 3. The entire difficulty of the program comes down to how that expectation is evaluated. We argue this is also why the simulation-free Schrödinger-bridge line has been stuck between parametric potentials (Korotin et al., 2024) and MCMC (Mokrov et al., 2024). The exact identity equation 3 writes $\psi_0(x)$ as an expectation over $z\sim\mathcal{N}(0,I)$. Its obvious unbiased estimator draws K i.i.d. samples and averages the integrand,

$$\hat{\psi}_0^{\text{naive}}(x) = \tfrac{1}{K}\sum_{k=1}^{K}\exp\big(-u_\theta(x+\sqrt{2\varepsilon}\,z_k)\big), \qquad z_k\sim\mathcal{N}(0,I), \tag{8}$$

which is consistent and unbiased for every ε and every θ. The draws z_k place proposal points at $x+\sqrt{2\varepsilon}\,z_k$, i.e. they sample the ball $\mathcal{N}(x,2\varepsilon I)$; the entire obstruction below is that this ball is in the wrong place.

Proposition 3.7 (Exponential sample complexity; Laplace-asymptotic, uses (A3))**.** *Let $s^2 := \mathrm{Var}_z[u_\theta(x+\sqrt{2\varepsilon}z)]$. The relative variance of equation 8 is $(e^{s^2}-1)/K$ under a log-normal approximation, and to first order $s^2 \approx \|\nabla\varphi_\theta(x)\|^2/2\varepsilon$, so $K_{\text{naive}} \gtrsim \exp\big(\|\nabla\varphi_\theta\|^2/2\varepsilon\big)$. (Derivation: Appendix G.1.)*

Proposition 3.8 (Variance deficit of the mean map; Laplace-asymptotic, uses (A2), (A5))**.** *The bridge kernel's conditional covariance is $\mathrm{Cov}(X_1|x_0) \approx 2\varepsilon(I+\nabla^2\varphi_\theta(y^*))^{-1}$, so the deterministic map under-disperses relative to $\hat{\rho}_1$ by a relative deficit $\approx 2\varepsilon/(1+\lambda)$: usable generation demands $\varepsilon \lesssim 10^{-2}$. (Appendix G.2.)*

Together these two propositions close like a vice. Transport requires a displacement $\|T(x_0)-x_0\| = O(\sqrt{d})$, so $\|\nabla\varphi_\theta\|^2 = O(d)$, and therefore $K_{\text{naive}} \gtrsim e^{cd/2\varepsilon}$. Proposition 3.7 then demands ε large, while Proposition 3.8 demands ε small. Beyond toy dimension the two ranges do not overlap, so no practical operating point exists for the naive estimator. The argument is heuristic (log-normal and local Taylor), not a formal impossibility proof. The two demands are also coupled: every unit of blur removed by sharpening φ_θ is paid back in estimator variance. The Cole–Hopf change of variables has not made the problem easy. It has only relocated the hardness, out of the time axis and into one Gaussian convolution.

The escape hatch is geometric. In y-space, $\psi_0(x) \propto \int e^{-F_x(y)/2\varepsilon}\,\mathrm{d}y$ is a partition function at inverse temperature $1/2\varepsilon$. Its mass concentrates at $y^*(x) = \mathrm{prox}_{\varphi_\theta}(x)$, which sits a full transport displacement, $O(\sqrt{d})$, away from x. The naive proposal $\mathcal{N}(x,2\varepsilon I)$ is a ball of radius $O(\sqrt{2\varepsilon d})$ centred at x. The mismatch is $O(\sqrt{d/2\varepsilon})$ proposal standard deviations, so the estimator essentially

never sees the mode. The key point is that the barrier is a property of the proposal, not of the integral. Nothing in Schrödinger-bridge theory forces the proposal $\mathcal{N}(x, 2\varepsilon I)$. That single observation is the whole of the next subsection.

3.5 THE REPAIR: PROX-TILTED ESTIMATION AND THE VARIANCE REVERSAL

Importance sampling is exactly unbiased for any proposal q_x covering the integrand,

$$\psi_0(x) = \mathbb{E}_{y\sim q_x}[G_{2\varepsilon}(y-x)e^{-u_\theta(y)}/q_x(y)] \tag{9}$$

We place the proposal where the mass concentrates. With $F_x(y) = \varphi_\theta(y) + \frac{1}{2}\|y-x\|^2$, $y^* = \mathrm{prox}_{\varphi_\theta}(x)$, $H = I + \nabla^2\varphi_\theta(y^*)$, and $D = \mathrm{diag}(H)$, the diagonal Laplace proposal is

$$q_x = \mathcal{N}\big(m(x), 2\varepsilon S(x)\big), \quad m(x) \approx y^*, \quad S(x)^{-1} \approx D. \tag{10}$$

Diagonal matching cannot cancel off-diagonal curvature; assumption (A6) (Appendix B) makes this precise, and under it the log-weight admits an exact cancellation identity whose remainder is $O(\varepsilon)$ (Appendix H.1).

Theorem 3.9 (Variance reversal; Laplace-asymptotic)**.** *If $\mathbb{E}_z[R_x(\sqrt{2\varepsilon}D^{-1/2}z)^2] \le 4\varepsilon^3 C_R$ with C_R independent of ε, then for the centred diagonal proposal*

$$s^2_{\mathrm{res}} := \mathrm{Var}_{q_x}(\log w) \le \varepsilon\big(\sqrt{C_{\mathrm{off}}/2} + \sqrt{C_R}\big)^2 = O(\varepsilon). \tag{11}$$

The familiar $O(dM_3^2\varepsilon)$ scaling follows when $C_{\mathrm{off}} + C_R = O(dM_3^2)$; it is not automatic from third-derivative boundedness (Appendix H.2).

The regularization ε moves from the denominator of the naive exponent to the numerator: colder bridges are more Laplace-like, so the matched proposal fits better. With $p_{\varepsilon,x} = G_{2\varepsilon}(\cdot - x)e^{-u_\theta}/\psi_0$, the K-sample relative variance is $\chi^2(p_{\varepsilon,x}\|q_x)/K$, and (A6) gives $\chi^2(p_{\varepsilon,x}\|q_x) \le C_{\mathrm{IS}}\varepsilon$, so the budget $K^\star_{\mathrm{tilted}}(\delta) \le \max\{1, \lceil C_{\mathrm{IS}}\varepsilon/\delta^2\rceil\} \to 1$ as $\varepsilon \to 0$ (Appendix H.2). Both ingredients are needed: recentring buys the ε-reversal, diagonal scaling the d-scaling; recentring alone leaves an ε-independent floor $\sim \frac{1}{2}d\bar{\lambda}^2$ (Appendix H.3). The constraint now depends on the controllable smoothness M_3, not the transport magnitude $\|\nabla\varphi\| = O(\sqrt{d})$. In practice a defensive mixture $q_x^{\mathrm{def}} = (1-\alpha)q_x + \alpha\mathcal{N}(x, 2\varepsilon I)$ bounds the weights; $\chi^2(p_{\varepsilon,x}\|q_x^{\mathrm{def}}) \le (\chi^2(p_{\varepsilon,x}\|q_x) + \alpha)/(1-\alpha)$, and $\alpha = O(\varepsilon)$ preserves the $O(\varepsilon)$ bound (Appendix H.4).

3.6 THE PROPOSAL IS THE GENERATOR

The proposal mean in equation 10 must be produced by a network, $m_\eta(x) \approx \mathrm{prox}_{\varphi_\theta}(x)$. This looks like an extra cost of the repair. It is not. That network is the generator itself. Three identities, each established above, chain together:

(a) Moreau equation 4:	$\nabla\varphi_0(x) = x - \mathrm{prox}_{\varphi_\theta}(x)$	(up to the $O(\varepsilon)$ correction gradient)
(b) Tweedie + prox equation 6:	$T(x_0) = \mathrm{prox}_{\varphi_\theta}(x_0)$	
(c) Laplace proposal equation 10:	$m(x) \approx \mathrm{prox}_{\varphi_\theta}(x)$	

$$\boxed{m_\eta(x_0) \;\approx\; \mathrm{prox}_{\varphi_\theta}(x_0) \;\approx\; T(x_0) \;=\; \text{the one-step generator.}} \tag{12}$$

The approximations are the two $\approx$ signs and the Laplace expansion of equation 4, and we keep them explicit. The deployed sampler is m_η, not the exact bridge mean T_ε; their gap decomposes as

$$\underbrace{m_\eta(x)}_{\text{deployed}} \xrightarrow{\delta_\eta} \underbrace{\mathrm{prox}_{\varphi_\theta}(x)}_{\text{prox}} \xrightarrow{\delta_\varepsilon} \underbrace{T_\varepsilon(x)}_{\text{exact bridge mean}}, \qquad \|m_\eta - T_\varepsilon\| \le \delta_\eta + \delta_\varepsilon,$$

with $\delta_\varepsilon = \|\mathrm{prox}_{\varphi_\theta} - T_\varepsilon\| = \varepsilon\|\nabla \log\det \nabla^2 F_x\|$ the finite-viscosity term, the gradient of the Laplace correction in equation 4 (not its $O(\varepsilon d)$ value), vanishing as $\varepsilon \to 0$ and $\delta_\eta = \|m_\eta - \mathrm{prox}_{\varphi_\theta}\|$ the learned-prox error.The bridge mean itself is exact (Theorem 3.3), so no error enters there. The

learned term δ_η is bounded by the measurable optimality residual $r_\eta(x) = \nabla\varphi_\theta(m_\eta) + m_\eta - x$ (Appendix E.1), and is empirically small: the deployed generator matches the reference prox closely across all ε (Table 7), so (A4) holds at the operating point. Guarantees for the implemented sampler therefore carry δ_η and δ_ε in addition to the kernel-spread term of Proposition 3.4, not that term alone.

The variance-reduction device and the generator are the same network. This one fact organizes the whole method, in three consequences. First, inference is strictly one network evaluation. Sample $x_0 \sim \mathcal{N}(0, I)$ and output $m_\eta(x_0)$. There are no K evaluations, no autograd at sampling time, and no distillation stage. Second, the K tilted proposals are needed only during training, and only to estimate the potential loss. Third, the generator trains by a plain regression on the prox first-order condition (§3.7). There is no adversary, no teacher, and no minimax. The free-form one-step optimal-transport generator that ICNN-based methods (Makkuva et al., 2020; Amos et al., 2017) could reach only through an unstable dual game arrives here as the byproduct of a variance fix. Each of the three lemmas is classical. The contribution is their composition, and its use as a training scheme.

3.7 The complete algorithm: training, sampling, guidance

Training. We use two networks, or one two-headed network: the potential $u_\theta(x, c)$ and the generator/proposal pair (m_η, s_η). Here s_η is a diagonal log-scale head that implements S in equation 10. The two networks alternate on two timescales.

The potential minimizes the flat objective equation 7, with ψ_0 evaluated by the tilted estimator at frozen η (stop-gradient $\bar{\eta}$), in the log domain:

$$\mathcal{L}_\theta = \mathbb{E}_{x_0\sim\rho_0}\big[\log\hat{\psi}_0^{\mathrm{tilt}}(x_0)\big] + \mathbb{E}_{(x_1,c)\sim\rho_1}\big[u_\theta(x_1, c)\big], \tag{13}$$

$$\log\hat{\psi}_0^{\mathrm{tilt}}(x_0) = \operatorname*{logsumexp}_{k\le K}\Big[\log G_{2\varepsilon}(y_k - x_0) - u_\theta(y_k, c) - \log q_{\bar{\eta}}^{\mathrm{def}}(y_k|x_0)\Big] - \log K, \tag{14}$$

where the y_k are drawn from the defensive proposal centred at $m_{\bar{\eta}}$, and the conditioning c is dropped to $\varnothing$ with probability p_{uncond}. One rule is load-bearing: the potential loss is always evaluated at guidance weight zero. The data-side objective is valid only against real data, so guidance (below) lives solely in the generator's conditioning and at sampling time. By Jensen, equation 14 is an IWAE-type lower bound (Burda et al., 2016). Its gap is $O(dM_3^2\varepsilon/K)$, priced by tilt quality. Although θ is incentivized to inflate this lower bound, the η-updates bound the gap independently of $\mathcal{L}_\theta$ (Appendix I).

The generator regresses the prox first-order condition $\nabla\varphi_\theta(y^*) + y^* - x_0 = 0$ directly, at frozen θ:

$$\mathcal{L}_\eta = \mathbb{E}_{x_0,c,w}\Big\|\nabla\varphi_{\bar{\theta}}^w\big(m_\eta(x_0, c, w)\big) + m_\eta(x_0, c, w) - x_0\Big\|^2 \;+\; \lambda_{\mathrm{scale}}\,\mathcal{L}_\eta^{\mathrm{scale}}. \tag{15}$$

Under (A2) its unique global minimizer is the true guided prox. The diagonal log-scale head s_η is trained by a reparameterized reverse-KL fit: with $y = m_\eta + \sqrt{2\varepsilon}\,e^{s_\eta/2}z$ (m_η and θ detached), it minimizes $-\varepsilon\sum_i s_i + \mathbb{E}_q[\frac{1}{2}\|y - x_0\|^2 + \varphi_\theta(y)]$, whose minimizer is $S^{-1} = I + \nabla^2\varphi_\theta(m_\eta)$ (Appendix I). One scalar certifies estimator health per batch: the normalized effective sample size ESS/K of the importance weights. It gates the ε-annealing schedule and empirically certifies the tracking assumption (A4). Appendix K gives the full pseudocode, schedules, and monitoring protocol.

Sampling. Generation is a single forward pass: output $\hat{x}_1 = m_\eta(x_0, c, w)$ (strict 1-NFE). By Proposition 3.4 this lands within $W_2 = O(\sqrt{\varepsilon d})$ of the model law, with no autograd, ODE solve, or distillation at inference time.

Classifier-free guidance. Guidance is pure inference-time potential arithmetic (Ho & Salimans, 2022):

$$\varphi_\theta^w(x, c) = (1 + w)\,\varphi_\theta(x, c) - w\,\varphi_\theta(x, \varnothing) \iff \psi_1^w = (\psi_1^c)^{1+w}(\psi_1^\varnothing)^{-w}. \tag{16}$$

Proposition 3.10 (Guided bridge)**.** *For any w such that φ^w satisfies (A1)-(A2), ψ_1^w is a valid data-side potential, and every result of §3.1–§3.5 applies verbatim with $\varphi_\theta \to \varphi^w$. The guided sampler approximates the prox sampler of the Schrödinger bridge onto the guided marginal $\rho_1^w := \hat{\psi}_1^w\psi_1^w$, up to the same $O(\varepsilon)$ and network-fit terms. (Appendix J.)*

Table 1: **FID ablations, 1-NFE ImageNet** 256×256**.** 138 M params, batch 128, 5000 iters from scratch. Each panel varies one factor; base config in **bold**. FID-50K.

K	FID
4	92.4
8	**87.3**
16	94.7
32	96.1

(a) draws K

α_{def}	FID
0.0	91.3
0.01	90.4
0.05	89.9
0.1	**86.4**

(b) α_{def}

w	FID
1.0	90.1
1.25	**89.1**
1.5	96.4
2.0	102.5

(c) CFG w

sched.	FID
const	99.9
linear	109.2
cosine	**87.1**
exp	92.1

(d) ε sched.

Table 2: **Proposal ablation.** ESS/K of the estimator equation 9 vs. viscosity ε, for four proposals of increasing structure: naive $\mathcal{N}(x_0, 2\varepsilon I)$; centered $\mathcal{N}(m_\eta, 2\varepsilon I)$; centered-diagonal $\mathcal{N}(m_\eta, 2\varepsilon D^{-1})$, $D{=}\mathrm{diag}(I{+}\nabla^2\varphi_\theta)$; and the defensive mixture q^{def}. Center m_η (deployed 1-NFE map). Using the reference prox $y^*{=}\,\mathrm{prox}_{\varphi_\theta}$ instead of m_η changes ESS/K only slightly throughout, confirming $m_\eta{\approx}y^*$ (A4). Naive is center-independent by construction. $N{=}512$, $K{=}64$, $M{=}5$ seeds. Operating point $\varepsilon{=}0.01$.

ε	naive	$\mathcal{N}(m, 2\varepsilon I)$	$\mathcal{N}(m, 2\varepsilon D^{-1})$	q^{def}
0.20	0.2698	0.9155	0.9164	0.9490
0.10	0.1414	0.9162	0.9459	0.9380
0.05	0.0416	0.8158	0.9158	0.9641
0.01	0.0768	0.9158	0.9758	**0.9895**
0.005	0.0312	0.8162	0.8359	0.7913

Two sentences delimit the claim. What PT-Flow's guidance avoids, unlike ODE-based CFG, is discretization error: there is no ODE and no step size, though the guided prox sampler carries the same $O(\varepsilon)$ and network-fit approximation as the unguided one. What it shares with ODE-based CFG is the target: ρ_1^w is a controlled w-sharpening of the conditional model. It agrees with the classical tilt $\rho^{1+w}\rho^{-w}$ to first order in w and in the $\varepsilon \to 0$ limit, but it is not identical at finite ε, because the conjugate factor $\hat{\psi}_1^w$ re-equilibrates through the bridge (Appendix J).

4 Experiment

4.1 Experiment Setting

We evaluate PT-Flow on class-conditional ImageNet (Deng et al., 2009) at 256×256. Following common practice (Peebles & Xie, 2023; Frans et al., 2025; Geng et al., 2025a), ImageNet generation is performed in the latent space of a pretrained SD-VAE tokenizer (Rombach et al., 2022), which maps each 256×256 image to a $4\times32\times32$ latent. All models are trained *from scratch*, with no pretraining, distillation, or teacher. We assess sample quality with the Fréchet Inception Distance (FID) (Heusel et al., 2017) computed over 50,000 generated samples (FID-50K), report the generator parameter count, and measure compute by the number of function evaluations (NFE), focusing on single-step (1-NFE) generation by default. Implementation details, architectures, and hyperparameters are in Appendix L.

4.2 Ablation Study

We ablate PT-Flow's hyperparameters in Table 1 (B/2 model, 5000 iters from scratch, one factor at a time; used to pick defaults, not to report final quality). Two trends hold above the run-to-run spread: guidance has a clear optimum ($w{\approx}1.2$ best, 102.5 at $w{=}2.0$), and the cosine ε anneal beats constant or linear (87.1 vs. 99.9, 109.2). The draw count K has a smaller effect, with $K{=}8$ best. We use these settings for all main runs.

The variance reversal, empirically. Table 2 evaluates the estimator on the frozen networks as ε cools (Figure 2a, Table 7). Three findings. (i) The naive proposal degrades sharply (ESS/K from 0.27 to below 0.05), reproducing the barrier of §3.4 as an observed effect, not a formal impossibility. (ii) Centering at the prox alone recovers most of the health ($\mathrm{ESS}/K \approx 0.82$–$0.92$); diagonal scaling adds a smaller further gain, so it is beneficial rather than strictly necessary at this scale. (iii) The deployed m_η closely matches the reference prox y^* across all ε (Table 7), the empirical certificate of (A4). Health peaks near the operating point $\varepsilon{=}0.01$; below it the fixed defensive component dominates and ESS/K falls, consistent with operating at moderate viscosity.

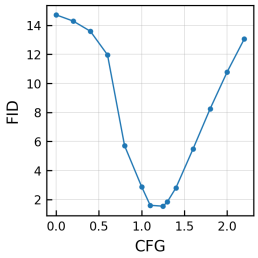
4.3 Comparison With Other Works

One-step generation. Using its best guidance setting, PT-Flow reaches an FID-50K of **1.54** with a single function evaluation (1-NFE). This follows a comparable setup to the other from-scratch one-step models in Table 3: a DiT-L/2 backbone in the same SD-VAE latent space, trained from scratch.

Table 3: **Class-conditional generation on ImageNet** 256×256**.** FID is computed on 50K images. All methods are reported with CFG if applicable. "#Params" denotes "generator + decoder". "×2" denotes an NFE of 2 per sampling step incurred by CFG. All generators are trained from scratch without distilling from a teacher.

method	#Params	NFE	FID↓
Multi-step Diffusion/Flows			
DiT-XL/2 (Peebles & Xie, 2023)	675M+49M	250×2	2.27
SiT-XL/2 (Ma et al., 2024)	675M+49M	250×2	2.06
SiT-XL/2+REPA (Yu et al., 2025)	675M+49M	250×2	**1.42**
Masking & Autoregressive			
MaskGIT (Chang et al., 2022)	227M+31M	8	6.18
VAR-*d*30 (Tian et al., 2024)	2B+61M	10×2	1.92
MAR-H (Li et al., 2024)	943M+41M	256×2	**1.55**
GANs			
BigGAN (Brock et al., 2019)	112M	1	6.95
GigaGAN (Kang et al., 2023)	569M	1	3.45
StyleGAN-XL (Sauer et al., 2022)	166M	1	**2.30**

method	#Params	NFE	FID↓
1-NFE Diffusion/Flows (from scratch)			
iCT-XL/2 (Song & Dhariwal, 2024)	675M+49M	1	34.24
Shortcut-XL/2 (Frans et al., 2025)	675M+49M	1	10.60
iMeanFlow-XL/2 (Geng et al., 2025b)	610M	1	1.72
MeanFlow-B/2 (Geng et al., 2025a)	131M	1	6.17
MeanFlow-L/2	459M	1	3.84
MeanFlow-XL/2	676M	1	3.43
SoFlow-B/2 (Luo et al., 2026)	131M	1	4.85
SoFlow-L/2	459M	1	3.20
SoFlow-XL/2	676M	1	2.96
α-Flow-XL/2+ (Zhang et al., 2026)	676M+49M	1	2.58
W-Flow-XL/2 (Han et al., 2026)	679M+49M	1	**1.29**
Schrödinger Bridge (ours)			
PT-Flow-L/2	462M	1	1.54



(a) FID-50K vs. CFG scale w.



(b) Samples at $w \in \{0,\ 0.6,\ 1.25,\ 2.0,\ 2.2\}$ (left to right).

Figure 1: **Classifier-free guidance for PT-Flow-L/2 on ImageNet** 256×256 **(1-NFE, Mode A).** *Left:* FID is U-shaped in w, optimal near $w^\star \approx 1.25$. *Right:* curated samples for classes 817 (sports car, top) and 323 (monarch butterfly, bottom), w increasing left to right; $w{=}0$ is washed out, $w{\geq}2.0$ oversaturates and adds artifacts.

At this scale, 1.54 is well below the best published 1-NFE results, namely MeanFlow-L/2 at 3.84 (Geng et al., 2025a) and SoFlow-L/2 at 3.20 (Luo et al., 2026). At sampling time, PT-Flow runs one forward pass of the prox network. It needs no autograd, no ODE solver, and no distillation step. PT-Flow-L/2 (462M) generates 165.28 images/s on a single H100, faster than the SoFlow (154.35), iMeanFlow (130.64), and WFlow (145.78) at 1-NFE (Table 6).

Classifier-free guidance. In PT-Flow, guidance is closed-form arithmetic on a valid data-side potential (Prop. 3.10), so one trained model works for every scale w. Figure 1a sweeps w at 1-NFE: the FID curve is U-shaped, dropping from about 14.7 at $w{=}0$ to a low near $w{=}1.1$–1.25 and rising past 13 by $w{=}2.2$, so the optimum lies in a narrow range. The samples in Figure 1b match: low-contrast at $w{\leq}0.6$, sharpest and best-matched at $w{=}1.25$, oversaturated with artifacts at $w{\geq}2.0$.

Limitations. PT-Flow has two limitations. First, training is expensive: each step evaluates the tilted importance estimator over K proposal draws and the prox and scale computations, which cost more per step than a plain flow-matching update. This cost is confined to training; sampling stays cheap (Table 6): generation is a single forward pass of the prox network. Second, we have not evaluated PT-Flow at the largest scales: our results use a 462M-parameter backbone, and whether quality and stability keep improving at XL and above is left to future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, writing refinement, and polishing the presentation of the manuscript. We also used generative AI tools to help draft portions of the text, which were subsequently reviewed, revised, and edited by the authors. We did not use generative AI tools for generating research results, conducting experiments, analyzing data, formulating scientific claims, or producing final figures and evaluation results. These aspects of the work were carried out and verified by the authors. All AI-assisted content was carefully reviewed for accuracy, clarity, and consistency with the scientific contributions of this work. The authors take full responsibility for the final content of the paper, including all text, claims, analyses, results, and artifacts produced with the aid of generative AI tools.

REFERENCES

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, 2017.

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

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Julian D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3):225–236, 1951.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied Mathematics and Optimization*, 23(1):313–329, 1991.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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

Roman Dyachenko, Nikita Gushchin, Kirill Sokolov, Petr Mokrov, Evgeny Burnaev, and Alexander Korotin. Variational entropic optimal transport. In *International Conference on Machine Learning*, 2026. arXiv:2602.02241.

Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2006.

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

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=uWj4s7rMnR.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Eberhard Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3):201–230, 1950.

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

Alexander Korotin, Nikita Gushchin, and Evgeny Burnaev. Light Schrödinger bridge. In *International Conference on Learning Representations*, 2024.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Tianze Luo, Haotian Yuan, and Zhuang Liu. Soflow: Solution flow models for one-step generative modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cjb03GNqYw.

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

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, 2020.

Petr Mokrov, Alexander Korotin, Alexander Kolesov, Nikita Gushchin, and Evgeny Burnaev. Energy-guided entropic neural optimal transport. In *International Conference on Learning Representations*, 2024.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH*, 2022.

Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*, pp. 144–153, 1931.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, 2023.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, 2024.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. Alphaflow: Understanding and improving meanflow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=adacb4JTIv`.

Xinxi Zhang, Shiwei Tan, Quang Nguyen, Quan Dao, Ligong Han, Xiaoxiao He, Tunyu Zhang, Alen Mrdovic, and Dimitris Metaxas. Flow straighter and faster: Efficient one-step generative modeling via MeanFlow on rectified trajectories. *arXiv preprint arXiv:2511.23342*, 2025.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *International Conference on Machine Learning*, 2025.

APPENDIX

A RELATED WORK

One-step generation by collapsing a flow. Diffusion (Ho et al., 2020; Song et al., 2021) and flow-matching (Lipman et al., 2023; Liu et al., 2023) models generate by integrating an SDE/ODE over many steps. A large body of work compresses this trajectory into one or few steps, either by distilling a pretrained teacher or by training from scratch with self-consistency constraints (Song et al., 2023; Frans et al., 2025; Zhou et al., 2025). MeanFlow (Geng et al., 2025a) recasts this as regressing an interval-*averaged* velocity and has become the dominant from-scratch one-step baseline; subsequent work refines it along several axes. AlphaFlow (Zhang et al., 2026) decomposes the MeanFlow objective into trajectory flow-matching and trajectory-consistency terms whose gradients conflict, and anneals between them by curriculum; SoFlow (Luo et al., 2026) instead learns the ODE *solution operator* directly, avoiding the Jacobian–vector products MeanFlow requires; and Re-MeanFlow (Zhang et al., 2025) trains MeanFlow on rectified, low-curvature trajectories for faster convergence. These methods share one training signal—an ODE velocity field originally designed for iterative sampling—and are assessed purely by sample fidelity (FID).

Evolving the pushforward at training time. A separate line abandons the ODE and instead *evolves the generated distribution during training*. Drifting Models (Deng et al., 2026) move generated particles along a hand-constructed field of kernel attraction to data and repulsion from generated samples, achieving strong one-step FID with no SDE/ODE and no JVP; W-Flow (Han et al., 2026) places this idea on a principled footing by deriving the update as the Wasserstein gradient flow of the *Sinkhorn divergence*, and sets a new one-step state of the art. W-Flow is the closest prior work to ours in machinery: it inherits the entropic-OT, quadratic-cost, ε-regularized geometry, and its velocity is a difference of entropic-OT barycentric projections—the same proximal map our generator realizes. Both methods, however, evolve the pushforward through a training-time field rather than solving the bridge in closed form, and neither certifies optimal-transport consistency of the learned coupling.

Simulation-free Schrödinger bridges and entropic OT. PT-Flow belongs instead to the line of *simulation-free* entropic-OT and Schrödinger-bridge solvers, which has historically forked between two costly options: restricting the potential to a tractable parametric family, as in LightSB (Korotin et al., 2024), or sampling the bridge kernel by MCMC, as in EgNOT (Mokrov et al., 2024). Matching-based solvers such as DSBM (Shi et al., 2023) and SF^{2}M (Tong et al., 2024) remove simulation from training but still integrate over many steps at inference, and free-form neural-OT maps have otherwise been reachable only through an unstable dual game (Makkuva et al., 2020). We trace this fork to a single variance barrier in the one Gaussian expectation the program reduces to: Cole–Hopf linearization supplies a closed-form solution operator, a prox-tilted importance estimator whose variance *reverses* in the viscosity ε removes the barrier, and the proposal-is-generator identity turns the fix into a one-step map. The concurrent VarEOT (Dyachenko et al., 2026) independently handles the entropic normalizer with a variational auxiliary network but leaves the inner variance untouched, and the two treatments compose. Among these, PT-Flow is simultaneously teacher-free, minimax-free, optimal-transport-consistent at convergence.

B NOTATION AND STANDING ASSUMPTIONS

Standing assumptions. Each result in the main text cites the labels it uses; nothing else is assumed anywhere.

- **(A0) Bridge well-posedness.** ρ_1 has finite second moment and $\mathrm{KL}(\rho_1\|\mathcal{N}) < \infty$ for some Gaussian; the Schrödinger system equation 1 then has a solution, unique up to the scalar gauge $(\psi \to c\psi, \hat{\psi} \to \hat{\psi}/c)$ (Léonard, 2014).
- **(A1) Coercivity/integrability.** $\varphi_\theta(y) + \frac{1-\delta}{2}\|y\|^2 \to +\infty$ for some $\delta \in (0,1)$. This single condition makes $\psi_0 = G_{2\varepsilon} * \psi_1^\theta$ finite everywhere, licenses the derivative–integral interchanges in Theorem 3.3, and keeps $\hat{\psi}_1$ finite. It is implied by (A2).

Symbol	Definition	Lives at
$t \in [0,1]$	generative time; t=0 noise, t=1 data	—
$s = 1 - t$	smoothing time; s=0 data side, s=1 noise side	—
$\rho_0 = \mathcal{N}(0, I_d)$, $\rho_1 = \rho_{\text{data}}$	marginals	t=0, t=1
$u_\theta(x)$	learned scalar log-potential (raw network output)	s=0
$\varphi_\theta = 2\varepsilon\, u_\theta$	potential in physical units	s=0
$\psi_1^\theta = e^{-u_\theta}$	data-side Schrödinger potential	s=0
$\psi^{(s)}$	ψ propagated to smoothing time s	s
$\psi_0 = \psi^{(1)}$, $\varphi_0 = -2\varepsilon \log \psi_0$	noise-side potential, smoothed potential	s=1
$\hat{\psi}_0, \hat{\psi}_1$	conjugate potentials	—
$F_x(y) = \varphi_\theta(y) + \frac{1}{2}\|x - y\|^2$	Laplace exponent; $y^*(x) = \arg\min F_x = \text{prox}_{\varphi_\theta}(x)$	—
$G_{\sigma^2}(z)$	Gaussian kernel $(2\pi\sigma^2)^{-d/2} e^{-\|z\|^2/2\sigma^2}$	—
ε	viscosity = half the Brownian variance rate	—

Table 4: Notation used throughout. Smoothing time runs from the data side to the noise side; generation runs in the opposite direction. The generator acts on noise through φ_0, whereas the training loss evaluates φ_θ directly on data.

- **(A2) Weak convexity.** $\nabla^2 \varphi_\theta \succeq -\lambda I$ with $\lambda < 1$; then F_x is $(1-\lambda)$-strongly convex, so the prox is unique and the Laplace point nondegenerate. This is a genuine modeling assumption on a nonconvex network; its observable failure signature is a multi-modal prox residual, and it is monitored and softly enforced as described in Appendix K. Guidance tightens it: $(1+w)\lambda_{\min}^{(c)} - w\lambda_{\max}^{(\varnothing)} > -1$ is required at guidance weight w.
- **(A3) Smoothness.** $\varphi_\theta \in C^3$ near prox points with $M_3 = \|\nabla^3 F_x(y^*)\|$ finite; the heuristic variance constants additionally take the diagonal third-order contribution as dominant (what survives without this is noted in Appendix H.2).
- **(A4) Tracking.** During alternating training, m_η stays within the Laplace basin of $\text{prox}_{\varphi_\theta}$ (two-timescale condition); the normalized ESS is its empirical certificate.
- **(A5) Unit-scaled data** for the blur calibration in Proposition 3.8—a preprocessing convention.
- **(A6) Asymptotic diagonal matching.** With $B_x = D^{-1/2}(H - D)D^{-1/2}$ where $H = I + \nabla^2\varphi_\theta(y^*)$ and $D = \text{diag}(H)$, the off-diagonal mismatch satisfies $\|B_x\|_F^2 \le C_{\text{off}}\,\varepsilon$ uniformly, with C_{off} independent of ε (exactly diagonal H gives $C_{\text{off}} = 0$). This is a modeling assumption on the trained potential, not a guarantee; we verify its consequence empirically rather than assume it a priori. Its consequence is examined in Tables 2 and 7: ESS/K stays high where it holds and degrades below $\varepsilon \approx 10^{-2}$.

C Derivation of the governing law: from potential flow to the coupled PDE system

This appendix derives the pair of equations imposed in §3.1 from first principles, in four short steps. Nothing here is assumed that is not either a modeling choice stated explicitly or a consequence of one. The endpoint is the coupled system equation 30. Two time variables are used throughout: generative time $t \in [0,1]$ (noise at t=0, data at t=1) and smoothing time $s = 1 - t$, so $\partial_s = -\partial_t$.

C.1 The velocity is a gradient

We model generation as the flow of a pressureless fluid whose velocity field is curl-free. Equivalently, the velocity is the (negative) gradient of a scalar potential φ,

$$v(x,t) \;=\; -\nabla\varphi(x,t). \tag{17}$$

This is imposed here for concreteness, but C.3 shows it is not a genuine restriction: the least-effort transport principle selects a gradient velocity on its own, so equation 17 is recovered as an optimality condition rather than an assumption.

C.2 Mass conservation gives Fokker–Planck

Probability is conserved locally: the mass in any region changes only through flux across its boundary. For a probability current J this is the continuity equation

$$\partial_t \rho + \nabla \cdot J \;=\; 0. \tag{18}$$

The current has two independent parts. Drift transports mass bodily at velocity v, contributing ρv. Diffusion, generated by the Brownian jitter of the bridge particles (variance rate 2ε), adds a Fickian flux down the density gradient, $-\varepsilon \nabla \rho$. Hence

$$J \;=\; \underbrace{\rho\, v}_{\text{drift}} \;-\; \underbrace{\varepsilon \nabla \rho}_{\text{diffusion}}\,. \tag{19}$$

Substituting equation 19 into equation 18 and using $\nabla \cdot \nabla \rho = \Delta \rho$,

$$\boxed{\partial_t \rho + \nabla \cdot (\rho v) \;=\; \varepsilon\, \Delta \rho} \qquad \text{(Fokker–Planck).} \tag{20}$$

With the gradient drift equation 17 this reads $\partial_t \rho - \nabla \cdot (\rho \nabla \varphi) = \varepsilon \Delta \rho$: mass flows downhill on φ. At $\varepsilon = 0$ the diffusive flux vanishes and equation 20 reduces to pure continuity $\partial_t \rho + \nabla \cdot (\rho v) = 0$.

C.3 Least-effort transport gives the inviscid Hamilton–Jacobi equation

Equation equation 20 is kinematics: it holds for any velocity field and selects none. To pick the transport we minimize the Benamou–Brenier kinetic action (Benamou & Brenier, 2000),

$$\mathcal{A}[\rho, v] \;=\; \int_0^1 \!\!\int \tfrac{1}{2}\|v\|^2 \,\rho \;\mathrm{d}x\;\mathrm{d}t, \qquad \text{s.t.} \quad \partial_t \rho + \nabla \cdot (\rho v) = 0,\;\; \rho(\cdot, 0) = \rho_0,\;\; \rho(\cdot, 1) = \rho_{\text{data}}. \tag{21}$$

Enforce the continuity constraint with a space-time multiplier field, which we name φ:

$$\mathcal{L} \;=\; \int_0^1 \!\!\int \Big[\tfrac{1}{2}\|v\|^2 \rho \;-\; \varphi\big(\partial_t \rho + \nabla \cdot (\rho v)\big) \Big] \,\mathrm{d}x\;\mathrm{d}t. \tag{22}$$

Integrating by parts to move derivatives off ρ and v onto φ (endpoint terms in t vanish because ρ is fixed at $t{=}0, 1$; spatial terms vanish at infinity) yields

$$\mathcal{L} \;=\; \int_0^1 \!\!\int \Big[\tfrac{1}{2}\|v\|^2 \rho + \rho\, \partial_t \varphi + \rho\, v \cdot \nabla \varphi \Big] \,\mathrm{d}x\;\mathrm{d}t \;+\; \text{const.} \tag{23}$$

The two stationarity conditions are:

(i) **Vary v (at fixed ρ):** $\rho v + \rho \nabla \varphi = 0 \;\Rightarrow\; v = -\nabla \varphi$. This derives the gradient structure posited in equation 17; the multiplier φ is the potential.

(ii) **Vary ρ (at fixed v):** $\tfrac{1}{2}\|v\|^2 + \partial_t \varphi + v \cdot \nabla \varphi = 0$.

Substituting the optimal $v = -\nabla \varphi$ into (ii), with $\tfrac{1}{2}\|v\|^2 = \tfrac{1}{2}\|\nabla \varphi\|^2$ and $v \cdot \nabla \varphi = -\|\nabla \varphi\|^2$,

$$\tfrac{1}{2}\|\nabla \varphi\|^2 + \partial_t \varphi - \|\nabla \varphi\|^2 = 0 \;\Longrightarrow\; \partial_t \varphi - \tfrac{1}{2}\|\nabla \varphi\|^2 = 0. \tag{24}$$

Passing to smoothing time $s = 1 - t$ ($\partial_s = -\partial_t$),

$$\boxed{\partial_s \varphi + \tfrac{1}{2}\|\nabla \varphi\|^2 = 0} \qquad \text{(inviscid Hamilton–Jacobi).} \tag{25}$$

C.4 The stochastic lift adds the viscosity term

The deterministic principle of Step 3 gives the $\varepsilon = 0$ equation. The $\varepsilon \Delta \varphi$ term appears the moment the transport is a controlled diffusion rather than an ODE flow. Consider

$$\mathrm{d}X_t = u\;\mathrm{d}t + \sqrt{2\varepsilon}\;\mathrm{d}W_t, \qquad V(x, t) = \min_u \mathbb{E}\left[\int_t^1 \tfrac{1}{2}\|u\|^2 \;\mathrm{d}r + g(X_1) \;\middle|\; X_t = x \right], \tag{26}$$

whose value function V we identify with the potential, $\varphi := V$. Bellman's principle over $[t, t+\,\mathrm{d}t]$ is $V(x,t) = \min_u \mathbb{E}\big[\frac{1}{2}\|u\|^2\,\mathrm{d}t + V(X_{t+\,\mathrm{d}t}, t+\,\mathrm{d}t)\big]$. Expanding the second term with Itô's lemma — whose second-order term is generated by the quadratic variation of $\sqrt{2\varepsilon}\;\mathrm{d}W$ — gives

$$\mathbb{E}[\,\mathrm{d}V] = \Big(\partial_t V + u\cdot\nabla V + \underbrace{\varepsilon\Delta V}_{\text{from } \mathrm{d}W^2}\Big)\,\mathrm{d}t, \tag{27}$$

so Bellman becomes $0 = \min_u \big[\frac{1}{2}\|u\|^2 + \partial_t V + u\cdot\nabla V + \varepsilon\Delta V\big]$. Minimizing over u gives the optimal control $u^* = -\nabla V$; substituting it back,

$$0 = \tfrac{1}{2}\|\nabla V\|^2 + \partial_t V - \|\nabla V\|^2 + \varepsilon\Delta V = \partial_t V - \tfrac{1}{2}\|\nabla V\|^2 + \varepsilon\Delta V. \tag{28}$$

With $\varphi = V$, $v = u^* = -\nabla\varphi$, and $\partial_s = -\partial_t$,

$$\boxed{\partial_s\varphi + \tfrac{1}{2}\|\nabla\varphi\|^2 = \varepsilon\,\Delta\varphi} \qquad \text{(viscous Hamilton–Jacobi).} \tag{29}$$

This is equation 25 with the viscosity term restored, and the two must agree as $\varepsilon \to 0$, which fixes the sign of the quadratic term as $+\frac{1}{2}\|\nabla\varphi\|^2$.

What the control problem does and does not fix. Problem equation 26 is the stochastic-control form of the Schrödinger bridge with Brownian reference of variance rate 2ε; the identification $\varphi = V$ is Fleming's logarithmic transformation, which linearizes equation 29 into the heat equation of Appendix D.1 (Dai Pra, 1991; Fleming & Soner, 2006). Two roles must be kept separate. The control principle fixes the form of the PDE equation 29. It does not by itself fix the boundary data: the terminal cost g (equivalently the data-side potential $\psi_1 = e^{-\varphi_\theta/2\varepsilon}$) that makes the controlled process actually land on ρ_{data} is determined by the Schrödinger factorization solved during training (§3.3), not by equation 26 alone.

C.5 The governing law assembled

Collecting C.2 and C.4 with the shared relation $v = -\nabla\varphi$:

$$\underbrace{\partial_t\rho + \nabla\cdot(\rho v) = \varepsilon\Delta\rho}_{\text{mass (constraint)}}, \qquad \underbrace{\partial_s\varphi + \tfrac{1}{2}\|\nabla\varphi\|^2 = \varepsilon\Delta\varphi}_{\text{potential (optimality)}}, \qquad v = -\nabla\varphi, \tag{30}$$

with boundary conditions $\rho(\cdot,0) = \rho_0 = \mathcal{N}(0, I_d)$ and $\rho(\cdot,1) = \rho_{\mathrm{data}}$. The first equation is kinematics — conservation of probability — and the second is the optimality condition selecting the least-effort transport among all flows satisfying it. At $\varepsilon = 0$ the pair is the Benamou–Brenier optimality system for W_2 optimal transport; at $\varepsilon > 0$ it is the Schrödinger bridge, whose boundary machinery we use in §3.3. The nonlinearity in the potential equation is removed exactly by the Cole–Hopf transform in Appendix D.1, which is the same logarithmic transformation that appears in equation 26.

D Proofs for Section 3.1

D.1 Theorem 3.1 (Cole–Hopf)

Write $\varphi = -2\varepsilon\log\psi$ and compute the three derivatives:

$$\partial_s\varphi = -2\varepsilon\frac{\partial_s\psi}{\psi}; \qquad \nabla\varphi = -2\varepsilon\frac{\nabla\psi}{\psi} \;\Rightarrow\; \tfrac{1}{2}\|\nabla\varphi\|^2 = 2\varepsilon^2\frac{\|\nabla\psi\|^2}{\psi^2};$$

$$\Delta\varphi = -2\varepsilon\Big[\frac{\Delta\psi}{\psi} - \frac{\|\nabla\psi\|^2}{\psi^2}\Big] \;\Rightarrow\; \varepsilon\Delta\varphi = -2\varepsilon^2\frac{\Delta\psi}{\psi} + 2\varepsilon^2\frac{\|\nabla\psi\|^2}{\psi^2}.$$

Substitute these into equation 2. The terms $2\varepsilon^2\|\nabla\psi\|^2/\psi^2$ appear on both sides with the same sign and cancel exactly. This cancellation is the entire content of the theorem. Dividing what remains by $-2\varepsilon/\psi$ gives $\partial_s\psi = \varepsilon\Delta\psi$. Both directions of the equivalence follow, since the change of variables is a bijection on positive ψ. □

D.2 COROLLARY 3.2 AND THE HOPF–LAX EXPANSION EQUATION 4

From the heat equation to the Gaussian expectation. By Theorem 3.1, ψ solves the linear heat equation $\partial_s \psi = \varepsilon \Delta \psi$ on $s \in [0,1]$ with initial datum $\psi^{(0)} = \psi_1^\theta$. This equation is solved in closed form by the heat semigroup: convolution with the Gaussian kernel whose variance grows at rate 2ε,

$$\psi^{(s)}(x) \;=\; \big(G_{2\varepsilon s} * \psi^{(0)}\big)(x) \;=\; \int G_{2\varepsilon s}(x-y)\, \psi_1^\theta(y)\, \mathrm{d}y. \tag{31}$$

The variance is $2\varepsilon s$ because the generator is $\varepsilon\Delta$: the kernel $G_{2\varepsilon s}$ satisfies $\partial_s G_{2\varepsilon s} = \varepsilon \Delta G_{2\varepsilon s}$ with $G_{2\varepsilon s} \to \delta$ as $s \to 0$, so equation 31 has the correct initial datum and solves the equation. Finiteness of the integral is guaranteed by (A1).

Specializing to the noise side $s = 1$ and using $\psi_1^\theta = e^{-u_\theta}$,

$$\psi_0(x) \;=\; \big(G_{2\varepsilon} * \psi_1^\theta\big)(x) \;=\; \int G_{2\varepsilon}(x-y)\, e^{-u_\theta(y)}\, \mathrm{d}y. \tag{32}$$

Write $y = x + \sqrt{2\varepsilon}\, z$. Then $G_{2\varepsilon}(x-y)\, \mathrm{d}y$ is exactly the law of $z \sim \mathcal{N}(0, I)$, since $G_{2\varepsilon}$ has covariance $2\varepsilon I$ and the substitution has unit Jacobian in z. Hence the convolution is a standard Gaussian expectation,

$$\psi_0(x) \;=\; \mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[e^{-u_\theta(x+\sqrt{2\varepsilon}\, z)}\big], \tag{33}$$

which is equation 3. This is the form the estimator of §3.5 evaluates.

The Hopf–Lax expansion. To obtain equation 4, write the kernel normalization explicitly:

$$\varphi_0(x) = -2\varepsilon \log \int (4\pi\varepsilon)^{-d/2} e^{-F_x(y)/2\varepsilon}\, \mathrm{d}y, \qquad F_x(y) = \varphi_\theta(y) + \tfrac{1}{2}\|x-y\|^2.$$

Under (A1)–(A2), F_x has a unique nondegenerate minimizer $y^*(x)$, and Laplace's method gives

$$\int e^{-F_x/2\varepsilon}\, \mathrm{d}y \;\approx\; (4\pi\varepsilon)^{d/2} \det\big(\nabla^2 F_x(y^*)\big)^{-1/2} e^{-F_x(y^*)/2\varepsilon}.$$

The prefactor $(4\pi\varepsilon)^{-d/2}$ cancels the $(4\pi\varepsilon)^{d/2}$ exactly, so no $\varepsilon \log \varepsilon$ terms survive. Taking $-2\varepsilon \log$ then yields

$$\varphi_0(x) = M_{\varphi_\theta}(x) + \varepsilon \log\det \nabla^2 F_x(y^*) + O(\varepsilon^2 \text{ per dim}).$$

The correction term does not vanish. Writing the determinant as a sum over eigenvalues, $\varepsilon \log\det \nabla^2 F_x(y^*) = \varepsilon \sum_{i=1}^d \log \lambda_i$, shows it is a sum of d order-one contributions, hence $O(\varepsilon d)$. The Hopf–Lax limit therefore requires $\varepsilon d \to 0$, not merely $\varepsilon \to 0$. Finally, the minimizer is by definition $y^*(x) = \mathrm{prox}_{\varphi_\theta}(x)$, and the Moreau-envelope gradient identity $\nabla M_{\varphi_\theta}(x) = x - \mathrm{prox}_{\varphi_\theta}(x)$ is classical (Moreau, 1965). □

E PROOFS FOR SECTION 3.2

Scope of exactness (referenced from §3.2). Four objects are kept distinct. (i) The bridge identities closed-form propagation equation 3 and the Tweedie mean equation 5 are exact. (ii) The identification of that mean with the proximal map holds up to the gradient of the Laplace correction equation 4 (order ε, not its $O(\varepsilon d)$ value), i.e. only as $\varepsilon \to 0$. (iii) The deployed generator m_η approximates the prox to within a measurable first-order residual. (iv) The model density is exactly normalized (Theorem 3.5). Claims of exactness in the main text refer to (i) and the normalization in (iv); (ii)–(iii) are quantified approximations.

E.1 THEOREM 3.3 (TWEEDIE IDENTITY)

Step 1 (the bridge conditional). Under the static factorization equation 1, the joint law is $\pi^*(x_0, x_1) = \hat{\psi}_0(x_0)\, G_{2\varepsilon}(x_1 - x_0)\, \psi_1(x_1)$, and the x_0-marginal is

$$p(x_0) = \hat{\psi}_0(x_0) \int G_{2\varepsilon}(x_1 - x_0)\, \psi_1(x_1)\, \mathrm{d}x_1 = \hat{\psi}_0(x_0)\, \psi_0(x_0),$$

using $\psi_0 = G_{2\varepsilon} * \psi_1$ from Corollary 3.2. Dividing, the $\hat{\psi}_0(x_0)$ factors cancel and

$$p(x_1 \mid x_0) = \frac{G_{2\varepsilon}(x_1 - x_0)\,\psi_1(x_1)}{\psi_0(x_0)}. \tag{34}$$

Step 2 (a gradient identity for the Gaussian kernel). Write $z = x_1 - x_0$. Since $G_{2\varepsilon}(z) \propto e^{-\|z\|^2/4\varepsilon}$ and $\partial z/\partial x_0 = -I$, the two sign flips cancel and

$$\nabla_{x_0} G_{2\varepsilon}(x_1 - x_0) = \frac{x_1 - x_0}{2\varepsilon}\, G_{2\varepsilon}(x_1 - x_0).$$

Rearranging isolates the quantity we need to integrate:

$$x_1\, G_{2\varepsilon}(x_1 - x_0) = x_0\, G_{2\varepsilon}(x_1 - x_0) + 2\varepsilon\, \nabla_{x_0} G_{2\varepsilon}(x_1 - x_0). \tag{35}$$

Step 3 (assemble the conditional mean). By equation 34, $\mathbb{E}[X_1 \mid x_0] = \psi_0(x_0)^{-1} \int x_1\, G_{2\varepsilon}(x_1 - x_0)\,\psi_1(x_1)\,\mathrm{d}x_1$. Substitute equation 35 into the integrand and split the two terms:

$$\int x_1 G_{2\varepsilon}\psi_1\,\mathrm{d}x_1 = x_0 \underbrace{\int G_{2\varepsilon}\psi_1\,\mathrm{d}x_1}_{=\psi_0(x_0)} + 2\varepsilon \int \big(\nabla_{x_0} G_{2\varepsilon}\big)\psi_1\,\mathrm{d}x_1.$$

Interchange ∇_{x_0} with the integral in the second term; this is justified by (A1) together with dominated convergence, since the Gaussian decay dominates ψ_1. The second integral becomes $\nabla_{x_0} \int G_{2\varepsilon}\psi_1\,\mathrm{d}x_1 = \nabla\psi_0(x_0)$. Hence

$$\mathbb{E}[X_1 \mid x_0] = \frac{x_0\,\psi_0(x_0) + 2\varepsilon\,\nabla\psi_0(x_0)}{\psi_0(x_0)} = x_0 + 2\varepsilon\,\frac{\nabla\psi_0(x_0)}{\psi_0(x_0)} = x_0 + 2\varepsilon\,\nabla \log \psi_0(x_0),$$

which is the stated identity.

Step 4 (potential and prox forms). Since $\varphi_0 = -2\varepsilon \log \psi_0$, we have $2\varepsilon\nabla \log \psi_0 = -\nabla\varphi_0$, so $T_\varepsilon(x_0) = x_0 - \nabla\varphi_0(x_0)$, which is equation 5. Applying the Moreau identity $\nabla\varphi_0(x) = x - \mathrm{prox}_{\varphi_\theta}(x)$ from Appendix D.2 gives $T(x_0) = \mathrm{prox}_{\varphi_\theta}(x_0)$, which is equation 6, exact up to $\varepsilon\nabla \log\det \nabla^2 F_x$, the gradient of the log-det correction. □

The learned-prox error is measurable. Under (A2), $F_x(y) = \varphi_\theta(y) + \frac{1}{2}\|y - x\|^2$ is $(1 - \lambda)$-strongly convex, and $\mathrm{prox}_{\varphi_\theta}(x)$ is its unique minimizer with $\nabla F_x(\mathrm{prox}_{\varphi_\theta}) = 0$. For the deployed m_η, the residual $r_\eta(x) = \nabla F_x(m_\eta) = \nabla\varphi_\theta(m_\eta) + m_\eta - x$ then bounds the distance to the prox,

$$\delta_\eta(x) = \|m_\eta(x) - \mathrm{prox}_{\varphi_\theta}(x)\| \le \frac{\|r_\eta(x)\|}{1 - \lambda}.$$

This connects the guarantee to a quantity computable from the trained model; we report it in §4.2 (Table 7).

E.2 Proposition 3.4 (sampler–marginal gap)

Draw $X_0 \sim \rho_0$ and $X_1 \sim p(\cdot \mid X_0)$ from the bridge conditional equation 34, and form the pair $(T(X_0), X_1)$. Its first marginal is $T_\#\rho_0$ and its second is $\hat{\rho}_1$, so $(T(X_0), X_1)$ is an admissible coupling of these two distributions.

Because $T(X_0) = \mathbb{E}[X_1 \mid X_0]$ by Theorem 3.3, the squared cost of this coupling is the conditional variance:

$$\mathbb{E}\|X_1 - T(X_0)\|^2 = \mathbb{E}\big\|X_1 - \mathbb{E}[X_1 \mid X_0]\big\|^2 = \mathbb{E}\big[\,\mathrm{tr}\,\mathrm{Cov}(X_1 \mid X_0)\big].$$

The squared Wasserstein distance is the infimum of $\mathbb{E}\|X_1 - T(X_0)\|^2$ over all couplings, so any single coupling upper-bounds it:

$$W_2^2\big(T_\#\rho_0, \hat{\rho}_1\big) \le \mathbb{E}\big[\,\mathrm{tr}\,\mathrm{Cov}(X_1 \mid X_0)\big].$$

A Laplace expansion of the conditional equation 34 about the prox point y^* (Appendix G.2) gives

$$\mathrm{Cov}(X_1 \mid x_0) \approx 2\varepsilon\big(I + \nabla^2\varphi_\theta(y^*)\big)^{-1},$$

whose trace is $\approx 2\varepsilon d/(1 + \bar{\lambda})$, with $\bar{\lambda}$ the mean curvature of φ_θ at the prox points. Taking square roots gives the stated bound. □

Consequence. The deterministic sampler's distribution converges to the model marginal $\hat{\rho}_1$, not to $T_\#\rho_0$, at rate $\sqrt{\varepsilon d}$. Inside the operating window $\varepsilon \lesssim 1/(dM_3^2)$ this is $O(M_3^{-1}/\sqrt{d})$

F PROOFS FOR SECTION 3.3

F.1 THEOREM 3.5 (AUTOMATIC NORMALIZATION)

Gaussian convolution is self-adjoint, $\int (G*f)\, g = \int f\, (G*g)$, by Fubini and the symmetry $G(z) = G(-z)$. Fubini is licensed by (A1), since all integrands are nonnegative. Then

$$\int \hat{\psi}_1 \psi_1 = \int (G_{2\varepsilon} * \hat{\psi}_0)\, \psi_1 = \int \hat{\psi}_0\, (G_{2\varepsilon} * \psi_1) = \int \hat{\psi}_0\, \psi_0 = \int \rho_0 = 1. \qquad \square$$

F.2 THEOREM 3.6(I) (STATIONARITY)

We vary ψ_1 at a point y and set the first variation of equation 7 to zero.

For the first term, recall $\psi_0(x_0) = \int G_{2\varepsilon}(y' - x_0)\, \psi_1(y')\ \mathrm{d}y'$, so $\delta\psi_0(x_0)/\delta\psi_1(y) = G_{2\varepsilon}(y - x_0)$. Differentiating through the logarithm by the chain rule,

$$\frac{\delta}{\delta\psi_1(y)}\, \mathbb{E}_{\rho_0}[\log \psi_0] = \int \rho_0(x_0)\, \frac{1}{\psi_0(x_0)}\, \frac{\delta\psi_0(x_0)}{\delta\psi_1(y)}\ \mathrm{d}x_0$$

$$= \int \rho_0(x_0)\, \frac{G_{2\varepsilon}(y - x_0)}{\psi_0(x_0)}\ \mathrm{d}x_0 = \int \hat{\psi}_0(x_0)\, G_{2\varepsilon}(y - x_0)\ \mathrm{d}x_0 = \hat{\psi}_1(y),$$

where the last two equalities use $\hat{\psi}_0 = \rho_0/\psi_0$ and $\hat{\psi}_1 = G_{2\varepsilon} * \hat{\psi}_0$.

For the second term, $\mathbb{E}_{\rho_1}[u_\theta] = -\mathbb{E}_{\rho_1}[\log \psi_1]$, whose variation is $-\rho_1(y)/\psi_1(y)$. Setting the sum of the two variations to zero gives $\hat{\psi}_1(y) = \rho_1(y)/\psi_1(y)$, that is $\hat{\psi}_1\psi_1 = \rho_1$. The noise-side constraint already holds by construction, so this is the one boundary condition left to enforce. Stationarity of the scalar loss is that boundary condition, and minimizing equation 7 solves the Schrödinger system rather than a surrogate for it. $\square$

F.3 THEOREM 3.6(II) (CONVEXITY)

Let $u = \log \psi_1$. For each fixed x_0, the map $u \mapsto \log \psi_0(x_0) = \log \int G_{2\varepsilon}(\cdot - x_0)\, e^u$ is a log-integral-exp functional, hence convex in u. Averaging these functionals over $x_0 \sim \rho_0$ preserves convexity, and the data term $-\mathbb{E}_{\rho_1}[u]$ is linear in u. Their sum $\mathcal{L}$ is therefore convex in u.

The only flat direction is the gauge $u \to u + c$, which shifts the two terms oppositely and leaves $\mathcal{L}$ invariant. This is the harmless scalar gauge of (A0). One caveat: this convexity holds in function space. It does not by itself yield convergence rates for a nonconvex network parameterization, a limitation we state explicitly. $\square$

F.4 THEOREM 3.6(III) (SEMI-DUAL IDENTITY)

Take quadratic cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ and regularization 2ε. Partial maximization of the entropic-OT dual over the noise-side potential is a closed-form (c, ε)-transform, available because ρ_0 is a known Gaussian. The surviving semi-dual functional is

$$J(g) = \mathbb{E}_{\rho_1}[g] - 2\varepsilon\, \mathbb{E}_{\rho_0} \log \int e^{(g(y) - c(x,y))/2\varepsilon}\ \mathrm{d}y.$$

We now evaluate J at $g = -\varphi_\theta = -2\varepsilon\, u_\theta$. The data term is $\mathbb{E}_{\rho_1}[-2\varepsilon\, u_\theta] = -2\varepsilon\, \mathbb{E}_{\rho_1}[u_\theta]$. In the second term the exponent is

$$\frac{g(y) - c(x, y)}{2\varepsilon} = -u_\theta(y) - \frac{\|x - y\|^2}{4\varepsilon},$$

so the inner integral is $\int e^{(g-c)/2\varepsilon}\ \mathrm{d}y = (4\pi\varepsilon)^{d/2} \psi_0(x)$ by equation 3. Hence

$$-2\varepsilon\, \mathbb{E}_{\rho_0} \log \int e^{(g-c)/2\varepsilon}\ \mathrm{d}y = -2\varepsilon\, \mathbb{E}_{\rho_0}[\log \psi_0] - 2\varepsilon \cdot \tfrac{d}{2} \log(4\pi\varepsilon).$$

Collecting the two terms,

$$J(-\varphi_\theta) = -2\varepsilon\big(\mathbb{E}_{\rho_1}[u_\theta] + \mathbb{E}_{\rho_0}[\log \psi_0]\big) + \text{const} = -2\varepsilon\,\mathcal{L}(\theta) + \text{const},$$

where the constant $-\varepsilon d \log(4\pi\varepsilon)$ does not depend on θ. So minimizing equation 7 is exactly maximizing the semi-dual.

Three consequences follow.

(a) The objective is the same family that LightSB optimizes, but over a conjugate parametric class. This makes the lineage statement exact rather than analogical.

(b) The alternating θ/η optimization is a function-space half-Sinkhorn: the noise side is transformed exactly and analytically, and the data side is fit by gradient steps.

(c) Standard entropic-OT semi-dual results, such as strong duality and uniqueness up to gauge, transfer directly.

□

F.5 Why the flat form, not the nested objective

Directly maximizing $\log \hat{\rho}_1$ would give $\mathcal{L}_{\text{nest}} = -\mathbb{E}_{\rho_1}[\log \hat{\psi}_1 - u_\theta]$. This is correct but numerically poor. It places an estimator inside a ratio, inside a convolution, inside a logarithm, which produces two nested Monte-Carlo levels with opposite-signed Jensen biases, and a flexible u_θ can exploit that gap. The flat form equation 7 avoids this. It has a single Monte-Carlo level, and its data term needs no estimator at all. By Theorem 3.6(i) it is also not a surrogate: its stationary point is exactly the missing boundary condition.

G Derivations for Section 3.4

G.1 Proposition 3.7 (variance barrier)

Fix x and write the single-sample summand of equation 8 as $e^{u(z)}$, where

$$u(z) := -u_\theta\big(x + \sqrt{2\varepsilon}\, z\big), \qquad z \sim \mathcal{N}(0, I).$$

Taylor-expand u_θ about x to first order. Since the step is $\sqrt{2\varepsilon}\, z$,

$$u(z) \approx -u_\theta(x) - \nabla u_\theta(x)\cdot\sqrt{2\varepsilon}\, z = u(0) - \frac{\nabla\varphi_\theta(x)}{2\varepsilon}\cdot\sqrt{2\varepsilon}\, z = u(0) - \frac{\nabla\varphi_\theta(x)\cdot z}{\sqrt{2\varepsilon}},$$

using $u_\theta = \varphi_\theta/2\varepsilon$, hence $\nabla u_\theta = \nabla\varphi_\theta/2\varepsilon$. Thus $u(z)$ is (to first order) an affine function of the Gaussian z, so it is itself Gaussian with mean $u(0)$ and variance

$$s^2 := \mathrm{Var}_z[u(z)] = \frac{1}{2\varepsilon}\,\mathbb{E}_z\big[(\nabla\varphi_\theta\cdot z)^2\big] = \frac{\|\nabla\varphi_\theta(x)\|^2}{2\varepsilon}, \tag{36}$$

where $\mathbb{E}_z[(a\cdot z)^2] = \|a\|^2$ for $z \sim \mathcal{N}(0, I)$.

The summand $e^{u(z)}$ is then log-normal. For a Gaussian $u \sim \mathcal{N}(\mu, s^2)$ the moment generating function gives $\mathbb{E}[e^u] = e^{\mu + s^2/2}$ and $\mathbb{E}[e^{2u}] = e^{2\mu+2s^2}$, so

$$\mathrm{Var}\big(e^u\big) = \mathbb{E}[e^{2u}] - \mathbb{E}[e^u]^2 = e^{2\mu+2s^2} - e^{2\mu+s^2} = e^{2\mu+s^2}\big(e^{s^2} - 1\big).$$

Dividing by $\mathbb{E}[e^u]^2 = e^{2\mu+s^2}$ gives the relative (squared-coefficient-of-variation) form

$$\frac{\mathrm{Var}(e^u)}{\mathbb{E}[e^u]^2} = e^{s^2} - 1. \tag{37}$$

The estimator equation 8 averages K independent copies, which divides the relative variance by K:

$$\frac{\mathrm{Var}\big(\hat{\psi}_0^{\text{naive}}\big)}{\mathbb{E}[\hat{\psi}_0^{\text{naive}}]^2} = \frac{e^{s^2} - 1}{K}.$$

To keep this relative error at $O(1)$ we need $K \gtrsim e^{s^2} - 1$, and substituting equation 36,

$$K_{\text{naive}} \;\gtrsim\; \exp\Big(\frac{\|\nabla\varphi_\theta(x)\|^2}{2\varepsilon}\Big). \tag{38}$$

Remark (what is heuristic, what is robust). Two approximations entered: the first-order truncation of u_θ in Step 1, and the exact log-normality it implies in Step 2. Both are heuristic. The robust content survives them: s^2 in equation 36 is exactly the variance of the log-integrand $u(z)$ regardless of any Taylor step (it is a definition), and for any integrand whose log has variance s^2 the relative variance grows like e^{s^2} up to constants, which is the standard concentration behavior of near-log-normal partition estimators. *Why fatal rather than inconvenient.* Carrying noise onto data requires a displacement $\|T(x_0) - x_0\| = O(\sqrt{d})$. By equation 5, $T(x_0) - x_0 = -\nabla\varphi_0(x_0)$, so $\|\nabla\varphi_0\| = O(\sqrt{d})$; identifying $\nabla\varphi_\theta(x)$ with the smoothed gradient $\nabla\varphi_0$ then gives $\|\nabla\varphi_\theta\|^2 = O(d)$. This identification is heuristic and holds only as $\varepsilon \to 0$: at finite ε the raw gradient at the proposal center x is not the transport displacement, which is set at the prox point a full $O(\sqrt{d})$ away Substituting into equation 38 gives

$$K_{\text{naive}} \;\gtrsim\; e^{\,c\,d/2\varepsilon}.$$

Two numbers make the point. At $d = 2,\ \varepsilon = 0.02$ this is already about 10^{21}. At latent-image scale, $d \approx 4096$, it exceeds 10^{1000} for any useful ε. Note the direction of the dependence: the requirement grows as ε shrinks. That is the opposite of what sharp generation needs, and it is the first jaw of the vice.

G.2 Proposition 3.8 (barycentric collapse)

From the bridge conditional equation 34,

$$\log p(x_1 \mid x_0) = -\frac{\|x_1 - x_0\|^2}{4\varepsilon} - u_\theta(x_1) + \text{const} = -\frac{\|x_1 - x_0\|^2}{4\varepsilon} - \frac{\varphi_\theta(x_1)}{2\varepsilon} + \text{const}.$$

Its negative Hessian in x_1 is

$$-\nabla^2_{x_1} \log p(x_1 \mid x_0) = \frac{1}{2\varepsilon} I + \frac{1}{2\varepsilon}\nabla^2\varphi_\theta(x_1) = \frac{1}{2\varepsilon}\big(I + \nabla^2\varphi_\theta(x_1)\big).$$

A Laplace (Gaussian) approximation about the mode $y^* = \text{prox}_{\varphi_\theta}(x_0)$ has covariance equal to the inverse of this Hessian,

$$\text{Cov}(X_1 \mid x_0) \;\approx\; 2\varepsilon\big(I + \nabla^2\varphi_\theta(y^*)\big)^{-1}, \tag{39}$$

which is the same $(I + \nabla^2\varphi_\theta)$ window that governs the Hopf–Lax correction in equation 4.

The one-step sampler outputs the conditional mean $T(x_0) = \mathbb{E}[X_1 \mid x_0]$, a single point, whereas the true bridge endpoint $X_1 \mid x_0$ is spread with covariance equation 39. The law of total variance decomposes the model marginal's covariance into a between-x_0 part and a within-x_0 part,

$$\underbrace{\text{Cov}(X_1)}_{\text{true marginal } \hat{\rho}_1} = \underbrace{\text{Cov}\big(\mathbb{E}[X_1 \mid X_0]\big)}_{=\,\text{Cov}(T_\#\rho_0)} + \underbrace{\mathbb{E}\big[\text{Cov}(X_1 \mid X_0)\big]}_{\text{lost by the mean map}}.$$

Collapsing the kernel to its mean discards exactly the second term. So the deterministic pushforward $T_\#\rho_0$ under-disperses relative to $\hat{\rho}_1$, and the missing trace is

$$\text{tr}\,\mathbb{E}\big[\text{Cov}(X_1 \mid X_0)\big] \approx \mathbb{E}_{x_0}\,\text{tr}\Big[2\varepsilon\big(I + \nabla^2\varphi_\theta(y^*)\big)^{-1}\Big] \approx \frac{2\varepsilon d}{1 + \bar{\lambda}},$$

where $\bar{\lambda}$ is the mean eigenvalue of $\nabla^2\varphi_\theta$ at the prox points. On unit-scaled data (A5), $\text{tr}\,\text{Cov}(X_1) = O(d)$, so the relative deficit is

$$\frac{\text{tr}\,\mathbb{E}[\text{Cov}(X_1 \mid X_0)]}{\text{tr}\,\text{Cov}(X_1)} \approx \frac{2\varepsilon}{1 + \bar{\lambda}}. \tag{40}$$

The deficit equation 40 depends only on ε and the curvature of φ_θ, not on how well the network is fit. No architecture or amount of training removes it; only cooling ε does, which is why usable generation demands $\varepsilon \lesssim 10^{-2}$. Drawing from the kernel properly rather than taking its mean would require sampling $p(x_1 \mid x_0) \propto G_{2\varepsilon}(x_1 - x_0)\psi_1(x_1)$, an energy-based draw that needs MCMC and destroys the one-step property. This is the second jaw of the vice.

G.3 THE BARRIER AS COMPETING CONSTRAINTS

Collect the two constraints. Proposition 3.8 demands $\varepsilon \lesssim 10^{-2}$, from the blur deficit equation 40. Proposition 3.7 demands $\varepsilon \gtrsim d/\log K$, obtained by inverting equation 38: a computable budget K caps $s^2 \lesssim \log K$, and with $s^2 = \|\nabla\varphi_\theta\|^2/2\varepsilon = O(d/2\varepsilon)$ this reads $\varepsilon \gtrsim d/\log K$.

Beyond toy dimension these two ranges do not overlap:

$$\{\varepsilon \gtrsim d/\log K\} \cap \{\varepsilon \lesssim 10^{-2}\} = \varnothing \quad \text{for any realistic } d.$$

We state this as a heuristic barrier, not a theorem: both constraints come from Laplace/log-normal approximations, so this is a strong practical obstruction rather than a formal impossibility. No practical operating point exists for the naive method, and the two demands are coupled on top of being opposed. Sharpening φ_θ raises $\bar{\lambda}$ and reduces the blur equation 40, but it also raises $\|\nabla\varphi_\theta\|$, which by equation 38 makes the estimator exponentially worse. Every unit of blur removed is paid back in estimator variance.

This is what forced the simulation-free Schrödinger-bridge line to fork. One branch keeps the convolution closed-form by restricting to conjugate parametric potentials (Korotin et al., 2024), at the cost of expressivity. The other keeps a free-form potential but samples the kernel by MCMC (Mokrov et al., 2024), at the cost of the one-step property. The obstruction in both cases is the same estimator equation 8. Section 3.5 removes it.

H DERIVATIONS FOR SECTION 3.5

Logical status of the results. The estimator equation 9 is unbiased for every ε, every θ, and every x, requiring only $q_x > 0$ on the support of the integrand; the defensive mixture of Appendix H.4 guarantees this coverage unconditionally, so unbiasedness is never in question. Only the variance magnitude is conditional: the cancellation identity is algebraic and needs only a nondegenerate critical point y^*, while the bound equation 11 additionally uses (A6) and the remainder bound of Theorem 3.9, and the sample budget equation 42 uses the stated weight-moment conditions. Outside these, the estimator stays correct in expectation but its variance is no longer controlled by equation 11.

H.1 THE EXACT CANCELLATION

From equation 9, $\psi_0(x) = (4\pi\varepsilon)^{-d/2}\int e^{-F_x(y)/2\varepsilon}\,\mathrm{d}y$, so the integrand is $e^{-F_x(y)/2\varepsilon}$ up to a constant. Sampling $y \sim q_x = \mathcal{N}(m, 2\varepsilon S)$ with density $(4\pi\varepsilon)^{-d/2}\det(S)^{-1/2}\exp\big(-\frac{1}{4\varepsilon}(y-m)^\top S^{-1}(y-m)\big)$, the log-weight is

$$\log w(y) = -\frac{1}{2\varepsilon}F_x(y) + \frac{1}{4\varepsilon}(y-m)^\top S^{-1}(y-m) + \tfrac{1}{2}\log\det S, \tag{41}$$

the $(4\pi\varepsilon)^{-d/2}$ factors having cancelled. Expand F_x about the mode with $h = y - y^*$ and $\nabla F_x(y^*) = 0$: $F_x(y) = F_x(y^*) + \frac{1}{2}h^\top H h + R_x(h)$, where R_x is the third- and higher-order remainder. Insert the diagonal match $m = y^*$, $S = D^{-1}$ (so $S^{-1} = D$) into equation 41. The two quadratic terms combine as

$$-\frac{1}{4\varepsilon}h^\top H h + \frac{1}{4\varepsilon}h^\top D h = -\frac{1}{4\varepsilon}h^\top(H-D)h = -\tfrac{1}{2}\,z^\top B_x z,$$

using $h = \sqrt{2\varepsilon}\,D^{-1/2}z$ and $B_x = D^{-1/2}(H-D)D^{-1/2}$. Collecting the z-independent terms into $c(x) = -\frac{1}{2\varepsilon}F_x(y^*) + \frac{1}{2}\log\det S$ gives exactly. Two remarks. First, $\mathrm{tr}(B_x) = \mathrm{tr}(D^{-1}(H-D)) = \sum_i H_{ii}/D_{ii} - d = 0$, so the diagonal quadratic part is removed entirely and only the off-diagonal fluctuation $\frac{1}{2}z^\top B_x z$ (mean zero) survives. Second, for full covariance matching $S = H^{-1}$ the quadratic terms cancel to $-\frac{1}{4\varepsilon}h^\top(H-H)h = 0$, so B_x disappears and $\log w = c(x) - R_x(h)/2\varepsilon$ exactly. □

H.2 THE VARIANCE REVERSAL AND THE SAMPLE BUDGET

Variance bound. $c(x)$ is constant in z, so $s^2_{\mathrm{res}} = \mathrm{Var}\big(\frac{1}{2}z^\top B_x z + R_x(h)/2\varepsilon\big)$, and by the triangle inequality for the standard deviation, $s_{\mathrm{res}} \le \mathrm{sd}(\frac{1}{2}z^\top B_x z) + \mathrm{sd}(R_x/2\varepsilon)$. For $z \sim \mathcal{N}(0, I)$ and symmetric B_x, $\mathrm{Var}(z^\top B_x z) = 2\|B_x\|_F^2$, hence $\mathrm{Var}(\frac{1}{2}z^\top B_x z) = \frac{1}{2}\|B_x\|_F^2 \le \frac{1}{2}C_{\mathrm{off}}\varepsilon$ by (A6), giving

$\mathrm{sd}(\frac{1}{2}z^\top B_x z) \le \sqrt{\varepsilon}\,\sqrt{C_{\mathrm{off}}/2}$. For the remainder, $\mathrm{Var}(R_x/2\varepsilon) \le \mathbb{E}[R_x^2]/(4\varepsilon^2) \le 4\varepsilon^3 C_R/(4\varepsilon^2) = \varepsilon C_R$, so $\mathrm{sd}(R_x/2\varepsilon) \le \sqrt{\varepsilon}\,\sqrt{C_R}$. Adding the two,

$$s_{\mathrm{res}} \le \sqrt{\varepsilon}\big(\sqrt{C_{\mathrm{off}}/2} + \sqrt{C_R}\big), \qquad \text{i.e.} \qquad s_{\mathrm{res}}^2 \le \varepsilon\big(\sqrt{C_{\mathrm{off}}/2} + \sqrt{C_R}\big)^2 = O(\varepsilon),$$

which is equation 11. The reversal is thus driven by two independent facts: the off-diagonal residual is $O(\varepsilon)$ under (A6), and the Taylor remainder, being third-order in $h = O(\sqrt{\varepsilon})$, contributes variance $O(\varepsilon)$ after the $1/2\varepsilon$ scaling. Neither constant is used elsewhere; the paper uses only the sign of the ε-dependence. The familiar $O(dM_3^2\varepsilon)$ form is recovered when both constants scale as $C_{\mathrm{off}} + C_R = O(dM_3^2)$, which holds when the off-diagonal Hessian mismatch and the third derivatives are governed by the operator norm $M_3 = \|\nabla^3 F_x(y^*)\|$; it does not follow from $M_3 < \infty$ alone.

Sample budget. Writing the single-draw weight $w(y) = G_{2\varepsilon}(y - x)e^{-u_\theta(y)}/q_x(y)$, one has $\mathbb{E}_{q_x}[w] = \psi_0(x)$ and, with $p_{\varepsilon,x}$ = the normalized integrand, $\mathbb{E}_{q_x}[w^2]/\psi_0(x)^2 = 1 + \chi^2(p_{\varepsilon,x}\|q_x)$. Hence the relative variance of one draw is $\chi^2(p_{\varepsilon,x}\|q_x)$, and of the K-sample average $\chi^2(p_{\varepsilon,x}\|q_x)/K$. A Laplace expansion of the first two weight moments about y^*, valid under the uniform smoothness and second-moment tail conditions, yields $\chi^2(p_{\varepsilon,x}\|q_x) \le C_{\mathrm{IS}}\varepsilon$ for small ε. Requiring relative RMS error δ gives

$$K^\star_{\mathrm{tilted}}(\delta) = \max\left\{1, \left\lceil \frac{\chi^2(p_{\varepsilon,x}\|q_x)}{\delta^2} \right\rceil\right\} \le \max\left\{1, \left\lceil \frac{C_{\mathrm{IS}}\varepsilon}{\delta^2} \right\rceil\right\}, \tag{42}$$

so $K^\star_{\mathrm{tilted}}(\delta) \to 1$ as $\varepsilon \to 0$. So $K^\star_{\mathrm{tilted}}(\delta) \to 1$ as $\varepsilon \to 0$. This argument uses the actual second moment, not a log-normal surrogate. □

H.3 The role of diagonal preconditioning

Drop the curvature match and recentre only, $m = y^*$, $S = I$. Then equation 41's quadratic terms give $-\frac{1}{4\varepsilon}h^\top(H - I)h = -\frac{1}{4\varepsilon}h^\top\nabla^2\varphi_\theta(y^*)h$ with $h = \sqrt{2\varepsilon}\,z$, i.e. a second-order remainder $-\frac{1}{2}z^\top\nabla^2\varphi_\theta(y^*)z$. Its variance is

$$\mathrm{Var}\big(\tfrac{1}{2}z^\top\nabla^2\varphi_\theta z\big) = \tfrac{1}{2}\|\nabla^2\varphi_\theta(y^*)\|_F^2 \sim \tfrac{1}{2}\,d\,\bar{\lambda}^2,$$

which is ε-independent and grows with d: recentring alone recovers the d-problem it was meant to solve. Recentring buys the ε-reversal (it removes the first-order misplacement); diagonal scaling drops the surviving remainder from second to third order and buys the d-scaling. What a diagonal S cannot absorb is the off-diagonal Hessian structure, which re-enters as the $\frac{1}{2}\|B_x\|_F^2$ term of; (A6) is exactly the assumption that this term is $O(\varepsilon)$, and it is monitored empirically through the effective sample size. The essential change is what the constraint depends on: the controllable smoothness M_3 (via architecture, weight decay, spectral normalization) rather than the transport magnitude $\|\nabla\varphi\|$, which must be $O(\sqrt{d})$ for generation to occur. □Empirically (Table 2), centering alone already reaches $\mathrm{ESS}/K \approx 0.9$; diagonal scaling adds a smaller gain. We present it as a consistent improvement, not a strict necessity.

H.4 The defensive mixture

Let $q_x^{\mathrm{def}} = (1 - \alpha)q_x + \alpha\, r$ with $r = \mathcal{N}(x, 2\varepsilon I)$. Since $q_x^{\mathrm{def}} \ge (1 - \alpha)q_x$ pointwise,

$$\int \frac{p_{\varepsilon,x}^2}{q_x^{\mathrm{def}}} \le \frac{1}{1 - \alpha}\int \frac{p_{\varepsilon,x}^2}{q_x} = \frac{1 + \chi^2(p_{\varepsilon,x}\|q_x)}{1 - \alpha},$$

so subtracting 1 gives $\chi^2(p_{\varepsilon,x}\|q_x^{\mathrm{def}}) \le (\chi^2(p_{\varepsilon,x}\|q_x) + \alpha)/(1 - \alpha)$. Substituting the bound $\chi^2(p_{\varepsilon,x}\|q_x) \le C_{\mathrm{IS}}\varepsilon$ replaces $C_{\mathrm{IS}}\varepsilon$ by $(C_{\mathrm{IS}}\varepsilon + \alpha)/(1 - \alpha)$ in equation 42. Choosing $\alpha = O(\varepsilon)$ preserves the $O(\varepsilon)$ relative variance; a fixed α leaves an $O(\alpha)$ floor. Three consequences hold regardless of α: the mixture upper-bounds the importance weights (the reference component keeps q_x^{def} from collapsing where q_x is small), it reduces to the naive estimator when m_η is uninformative, and it costs one extra density evaluation and no extra network calls. □

I TRAINING DETAILS AND GUARANTEES

I.1 THE POTENTIAL STEP NEVER SEES THE GUIDANCE WEIGHT

Theorem 3.6(i) characterizes the boundary condition for the potential that actually appears in the loss. Suppose instead we fed a w-tilted potential φ^w into equation 13 while the data term still used $x_1 \sim \rho_1$. This would ask the network to make a w-indexed family of wrong bridges fit one dataset at once, which silently corrupts the training objective. The rule is therefore simple. The guidance weight exists only in two places: the generator's conditioning and loss, and sampling. The potential is always trained at $w = 0$. Conditioning dropout with $p_{\text{uncond}} \approx 0.1$ trains the conditional and unconditional potentials jointly against real data, which is the only regime in which the data-side objective is valid.

I.2 BIAS DIRECTION OF THE TILTED OBJECTIVE AND ITS CONTROL

By Jensen, $\mathbb{E}[\log \hat{\psi}_0^{\text{tilt}}] \leq \log \psi_0$, so equation 14 is an IWAE-type lower bound, monotone in K (Burda et al., 2016). θ is incentivized to inflate this gap; two mechanisms keep it bounded. First, the gap is $O(dM_3^2\varepsilon/K)$, so it is priced by tilt quality rather than free. Second, η trains on equation 15, which never references the value of $\mathcal{L}_\theta$, so, despite θ's incentive, the gap is bounded by the η-updates independently of $\mathcal{L}_\theta$. We report the measured gap at large evaluation K as a diagnostic.

I.3 THE GENERATOR AND SCALE LOSSES

The prox is characterized by the first-order condition $\nabla\varphi_\theta(y^*) + y^* - x_0 = 0$. Equation equation 15 regresses this residual directly, at the guided gradient $\nabla\varphi^w = (1+w)\nabla\varphi(\cdot, c) - w\nabla\varphi(\cdot, \varnothing)$, with $w \sim p(w)$ on $[0, w_{\max}]$. The generator, unlike the potential, is legitimately w-aware. It approximates the guided prox, which is a derived object, not the training target. Under (A2) its unique global minimizer is the true guided prox.

The generator step is second-order. Because the residual $r = \nabla\varphi_\theta^w(m_\eta) + m_\eta - x_0$ contains the potential gradient evaluated at m_η, its exact derivative in η carries the potential's Hessian,

$$\frac{\partial \mathcal{L}_\eta}{\partial \eta} = 2\, r^\top \big(I + \nabla^2\varphi_\theta^w(m_\eta)\big)\, \frac{\partial m_\eta}{\partial \eta}.$$

We compute this exact gradient; the potential parameters are frozen ($\bar{\theta}$) during the generator update, but differentiation through m_η is retained. The Hessian is never formed explicitly: back-propagating through the input gradient $\nabla\varphi_\theta^w(m_\eta)$ costs one Hessian–vector product (HVP) per step via double-backward. Under guidance the HVP differentiates both potential branches, $(1+w)\nabla^2\varphi_\theta(\cdot, c) - w\nabla^2\varphi_\theta(\cdot, \varnothing)$. Because double-backward is unsupported by fused attention kernels, generator steps run on the non-fused (math) attention path. The generator warm-starts at the identity by zero-initializing the residual head.

A stop-gradient variant, which detaches the target $x_0 - \nabla\varphi_\theta^w(m_\eta)$ and drops the Hessian factor, has the same stationary points ($r = 0$) under (A2) and costs one input gradient instead of one HVP; all runs reported here use the exact (Hessian-retaining) form.

The scale head is first-order. The diagonal log-scale head s_η is fitted by a reparameterized reverse-KL objective: with $y = m + \sqrt{2\varepsilon}\, e^{s/2} z$ at detached m and frozen θ, gradients flow through $y(s)$, and its optimum matches e^{-s_η} to $\operatorname{diag}(I + \nabla^2\varphi_\theta^w(m_\eta))$. This target is realized through the reparameterization alone, without forming or probing the Hessian, so the scale fit adds no Hessian-vector products.

Per-step cost. Each generator update therefore costs a forward pass, one input gradient, and one HVP through φ_θ (from the prox residual; the scale head adds none), on the non-fused attention path-more than a plain flow-matching update, and the reason generator training dominates step time. This cost is confined to training: sampling is a single forward pass of m_η with no autograd (§3.7).

I.4 Estimator health and the alternating scheme

One scalar certifies the scheme per batch: the normalized effective sample size

$$\mathrm{ESS}/K = \frac{(\sum_k w_k)^2}{K \sum_k w_k^2} \in (0, 1],$$

logged every step. It is healthy above 0.3. It is degrading in $[0.05, 0.3]$, where we slow the ε-anneal, raise α_{def}, and add η steps. It is broken below 0.05, where we freeze θ, retrain η, and resume. ESS is the empirical certificate of (A4). We keep three secondary monitors: the measured Jensen gap versus K; the prox residual $\|\nabla\varphi(m) + m - x_0\|$, whose multi-modal blow-up is the alarm for an (A2) violation; and antithetic pairs $\pm z_k$ throughout, for free variance reduction.

We state convergence informally, in the two-timescale sense. Suppose each round's η-update reaches prox accuracy δ and ESS stays bounded below. Then the θ-gradient is a biased stochastic gradient of the function-space-convex $\mathcal{L}$, with bias $O(\delta + dM_3^2\varepsilon/K)$. Standard biased-SGD analysis then gives convergence to a neighborhood of the optimum, whose radius scales with the bias. A full rate theorem for the coupled nonconvex parameterization remains open. The semi-dual identity frames the scheme as a function-space half-Sinkhorn, which is the natural route for future analysis. In practice we take one to two η steps per θ step, doubled whenever ESS degrades.

J Classifier-free guidance: proof and delimitation

Proof of Proposition 3.10. Write $\psi_1^w = (\psi_1^c)^{1+w}(\psi_1^\varnothing)^{-w} = e^{-\varphi^w/2\varepsilon}$. This is positive whenever the potentials are finite, and whenever φ^w satisfies (A1)–(A2) it is an admissible data-side potential. Every result of §3.1–§3.5 was proved for an arbitrary admissible potential, so each applies verbatim with $\varphi_\theta \to \varphi^w$. This covers propagation equation 3, the Tweedie/prox sampler, the tilted estimator, and the variance reversal. The guided sampler is therefore the prox sampler of the Schrödinger bridge from ρ_0 onto $\rho_1^w = \hat{\psi}_1^w \psi_1^w$, carrying the same $O(\varepsilon)$ and network-fit approximation as the unguided sampler. □

What is not claimed. The guided marginal ρ_1^w is not exactly the classical CFG tilt $\propto \rho_1(x|c)^{1+w}\rho_1(x)^{-w}$. The tilt acts on the potential factor ψ_1, but the conjugate factor $\hat{\psi}_1^w$ is derived from ψ_1^w through the bridge, rather than tilted in lockstep with it. So ρ_1^w is a w-sharpened reweighting of the same character. It agrees with the classical tilt to first order in w and in the $\varepsilon \to 0$ limit.

This comparison should be stated in its survivable form, because diffusion CFG does not sample its nominal tilted density either. Its vector-field combination is a heuristic that carries additional discretization error (Ho & Salimans, 2022). So the honest claim is two-sided. What PT-Flow avoids is discretization error: the guided object is a bridge with no ODE and no step size, though the prox sampler carries the same $O(\varepsilon)$ and network-fit approximation as in the unguided case. What it shares with ODE-CFG is that the target is a controlled sharpening, not literally $\rho^{1+w}\rho^{-w}$.

Operational rules. The generator is conditioned on (c, w) and trained with $w \sim U[0, w_{\max}]$, so one trained model serves every guidance weight. Assumption (A2) tightens as w grows, and large w can break prox uniqueness, so the practical range is $w \in [0, 3]$, with the prox residual serving as the early-warning diagnostic.

K Algorithm

Throughout, $\overline{(\cdot)}$ denotes the batch mean, and $g_w(\cdot)$ is the guided log-potential gradient evaluated at its argument,

$$g_w(y) := (1+w)\,\nabla u_{\bar{\theta}}(y, c) - w\,\nabla u_{\bar{\theta}}(y, \varnothing), \qquad \text{so } 2\varepsilon\, g_w(y) = \nabla\varphi^w(y),$$

since $\varphi = 2\varepsilon u$. The defensive proposal is $q^{\mathrm{def}}(\cdot \mid x_0) = (1 - \alpha_{\mathrm{def}})\mathcal{N}(m, 2\varepsilon\,\mathrm{diag}\, e^s) + \alpha_{\mathrm{def}}\mathcal{N}(x_0, 2\varepsilon I)$, always conditioned on x_0 through $m = m_\eta(x_0, c, w)$ and $s = s_\eta(x_0, c, w)$. Antithetic sampling uses $K/2$ pairs $\pm z_k$ (so K is even).

All computations run in the log domain via `logsumexp`, with EMA on both networks. The prox residual is written in φ-units ($2\varepsilon\,\nabla u$), so its scale is transport displacement and is invariant under

Algorithm 1 PT-Flow training (alternating two-timescale scheme)

Networks: $u_\theta(x, c) \to \mathbb{R}$ (log-potential; $\varphi = 2\varepsilon u$); $m_\eta(x_0, c, w) \to \mathbb{R}^d$ (prox guess = generator; residual head, zero-init); $s_\eta(x_0, c, w) \to \mathbb{R}^d$ (diagonal log-scale).
Hyperparameters: K (even); n_η base inner-loop count; ESS thresholds $\tau_{\text{ok}} = 0.3$, $\tau_{\text{bad}} = 0.05$; defensive weight α_{def}; scale weight λ_{scale}; annealing range $\varepsilon_{\max} \to \varepsilon_{\min}$ (cosine).
for each step **do**
  Sample $(x_1, c) \sim$ data; drop $c \to \varnothing$ w.p. p_{uncond}; sample $x_0 \sim \mathcal{N}(0, I)$.
  (A) Generator steps (η; θ frozen). repeat n_η times:
    Sample $w \sim p(w)$ on $[0, w_{\max}]$; set $m = m_\eta(x_0, c, w)$, $s = s_\eta(x_0, c, w)$.
    Prox residual (in φ-units): $\mathcal{L}_\eta^{\text{loc}} = \left\| 2\varepsilon\, g_w(m) + m - x_0 \right\|^2$.
    fit (reparam. reverse-KL): $y = m + \sqrt{2\varepsilon} e^{s/2} z$; $\mathcal{L}_\eta^{\text{scale}} = -\varepsilon \sum_i s_i + \overline{\frac{1}{2}\|y - x_0\|^2 + \varphi_{\bar\theta}(y)}$
    Update η on $\mathcal{L}_\eta^{\text{loc}} + \lambda_{\text{scale}} \mathcal{L}_\eta^{\text{scale}}$.
  (B) Potential step (θ; η frozen; $w \equiv 0$). Set $m_0 = m_{\bar\eta}(x_0, c, 0)$, $s_0 = s_{\bar\eta}(x_0, c, 0)$.
    Draw $K/2$ antithetic pairs $\pm z_k$; form $y_k = m_0 + \sqrt{2\varepsilon}\, e^{s_0/2} \odot z_k$, replacing each by a draw from $\mathcal{N}(x_0, 2\varepsilon I)$ w.p. α_{def}.
    $\log w_k = \log G_{2\varepsilon}(y_k - x_0) - u_\theta(y_k, c) - \log q^{\text{def}}(y_k \mid x_0)$.
    $\log \hat\psi_0 = \text{logsumexp}_k \log w_k - \log K$.
    $\mathcal{L}_\theta = \overline{\log \hat\psi_0 + u_\theta(x_1, c)}$; update θ.
  (C) Monitor and anneal. Compute $\text{ESS}/K = (\sum_k w_k)^2 / (K \sum_k w_k^2)$ from the step-(B) weights.
    if $\text{ESS}/K < \tau_{\text{bad}}$: freeze θ, run extra η steps until recovered, then resume.
    else if $\text{ESS}/K < \tau_{\text{ok}}$: raise α_{def}, double n_η, hold ε.
    else: anneal ε toward $\varepsilon_{\min}$ (cosine from $\varepsilon_{\max}$); reset n_η to base.
end for

the anneal. The anneal floor is a tuned hyperparameter, $\varepsilon_{\min} = 0.01$ (cosine from $\varepsilon_{\max} = 0.2$); Table 2 shows the estimator is healthiest near this value. The dimension-dependent floor $c_\varepsilon/(dM_3^2)$ suggested by Theorem 3.9 lies far below $\varepsilon_{\min} = 0.01$ at ImageNet scale, so the runs operate outside the strict Laplace window ($\varepsilon d \approx 41$ at d=4096); we use the small-ε analysis to motivate the proposal construction and verify finite-ε behavior empirically (§4.2). Weak convexity (A2) is monitored by Hutchinson probes of directional curvature at proposal points, softly enforced by a hinge penalty on curvature violations, with an architectural fallback ($\varphi_\theta = \frac{\tau}{2}\|x\|^2 + h_\theta$, spectrally-normalized ∇h, which ensures $\nabla^2 \varphi_\theta \succeq (\tau - \|\nabla^2 h\|) I$) if monitoring and penalty fail.

Operating regime. The small-viscosity results of §3.1 and §3.5—the Hopf–Lax expansion equation 4 and the $O(\varepsilon)$ variance reversal (Theorem 3.9)—are asymptotic in the joint limit $\varepsilon \to 0$, $d \to \infty$ with $\varepsilon d \to 0$. At ImageNet latent scale $d = 4096$, our operating point $\varepsilon = 0.01$ gives $\varepsilon d \approx 41$, so the experiments lie outside this asymptotic window. We therefore use the asymptotic analysis to motivate the proposal construction (prox centering and diagonal curvature scaling), and verify its finite-ε behavior empirically: Table 2 shows the estimator remains healthy ($\text{ESS}/K \approx 0.99$) at $\varepsilon = 0.01$, and the sample–marginal gap of Proposition 3.4 is $O(\sqrt{\varepsilon d})$ at this operating point. The results that hold at every ε—the exact potential loss (Theorem 3.6) and the exact normalization (Theorem 3.5)—apply to the ImageNet model directly; the $\varepsilon \to 0$ results are limits it approaches, not certificates at $\varepsilon = 0.01$.

Algorithm 2 PT-Flow sampling (strict 1-NFE)

$x_0 \sim \mathcal{N}(0, I)$; return $\hat x_1 = m_\eta(x_0, c, w)$.
Decode $\hat x_1$ through the VAE if operating in latent space.

Table 5: Configurations of PT-Flow on ImageNet 256×256.

Configuration	Value
Model	
Trainable parameters	462.91M
Image representation	$32\times32\times4$ latent (decoded to 256×256)
Image categories	1,000 ImageNet classes
Transformer layers	24
Hidden width	1,024
Attention heads	16
Patch size	2×2
Conditioning width	1,024
Training	
Precision	BF16
Batch size	128
Gradient accumulation	8
Iterations	160k
Inference FLOPs	
1 NFE	785.71 GFLOPs

L Implementation Details

L.1 Datasets.

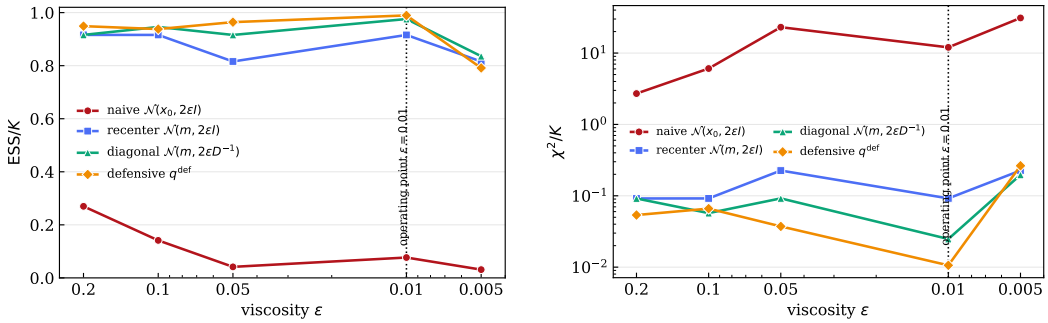
We train on class-conditional ImageNet-1k (Deng et al., 2009) at 256×256 resolution, which contains 1,281,167 training images across 1,000 classes. Each image is center-cropped and resized to 256×256 and encoded once by the pretrained SD-VAE tokenizer (Rombach et al., 2022) into a $32\times32\times4$ latent. Following common practice, we precompute and cache the latent *moments* (mean and standard deviation) for every image and its horizontal flip; during training we draw a latent by sampling from the cached moments, which provides horizontal-flip augmentation without rerunning the encoder. The generator is trained in this latent space, and samples are decoded back to pixels by the VAE decoder. We report FID-50K (Heusel et al., 2017), computed between 50,000 generated samples and the ImageNet training-set reference statistics.

L.2 Hyperparameters

Table 1 studies the five hyperparameters listed below. For each one, the bracket points to the method section where it appears, and we give a plain description plus the default we use in all main runs.

- **Viscosity floor** $\varepsilon_{\min}$ (§3.5). The lowest value ε drops to during annealing. It controls how sharp the bridge gets, and therefore how closely the proposal has to fit. Default: 0.01.
- **Training draws** K (§3.7). How many proposal samples we use to estimate the potential at each step. A larger K gives a lower-variance estimate but costs more per step. Default: 8.
- **Defensive weight** α_{def} (§3.5). How much weight we put on the backup proposal $\mathcal{N}(x, 2\varepsilon I)$. This backup keeps the importance weights from blowing up early in training. Default: 0.1.
- **Guidance scale** w (§3.7). The strength of classifier-free guidance used when sampling. Default: 1.25.
- ε **schedule** (§3.7). How ε is annealed from $\varepsilon_{\max}$=0.2 down to $\varepsilon_{\min}$. Default: cosine.

.



(a) Effective sample size ESS/K (higher is better). (b) Relative variance χ^2/K (lower is better).

Figure 2: **Proposal ablation on the frozen checkpoint (PT-Flow-L/2, generator center).** As ε cools, the naive proposal collapses (ESS/K from 0.27 to below 0.05; χ^2/K up to $\sim$31) while prox-centering and diagonal curvature scaling keep the estimator healthy. Centering alone already reaches $\mathrm{ESS}/K \approx 0.9$; diagonal scaling adds a smaller gain. Both metrics are best near the operating point ε=0.01 and degrade below it. The two panels show the same information (χ^2/K is computed from ESS/K). N=512, K_{eval}=64, M=5 seeds; full generator/oracle numbers in Table 7.

L.3 Training Configurations and Extended Results

We report the configurations for our PT-Flow model on the ImageNet 256×256 dataset in Table 5. The generator follows a DiT-L/2 backbone operating in the 32×32×4 SD-VAE latent space, and its one-step output is decoded to a 256×256 image by the pretrained VAE decoder. All models are trained from scratch.

Table 6: Single-H100 1-NFE sampling throughput (images/s, batch-matched, latent generation including VAE decoding). Higher is better.

Method	#Params	Throughput (img/s) ↑
iMeanFlow (Geng et al., 2025b)	610M	130.64
SoFlow (Luo et al., 2026)	676M	154.35
WFlow (Han et al., 2026)	679M	145.78
PT-Flow (ours)	**462M**	**165.28**

L.4 Additional Qualitative Samples

Figure 3 shows additional 1-NFE samples from PT-Flow-L/2 on ImageNet 256×256, generated at guidance scale w=1.25. The classes are drawn at random and the samples are uncurated. Figure 4 shows uncurated 1-NFE samples grouped by class, eight per class, from PT-Flow-L/2 on ImageNet 256×256 at guidance scale w=1.25.

Table 7: **Full proposal ablation.** Complete version of Table 2 with both centers: the deployed generator m_η and the reference prox $y^* = \mathrm{prox}_{\varphi_\theta}(x_0)$, obtained by iterating the first-order condition on the frozen potential (an evaluation oracle, not a sampler). $\mathrm{ESS}/K \in (0, 1]$ of the importance estimator equation 9 for four proposals of increasing structure: naive $\mathcal{N}(x_0, 2\varepsilon I)$; centered $\mathcal{N}(m, 2\varepsilon I)$; centered-diagonal $\mathcal{N}(m, 2\varepsilon D^{-1})$ with $D = \mathrm{diag}(I + \nabla^2\varphi_\theta(y^*))$; and the defensive mixture q^{def}. The two centers agree closely at every ε-the empirical certificate of the tracking assumption (A4). Naive is center-independent by construction. Networks frozen; only the proposal is ablated. N=512, K=64, mean over M=5 seeds. Operating point ε=0.01.

ε	center m	naive	$\mathcal{N}(m, 2\varepsilon I)$	$\mathcal{N}(m, 2\varepsilon D^{-1})$	q^{def}
0.20	m_η	0.2698	0.9155	0.9164	0.9490
	y^*	0.2698	0.9262	0.9108	0.9359
0.10	m_η	0.1414	0.9162	0.9459	0.9380
	y^*	0.1414	0.9344	0.9336	0.9387
0.05	m_η	0.0416	0.8158	0.9158	0.9641
	y^*	0.0416	0.8386	0.9217	0.9646
0.01	m_η	0.0768	0.9158	0.9758	**0.9895**
	y^*	0.0768	0.8955	0.9881	**0.9949**
0.005	m_η	0.0312	0.8162	0.8359	0.7913
	y^*	0.0312	0.8273	0.8351	0.7798



Figure 3: **Additional 1-NFE samples from PT-Flow-L/2 on ImageNet** 256×256 (w=1.25). Classes are randomly selected and samples are uncurated.



class 094: hummingbird



class 985: daisy



class 970: alp



class 817: sports car



class 628: liner, ocean liner



class 562: fountain



class 468: cab, taxi



class 130: flamingo

Figure 4: *Uncurated* 1-NFE class-conditional samples of PT-Flow-L/2 on ImageNet 256×256.