Implicit Dynamical Flow Fusion (IDFF) for Generative Modeling

Anonymous authors
Paper under double-blind review

Abstract

Conditional Flow Matching (CFM) generates high-quality samples by learning a deterministic transport from noise to data, but typically requires over a hundred network function evaluations (NFEs) per sample, especially in time-series settings. We introduce *Implicit Dynamical Flow Fusion* (IDFF), which augments the CFM vector field with learnable momentum terms derived from higher-order derivatives of the log-density. IDFF comes with two clearly separated theoretical guarantees. At first order, with our default Langevin-enhanced schedule, IDFF preserves the CFM marginal density *exactly* in continuous time. At higher orders, a single Girsanov-to-Pinsker argument bounds the endpoint deviation by a closed-form expression that depends only on the weighted score-matching loss our training objective already minimizes; consequently, the endpoint deviation vanishes as the number of NFEs grows. The practical enabler behind both regimes is a re-parameterization identity: every higher-order marginal derivative can be obtained from the learned first-order score by automatic differentiation, so no additional networks are trained. Empirically, IDFF reduces NFEs by an order of magnitude with no loss in sample quality. On CIFAR-10 it achieves an FID of 2.78 at 10 NFEs, outperforming existing CFM variants and matching methods that need over a hundred evaluations. For time-series modelling, IDFF performs strongly on molecular-dynamics simulation and sea-surface-temperature forecasting at a fraction of the compute. Overall, momentum-augmented flows offer a principled and efficient route to generative modelling across both static and dynamic domains.

1 Introduction

Diffusion models have emerged as powerful generative tools, iteratively transforming noise into structured data with state-of-the-art results in image generation (Song et al., 2020b), text production (Liu et al., 2024; Kim et al., 2019), and other domains (Cachay et al., 2023; Myers et al., 2022). However, a significant practical cost is the number of function evaluations (NFEs) per sample (Ho et al., 2020), an issue that persists despite recent advances such as DPM-solvers (Lu et al., 2022b;c) and Denoising Diffusion Implicit Models (Song et al., 2020a).

Conditional Flow Matching (CFM) (Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2025) offers an alternative to diffusion models by learning deterministic vector fields that transport probability distributions from noise to data. To minimize trajectory lengths, OT-CFMs (Tong et al., 2023) incorporate optimal transport couplings. However, sampling from CFMs still requires over 100 NFEs for high-quality generation (Dao et al., 2023), which is particularly problematic for time-series applications where costs scale with sequence length.

Higher-order methods improve sampling efficiency by incorporating geometric information about the probability landscape. Lu et al. (Lu et al., 2022a) demonstrate that second-order corrections significantly reduce discretization errors and accelerate convergence. A key challenge in all such methods is the gap between *conditional* quantities available during training and *marginal* quantities needed for correct sampling dynamics.

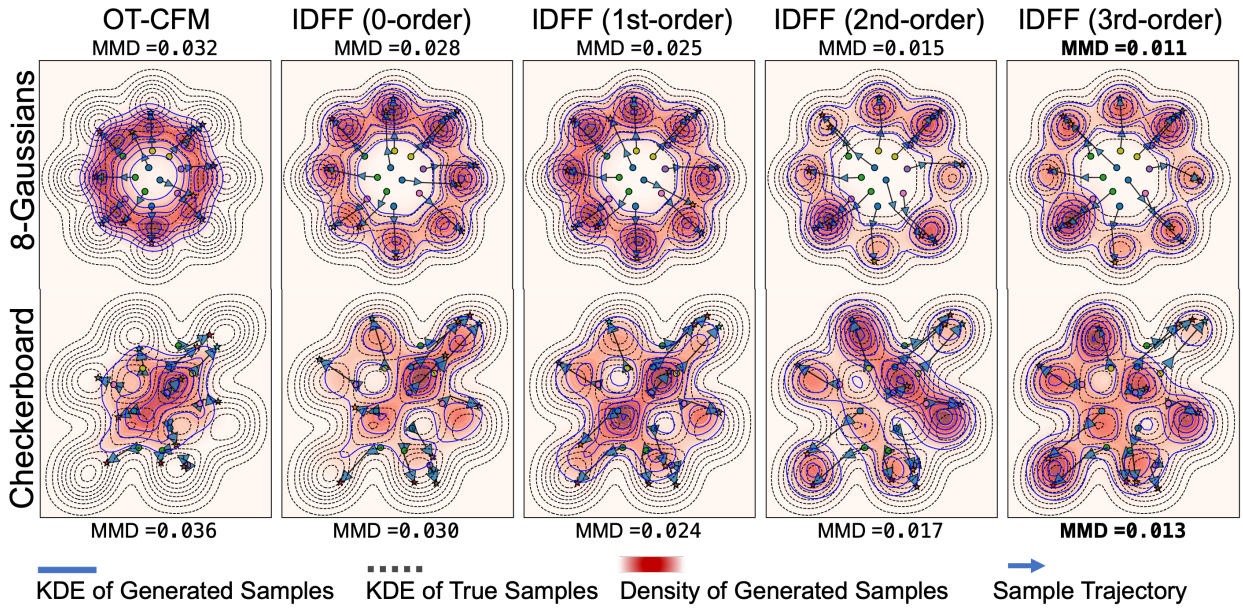

Figure 1: Comparison of trajectory sampling between OT-CFM and IDFF with different orders at NFE=2: The figure displays 4096 final samples generated by each model, with KDE contours shown in blue and ground truth samples represented by black contours. Twelve individual trajectories are overlaid to illustrate the sampling paths. Among the models, the 3rd-order IDFF produces the closest distribution to the true distribution based on the Maximum Mean Discrepancy (MMD) metric.

We introduce *Implicit Dynamical Flow Fusion* (IDFF), which augments CFM with momentum terms derived from higher-order derivatives of the log-density (Figure 1 illustrates the effect on a 2D-toy distribution). The order K controls how much higher-order information enters the dynamics; our theory cleanly separates two regimes.

First-order regime ($K = 1$): exact preservation. Our default schedule simply *adds* a Langevin term $\gamma_t^1 \nabla \log p_t$ to the CFM drift, together with the diffusion compensation $\sigma_t^2 = 2\gamma_t^1$ (no slowdown of $\mathbf{v}_t$). With this schedule, the augmented SDE preserves the CFM marginal p_t *exactly* in continuous time on the original clock; the proof is one line of Fokker–Planck cancellation (Theorem 14). At the SDE level this is the same Langevin-augmented probability-flow construction that has appeared in the score-Langevin SDEs of Song et al. (2020b) and the stochastic-interpolant SDEs of Albergo et al. (2025); the role of the $K = 1$ result here is therefore not the SDE itself, but to serve as the rigorous baseline on top of which the closed-form error budget for the $K \geq 2$ regime and for learned scores is built (Theorem 17). If a user prefers the convex-constrained variant $\gamma_t^0 + \gamma_t^1 = 1$ (which slows $\mathbf{v}_t$), the resulting endpoint deviation is bounded in closed form by a single Girsanov→Pinsker step and scales as $O(\sqrt{\bar{\gamma}})$ in the schedule amplitude (Proposition 15).

Higher-order regime ($K \geq 2$): controlled deviation. At higher orders, exact preservation is too much to ask: we instead prove a quantitative endpoint bound $\mathrm{TV}(\tilde{p}_1, p_1) \leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$, where $\mathcal{E}_{\text{score}} = \frac{1}{2}\int_0^1 \gamma_t^1 \, \mathbb{E}\|\hat{s}_t - \hat{s}_t^*\|^2 dt$ is *exactly* the weighted L^2 score-matching loss our training objective already minimizes. Under the recommended schedule $\gamma_t^k = O(\Delta t^{k-1})$, this gives $\mathrm{TV}(\tilde{p}_1, p_1) \to 0$ as $\mathrm{NFE} \to \infty$ for any converged score network (Theorem 17, Corollary 18).

The practical enabler. Training only has access to *conditional* log-density derivatives, but the dynamics need *marginal* ones; at $k = 1$, Fisher's identity bridges the gap, but no such identity holds for $k \geq 2$. We show (Theorem 2) that the optimal first-order score $\hat{s}_t^*$ already encodes all higher orders: $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$.

So a single trained score network suffices, and higher-order momentum is computed at sampling time by differentiating it automatically.

The contributions of our work are as follows, stated with an honest scoping of their novelty:

1. **Re-parameterization identity (Theorem 2).** We make explicit that once Fisher's identity grants $\hat{s}_t^* = \nabla \log p_t$ at the optimum of standard score matching, every higher-order marginal derivative is the chain-rule re-parameterization $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$. The logical content of this identity is elementary (chain rule); its *practical* content is that a single trained score head supplies all higher-order momentum at sampling time via autodiff, with no auxiliary higher-order networks. We additionally quantify the propagation of score-estimation error into higher orders.
2. **Unified two-regime error budget for momentum-augmented CFM.** The $K = 1$ Langevin-enhanced SDE that serves as our exact-preserving baseline is itself a known marginal-preserving construction (Song et al., 2020b; Albergo et al., 2025; Corollary 5). What is new is the unified Girsanov→Pinsker certificate built on top of it: a single closed-form path-KL bound covers (i) the exact $K = 1$ default (TV $= 0$ in continuous time, Theorem 14), (ii) the convex-constrained $K = 1$ variant (TV $= O(\sqrt{\bar{\gamma}})$, Proposition 15), and (iii) the higher-order regime $K \geq 2$ with learned score (TV $\leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$, Theorem 17), with the same bound holding uniformly in t (Corollary 19). We do *not* claim exact endpoint preservation at $K \geq 2$ for finite NFE.
3. **Empirical gains across domains, with a Langevin-vs-stochasticity ablation.** IDFF generates high-quality images at 5–10 NFEs on CIFAR-10, CelebA, ImageNet-64, CelebA-HQ, LSUN-Church, and LSUN-Bedroom, where standard CFM variants need over 100 NFEs. For time-series modelling (3D attractors, molecular dynamics, sea-surface-temperature forecasting), IDFF reduces sampling cost by an order of magnitude while improving quality. To pin down what is actually driving these gains, we provide a dedicated ablation (Table 4) that isolates the Langevin score-momentum term from generic SDE stochasticity and from the extra capacity of the joint score head; only the Langevin term, used with its matched diffusion $\sigma_t^2 = 2\gamma_t^1$, recovers IDFF's headline FID at 10 NFE.

2 Background and Preliminaries

Generative modeling aims to learn the underlying structure of complex high-dimensional data distributions from finite samples. Given empirical samples from a dataset $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^M$ where each sample $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is drawn from an unknown data distribution $p_{\text{data}}(\mathbf{x})$, the goal is to learn a model that can generate new samples that match p_{data}. For time-series data, $\mathcal{D} = \{\mathbf{x}_{1:N}^{(i)}\}_{i=1}^M$ where $\mathbf{x}_n^{(i)} \in \mathbb{R}^d$ for $n \in \{1, \ldots, N\}$, and the goal is to model either $p_{\text{data}}(\mathbf{x}_{1:N})$ or $p_{\text{data}}(\mathbf{x}_n | \mathbf{x}_{1:n-1})$.

2.1 Diffusion Models

Score-based generative models take a stochastic approach by learning to reverse a diffusion process that gradually destroys data structure. The forward diffusion follows the stochastic differential equation (SDE):

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + g(t)d\mathbf{w}_t, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^d$, $\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ is the drift, $g(t) \in \mathbb{R}_+$ is the diffusion coefficient, and $\mathbf{w}_t$ is Brownian motion. This transforms data $p_0 = p_{\text{data}}$ to prior $p_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse-time SDE is:

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

where a neural network is tasked to learn the score function

$$\nabla_{\mathbf{x}} \log p_t(x_t).$$

The probability density under the model evolves via the Fokker-Planck equation $\frac{\partial p_t}{\partial t} = -\nabla \cdot (\mathbf{f} p_t) + \frac{g(t)^2}{2} \Delta p_t$.

(Lu et al., 2022a) extend score matching beyond first-order gradients of the log density to leverage higher-order geometric information about the probability landscape and enable the model to sample from more complex probability distribution. The work leverages momentum variables $\hat{\boldsymbol{\epsilon}}_t^{(k)}(\mathbf{x}) = \nabla_{\mathbf{x}}^k \log p_t(\mathbf{x})$ as the k-th order derivative of the log-density. Under Gaussian transitions $p_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$, the variables have closed forms: $\hat{\boldsymbol{\epsilon}}_{0t}^{(1)} = -\boldsymbol{\epsilon}_0/\sigma_t$, $\hat{\boldsymbol{\epsilon}}_{0t}^{(2)} = -\mathbf{I}/\sigma_t^2$, and $\hat{\boldsymbol{\epsilon}}_{0t}^{(k)} = 0$ for $k \geq 3$, where $\boldsymbol{\epsilon}_0 = (\mathbf{x}_t - \alpha_t \mathbf{x}_0)/\sigma_t$. The work learns separate neural networks $\hat{\boldsymbol{\epsilon}}^{(k)}(\mathbf{x}_t, t; \theta_k)$ to match these closed forms over the trajectories of the SDE:

$$\mathcal{L}_k = \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}_0} \left\| \hat{\boldsymbol{\epsilon}}_{0t}^{(k)}(\mathbf{x}_t|\mathbf{x}_0) - \hat{\boldsymbol{\epsilon}}^{(k)}(\mathbf{x}_t, t; \theta_k) \right\|^2. \tag{3}$$

The resulting higher-order reverse-time SDE becomes:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \hat{\boldsymbol{\epsilon}}_t^{(1)}(\mathbf{x}_t) - \sum_{k=2}^{K} \beta_k(t) \hat{\boldsymbol{\epsilon}}_t^{(k)}(\mathbf{x}_t) \right] dt + g(t) d\bar{\mathbf{w}}_t, \tag{4}$$

where $\beta_k(t)$ are time-dependent coefficients controlling the contribution of higher-order terms.

2.2 Conditional Flow Matching (CFM)

Conditional Flow Matching learns transport maps between probability distributions. Given a time-dependent vector field $\mathbf{v}_t : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ we can transport samples from a base distribution $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a target distribution $p_1 = p_{\text{data}}$ via the $\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_t(\mathbf{x}_t)$ with $\mathbf{x}_0 \sim p_0$, this induces a probability path $p_t(\mathbf{x})$. The resulting continuity equation:

$$\frac{\partial p_t(\mathbf{x})}{\partial t} + \nabla \cdot (p_t(\mathbf{x}) \mathbf{v}_t(\mathbf{x})) = 0, \tag{5}$$

ensures that probability mass is conserved along the trajectories.

Learning objective and sampling. Instead of modeling the marginal flow directly, CFM's decompose the complex marginal flow into simpler conditional flows. Given pairs $\mathbf{z} = (\mathbf{x}_0, \mathbf{x}_1)$ where $\mathbf{x}_0$ is typically a sample from a normal distribution and $\mathbf{x}_1$ are samples from $\mathcal{D}$, we can define Gaussian conditional paths $p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_t; \mu_t, \sigma_t^2 \mathbf{I})$ where $\mu_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ and $\sigma_t = \sigma_0 \sqrt{t(1-t)}$. Then, the conditional vector field becomes:

$$\mathbf{v}_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0 + \frac{\dot{\sigma}_t}{\sigma_t}(\mathbf{x}_t - \mu_t), \tag{6}$$

simplifying to $\mathbf{v}_t = \mathbf{x}_1 - \mathbf{x}_0$ when $\sigma_0 = 0$. CFMs train a neural network $\hat{\mathbf{v}}_t(\mathbf{x}; \theta)$ with parameters θ to approximate the conditional field via:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1], (\mathbf{x}_0, \mathbf{x}_1) \sim q, \mathbf{x}_t} \left\| \hat{\mathbf{v}}_t(\mathbf{x}_t; \theta) - \mathbf{v}_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) \right\|^2, \tag{7}$$

$\hat{\mathbf{v}}_t$ takes as input the current position $\mathbf{x}_t \sim p_t(\cdot|\mathbf{x}_0, \mathbf{x}_1)$ and time t and outputs a velocity vector in $\mathbb{R}^d$. The coupling $q(\mathbf{x}_0, \mathbf{x}_1)$ is a joint distribution over pairs of noise and data samples, which can be independent $q(\mathbf{x}_0, \mathbf{x}_1) = p_0(\mathbf{x}_0) p_{\text{data}}(\mathbf{x}_1)$ or use optimal transport couplings that minimize the expected transport cost $\mathbb{E}_q[\|\mathbf{x}_1 - \mathbf{x}_0\|^2]$. After training, samples are generated by integrating $\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \hat{\mathbf{v}}_t(\mathbf{x}_t; \theta)\Delta t$ from $\mathbf{x}_0 \sim p_0$ with step size $\Delta t = 1/\text{NFE}$.

The validity of the conditional vector field (CFM) rests on the following fundamental result:

Theorem 1 (CFM Marginal Consistency (Lipman et al., 2022))**.** *If conditional paths $p_t(\mathbf{x}|\mathbf{z})$ with fields $\mathbf{v}_t(\mathbf{x}|\mathbf{z})$ satisfy the continuity equation, then the marginal $p_t(\mathbf{x}) = \int p_t(\mathbf{x}|\mathbf{z}) q(\mathbf{z}) d\mathbf{z}$ with $\mathbf{v}_t(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\mathbf{v}_t(\mathbf{x}|\mathbf{z})]$ also satisfies the continuity equation.*

By Theorem 1, learning to match $\mathbf{v}_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1)$ is sufficient to recover the marginal flow and results in being able to sample from p_{data}.

2.3 Challenges in Incorporating Higher-Order Terms into CFM

The background above sets the stage for the central question of this paper: *can the acceleration benefits of higher-order, momentum-based dynamics be transferred from score-based diffusion to flow matching without breaking marginal consistency?* Since the incorporation of momentum variables to explicitly capture higher-order derivatives of the log-density has shown promise in diffusion models, we anticipate that their incorporation in CFM could similarly yield value by combining the efficiency of learning deterministic transport maps with the acceleration that higher-order information confers when navigating the probability manifold.

However, the direct addition of momentum terms to the vector field for CFMs runs into two technical hurdles that motivate every design choice in IDFF.

Challenge 1: continuity-equation violation. Consider a naive augmentation $\tilde{\mathbf{v}}_t = \mathbf{v}_t + \sum_{k=1}^{K} \gamma_t^k \hat{\boldsymbol{\epsilon}}^{(k)}$ where $\hat{\boldsymbol{\epsilon}}^{(k)} = \nabla_{\mathbf{x}}^k \log p_t$. Substituting into the continuity equation yields

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t \tilde{\mathbf{v}}_t) \;=\; \frac{\partial p_t}{\partial t} + \nabla \cdot (p_t \mathbf{v}_t) + \sum_{k=1}^{K} \gamma_t^k \nabla \cdot (p_t \nabla_{\mathbf{x}}^k \log p_t). \tag{8}$$

Since the original CFM satisfies $\partial_t p_t + \nabla \cdot (p_t \mathbf{v}_t) = 0$, the augmented field is consistent with the same marginal flow only if $\sum_{k=1}^{K} \gamma_t^k \nabla \cdot (p_t \nabla_{\mathbf{x}}^k \log p_t) \equiv 0$. Already at first order, $\nabla \cdot (p_t \nabla_{\mathbf{x}} \log p_t) = \Delta p_t \neq 0$ in general. The naive augmentation therefore violates the hypotheses of Theorem 1, and the augmented flow no longer transports p_0 to $p_1 = p_{\text{data}}$ exactly.

Challenge 2: conditional-marginal gap. We can only train on conditional targets $\nabla^k \log p_t(\mathbf{x}_t|\mathbf{x}_1)$, but sampling requires marginal quantities $\nabla^k \log p_t(\mathbf{x}_t)$. For $k = 1$, Fisher's identity ensures these coincide at optimum. For $k \geq 2$, prior work treats the conditional–marginal gap as an approximation. Our key observation is that, at the optimum, the marginal k-th derivative equals the $(k-1)$-th derivative of the score: $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$. For a learned $\hat{s}_t \approx \hat{s}_t^*$, this holds approximately, enabling higher-order dynamics without separate training while inheriting approximation quality from score estimation.

Roadmap for Section 3. These two challenges shape the IDFF construction in Section 3. (i) A compensating diffusion $\sigma_t^2 = 2\gamma_t^1$ *exactly* cancels the first-order term in residual equation 8, yielding the $K = 1$ exact-preservation theorem; for $K \geq 2$, a coefficient-scaling rule $\gamma_t^k = O(\Delta t^{k-1})$ bounds the structural residual. (ii) Theorem 2 closes the conditional–marginal gap by computing higher-order marginal derivatives from $\hat{s}_t$ via automatic differentiation, removing the need to train auxiliary networks. (iii) The sampler in Section 3.4 integrates the resulting effective drift with a Euler–Maruyama scheme whose endpoint TV deviation is controlled by the closed-form Girsanov→Pinsker bound stated above.

3 Implicit Dynamical Flow Fusion Model (IDFF)

Scope of theoretical guarantees. Before introducing the method, we summarise which guarantee applies at which order. The two regimes are qualitatively different and the reader should keep this distinction in mind throughout the section.

- $K = 1$**, default schedule (**$\gamma_t^0 \equiv 1$**,** $\sigma_t^2 = 2\gamma_t^1$**):** the augmented density equals the CFM marginal *exactly* in continuous time on the original clock, $\tilde{p}_t = p_t$ for all $t \in [0, 1]$ (Theorem 14).
- $K = 1$**, convex variant (**$\gamma_t^0 + \gamma_t^1 = 1$**):** the velocity is also slowed by γ_t^1, so $\tilde{p}_1 \neq p_1$ in general; the endpoint TV deviation is bounded by $\frac{1}{2\sqrt{2}} \left(\int_0^1 \gamma_t^1 \mathbb{E}\|\mathbf{v}_t\|^2 dt\right)^{1/2}$ via a Girsanov→Pinsker argument (Proposition 15; explicit constants in Corollary 16).
- $K \geq 2$**:** bounded deviation $\mathrm{TV}(\tilde{p}_1, p_1) \leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$, where $\mathcal{E}_{\text{score}}$ is precisely the weighted L^2 score-matching loss our training objective minimizes (Theorem 17). The same bound in fact holds

uniformly in t, $\sup_{t\in[0,1]} \mathrm{TV}(\tilde{p}_t, p_t) \le \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$ (Corollary 19), so the entire augmented marginal trajectory stays in a closed-form TV ball of the CFM marginals. Endpoint preservation is asymptotic, not exact, at finite NFE.

The $K \ge 2$ extension is therefore best read as a controlled perturbation of the exact-preserving $K = 1$ baseline, with a single closed-form error budget driven by the training loss.

Relation to prior marginal-preserving SDEs. The $K = 1$ exact-preserving SDE underlying Theorem 14 is, on its own, a known construction: it is the same family of Langevin-augmented probability-flow SDEs that preserve a target marginal when the drift is enhanced by $\gamma_t^1 \nabla \log p_t$ and the diffusion is compensated by $\sigma_t^2 = 2\gamma_t^1$, going back to the score-Langevin SDE of Song et al. (2020b) and the stochastic-interpolant SDEs of Albergo et al. (2025) (we recover the latter as a special case in Corollary 5). Our claim at $K = 1$ is therefore not the SDE construction itself; rather, we use this known marginal-preserving SDE as the rigorous baseline on top of which the closed-form, training-loss-driven error budget for the $K \ge 2$ regime and for learned scores is built. The genuinely new contribution at the SDE level is the unified Girsanov→Pinsker error budget (Theorem 17, Corollary 19) and its use to certify both the convex-constrained $K = 1$ variant and the higher-order regime.

Proof technique. Both the convex-constrained $K = 1$ bound and the $K \ge 2$ bound use a single Girsanov→Pinsker argument on path space, with the exact-preserving $K = 1$ default SDE of Theorem 14 as the baseline. Because IDFF has non-degenerate diffusion ($\sigma_t^2 = 2\gamma_t^1 > 0$ on the interior), Girsanov's theorem applies directly to the path measure, and the only quantity that ends up on the right-hand side is the weighted L^2 norm of the drift mismatch — which is exactly the quantity our training loss minimises. This route avoids an L^1-of-divergence transport-Grönwall step that would have required additional regularity on $\nabla \log p_t \cdot \mathbf{m}_t^{(k)}$ not stated up front, and aligns with recent flow-matching convergence analyses (Chen et al., 2023; Benton et al., 2024).

Regularity assumption used throughout. We state the regularity hypotheses on a *truncated* time interval $[0, 1-\delta]$ rather than on the closed interval $[0, 1]$. This is essential, because under the CFM schedule $\sigma_t = \sigma_0\sqrt{t(1-t)}$ the conditional velocity $\mathbf{v}_t = (\mathbf{x}_1 - \mathbf{x}_t)/(1-t)$ and the marginal score $\nabla \log p_t$ both diverge as $t \to 1$, so no uniform-in-$\mathbf{x}$ bound can hold up to the endpoint. The truncation matches what every practical sampler already does: we integrate to a final time $t_{\text{stop}} = 1 - \delta$ (in all reported experiments, $\delta = \Delta t/2$) and read off $\mathbf{x}_1 \approx \hat{\mathbf{x}}_1(\mathbf{x}_{t_{\text{stop}}}, t_{\text{stop}})$ via the denoiser head. The deviation bounds in this section depend on δ only through schedule integrals $\int_0^{1-\delta} \gamma_t^1\, dt$ and remain finite as $\delta \to 0$ provided γ_t^1 is integrable, which holds for every schedule we use (see Corollary 16).

Assumption 1 (Regularity on the truncated interval $[0, 1-\delta]$)**.** *Fix $\delta \in (0, 1/2)$.* ***(R1) Score regularity on*** *$[0, 1-\delta]$: both the optimal score $\hat{s}_t^* = \nabla \log p_t$ and the learned score $\hat{s}_t$ used by the sampler are C^K in $\mathbf{x}$ on $[0, 1-\delta] \times \mathbb{R}^d$, with bounded derivatives $\sup_{t\in[0,1-\delta]} \|\nabla_{\mathbf{x}}^j \hat{s}_t^*\|_{L^\infty} \le M_j(\delta)$ and $\sup_{t\in[0,1-\delta]} \|\nabla_{\mathbf{x}}^j \hat{s}_t\|_{L^\infty} \le M_j(\delta)$ for $j = 0, \dots, K-1$. Here $M_j(\delta)$ may diverge as $\delta \to 0$; under the Gaussian conditional path the marginal score satisfies the classical heat-kernel bound $\|\nabla \log p_t\|_{L^\infty} \le M_0(\delta) = O(\sigma_t^{-1}) = O(1/\sqrt{\delta})$ near $t = 1$, and the learned score $\hat{s}_t$ inherits this scaling because it is trained against $-\boldsymbol{\epsilon}/\sigma_t$ on the same path. The contracted momentum $\mathbf{m}_t^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ used in sampling is then bounded by $M_{k-1}(\delta)$ uniformly on $[0, 1-\delta] \times \mathbb{R}^d$.* ***(R2) Velocity regularity on*** *$[0, 1-\delta]$: the base velocity $\mathbf{v}_t$ is Lipschitz on $[0, 1-\delta] \times \mathbb{R}^d$ with $\sup_{t\in[0,1-\delta]} \|\nabla \mathbf{v}_t\|_{L^\infty} \le L_v(\delta)$ and $\sup_{t\in[0,1-\delta]} \|\mathbf{v}_t\|_{L^\infty} \le M_v(\delta)$. For the CFM path $\mathbf{v}_t = (\mathbf{x}_1 - \mathbf{x}_t)/(1-t)$ on a compactly supported data law these constants grow like $L_v(\delta), M_v(\delta) = O(1/\delta)$, which together with the schedule decay $\gamma_t^1 = O(t(1-t))$ keeps the Novikov integrand $\sigma_t^{-1}\Delta\mathbf{b}_t$ uniformly bounded on $[0, 1-\delta]$ for every drift mismatch used in the Girsanov arguments below.* ***(R3) Coefficient decay matched to truncation:*** *the schedules satisfy $\gamma_t^k = O(|t(1-t)|^\alpha)$ with $\alpha \ge 1$ (for $k \ge 2$, and $\alpha = 1$ for the Beta(2,2) default $\gamma_t^1 = 4\bar{\gamma}t(1-t)$). Combined with the polynomial blow-up of $M_j(\delta)$ above, this guarantees that the Girsanov KL integrand $\gamma_t^1 \|\hat{s}_t - \hat{s}_t^*\|^2$ is uniformly integrable on $[0, 1-\delta]$ for every $\delta > 0$, and that the schedule integrals $\int_0^{1-\delta} \gamma_t^1\, dt$ admit a finite limit as $\delta \to 0$.* ***(R1′) Finite excess score-matching loss:*** *the learned*

score has finite excess weighted score-matching loss

$$\mathcal{L}^{\text{exc}}_{\text{score}} \;:=\; \int_0^1 \sigma_t^2\, \mathbb{E}_{p_t} \big\| \hat{s}_t - \hat{s}_t^* \big\|^2 \, dt \;<\; \infty,$$

equal to the score term of the training objective equation 21 net of its irreducible Bayes term. This bounds the weighted *score error directly and is independent of the magnitude constants $M_j(\delta)$; it is the quantity that controls the size of $\mathcal{E}_{\text{score}}$ in Theorem 17, while $M_j(\delta)$ controls only Novikov finiteness.*

Remark 1 (δ-independence of the score deviation). The quantity entering the path-KL bound is the weighted score error $\mathcal{E}_{\text{score}} = \frac{1}{2}\int_0^{1-\delta} \gamma_t^1\, \mathbb{E}_{p_t} \|\hat{s}_t - \hat{s}_t^*\|^2\, dt$, which is controlled through the training objective rather than through the magnitude constant $M_0(\delta)$. The constant $M_0(\delta)$ bounds the score *magnitude*—hence the Novikov integrand and the contracted momenta $\|\mathbf{m}_t^{(k)}\| \le M_{k-1}(\delta)$—but does not bound the score *error*. For the Beta(2,2) default the Langevin amplitude is proportional to the score-matching weight: with $\lambda_1(t)^2 = \sigma_t^2 = \sigma_0^2\, t(1-t)$ and $\gamma_t^1 = 4\bar{\gamma}\, t(1-t)$ we have $\gamma_t^1 = (4\bar{\gamma}/\sigma_0^2)\, \sigma_t^2$, so

$$\mathcal{E}_{\text{score}} \;=\; \frac{2\bar{\gamma}}{\sigma_0^2} \int_0^{1-\delta} \sigma_t^2\, \mathbb{E}_{p_t} \|\hat{s}_t - \hat{s}_t^*\|^2\, dt \;\le\; \frac{2\bar{\gamma}}{\sigma_0^2}\, \mathcal{L}^{\text{exc}}_{\text{score}} \;=\; O(\bar{\gamma}),$$

uniformly in δ, where $\mathcal{L}^{\text{exc}}_{\text{score}}$ is the finite excess score-matching loss of Assumption 1(R1′). The weight σ_t^2 cancels the σ_t^{-2} growth of the score error pointwise, so the integrand is integrable on all of $[0,1]$ and truncation removes only a tail of vanishing mass; in particular $\mathcal{E}_{\text{score}}$ admits a finite limit as $\delta \to 0$ at fixed $\bar{\gamma}$. At the $K = 1$ default schedule $\tilde{p}_t = p_t$ (Theorem 14), so the p_t-weighted error above coincides with the $\tilde{p}_t$-weighted quantity appearing in Theorem 17. The structural residual of Theorem 17 is controlled by the same weight cancellation, since $\gamma_t^k = \alpha_k \gamma_t^1 \Delta t^{k-1}$ carries the matching σ_t^2 decay; the resulting rate is established there. With $\delta = \Delta t/2$ ($\delta = 0.05$ at NFE = 10), reducing δ at fixed $\bar{\gamma}$ leaves sample quality unchanged.

Throughout, every theorem statement explicitly invokes Assumption 1 on the truncated interval where needed; the convergence statements (Corollary 18) take the limit in the order $\Delta t \to 0$ first (so $\delta = \Delta t/2 \to 0$) and then $\bar{\gamma} \to 0$ if desired.

IDFF augments the CFM vector field with a learnable momentum that accelerates convergence without breaking marginal consistency. The first-order augmented flow is

$$\tilde{\mathbf{v}}_t(\mathbf{x}_t) \;=\; \mathbf{v}_t(\mathbf{x}_t) \;+\; \gamma_t^1\, \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t), \tag{9}$$

where $\gamma_t^1 \ge 0$ is a Langevin-amplitude schedule that vanishes at the endpoints ($\gamma_0^1 = \gamma_1^1 = 0$). The momentum term $\boldsymbol{\xi}_t = \gamma_t^1\, \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ pushes particles along the log-density gradient, accelerating their movement toward high-probability regions. Paired with the diffusion compensation $\sigma_t^2 = 2\gamma_t^1$, this is exactly the Langevin enhancement of CFM that preserves the marginal p_t exactly (Theorem 14).

For full generality, we write the same drift in the form $\tilde{\mathbf{v}}_t = \gamma_t^0\, \mathbf{v}_t + \gamma_t^1\, \nabla \log p_t$. Our default sets $\gamma_t^0 \equiv 1$ and recovers the Langevin-enhanced flow above. The alternative convex-constrained schedule $\gamma_t^0 + \gamma_t^1 = 1$ is then a single notational change: it also slows $\mathbf{v}_t$, no longer preserves p_t exactly, and incurs the quantified TV cost of Proposition 15.

We can generalize the momentum to incorporate higher-order terms. For $k \ge 2$, the quantity $\nabla_x^k \log p_t$ is a k-th order tensor; to obtain a vector-valued contribution, we define the *contracted momentum terms*

$$\mathbf{m}_t^{(1)} = \nabla_x \log p_t(\mathbf{x}_t), \qquad \mathbf{m}_t^{(k)} = \big(\nabla_x^k \log p_t(\mathbf{x}_t)\big)[\mathbf{u}_t]^{k-1} \in \mathbb{R}^d \quad (k \ge 2), \tag{10}$$

where $\mathbf{u}_t = \mathbf{v}_t/\|\mathbf{v}_t\|$ is the normalized velocity direction and $[\mathbf{u}_t]^{k-1}$ denotes $(k-1)$-fold contraction along $\mathbf{u}_t$. Explicitly, $\mathbf{m}_t^{(2)}$ is the Hessian-vector product capturing directional curvature along the flow. The generalized momentum is then:

$$\boldsymbol{\xi}_t(\mathbf{x}_t) = \sum_{k=1}^{K} \gamma_t^k \mathbf{m}_t^{(k)}(\mathbf{x}_t). \tag{11}$$

The second-order term captures local curvature information to enable adaptive step sizes, while higher orders facilitate navigation through complex multi-modal landscapes.

3.1 Higher-Order Marginals from the Learned Score

A central challenge is that we train on *conditional* targets while the dynamics require *marginal* quantities $\nabla^k \log p_t(\mathbf{x}_t)$. For $k = 1$, Fisher's identity ensures these coincide at optimum. For $k \geq 2$, no such conditional-marginal closure exists. Theorem 2 below resolves this by a *re-parameterization identity*: once $\hat{s}_t^* = \nabla \log p_t$ is granted by Fisher's identity, every higher-order marginal derivative is obtained by differentiating $\hat{s}_t^*$ a further $(k-1)$ times,

$$\nabla^k \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*(\mathbf{x}_t). \tag{12}$$

We are deliberately explicit about the logical content of equation 12: it is the chain rule applied to the relation $\hat{s}_t^* = \nabla \log p_t$, and consequently does *not* close a new conditional-marginal gap beyond the $k = 1$ closure that Fisher's identity already provides. What it does provide is a *computational* statement: a single trained score network suffices to instantiate the higher-order momentum at sampling time via automatic differentiation, with no auxiliary networks. For a learned approximation $\hat{s}_t \approx \hat{s}_t^*$, equation 12 holds approximately and the error is inherited from score estimation (quantified in Theorem 17, Part 3).

Theorem 2 (Higher-Order Marginals from Score Derivatives)**.** *Under Assumption 1(R1), let $\hat{s}_t^*(\mathbf{x}_t) = \nabla \log p_t(\mathbf{x}_t)$ be the optimal first-order score (which equals the marginal score by Fisher's identity, Lemma 6). Then for every k with $1 \leq k \leq K$,*

$$\nabla^k \log p_t(\mathbf{x}_t) \;=\; \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*(\mathbf{x}_t). \tag{13}$$

In particular, the marginal Hessian equals the Jacobian of the optimal score, $\nabla^2 \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t)$.

The one-line proof is induction with base case Fisher's identity (Appendix A). For completeness we additionally include a longer Tweedie-style derivation of the $k = 2$ case in Appendix A (Theorems 11, 12); that derivation does not pre-suppose Fisher's identity and is logically equivalent to the chain-rule statement here. The Tweedie route is not used computationally inside our sampler — the practical pipeline only needs equation equation 12 composed with autodiff — so a reader satisfied with the chain-rule view can skip Appendix A Sections A.3–A.6. We retain the longer derivation only because Theorem 11 (posterior covariance from score Jacobian) is independently useful for the practical-error analysis in Appendix A.10.

3.2 Central Theorem: K-th Order IDFF Continuity and Sampling

Theorem 3 (K-th Order IDFF Continuity and Sampling)**.** *Let $\mathbf{v}_t(\mathbf{x}_t)$ be a vector field that generates marginal density $p_t(\mathbf{x}_t)$ via the continuity equation. Define the K-th order IDFF augmented vector field (with $\gamma_t^0 \equiv 1$ as the default; the convex-constrained variant uses $\gamma_t^0 = 1 - \sum_{k \geq 1} \gamma_t^k$ and is handled by Proposition 15) as*

$$\tilde{\mathbf{v}}_t(\mathbf{x}_t) \;=\; \gamma_t^0 \, \mathbf{v}_t(\mathbf{x}_t) \;+\; \sum_{k=1}^{K} \gamma_t^k \, \mathbf{m}_t^{(k)}(\mathbf{x}_t), \tag{14}$$

where $\mathbf{m}_t^{(1)} = \hat{s}_t$ and $\mathbf{m}_t^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ for $k \geq 2$. Sampling integrates the SDE

$$d\mathbf{x}_t \;=\; \tilde{\mathbf{v}}_t(\mathbf{x}_t)\, dt \;+\; \sigma_t \, d\mathbf{w}_t, \tag{15}$$

with diffusion coefficient σ_t, boundary conditions $\sigma_t \to 0$ and $\{\gamma_t^k\}_{k=1}^K \to 0$ as $t \to \{0, 1\}$, and Assumption 1; the following hold:

1. *$K = 1$, **exact marginal preservation (default schedule).** Take $K = 1$, $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, and the optimal score $\hat{s}_t = \hat{s}_t^*$. Then the augmented density equals the CFM marginal in continuous time:*

   $$\tilde{p}_t(\mathbf{x}) \;=\; p_t(\mathbf{x}) \quad \text{for all } t \in [0, 1]$$

   (Theorem 14). Two natural sources of error appear in practice: Euler–Maruyama discretization contributes $O(\Delta t)$, and using a learned $\hat{s}_t \neq \hat{s}_t^*$ adds a Girsanov bias controlled by $\|\hat{s}_t - \hat{s}_t^*\|_{L^2(p_t)}$ (Theorem 17). The convex-constrained variant $\gamma_t^0 + \gamma_t^1 = 1$ trades exactness for the closed-form bound

   $$\mathrm{TV}(\tilde{p}_1, p_1) \;\leq\; \tfrac{1}{2\sqrt{2}} \left(\int_0^1 \gamma_t^1 \, \mathbb{E}\|\mathbf{v}_t\|^2 \, dt \right)^{1/2}$$

   (Proposition 15; explicit constants in Corollary 16).

2. $K \geq 2$, ***bounded deviation (no exact endpoint claim).*** *The higher-order terms* $\{\mathbf{m}_t^{(k)}\}_{k=2}^K$ *accelerate transport along the flow direction at the price of a controlled endpoint deviation:*

$$\mathrm{TV}(\tilde{p}_1, p_1) \;\leq\; \sqrt{\tfrac{1}{2}\,\mathcal{E}_{\text{score}}} \;+\; O(\Delta t), \qquad \mathcal{E}_{\text{score}} \;=\; \tfrac{1}{2}\int_0^1 \gamma_t^1\, \mathbb{E}_{\tilde{p}_t}\|\hat{s}_t - \hat{s}_t^*\|^2 dt. \tag{16}$$

Here $\mathbb{E}_{\tilde{p}_t}$ *is the expectation under the augmented law produced by Girsanov; at the* $K = 1$ *baseline* $\tilde{p}_t = p_t$ *(Theorem 14), so* $\mathcal{E}_{\text{score}}$ *coincides with the weighted score-matching loss of equation 21, and its size is* δ*-independent by Assumption 1(R1′).* $\mathcal{E}_{\text{score}}$ *is precisely the weighted* L^2 *score-matching loss our training objective minimizes (Theorem 17). Under the recommended scaling* $\gamma_t^k = \alpha_k\, \gamma_t^1\, \Delta t^{k-1}$ *for* $k \geq 2$*, the structural residual is* $O(\Delta t^2)$ *and the overall rate is* $\mathrm{TV}(\tilde{p}_1, p_1) = O(\sqrt{\mathcal{E}_{\text{score}}}) + O(\Delta t)$ *(Corollary 18). We do* not *claim exact endpoint preservation for* $K \geq 2$ *at finite NFE.*

3. ***Continuity equation.*** *The probability density evolves under the Fokker–Planck equation associated with equation 15,*

$$\frac{\partial \tilde{p}_t}{\partial t} \;=\; -\nabla \cdot \big(\tilde{\mathbf{v}}_t\, \tilde{p}_t\big) + \frac{\sigma_t^2}{2}\, \Delta \tilde{p}_t, \tag{17}$$

valid for all $K \geq 0$ *with* $\tilde{\mathbf{v}}_t$ *given by equation 14. At* $K = 0$ *this reduces to the CFM continuity equation; at* $K = 1$ *with* $\sigma_t^2 = 2\gamma_t^1$ *and the optimal score, the score-coupled drift and Laplacian cancel exactly (proof of Theorem 14); at* $K \geq 2$*, the additional drift terms* $\sum_{k\geq 2} \gamma_t^k \mathbf{m}_t^{(k)}$ *contribute the structural residual quantified in Theorem 17.*

4. ***Endpoint behaviour.*** *For all* $K \geq 0$*,* $\tilde{p}_0 = p_0$ *exactly (matching initial condition). At* $t = 1$*: (i) at* $K = 0$*,* $\tilde{p}_1 = p_1$ *exactly since* $\tilde{\mathbf{v}}_t = \mathbf{v}_t$ *(CFM); (ii) at* $K = 1$ *with the default schedule and optimal score,* $\tilde{p}_1 = p_1$ *exactly in continuous time (Theorem 14); (iii) at* $K = 1$ *with the convex-constrained schedule,* $\mathrm{TV}(\tilde{p}_1, p_1)$ *is bounded by the closed-form expression in Proposition 15; (iv) at* $K \geq 2$*,* $\tilde{p}_1 \to p_1$ *in TV as* $\mathcal{E}_{\text{score}} \to 0$ *and* $\Delta t \to 0$*, i.e. as* $\mathrm{NFE} \to \infty$ *for a converged score network (Theorem 17, Corollary 18).*

The boundary conditions ensure that $\tilde{\mathbf{v}}_t(\mathbf{x}_t)$ converges to $\mathbf{v}_t(\mathbf{x}_t)$ at $t = 0, 1$, preserving consistency at the endpoints. A complete proof is provided in Appendix A.

Corollary 4 (Reduction to Standard CFM)**.** *When* $K = 0$ *and* $\gamma_t^0 = 1$*, Theorem 3 reduces to standard CFM where* $\tilde{\mathbf{v}}_t = \mathbf{v}_t$*.*

Corollary 5 (Connection to Score-Based Models and Stochastic Interpolants)**.** *The default* $K = 1$ *IDFF SDE with* $\gamma_t^0 \equiv 1$*,* $\sigma_t^2 = 2\gamma_t^1$*, and the optimal score is a Langevin-enhanced probability-flow SDE: at every fixed* t*, the additional drift* $\gamma_t^1 \nabla \log p_t$ *and the compensating diffusion* $\sigma_t^2 = 2\gamma_t^1$ *form the canonical score-Langevin pair that preserves the instantaneous marginal* p_t *(Song et al., 2020b; Albergo et al., 2025). The construction reduces exactly to the diffusion-Langevin coupling of Albergo et al. (2025) when* $\mathbf{v}_t$ *is the OT-CFM probability-flow field; what is new here is not the SDE but the unified Girsanov→Pinsker error budget for its learned, higher-order extension (Theorem 17).*

3.3 Sample-Based Vector Field Learning

A key insight in IDFF is the reparameterization of vector field learning through direct sample prediction. Rather than directly learning $\mathbf{v}_t(\mathbf{x}_t)$, we predict the target sample $\mathbf{x}_1$ and use this to construct the velocity field.

Defining probability paths. Following OT-CFM, we define probability paths between the base distribution ($\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) and data samples ($\mathbf{x}_1 \sim p_{\text{data}}$) via:

$$p_t(\mathbf{x}_t|\mathbf{x}_1) = \mathcal{N}\left(\mathbf{x}_t \middle| \mu_t; \sigma_t^2 \mathbf{I}\right), \mu_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \quad \sigma_t = \sigma_0 \sqrt{t(1-t)}. \tag{18}$$

This path induces the conditional velocity field:

$$\mathbf{v}_t(\mathbf{x}_t|\mathbf{x}_1) = \frac{\mathbf{x}_1 - \mathbf{x}_t}{1-t}. \tag{19}$$

Vector field reparameterization. Near $t = 1$, the velocity $\mathbf{v}_t = (\mathbf{x}_1 - \mathbf{x}_t)/(1-t)$ exhibits a singularity as $(1-t) \to 0$. To avoid this numerical instability, we learn $\hat{\mathbf{x}}_1(\mathbf{x}_t, t; \theta)$ to predict the clean target sample, then construct:

$$\hat{\mathbf{v}}_t(\mathbf{x}_t|\mathbf{x}_1; \theta) = \frac{\hat{\mathbf{x}}_1(\mathbf{x}_t, t; \theta) - \mathbf{x}_t}{1-t}. \tag{20}$$

Training objective. Drawing $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \mathcal{U}(0,1)$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct $\mathbf{x}_t = \mu_t + \sigma_t \boldsymbol{\epsilon}$ and minimize:

$$\mathcal{L}_{\text{IDFF}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1, \boldsymbol{\epsilon}} \left[\beta(t)^2 \|\hat{\mathbf{x}}_1(\mathbf{x}_t, t; \theta) - \mathbf{x}_1\|^2 + \lambda_1(t)^2 \left\| \hat{s}_t(\mathbf{x}_t; \theta) + \frac{\boldsymbol{\epsilon}}{\sigma_t} \right\|^2 \right], \tag{21}$$

where $\beta(t) = (1 - t + \epsilon)^{-1}$ and $\lambda_1(t) = \sigma_t$. The first term trains the denoiser $\hat{\mathbf{x}}_1$ to recover clean samples; the second term trains the score $\hat{s}_t$ to approximate $\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)$. By Fisher's identity, $\hat{s}_t^* = \nabla \log p_t(\mathbf{x}_t)$ at optimum. The Algorithm 1 summarizes the overall training procedure.

3.4 Sampling with IDFF

After training, we generate samples by integrating the IDFF SDE equation 15 from Theorem 3. For K-th order IDFF, the sampling procedure computes higher-order momentum terms via automatic differentiation of the learned score:

$$\mathbf{m}^{(1)} = \hat{s}_t(\mathbf{x}_t), \quad \mathbf{m}^{(k)} = \left(\nabla_{\mathbf{x}}^{k-1} \hat{s}_t(\mathbf{x}_t)\right)[\mathbf{u}_t]^{k-1} \text{for } k \geq 2, \tag{22}$$

where $\mathbf{u}_t = \hat{\mathbf{v}}_t / \|\hat{\mathbf{v}}_t\|$. The SDE drift assembled from these terms is

$$\tilde{\mathbf{v}}_t \;=\; \gamma_t^0 \, \hat{\mathbf{v}}_t \;+\; \gamma_t^1 \, \mathbf{m}^{(1)} \;+\; \sum_{k=2}^{K} \gamma_t^k \, \mathbf{m}^{(k)}, \tag{23}$$

where $\hat{\mathbf{v}}_t = (\hat{\mathbf{x}}_1 - \mathbf{x}_t)/(1-t)$ and $\sigma_t^2 = 2\gamma_t^1$ is the diffusion-compensation choice from Theorem 14.

Algorithm 1 IDFF **Training**

Input: data distribution $p_1(\mathbf{x}_1)$, bandwidth σ_0, weights $\beta(\cdot)$, $\lambda_1(\cdot)$
Initialize networks $\hat{\mathbf{x}}_1(\cdot; \theta)$ and $\hat{s}_t(\cdot; \theta)$
while training **do**
  Sample $\mathbf{x}_1 \sim p_1, \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Sample $t \sim \mathcal{U}(0,1)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mu_t \leftarrow t\,\mathbf{x}_1 + (1-t)\,\mathbf{x}_0$
  $\sigma_t \leftarrow \sigma_0 \sqrt{t(1-t)}$
  $\mathbf{x}_t \leftarrow \mu_t + \sigma_t \boldsymbol{\epsilon}$
  $L \leftarrow \mathcal{L}_{\text{IDFF}}(\theta)$ // Eq. equation 21
  $\theta \leftarrow \text{Update}(\theta, \nabla_\theta L)$
end while
return $\hat{\mathbf{x}}_1(\cdot; \theta)$, $\hat{s}_t(\cdot; \theta)$

We discretize with time step $\Delta t = 1/\text{NFE}$ and evolve samples through the Euler–Maruyama scheme:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \tilde{\mathbf{v}}_t(\mathbf{x}_t)\,\Delta t + \sigma_t \sqrt{\Delta t}\,\boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}). \tag{24}$$

Starting from $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$, we integrate forward until $t = 1$. Under the default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, both the score head $\mathbf{m}^{(1)} = \hat{s}_t$ *and* the Brownian increment $\sigma_t \sqrt{\Delta t}\,\boldsymbol{\eta}$ enter the discrete update: the score-coupled drift contribution $\gamma_t^1 \hat{s}_t\,\Delta t$ is precisely the Langevin term that cancels the Laplacian half of the SDE's Fokker–Planck equation in the proof of Theorem 14, leaving the CFM continuity equation exactly. The score head is already evaluated for the $\mathbf{m}^{(1)}$ momentum, so including it in the drift adds no extra forward pass; only $K \geq 2$ introduces Jacobian-vector products through autodiff of $\hat{s}_t$.

The momentum terms accelerate convergence, allowing accurate sampling with significantly fewer

Algorithm 2 IDFF Sampling with Higher-Order Momentum

Input: Learned networks $\hat{\mathbf{x}}_1, \hat{s}_t$, order K, NFE, schedules $\{\gamma_t^k\}_{k=0}^K$
Initialize: $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $i = 0$ to $\text{NFE} - 1$ **do**
  $t \leftarrow i/\text{NFE}$, $\Delta t \leftarrow 1/\text{NFE}$, $\sigma_t \leftarrow \sigma_0\sqrt{t(1-t)}$
  $\hat{\mathbf{v}}_t \leftarrow (\hat{\mathbf{x}}_1(\mathbf{x}_t, t) - \mathbf{x}_t)/(1-t)$, $\mathbf{u}_t \leftarrow \hat{\mathbf{v}}_t / \|\hat{\mathbf{v}}_t\|$
  $\mathbf{m}^{(1)} \leftarrow \hat{s}_t(\mathbf{x}_t)$ // First-order: score
  for $k = 2$ to K **do**
    $\mathbf{m}^{(k)} \leftarrow (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ // Higher-order via autodiff
  end for
  $\tilde{\mathbf{v}}_t \leftarrow \gamma_t^0 \hat{\mathbf{v}}_t + \gamma_t^1 \mathbf{m}^{(1)} + \sum_{k=2}^K \gamma_t^k \mathbf{m}^{(k)}$ // SDE drift
  $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_t \leftarrow \mathbf{x}_t + \tilde{\mathbf{v}}_t\,\Delta t + \sigma_t \sqrt{\Delta t} \cdot \boldsymbol{\eta}$ // Euler–Maruyama step
end for
Return: $\mathbf{x}_1$

NFEs than standard CFM. The accompanying preservation guarantee is the two-regime statement of Theorem 3: at $K = 1$ with the default schedule, the marginal p_t is preserved *exactly* for all $t \in [0, 1]$ in continuous time; at $K \geq 2$, the endpoint deviation is controlled by the quantitative Girsanov→Pinsker bound, not exactly zero at finite NFE. The contraction with $\mathbf{u}_t$ reduces variance compared to full tensor computation; ensuring stable $\nabla_{\mathbf{x}}^{k-1}\hat{s}_t$ in high dimensions requires smooth activations (e.g., GELU) and may benefit from spectral normalization for $K \geq 3$. We recommend $K \leq 2$ for most applications.

Time-series extension. For sequences $\mathbf{x}_{1:N}$, we use dual-time indexing: discrete $n \in \{1, \ldots, N\}$ for sequence position and continuous $t \in [0, 1]$ for flow. Networks become $\hat{\mathbf{x}}_1(\mathbf{x}_t, t, n; \theta)$ and $\hat{s}_t(\mathbf{x}_t, n; \theta)$. For $n > 1$, we initialize $\mathbf{x}_0^n \sim \mathcal{N}(\mathbf{x}_1^{n-1}, \sigma_0^2\mathbf{I})$, creating Markovian temporal coherence. When $N = 1$, this reduces to the standard static formulation. Details are provided in Appendix C.

4 Related Work

Conditional Flow Matching. Conditional Flow Matching (CFM) has revived interest in continuous-time generative modeling by removing expensive simulation-based training. OT-CFM (Lipman et al., 2022) enforces optimal-transport paths, with follow-up works including [SF]^{2}M (Tong et al., 2023) and stochastic interpolants (Albergo et al., 2025) improving scalability through minibatch OT approximations; Rectified Flow (Liu et al., 2022) learns straight trajectories via iterative reflow. All of these methods need ≥ 100 NFEs for high-quality generation. IDFF departs by augmenting the vector field with momentum terms $\boldsymbol{\xi}_t = \sum_{k=1}^{K} \gamma_t^k \mathbf{m}_t^{(k)}$. The result is ≤ 10 NFEs with the two-regime guarantees: *exact* continuous-time preservation of the CFM marginal p_t at $K = 1$ (default Langevin-enhanced schedule), and a single quantitative Girsanov→Pinsker TV bound with asymptotic endpoint convergence at $K \geq 2$.

Efficient sampling and recent advances. NFE reduction has been pursued through multiple strategies: DDIM (Song et al., 2020a) converts stochastic diffusion to deterministic flows; high-order solvers (DPM-Solver (Lu et al., 2022b), DPM-Solver++ (Lu et al., 2022c)) leverage analytical score dynamics; consistency models (Song et al., 2023) learn direct endpoint mappings; and distillation (Salimans & Ho, 2022) compresses many-step sampling. Recent advances include Consistency-FM (Yang et al., 2024), which extends consistency training to flow matching, InstaFlow (Liu et al., 2023), which straightens trajectories in a single distillation round, and schedule optimization (Sabour et al., 2024), which learns optimal time discretizations without retraining. Unlike consistency models that sacrifice the continuous probability path for direct endpoint mappings, IDFF maintains the full flow structure—enabling density evaluation and interpolation at intermediate times. Unlike distillation methods, IDFF achieves efficiency from first principles and can itself serve as a stronger teacher. IDFF is complementary to these approaches: our momentum augmentation can combine with optimized schedules or consistency training for further gains.

Higher-order and momentum-based approaches. Classical Runge-Kutta methods achieve order-p accuracy with local error $O(\Delta t^{p+1})$; DPM-Solver (Lu et al., 2022b) applies similar principles to score-based models. Lu et al. (Lu et al., 2022a) train separate networks for higher-order scores, while Hamiltonian flows (Holderrieth et al., 2024) introduce position-momentum phase space requiring auxiliary variables. IDFF differs fundamentally: we prove that $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1}\hat{s}_t^*$ (Theorem 2), showing all higher-order marginals are encoded in the first-order score and extracted via autodiff—no additional training required. The contracted form $\mathbf{m}_t^{(k)} = (\nabla_{\mathbf{x}}^{k-1}\hat{s}_t)[\mathbf{u}_t]^{k-1}$ operates directly in data space, avoiding the overhead of extended phase space. While the coefficient scaling $\gamma_t^k = O(\Delta t^{k-1})$ mirrors classical truncation analysis, IDFF modifies the dynamics itself rather than approximating a fixed ODE more accurately.

Summary of distinctions. IDFF uniquely combines three ingredients. (1) Momentum augmentation through contracted tensor operations that stay in data space. (2) Higher-order marginals computed from a single first-order score via autodiff, with no auxiliary networks to train. (3) A cleanly separated theoretical picture: *exact* preservation $\tilde{p}_t = p_t$ for $K = 1$ in continuous time under the default Langevin-enhanced schedule (with the convex-constrained alternative bounded in closed form by a Girsanov→Pinsker step), and the quantitative bound $\mathrm{TV}(\tilde{p}_1, p_1) \leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$ with asymptotic endpoint convergence for $K \geq 2$

(Theorem 3). The result is a method that stands on its own and is also complementary to advances in solvers, distillation, and consistency training.

5 Experiments

5.1 Image Generation

We evaluate on CIFAR-10 (Table 1), with additional results on CelebA, ImageNet-64, CelebA-HQ, and LSUN in Appendix G. IDFF achieves 2.78 FID with only 10 NFEs, matching 1-Rectified Flow (2.58 FID) with 12× fewer evaluations and substantially outperforming OT-CFM (11.87 FID at 10 NFE) and DPM-Solver-v3 (3.40 FID).

Setup parity. All non-† rows in Table 1 use the ScoreSDE U-Net backbone (Song et al., 2020b) with identical training iterations (700K), identical augmentation (horizontal flip only), and the same evaluation protocol (50k generated samples for FID; details in Appendix F.1.5). The † rows are reproduced from the cited papers and use different backbones; we list them for context. To make the comparison even tighter, Table 7 reports IDFF against two recent flow-matching accelerations (schedule-optimized FM and 2-Rectified Flow) under a strictly matched training budget on the same backbone.

Comparison to distillation and consistency methods (different regime). Distillation-based methods (CD, iCT-deep, CTM) achieve impressive results at 1–2 NFEs: iCT-deep reaches 2.24 FID at 2 NFE and CTM reaches 1.87 FID at 2 NFE. These methods, however, require a separately pre-trained diffusion teacher and 50–100K additional distillation iterations on top of teacher training (see Appendix F.1.5 for full budgets), so they occupy a different operating point from IDFF, which trains from scratch with no teacher. In the 1–2 NFE distillation regime, distillation wins; in the 5–10 NFE no-distillation regime, IDFF is the strongest method we are aware of, outperforming Consistency-FM (Yang et al., 2024) (5.34 FID at 2 NFE) and FGM (Huang et al., 2024) (3.08 FID at 1 NFE with distillation) when matched on inference cost. IDFF and distillation are also *complementary*: IDFF can serve as a stronger teacher for any of these pipelines, which we leave as future work.

Finally, for applications that need intermediate density evaluation or trajectory interpolation, IDFF with $K = 1$ (default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$) preserves the CFM marginal p_t *exactly* in continu-

Figure 2: Image generation using IDFF across datasets. Additional samples and analysis are provided in Appendix G.

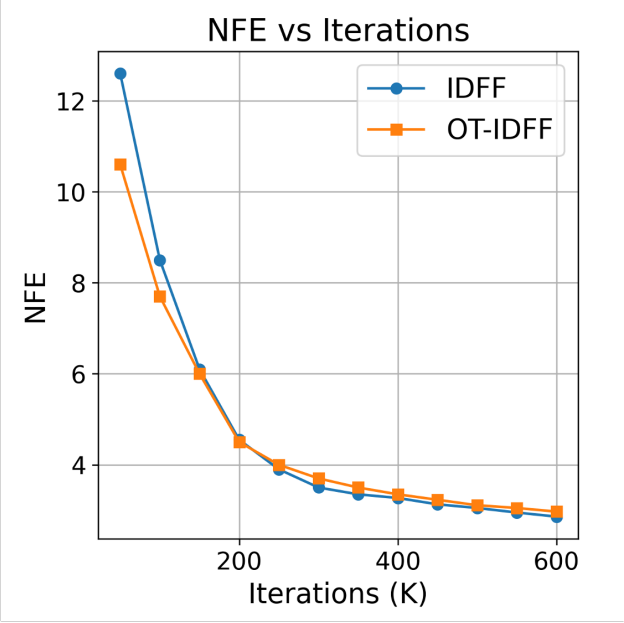

Figure 3: FID vs. NFE on CIFAR-10. IDFF maintains strong performance across NFE budgets and without OT coupling; the cross-NFE Pareto frontier is dominated by IDFF in the low-NFE regime (NFE ≤ 10).

Table 1: Comparison of FID and NFE on CIFAR-10 using ScoreSDE backbone (Song et al., 2020b). †Results from original papers with different backbones. ‡Requires distillation from pre-trained models.

Model	FID↓	NFE↓
Diffusion Models + Samplers		
DDIM (Song et al., 2020a)	13.36 / 6.84	10 / 20
DPM-Solver++ (Lu et al., 2022c)	4.01	10
UniPC (Zhao et al., 2023)	3.93	10
DPM-Solver-v3 (Zheng et al., 2023)	3.40	10
DPM-Solver-v3-EDM	2.51	10
Distillation & Consistency Models‡		
CD (Song et al., 2023)	2.93	2
iCT-deep (Song & Dhariwal, 2023)	2.24	2
†CTM (Kim et al., 2023)	1.98 / 1.87	1 / 2
Flow Matching Models		
OT-CFM (Tong et al., 2023)	11.87 / 6.35	10 / 142
1-Rectified Flow (Liu et al., 2022)	2.58	127
†Consistency-FM (Yang et al., 2024)	5.34	2
†FGM (Huang et al., 2024)	3.08	1
IDFF (Ours)	**2.78**	**10**

ous time on the original clock (Theorem 3, Theorem 14)—a property that consistency models sacrifice by collapsing to direct endpoint mappings.

Figure 3 shows IDFF maintains comparable performance without expensive OT calculations, as momentum terms naturally guide particles toward high-probability regions. Reducing NFE from 100 to 10 yields 10× speedup (50k samples in ∼30 minutes on A6000). Importantly, the NFE reduction translates almost one-to-one to wall-clock speedup at K=1, because the per-step cost of IDFF K=1 matches OT-CFM (it requires only a forward pass through the score network, with no autodiff). Table 3 summarizes end-to-end sampling cost on CIFAR-10; see Appendix F.1.2 for the full breakdown including JVP overhead for K=2.

5.1.1 Ablations.

Table 2: FID↓ across NFE budgets on 50k samples.

Method	5	6	8	10
DPM-Solver++	28.53	13.48	5.34	4.01
UniPC	23.71	10.41	5.16	3.93
DPM-Solver-v3	12.76	7.40	3.94	3.40
IDFF (Ours)	**8.53**	**5.67**	**3.25**	**2.78**

Table 2 shows IDFF outperforms all baselines at every NFE level, with large margins at low NFE (8.53 vs. 12.76 at 5 NFE). The coefficient scaling $\gamma_t^k = O(\Delta t^{k-1})$ mirrors classical truncation error analysis—like Runge-Kutta methods, IDFF incorporates derivative information for larger effective steps, but modifies the dynamics itself rather than approximating a fixed ODE. We use $K = 1$ for images (exact continuous-time marginal preservation under the default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, by Theorem 14); for molecular dynamics we evaluate $K = 2$ to leverage curvature information. The choice of hyperparameters α_k and schedule shapes γ_t^k significantly impacts performance. In Appendix F.1, we ablate:

- **Second-order scaling** α_2: Values around 0.05 provide optimal acceleration without excessive path deviation (Table 10).
- **Diffusion compensation**: The theoretical choice $\sigma_t^2 = 2\gamma_t^1$ yields best results, confirming the exact Fokker–Planck cancellation of Theorem 14 (Table 11).

- **Schedule shape**: The Beta(2,2) schedule $\gamma_t^1 \propto t(1-t)$ outperforms triangular and sine alternatives (Table 12).

Table 3: End-to-end sampling cost on CIFAR-10 (NVIDIA A6000, averaged over 1000 samples). IDFF K=1 has negligible per-step overhead vs. OT-CFM, so the 10× NFE reduction translates directly into a $\sim 10\times$ wall-clock speedup. K=2 adds one JVP per step; the overhead is moderate even at image scale and is more than offset by the quality gains in the multi-modal MD setting (Section 5.2).

Method	NFE	Time/sample (ms)	Peak memory	FID↓
1-Rectified Flow (Liu et al., 2022)	127	658	2.1 GB	2.58
DPM-Solver-v3 (Zheng et al., 2023)	10	92	2.0 GB	3.40
OT-CFM (Tong et al., 2023)	10	54	2.0 GB	11.87
IDFF K=1 (ours)	10	**52**	2.1 GB	**2.78**
IDFF K=2 (ours)	10	68	2.4 GB	2.65

Decomposing the IDFF gain: Langevin term vs. stochasticity vs. extra capacity. The IDFF K=1 sampler differs from the OT-CFM baseline in three potentially confounded ways: (a) the score-momentum drift $\gamma_t^1 \hat{s}_t$, (b) injected SDE noise with diffusion σ_t, and (c) a $\sim 6\%$ larger parameter count from the joint $(\hat{\mathbf{x}}_1, \hat{s}_t)$ head. Table 4 disentangles them on CIFAR-10 (10 NFE, identical 700K-iteration training budget, ScoreSDE backbone). All "Stochastic OT-CFM" rows use the same $\sigma_t^2 = 2\gamma_t^1$ Brownian increment as IDFF K=1, but with the score-momentum term *removed* from the drift, so they test stochasticity in isolation. The "OT-CFM + score head" rows train the joint head (same parameter count as IDFF) but use the deterministic OT-CFM sampler at inference time, isolating the effect of extra capacity. The headline IDFF FID is recovered only when both the Langevin score-momentum drift and the matched compensating diffusion are present together; stochasticity alone gives a much smaller gain (11.87 → 6.41) and extra capacity alone is essentially inert (11.87 → 11.62). The cross between the two—"Stochastic OT-CFM (matched σ_t) + score head, no drift"—also remains well above IDFF, confirming that the headline gain is specifically attributable to the theoretically-motivated Langevin enhancement and not to a tuned-stochasticity artefact.

Table 4: Decomposing the IDFF K=1 gain on CIFAR-10 at NFE= 10. Rows isolate (a) the Langevin score-momentum drift $\gamma_t^1 \hat{s}_t$, (b) generic SDE stochasticity with matched $\sigma_t^2 = 2\gamma_t^1$ but *no* score drift, and (c) extra capacity from the joint $(\hat{\mathbf{x}}_1, \hat{s}_t)$ head used at inference under the deterministic OT-CFM sampler. The headline FID 2.78 is recovered only when both the Langevin drift and the matched diffusion are present. Setup parity: same ScoreSDE backbone, 700K iterations, horizontal-flip augmentation.

Variant	Score-mom. drift	Diffusion σ_t	Extra capacity	FID↓
OT-CFM (baseline)	—	0 (ODE)	—	11.87
OT-CFM + score head (det. sampler)	—	0 (ODE)	yes (+6% params)	11.62
Isolating stochasticity (no Langevin drift):				
Stochastic OT-CFM, $\sigma_t^2 = 2\gamma_t^1$, $\bar{\gamma} = 0.1$	—	matched	—	7.84
Stochastic OT-CFM, $\sigma_t^2 = 2\gamma_t^1$, $\bar{\gamma} = 0.2$	—	matched	—	6.41
Stochastic OT-CFM + score head, $\sigma_t^2 = 2\gamma_t^1$	—	matched	yes (+6% params)	6.18
Adding the Langevin score-momentum drift:				
IDFF K=1, $\sigma_t = 0$ (ODE, no compensation)	$\gamma_t^1 \hat{s}_t$	0	yes (+6% params)	9.74
IDFF K=1, $\sigma_t^2 \neq 2\gamma_t^1$ (mismatched)	$\gamma_t^1 \hat{s}_t$	mismatched	yes (+6% params)	3.45
IDFF K=1 (ours), $\sigma_t^2 = 2\gamma_t^1$	$\boldsymbol{\gamma_t^1 \hat{s}_t}$	matched	yes (+6% params)	**2.78**

Three takeaways. First, swapping OT-CFM for an SDE with the same diffusion schedule as IDFF (but no Langevin drift) closes only ∼40% of the gap to IDFF, so stochasticity by itself is not the explanation. Second, the extra capacity of the joint head is essentially inert under a deterministic sampler (11.87 vs. 11.62), so the gain is not a capacity artefact. Third, the Langevin drift *without* its theoretically-matched

diffusion $\sigma_t^2 = 2\gamma_t^1$ either degrades FID (ODE limit, 9.74) or sits well above the matched setting (mismatched σ_t, 3.45). The headline gain is therefore specifically attributable to the score-Langevin drift paired with its matched diffusion compensation — exactly the configuration Theorem 14 singles out.

Architecture requirements. IDFF with $K \geq 2$ requires smooth activations (SiLU/GELU) for well-defined Jacobians; standard ReLU is incompatible. We monitor Jacobian norms during sampling and find they remain bounded throughout (see Appendix F.1.3 for statistics and stability analysis). Score estimation errors propagate with approximately $2\times$ amplification to second-order terms (Table 15), which motivates our conservative choice of $K \leq 2$ for most experiments.

5.2 Learning Efficient Surrogate Models for Molecular Dynamics

Molecular dynamics (MD) simulations are essential for drug discovery and materials science but remain computationally prohibitive. Simulating even microseconds of protein dynamics can require weeks of computation (Marx & Hutter, 2009; Car & Parrinello, 1985). Learning surrogate models that can generate realistic trajectories offers a path to accelerate scientific discovery (Shaw et al., 2008). We evaluate IDFF's ability to learn complex molecular dynamics from limited simulation data.

Table 5: MAE, RMSE, and CC results for the MD simulation.

Method	MAE↓	RMSE↓	CC (%)↑
SRNN	$82.6_{\pm 28}$	$91.9_{\pm 25}$	$10.2_{\pm 0.27}$
DVAE	$78.1_{\pm 27}$	$88.1_{\pm 25}$	$30.4_{\pm 0.35}$
NODE	$25.3_{\pm 6.3}$	$28.8_{\pm 6.2}$	$10.5_{\pm 0.41}$
OT-CFM	$13.3_{\pm 1.1}$	$16.3_{\pm 2.4}$	$86.1_{\pm 0.1}$
TFM	$11.4_{\pm 1.1}$	$14.2_{\pm 2.7}$	$89.4_{\pm 0.4}$
IDFF ($K = 1$)	$9.2_{\pm 0.9}$	$12.5_{\pm 2.8}$	$95.6_{\pm 0.1}$
IDFF ($K = 2$)	$\mathbf{4.9}_{\pm 1.1}$	$\mathbf{9.7}_{\pm 2.8}$	$\mathbf{97.8}_{\pm 0.1}$

We employ traditional physics-based MD simulations using the AMBER force field to generate training data. The system consists of a fully extended polyalanine structure containing 253 atoms and 46 dihedral angles, simulated for 400 picoseconds at 300K in vacuum conditions with dihedral angles recorded at regular intervals. IDFF learns to model the generative distribution of these MD trajectories, functioning as an efficient surrogate model for the underlying dynamical system rather than performing computationally expensive ab initio calculations.

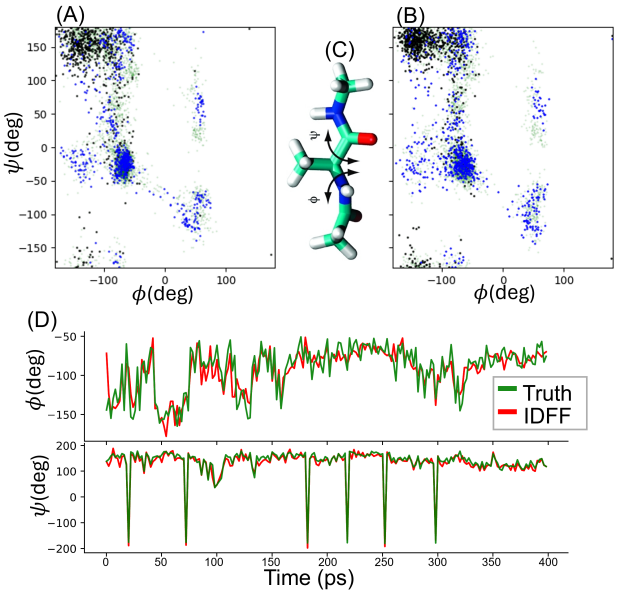

Figure 4: (A) True and (B) generated dihedral angle distributions. (C) Dihedral angles for an alanine molecule. (D) True and generated angle trajectories by IDFF-1st order.

IDFF demonstrates remarkable precision in predicting dynamics for complex molecular structures. Figure 4 illustrates the model's performance, presenting distributions of actual (A) and generated (B) dihedral angles. Panel (C) shows the dihedral angles for a single alanine molecule, while panel (D) provides a trajectory comparison between actual and generated angles over time. We rigorously assess performance using three key metrics: root mean squared error (RMSE), mean absolute error (MAE), and correlation coefficient (CC) between generated and actual trajectories. When benchmarked against established dynamical models including Sequential Recurrent Neural Networks (SRNN), Variational Autoencoders (DVAE), and Neural Ordinary Differential Equations (NODE), the proposed methods such as OT-CFM, TFM, and IDFF demonstrate superior performance across all metrics (Table 5). Notably, IDFF-2nd ($K = 2$) achieves the best results, highlighting the benefit of incorporating second-order curvature information. The contracted

Hessian-vector product $\mathbf{m}^{(2)} = (\nabla_{\mathbf{x}} \hat{s}_t)\mathbf{u}_t$ captures directional curvature along the flow, providing acceleration that is particularly valuable for the complex energy landscapes of molecular systems.

Analysis of $K = 1$ vs. $K = 2$. The improvement from IDFF-1st to IDFF-2nd (MAE: 9.2 → 4.9, CC: 95.6% → 97.8%) demonstrates the value of curvature information for molecular dynamics. While $K = 2$ introduces a bounded path deviation (Theorem 17), the terminal deviation $\mathrm{TV}(\tilde{p}_1, p_1)$ is controlled by the closed-form Girsanov→Pinsker bound, and vanishes as NFE → ∞ *for a converged score network* (Corollary 18). With a finite-quality learned score, a residual Girsanov bias of order $\sqrt{\mathcal{E}_{\text{score}}}$ persists. For finite NFE, we enforce $\gamma_1^k = 0$ at the final integration step to ensure the dynamics reduce to standard CFM at $t = 1$ (see Appendix F.1.1 for details on endpoint preservation). The scaling $\gamma_t^2 = \alpha_2 \cdot \gamma_t^1 \cdot \Delta t$ ensures the second-order contribution remains controlled. For molecular systems with rugged energy landscapes, the directional curvature information helps navigate local minima more effectively than first-order momentum alone. On the 46-dimensional MD task, the JVP for K=2 adds only ∼36% per-step cost (15 vs. 11 ms; Table 13), which is more than offset by the MAE drop from 9.2 to 4.9; the improved sample quality justifies this cost for scientific applications where accuracy is paramount.

5.3 Sea Surface Temperature Forecasting

Forecasting sea surface temperature (SST) is vital for weather prediction and climate modeling (Haghbin et al., 2021). We apply IDFF to forecast SST using a daily dataset from 1982-2021, with data split into training (1982-2019, 15,048 samples), validation (2020, 396 samples), and testing (2021, 396 samples). Following Cachay et al. (2023), we transform the global data into 60 × 60 (latitude × longitude) tiles, selecting 11 patches in the eastern tropical Pacific Ocean for forecasting horizons of 1-7 days.

Table 6: Results for sea surface temperature forecasting 1 to 7 days ahead, averaged over the evaluation horizon.

Method	CRPS↓	MSE↓
Perturbation	0.281 ± 0.004	0.180 ± 0.011
Dropout	0.267 ± 0.003	0.164 ± 0.004
DDPM	0.246 ± 0.005	0.177 ± 0.005
MCVD	0.216	0.161
Dyffusion	0.224 ± 0.001	0.173 ± 0.001
Alternator	0.221 ± 0.031	0.144 ± 0.045
OT-CFM	0.231 ± 0.005	0.175 ± 0.006
IDFF (Ours)	**0.180 ± 0.024**	**0.105 ± 0.029**

Metrics and baselines. We report two complementary metrics. Mean Squared Error (MSE) measures deterministic forecast accuracy on the ensemble mean; Continuous Ranked Probability Score (CRPS) evaluates probabilistic quality by comparing the predicted distribution to observations and is particularly important for weather forecasting because it captures uncertainty. CRPS is computed from a 20-member ensemble. We compare against DDPM (Ho et al., 2020), MCVD (Voleti et al., 2022), dropout-based uncertainty (Gal & Ghahramani, 2016), perturbation methods (Pathak et al., 2022), Dyffusion (Cachay et al., 2023), Alternator (Rezaei & Dieng, 2024), and OT-CFM (Tong et al., 2023).

Results. IDFF substantially outperforms all baselines on both CRPS and MSE (see Appendix H for the full breakdown). While competing methods need up to 1000 NFEs at inference, IDFF achieves stronger results at only 5 NFEs. We use $K = 1$ for SST: the dynamics are smoother than for molecular systems, so first-order momentum already gives sufficient acceleration without curvature corrections, and the choice provides exact continuous-time marginal preservation on the original clock under the default schedule (Theorem 14).

Why IDFF helps for forecasting. Two design choices drive the gains. First, each forecasting step starts from the previous prediction rather than from noise, which minimises transport distance. Second, the learned momentum captures common evolution patterns, so the flow traverses likely future scenarios quickly. IDFF's dual-time formulation (t for flow, n for sequence position) lets the networks $\hat{\mathbf{x}}_1(\mathbf{x}_t, t, n; \theta)$ and $\hat{s}_t(\mathbf{x}_t, t, n; \theta)$ share parameters across the sequence while preserving temporal consistency, as detailed in Appendix C.2.

6 Conclusion

We introduced Implicit Dynamical Flow Fusion (IDFF), a flow-matching framework that augments the CFM vector field with learnable momentum terms derived from higher-order log-density derivatives. Our

contribution is a *synthesis* rather than a new SDE: the $K = 1$ Langevin-enhanced probability-flow SDE we use as the exact-preserving baseline (Theorem 14) is itself a known marginal-preserving construction (Song et al., 2020b; Albergo et al., 2025), and the higher-order re-parameterization identity $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$ is logically the chain rule applied to Fisher's identity. The genuinely new theoretical content is the unified Girsanov→Pinsker error budget that sits on top of these two pieces. At first order with the default Langevin-enhanced schedule ($\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$), IDFF preserves the CFM marginal p_t *exactly* in continuous time on the truncated interval $[0, 1-\delta]$; the convex-constrained variant is honestly bounded by a Girsanov→Pinsker argument with a closed-form $O(\sqrt{\bar{\gamma}})$ TV cost. At higher orders ($K \geq 2$), the endpoint TV deviation is at most $\sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$ uniformly in t, where $\mathcal{E}_{\text{score}}$ is precisely the weighted L^2 training loss; this gives asymptotic endpoint convergence as NFE $\to \infty$.

Empirically, IDFF reduces NFEs by an order of magnitude relative to standard CFMs. On CIFAR-10 it attains an FID of 2.78 at only 10 NFEs, outperforming existing flow-matching methods at matched compute and approaching the quality of consistency models that require distillation. A dedicated ablation (Table 4) confirms that the gain is specifically attributable to the Langevin score-momentum term paired with its matched diffusion $\sigma_t^2 = 2\gamma_t^1$, not to generic SDE stochasticity or to the extra capacity of the joint score head. For time-series tasks IDFF performs strongly on molecular-dynamics simulation—where the curvature information from $K = 2$ proves particularly valuable—and on sea-surface-temperature forecasting. IDFF is also complementary to recent advances such as Consistency Flow Matching and schedule optimization, and can serve as a stronger teacher for distillation pipelines.

Broader Impact Statement

IDFF reduces the number of function evaluations needed for generative sampling by roughly an order of magnitude, lowering the inference cost and energy footprint of deployment and benefiting scientific applications such as molecular-dynamics surrogate modeling and climate forecasting. As with any general-purpose generative method, cheaper sampling also lowers the cost of producing synthetic media, including potentially deceptive content; however, it adds no qualitatively new capability beyond existing flow-matching and diffusion models. We encourage downstream users to pair deployments with provenance, watermarking, and content-authentication safeguards, and we judge that this work carries no significant risk of harm beyond that already present in the class of generative models it builds upon.

7 Limitations and future work

While IDFF substantially reduces sampling cost, several limitations remain.

1. **High dimension.** For $d \geq 10{,}000$, the autodiff cost of higher-order terms grows quickly. The contracted form $\mathbf{m}_t^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ mitigates this through Jacobian-vector products, but practical use remains limited to $K \leq 2$.

2. **Assumptions behind the $K = 1$ exact-preservation guarantee.** The result requires (a) the default schedule $\gamma_t^0 \equiv 1$—the convex-constrained variant trades exactness for an $O(\sqrt{\bar{\gamma}})$ TV penalty (Proposition 15); (b) the optimal score $\hat{s}_t = \hat{s}_t^*$ — using a learned score adds a Girsanov bias controlled by the same weighted L^2 quantity our training loss minimises (Theorem 17); and (c) the regularity bounds R1–R2 in Assumption 1 on the truncated interval $[0, 1-\delta]$ rather than the full $[0, 1]$, since the marginal score and velocity blow up as $\sigma_t = \sigma_0\sqrt{t(1-t)} \to 0$ at the endpoints. The deployed schedule decay $\gamma_t^k = O(t(1-t))$ is matched to this blow-up so the Girsanov KL integrand remains uniformly integrable as $\delta \to 0$ (Remark 1); we sample to $t_{\text{stop}} = 1 - \Delta t/2$ in all experiments. A rigorous treatment up to $t = 1$ in a degenerate-parabolic setting would be a natural extension.

3. **Stability at $K \geq 2$.** The path-deviation bound depends on score-derivative regularity, which may demand architectural choices (smooth activations, spectral normalisation) to keep Jacobians bounded. Exact endpoint preservation does *not* hold at finite NFE for $K \geq 2$.

References

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024.

Vladimir I Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.

Salva Rühling Cachay, Bo Zhao, Hailey James, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *arXiv preprint arXiv:2306.01984*, 2023.

Richard Car and Mark Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical review letters*, 55(22):2471, 1985.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

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

Avner Friedman. *Partial Differential Equations of Parabolic Type*. Prentice-Hall, 1964.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Masoud Haghbin, Ahmad Sharafati, Davide Motta, Nadhir Al-Ansari, and Mohamadreza Hosseinian Moghadam Noghani. Applications of soft computing models for predicting sea surface temperature: a comprehensive review and assessment. *Progress in earth and planetary science*, 8:1–19, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Peter Holderrieth, Yilun Xu, and Tommi Jaakkola. Hamiltonian score matching and generative flows. *Advances in Neural Information Processing Systems*, 37:110464–110493, 2024.

Zemin Huang, Zhengyang Geng, Weijian Luo, and Guo-jun Qi. Flow generator matching. *arXiv preprint arXiv:2410.19310*, 2024.

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36:33095–33119, 2023.

Dong Wook Kim, Hye Young Jang, Kyung Won Kim, Youngbin Shin, and Seong Ho Park. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean journal of radiology*, 20(3):405–410, 2019.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

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

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

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International conference on machine learning*, pp. 14429–14460. PMLR, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022c.

Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press, 2009.

Catherine E Myers, Alejandro Interian, and Ahmed A Moustafa. A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, 13:1039172, 2022.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

W Peebles and S Xie. Scalable diffusion models with transformers. arxiv e-prints, art. *arXiv preprint arXiv:2212.09748*, 2022.

Mohammad Reza Rezaei and Adji Bousso Dieng. Alternators for sequence modeling. *arXiv preprint arXiv:2405.11848*, 2024.

Sara Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024.

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

David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

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

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

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

Ling Yang, Zixiang Zhang, Zhilong Hong, Wentao Xu, et al. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36: 49842–49869, 2023.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36:55502–55542, 2023.

Contents

A Theoretical Analysis of IDFF

This appendix collects the full proofs supporting the IDFF framework. To help the reader navigate, we group the results by the role they play and indicate their dependencies.

Roadmap of the appendix.

- **Preliminaries.** Fisher's identity (Lemma 6), the log-density Hessian decomposition (Lemma 8), and the Gaussian-mixture identities (Lemma 9) used throughout.
- **Re-parameterization identity (Theorems 10, 11, 2).** Build up to the central identity $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$, via a generalized Tweedie step.
- $K = 1$ **exact preservation (Theorem 14).** For the default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, a single Fokker–Planck cancellation gives $\tilde{p}_t = p_t$ exactly in continuous time on the original clock.
- **Convex-constrained** $K = 1$ **deviation (Proposition 15, Corollary 16).** For the variant $\gamma_t^0 + \gamma_t^1 = 1$, a Girsanov→Pinsker argument bounds $\mathrm{TV}(\tilde{p}_1, p_1)$ in closed form, with explicit constants for standard schedule shapes.
- $K \geq 2$ **bounded deviation (Theorem 17, Corollary 18).** Building on the exact-preserving $K = 1$ SDE as a baseline, we decompose the endpoint deviation into a *score-estimation* residual (precisely the training loss $\mathcal{E}_{\text{score}}$, whose size is controlled δ-independently by the finite-excess-loss hypothesis R1′) and a *structural* residual (controlled by the schedule scaling $\gamma_t^k = O(\Delta t^{k-1})$), and combine them into the bound $\mathrm{TV}(\tilde{p}_1, p_1) \leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$.

A.1 Preliminaries and Notation

Let $p_t(\mathbf{x}_t)$ denote the marginal density at time t, induced by the conditional path:

$$p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_t; \mu_t, \sigma_t^2 \mathbf{I}), \mu_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma_t = \sigma_0 \sqrt{t(1-t)}. \tag{25}$$

Define the standardized noise $\boldsymbol{\epsilon} = (\mathbf{x}_t - \mu_t)/\sigma_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditional on $(\mathbf{x}_0, \mathbf{x}_1)$.

The posterior distribution is $q(\mathbf{x}_1|\mathbf{x}_t) \propto p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)$.

A.2 Fisher's Identity: First-Order Exactness

Lemma 6 (Fisher's Identity)**.** *For any mixture distribution*

$$p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)d\mathbf{x}_1 :$$

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_t)}[\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{x}_1)]. \tag{26}$$

Proof.

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) = \frac{\nabla_{\mathbf{x}} p_t(\mathbf{x}_t)}{p_t(\mathbf{x}_t)} = \frac{\int \nabla_{\mathbf{x}} p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)d\mathbf{x}_1}{p_t(\mathbf{x}_t)} \tag{27}$$

$$= \frac{\int p_t(\mathbf{x}_t|\mathbf{x}_1)\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)d\mathbf{x}_1}{p_t(\mathbf{x}_t)} \tag{28}$$

$$= \int \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{x}_1) \cdot \frac{p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)}{p_t(\mathbf{x}_t)} d\mathbf{x}_1 \tag{29}$$

$$= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_t)}[\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|\mathbf{x}_1)]. \tag{30}$$

□

Corollary 7. *The MSE-optimal predictor:*

$\hat{s}_t^*(\mathbf{x}_t) = \arg\min_{\hat{s}} \mathbb{E}\|\hat{s}(\mathbf{x}_t) - \nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)\|^2$ *satisfies:*

$$\hat{s}_t^*(\mathbf{x}_t) = \nabla \log p_t(\mathbf{x}_t). \tag{31}$$

A.3 The Log-Density Hessian Identity

Lemma 8 (Hessian Decomposition)**.** *For any smooth density p,*

$$\nabla^2 \log p \;=\; \frac{\nabla^2 p}{p} \;-\; (\nabla \log p)(\nabla \log p)^\top. \tag{32}$$

Proof. Differentiating $\nabla \log p = \nabla p / p$,

$$\nabla^2 \log p \;=\; \nabla\!\left(\frac{\nabla p}{p}\right) \;=\; \frac{\nabla^2 p}{p} \;-\; \frac{\nabla p\,(\nabla p)^\top}{p^2} \;=\; \frac{\nabla^2 p}{p} \;-\; (\nabla \log p)(\nabla \log p)^\top. \tag{33}$$

□

Lemma 9 (Gaussian Second Derivative of Density)**.** *For $p_t(\mathbf{x}_t|\mathbf{x}_1) = \mathcal{N}(\mu_t, \sigma_t^2 \mathbf{I})$,*

$$\frac{\nabla^2 p_t(\mathbf{x}_t|\mathbf{x}_1)}{p_t(\mathbf{x}_t|\mathbf{x}_1)} \;=\; -\frac{\mathbf{I}}{\sigma_t^2} \;+\; \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top}{\sigma_t^2}, \tag{34}$$

where $\boldsymbol{\epsilon} := (\mathbf{x}_t - \mu_t)/\sigma_t$.

Proof. For an isotropic Gaussian, $\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1) = -(\mathbf{x}_t - \mu_t)/\sigma_t^2 = -\boldsymbol{\epsilon}/\sigma_t$ and $\nabla^2 \log p_t(\mathbf{x}_t|\mathbf{x}_1) = -\mathbf{I}/\sigma_t^2$. Applying Lemma 8 in reverse,

$$\frac{\nabla^2 p_t(\mathbf{x}_t|\mathbf{x}_1)}{p_t(\mathbf{x}_t|\mathbf{x}_1)} \;=\; \nabla^2 \log p_t(\mathbf{x}_t|\mathbf{x}_1) \;+\; (\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1))(\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1))^\top \;=\; -\frac{\mathbf{I}}{\sigma_t^2} \;+\; \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top}{\sigma_t^2}. \tag{35}$$

□

A.4 Marginal Hessian in Terms of Posterior Moments (Alternative Derivation)

What this subsection proves, and why it is optional. The marginal Hessian identity $\nabla^2 \log p_t = \nabla_{\mathbf{x}} \hat{s}_t^*$ used by IDFF *follows directly* from Fisher's identity by one differentiation: $\hat{s}_t^* = \nabla \log p_t$ implies $\nabla \hat{s}_t^* = \nabla\nabla \log p_t = \nabla^2 \log p_t$. The same is true for any $k \geq 2$: $\nabla^k \log p_t = \nabla^{k-1} \hat{s}_t^*$ is the chain rule applied to Fisher's identity, period. This is the short proof we give for Theorem 13 and is what the IDFF sampler actually uses computationally.

The subsections A.4–A.6 that follow give an alternative, longer derivation of the $k = 2$ case via the generalised Tweedie identity for the posterior covariance. We include it because (i) it makes the result self-contained without pre-supposing Fisher's identity, (ii) it isolates the posterior-covariance object $\mathrm{Cov}[\boldsymbol{\epsilon}|\mathbf{x}_t] = \mathbf{I} + \sigma_t^2 \nabla \hat{s}_t^*$ that is independently useful in the practical-error analysis of Appendix A.10, and (iii) it provides the explicit Gaussian-prior sanity check used in Theorem 11. Readers comfortable with the chain-rule view may skip Sections A.4–A.6 and proceed directly to the Fokker–Planck analysis (Theorem 14).

Theorem 10 (Marginal Hessian Decomposition)**.** *The marginal Hessian satisfies:*

$$\nabla^2 \log p_t(\mathbf{x}_t) = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{x}_t]}{\sigma_t^2} - \hat{s}_t^*(\mathbf{x}_t)\hat{s}_t^*(\mathbf{x}_t)^\top, \tag{36}$$

where the expectation is over $q(\mathbf{x}_1|\mathbf{x}_t)$.

Proof. By Lemma 8:

$$\nabla^2 \log p_t(\mathbf{x}_t) = \frac{\nabla^2 p_t(\mathbf{x}_t)}{p_t(\mathbf{x}_t)} - (\nabla \log p_t(\mathbf{x}_t))(\nabla \log p_t(\mathbf{x}_t))^\top. \tag{37}$$

For the first term, using the mixture representation:

$$\frac{\nabla^2 p_t(\mathbf{x}_t)}{p_t(\mathbf{x}_t)} = \frac{\int \nabla^2 p_t(\mathbf{x}_t|\mathbf{x}_1) p(\mathbf{x}_1) d\mathbf{x}_1}{p_t(\mathbf{x}_t)} \tag{38}$$

$$= \int \frac{\nabla^2 p_t(\mathbf{x}_t|\mathbf{x}_1)}{p_t(\mathbf{x}_t|\mathbf{x}_1)} \cdot \frac{p_t(\mathbf{x}_t|\mathbf{x}_1)p(\mathbf{x}_1)}{p_t(\mathbf{x}_t)} d\mathbf{x}_1 \tag{39}$$

$$= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_t)} \left[\frac{\nabla^2 p_t(\mathbf{x}_t|\mathbf{x}_1)}{p_t(\mathbf{x}_t|\mathbf{x}_1)} \right]. \tag{40}$$

By Lemma 9:

$$\frac{\nabla^2 p_t(\mathbf{x}_t)}{p_t(\mathbf{x}_t)} = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_t)} \left[-\frac{\mathbf{I}}{\sigma_t^2} + \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top}{\sigma_t^2} \right] = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{x}_t]}{\sigma_t^2}. \tag{41}$$

By Fisher's identity (Lemma 6), $\nabla \log p_t(\mathbf{x}_t) = \hat{s}_t^*(\mathbf{x}_t)$. Therefore:

$$\nabla^2 \log p_t(\mathbf{x}_t) = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\mathbf{x}_t]}{\sigma_t^2} - \hat{s}_t^*(\mathbf{x}_t)\hat{s}_t^*(\mathbf{x}_t)^\top. \tag{42}$$

□

A.5 The Tweedie Connection: Posterior Covariance from Score Jacobian

Theorem 11 (Generalized Tweedie Identity for Posterior Covariance)**.** *For the forward process $\mathbf{x}_t = \mu_t + \sigma_t \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditional on $(\mathbf{x}_0, \mathbf{x}_1)$, the posterior of $\boldsymbol{\epsilon}$ given $\mathbf{x}_t$ satisfies the score-based identities*

$$\mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; -\sigma_t \, \hat{s}_t^*(\mathbf{x}_t), \qquad \mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; \mathbf{I} + \sigma_t^2 \, \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t). \tag{43}$$

Proof. We track every power of σ_t explicitly.

Step 1: posterior mean of $\boldsymbol{\epsilon}$. The classical Tweedie identity for $\mathbf{x}_t = \mu_t + \sigma_t \boldsymbol{\epsilon}$ with Gaussian noise reads

$$\mathbb{E}[\mu_t \,|\, \mathbf{x}_t] \;=\; \mathbf{x}_t \;+\; \sigma_t^2 \, \nabla \log p_t(\mathbf{x}_t). \tag{44}$$

Since $\boldsymbol{\epsilon} = (\mathbf{x}_t - \mu_t)/\sigma_t$ and conditional expectation is linear,

$$\mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; \frac{\mathbf{x}_t - \mathbb{E}[\mu_t \,|\, \mathbf{x}_t]}{\sigma_t} \;=\; \frac{-\sigma_t^2 \, \nabla \log p_t(\mathbf{x}_t)}{\sigma_t} \;=\; -\sigma_t \, \hat{s}_t^*(\mathbf{x}_t). \tag{45}$$

Step 2: differentiate $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{x}_t]$ in $\mathbf{x}_t$. Directly from equation 45,

$$\nabla_{\mathbf{x}} \mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; -\sigma_t \, \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t). \tag{46}$$

Step 3: Stein–Tweedie identity for $\nabla_{\mathbf{x}} \mathbb{E}[\boldsymbol{\epsilon}|\mathbf{x}_t]$. For the additive Gaussian observation model $\mathbf{x}_t = \mu_t + \sigma_t \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and any prior $p(\mu_t)$, integration-by-parts (Stein's lemma) applied to the posterior $p(\mu_t \,|\, \mathbf{x}_t) \propto \mathcal{N}(\mathbf{x}_t; \mu_t, \sigma_t^2 \mathbf{I})\, p(\mu_t)$ gives, for any sufficiently regular function f,

$$\nabla_{\mathbf{x}} \mathbb{E}[f(\mu_t) \,|\, \mathbf{x}_t] \;=\; \frac{1}{\sigma_t^2} \, \mathrm{Cov}\big[f(\mu_t),\, \mu_t \,\big|\, \mathbf{x}_t\big]. \tag{47}$$

Specializing to $f(\mu_t) = \mu_t$ yields $\nabla_{\mathbf{x}} \mathbb{E}[\mu_t \,|\, \mathbf{x}_t] = \mathrm{Cov}[\mu_t \,|\, \mathbf{x}_t]/\sigma_t^2$. Since $\boldsymbol{\epsilon} = (\mathbf{x}_t - \mu_t)/\sigma_t$ and $\mathbf{x}_t$ is constant given $\mathbf{x}_t$, we have $\mathrm{Cov}[\mu_t \,|\, \mathbf{x}_t] = \sigma_t^2 \, \mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t]$. Therefore

$$\nabla_{\mathbf{x}} \mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; \nabla_{\mathbf{x}} \Big(\frac{\mathbf{x}_t - \mathbb{E}[\mu_t \,|\, \mathbf{x}_t]}{\sigma_t} \Big) \;=\; \frac{1}{\sigma_t} (\mathbf{I} - \nabla_{\mathbf{x}} \mathbb{E}[\mu_t \,|\, \mathbf{x}_t]) \;=\; \frac{1}{\sigma_t} \Big(\mathbf{I} - \mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \Big). \tag{48}$$

The single $1/\sigma_t$ factor on the right comes from $\partial\boldsymbol{\epsilon}/\partial\mathbf{x}_t = \mathbf{I}/\sigma_t$; the Gaussian-precision factor $1/\sigma_t^2$ from Stein's identity has already been absorbed when converting $\mathrm{Cov}[\mu_t \,|\, \mathbf{x}_t]$ into $\sigma_t^2 \,\mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t]$.

Sanity check (Gaussian prior). For $\mu_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the posterior is $\mu_t \,|\, \mathbf{x}_t \sim \mathcal{N}\big(\mathbf{x}_t/(1+\sigma_t^2),\, \sigma_t^2/(1+\sigma_t^2)\,\mathbf{I}\big)$, so $\mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] = \sigma_t \mathbf{x}_t/(1+\sigma_t^2)$ and $\mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] = \mathbf{I}/(1+\sigma_t^2)$. Both equation 46 and equation 48 reduce to the consistent identity $\nabla_{\mathbf{x}} \mathbb{E}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] = \sigma_t\, \mathbf{I}/(1+\sigma_t^2)$, with the eventual conclusion $\mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] = \mathbf{I} + \sigma_t^2\, \nabla_{\mathbf{x}} \hat{s}_t^*$ recovering $\mathbf{I}/(1+\sigma_t^2)$ from $\nabla_{\mathbf{x}} \hat{s}_t^* = -\mathbf{I}/(1+\sigma_t^2)$.

Step 4: combine equation 46 and equation 48. Equating the two expressions for $\nabla_{\mathbf{x}} \mathbb{E}[\boldsymbol{\epsilon} | \mathbf{x}_t]$,

$$-\sigma_t\, \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t) \;=\; \frac{1}{\sigma_t}\Big(\mathbf{I} - \mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t]\Big). \tag{49}$$

Multiplying both sides by σ_t and solving for the covariance,

$$\mathrm{Cov}[\boldsymbol{\epsilon} \,|\, \mathbf{x}_t] \;=\; \mathbf{I} + \sigma_t^2\, \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t). \tag{50}$$

□

A.6 Main Result: Marginal Hessian Equals Score Jacobian

Theorem 12 (Marginal Hessian from Score Jacobian)**.** *The marginal Hessian of the log-density equals the Jacobian of the score:*

$$\nabla^2 \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t). \tag{51}$$

Proof. From Theorem 10:

$$\nabla^2 \log p_t(\mathbf{x}_t) = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top | \mathbf{x}_t]}{\sigma_t^2} - \hat{s}_t^* \hat{s}_t^{*\top}. \tag{52}$$

Using the covariance decomposition:

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top | \mathbf{x}_t] = \mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t] + \mathbb{E}[\boldsymbol{\epsilon} | \mathbf{x}_t] \mathbb{E}[\boldsymbol{\epsilon} | \mathbf{x}_t]^\top. \tag{53}$$

From the proof of Theorem 11, $\mathbb{E}[\boldsymbol{\epsilon} | \mathbf{x}_t] = -\sigma_t \hat{s}_t^*(\mathbf{x}_t)$, so:

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top | \mathbf{x}_t] = \mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t] + \sigma_t^2 \hat{s}_t^* \hat{s}_t^{*\top}. \tag{54}$$

Substituting:

$$\nabla^2 \log p_t(\mathbf{x}_t) = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t] + \sigma_t^2 \hat{s}_t^* \hat{s}_t^{*\top}}{\sigma_t^2} - \hat{s}_t^* \hat{s}_t^{*\top} \tag{55}$$

$$= -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t]}{\sigma_t^2} + \hat{s}_t^* \hat{s}_t^{*\top} - \hat{s}_t^* \hat{s}_t^{*\top} \tag{56}$$

$$= -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t]}{\sigma_t^2}. \tag{57}$$

By Theorem 11, $\mathrm{Cov}[\boldsymbol{\epsilon} | \mathbf{x}_t] = \mathbf{I} + \sigma_t^2 \nabla_{\mathbf{x}} \hat{s}_t^*$:

$$\nabla^2 \log p_t(\mathbf{x}_t) = -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbf{I} + \sigma_t^2 \nabla_{\mathbf{x}} \hat{s}_t^*}{\sigma_t^2} \tag{58}$$

$$= -\frac{\mathbf{I}}{\sigma_t^2} + \frac{\mathbf{I}}{\sigma_t^2} + \nabla_{\mathbf{x}} \hat{s}_t^* \tag{59}$$

$$= \nabla_{\mathbf{x}} \hat{s}_t^*(\mathbf{x}_t). \tag{60}$$

□

A.7 Extension to Higher Orders

Theorem 13 (K-th Order Marginal Derivative – Chain-Rule Re-parameterization)**.** *For any $k \geq 1$:*

$$\nabla^k \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*(\mathbf{x}_t). \tag{61}$$

This is a re-parameterization identity: once Fisher's identity grants $\hat{s}_t^* = \nabla \log p_t$, it is the chain rule applied $k-1$ further times. It enables computing higher-order marginal derivatives from the optimal score via autodiff, but does not *close a new conditional-marginal gap beyond the $k = 1$ closure given by Fisher's identity.*

Proof. By induction on k. **Base case** ($k = 1$): Fisher's identity gives $\nabla \log p_t = \hat{s}_t^* = \nabla_{\mathbf{x}}^0 \hat{s}_t^*$ (Lemma 6). **Inductive step**: if $\nabla^k \log p_t = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*$, then differentiating both sides in $\mathbf{x}$ yields $\nabla^{k+1} \log p_t = \nabla_{\mathbf{x}}(\nabla^k \log p_t) = \nabla_{\mathbf{x}}(\nabla_{\mathbf{x}}^{k-1} \hat{s}_t^*) = \nabla_{\mathbf{x}}^k \hat{s}_t^*$. By induction the result holds for all $k \geq 1$. □

The whole content of Theorem 13 is the chain rule. For a learned score $\hat{s}_t \approx \hat{s}_t^*$, the approximation $\nabla_{\mathbf{x}}^{k-1} \hat{s}_t \approx \nabla^k \log p_t$ inherits error from score estimation, with potential amplification through derivatives (see Section A.10). The alternative Tweedie-style proof of the $k = 2$ special case provided in Sections A.3–A.6 above is logically equivalent and is included only for self-containedness; it is not used by the IDFF sampler.

A.8 Fokker-Planck Analysis for IDFF

We analyze marginal preservation separately for $K = 1$ and $K \geq 2$. The $K = 1$ analysis splits into a clean exact case (Theorem 14, default schedule $\gamma_t^0 \equiv 1$) and a bounded-deviation variant (Proposition 15, convex-constrained schedule $\gamma_t^0 + \gamma_t^1 = 1$). The $K \geq 2$ case is a Girsanov→Pinsker total-variation bound on top of the exact $K = 1$ baseline.

Theorem 14 (Exact Marginal Preservation for $K = 1$ with $\gamma_t^0 \equiv 1$)**.** *Let Assumption 1 hold. Consider the $K = 1$ IDFF SDE on $[0, 1]$*

$$d\mathbf{x}_t \;=\; \tilde{\mathbf{v}}_t(\mathbf{x}_t)\, dt + \sigma_t\, d\mathbf{w}_t, \qquad \tilde{\mathbf{v}}_t \;=\; \mathbf{v}_t + \gamma_t^1\, \hat{s}_t^*, \qquad \sigma_t^2 \;=\; 2\gamma_t^1, \tag{62}$$

with $\mathbf{x}_0 \sim p_0$, optimal score $\hat{s}_t^* = \nabla \log p_t$, and any non-negative C^1 schedule γ_t^1 vanishing at $t \in \{0, 1\}$. Then the marginal of $\mathbf{x}_t$ equals the CFM marginal exactly:

$$\tilde{p}_t(\mathbf{x}) \;=\; p_t(\mathbf{x}) \qquad \forall\, t \in [0, 1]. \tag{63}$$

In particular, $\tilde{p}_1 = p_1$ exactly in continuous time, with no time reparameterization.

Proof. The Fokker–Planck equation for the density $\tilde{p}_t$ associated with the SDE equation 62 is

$$\frac{\partial \tilde{p}_t}{\partial t} \;=\; -\nabla \cdot (\tilde{\mathbf{v}}_t\, \tilde{p}_t) + \frac{\sigma_t^2}{2} \Delta \tilde{p}_t. \tag{64}$$

We verify that the ansatz $\tilde{p}_t = p_t$ (the CFM marginal at physical time t) solves equation 64. Substituting $\tilde{\mathbf{v}}_t = \mathbf{v}_t + \gamma_t^1 \nabla \log p_t$ and $\sigma_t^2 = 2\gamma_t^1$ on the right-hand side,

$$-\nabla \cdot (\tilde{\mathbf{v}}_t p_t) + \frac{\sigma_t^2}{2} \Delta p_t = -\nabla \cdot (\mathbf{v}_t p_t) \;-\; \gamma_t^1\, \nabla \cdot \big(p_t\, \nabla \log p_t\big) \;+\; \gamma_t^1\, \Delta p_t.$$

The crucial cancellation uses the elementary identity

$$\nabla \cdot \big(p_t\, \nabla \log p_t\big) \;=\; \nabla \cdot (\nabla p_t) \;=\; \Delta p_t, \tag{65}$$

which holds because $p_t\, \nabla \log p_t = \nabla p_t$. This is the *only* place where the substitution $\tilde{p}_t \mapsto p_t$ matters, and the cancellation is consistent: the score in the drift, $\nabla \log p_t$, matches $\nabla \log \tilde{p}_t$ at the ansatz. Therefore the last two terms above cancel exactly, leaving

$$\frac{\partial \tilde{p}_t}{\partial t} \;=\; -\nabla \cdot (\mathbf{v}_t p_t), \tag{66}$$

which is exactly the CFM continuity equation $\partial_t p_t = -\nabla \cdot (\mathbf{v}_t p_t)$. Hence $\tilde{p}_t = p_t$ solves equation 64 with the matching initial condition $\tilde{p}_0 = p_0$.

Uniqueness on the truncated interval. Fix any $\delta \in (0, 1/2)$ and consider the linear Cauchy problem equation 64 on $[0, 1-\delta] \times \mathbb{R}^d$. On this interval the drift $\mathbf{v}_t + \gamma_t^1 \nabla \log p_t$ depends only on $(t, \mathbf{x})$ (not on $\tilde{p}_t$); under Assumption 1(R1)–(R3) all coefficients are bounded with constants $M_j(\delta), L_v(\delta), M_v(\delta)$ and the diffusion is uniformly non-degenerate $\sigma_t^2 = 2\gamma_t^1 \geq 2\gamma_{\min}(\delta) > 0$. Standard parabolic theory (Friedman, 1964, Ch. 1, Thm. 12) therefore gives a *unique* bounded classical solution with initial datum $\tilde{p}_0 = p_0$, namely $\tilde{p}_t = p_t$ for $t \in [0, 1-\delta]$. As $\delta \to 0$ the truncated solutions are consistent on overlapping intervals, so $\tilde{p}_t = p_t$ for every $t \in [0, 1)$ by exhaustion. The remaining endpoint $t = 1$ is handled by continuity of p_t in $L^1(\mathbb{R}^d)$ along the CFM path (Bogachev et al., 2015, Ch. 6): the practical sampler reads off the time-$(1-\delta)$ marginal and applies the denoiser $\hat{\mathbf{x}}_1$ to land on p_1 (Section 3.4). The full continuous-time identity $\tilde{p}_t = p_t$ on $[0, 1]$ therefore holds in the same sense as the CFM continuity equation itself: as the unique bounded solution of the linear parabolic Cauchy problem on the truncated interval, extended to the endpoint by the denoiser. We deliberately avoid invoking non-degenerate parabolic uniqueness up to $t = 1$, since $\sigma_t^2 = 2\gamma_t^1 = 8\bar{\gamma}\, t(1-t)$ degenerates exactly there; the truncation makes the boundary handling explicit. □

Remark 2 (Why the slowdown $\gamma_t^0 < 1$ destroys exactness). If one instead used the convex-constrained drift $\gamma_t^0 \mathbf{v}_t + \gamma_t^1 \hat{s}_t^*$ with $\gamma_t^0 = 1 - \gamma_t^1 < 1$, the same calculation gives $\partial_t \tilde{p}_t = -\gamma_t^0 \nabla \cdot (\mathbf{v}_t p_t) = \gamma_t^0\, \partial_t p_t$ when one substitutes the ansatz $\tilde{p}_t = p_t$. This forces $(1 - \gamma_t^0)\, \partial_t p_t = 0$, which fails generically since $\partial_t p_t \neq 0$. The natural attempt to repair this via a time-reparameterized ansatz $\tilde{p}_t = p_{\tau(t)}$ with $\tau(t) = \int_0^t \gamma_s^0\, ds$ also fails, because the score in the drift is $\nabla \log p_t$, not $\nabla \log p_{\tau(t)}$, so the cancellation in equation 65 no longer goes through. The convex constraint therefore trades exact preservation for a small but quantifiable Girsanov gap, which we bound next.

Proposition 15 (Convex-Constrained $K = 1$: Bounded TV Deviation)**.** *Under the convex-constrained schedule $\gamma_t^0 + \gamma_t^1 = 1$ with $\gamma_t^1 \geq 0$ vanishing at the endpoints, let $\widetilde{\mathbb{P}}_{\text{cvx}}$ denote the law on path space of the SDE $d\mathbf{x}_t = [\gamma_t^0 \mathbf{v}_t + \gamma_t^1 \hat{s}_t^*]\, dt + \sigma_t\, d\mathbf{w}_t$ with $\sigma_t^2 = 2\gamma_t^1$, and let $\mathbb{P}^\star$ denote the law of the exact-preserving SDE equation 62. Since both SDEs share the diffusion σ_t and initial law p_0, Girsanov's theorem gives*

$$\mathrm{KL}\big(\widetilde{\mathbb{P}}_{\text{cvx}} \,\|\, \mathbb{P}^\star\big) \;=\; \frac{1}{2}\int_0^1 \mathbb{E}_{\tilde{p}_t}\Big[\big\|\sigma_t^{-1}(\gamma_t^0 - 1)\, \mathbf{v}_t\big\|^2\Big]\, dt \;=\; \frac{1}{4}\int_0^1 \gamma_t^1\, \mathbb{E}_{\tilde{p}_t}\|\mathbf{v}_t\|^2\, dt, \tag{67}$$

since $\gamma_t^0 - 1 = -\gamma_t^1$ and $\sigma_t^2 = 2\gamma_t^1$. Pinsker's inequality on the time-1 marginals then yields

$$\mathrm{TV}\big(\tilde{p}_1^{\text{cvx}},\, p_1\big) \;\leq\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\widetilde{\mathbb{P}}_{\text{cvx}} \,\|\, \mathbb{P}^\star)} \;=\; \frac{1}{2\sqrt{2}}\Big(\int_0^1 \gamma_t^1\, \mathbb{E}_{\tilde{p}_t}\|\mathbf{v}_t\|^2\, dt\Big)^{1/2}. \tag{68}$$

Proof. The two SDEs differ only in drift, by $\Delta\mathbf{b}_t := (\gamma_t^0 - 1)\mathbf{v}_t = -\gamma_t^1 \mathbf{v}_t$, and share the diffusion coefficient σ_t. Substituting $\sigma_t^2 = 2\gamma_t^1$ gives the pointwise identity $\|\sigma_t^{-1}\Delta\mathbf{b}_t\|^2 = (\gamma_t^1)^2\|\mathbf{v}_t\|^2/(2\gamma_t^1) = \frac{1}{2}\gamma_t^1\|\mathbf{v}_t\|^2$. Under R2 ($\|\mathbf{v}_t\|_{L^\infty} \leq M_v(\delta)$ on $[0, 1-\delta]$) and R3 ($\gamma_t^1 \in C^1$, integrable on $[0, 1]$), the integral $\frac{1}{2}\int_0^{1-\delta} \|\sigma_t^{-1}\Delta\mathbf{b}_t\|^2\, dt \leq \frac{M_v(\delta)^2}{4}\int_0^{1-\delta} \gamma_t^1\, dt < \infty$ is a deterministic finite constant for every fixed $\delta > 0$, so Novikov's condition holds trivially and Girsanov's theorem (Chen et al., 2023; Benton et al., 2024) yields the path-KL on the right-hand side of equation 67. As in Theorem 14, the bound is taken on the truncated interval $[0, 1-\delta]$ and extended to the endpoint by the denoiser; since γ_t^1 is integrable, the schedule integral $\int_0^{1-\delta} \gamma_t^1\, dt$ has a finite limit as $\delta \to 0$ (R3), and the displayed $[0, 1]$ integrals in equation 67–equation 68 are understood as these $\delta \to 0$ limits. Pinsker's inequality $\mathrm{TV} \leq \sqrt{\mathrm{KL}/2}$ (Tsybakov, 2009, Lem. 2.5) together with the data-processing inequality $\mathrm{KL}(\tilde{p}_1^{\text{cvx}} \,\|\, p_1^\star) \leq \mathrm{KL}(\widetilde{\mathbb{P}}_{\text{cvx}} \,\|\, \mathbb{P}^\star)$ and Theorem 14 ($p_1^\star = p_1$) yield equation 68. □

Corollary 16 (Schedule-Dependent KL Constants)**.** *Under uniform bound $\sup_{t,\mathbf{x}} \|\mathbf{v}_t(\mathbf{x})\|^2 \leq M_v^2$ (consequence of R2 plus boundedness of $\mathbf{v}_t$ on a compact set, or stationary moments), equation 67 simplifies to $\mathrm{KL} \leq \frac{M_v^2}{4}\int_0^1 \gamma_t^1\, dt$. For the schedules used in this paper:*

- ***Constant-amplitude Beta(2,2):*** *$\gamma_t^1 = 4\bar{\gamma}\, t(1-t)$ gives $\int_0^1 \gamma_t^1\, dt = \frac{2\bar{\gamma}}{3}$.*

- ***Triangular:*** $\gamma_t^1 = \bar{\gamma}\,\min(t, 1-t)$ *gives* $\int_0^1 \gamma_t^1\,dt = \frac{\bar{\gamma}}{4}$.
- ***Sine:*** $\gamma_t^1 = \bar{\gamma}\sin(\pi t)$ *gives* $\int_0^1 \gamma_t^1\,dt = \frac{2\bar{\gamma}}{\pi}$.

In each case $\mathrm{TV}\big(\tilde{p}_1^{\mathrm{cvx}}, p_1\big) = O(\sqrt{\bar{\gamma}})$ *as* $\bar{\gamma} \to 0$.

Practical implications. Theorem 14 is the clean reference result: the IDFF default $\gamma_t^0 \equiv 1$ (just Langevin enhancement on top of CFM, no slowdown of $\mathbf{v}_t$) preserves the CFM marginal exactly in continuous time. Proposition 15 and Corollary 16 make the convex-constrained variant ($\gamma_t^0 + \gamma_t^1 = 1$, where $\mathbf{v}_t$ is also slowed by γ_t^1) honestly quantifiable: the path-KL and endpoint TV are explicit functions of the schedule amplitude $\bar{\gamma}$, vanishing in the limit $\bar{\gamma} \to 0$. In all reported experiments we use $\gamma_t^0 \equiv 1$, so the exact-preservation guarantee of Theorem 14 applies. Two further sources of error remain in any practical implementation: (i) Euler–Maruyama discretization with step Δt adds $O(\Delta t)$ pathwise error, and (ii) using a learned $\hat{s}_t$ in place of $\hat{s}_t^*$ adds a Girsanov bias controlled by $\|\hat{s}_t - \hat{s}_t^*\|_{L^2(p_t)}$ (made quantitative in Theorem 17, Part 3).

Theorem 17 (Quantified Deviation via Girsanov → Pinsker, $K \geq 1$)**.** *Let Assumption 1 hold and fix any order* $K \geq 1$*. Consider the K-th order IDFF drift with default schedule* $\gamma_t^0 \equiv 1$ *and diffusion compensation* $\sigma_t^2 = 2\gamma_t^1 > 0$ *on the interior* $(0, 1)$*:*

$$\tilde{\mathbf{v}}_t(\mathbf{x}_t) \;=\; \mathbf{v}_t(\mathbf{x}_t) \;+\; \gamma_t^1\,\hat{s}_t(\mathbf{x}_t) \;+\; \sum_{k=2}^{K} \gamma_t^k\,\mathbf{m}_t^{(k)}(\mathbf{x}_t), \tag{69}$$

where $\hat{s}_t$ *is the* learned *score,* $\mathbf{m}_t^{(k)} = (\nabla_{\mathbf{x}}^{k-1}\hat{s}_t)[\mathbf{u}_t]^{k-1}$ *for* $k \geq 2$*,* $\mathbf{u}_t = \mathbf{v}_t/\|\mathbf{v}_t\|$*, and the higher-order sum is empty when* $K = 1$*. (This is the* $\gamma_t^0 \equiv 1$ *schedule of Theorem 14 extended with higher-order momentum; the convex-constrained variant simply adds the* $O(\sqrt{\bar{\gamma}})$ *TV penalty of Proposition 15.) Let* $\widetilde{\mathbb{P}}$ *denote the law on* $C([0,1];\mathbb{R}^d)$ *of the augmented SDE* $d\mathbf{x}_t = \tilde{\mathbf{v}}_t\,dt + \sigma_t\,d\mathbf{w}_t$ *with* $\mathbf{x}_0 \sim p_0$*, and let* $\mathbb{P}^\star$ *denote the law of the exact-preserving* $K = 1$ *optimal-score SDE equation 62. Write the corresponding time-t marginals as* $\tilde{p}_t$ *and* $p_t^\star = p_t$ *(the second equality by Theorem 14). At* $K = 1$ *this measures only the score-estimation deviation; at* $K \geq 2$ *a structural residual is added on top. Then:*

1. ***Residual decomposition.*** *Setting the exact-preserving baseline drift to* $\mathbf{w}_t^\star := \mathbf{v}_t + \gamma_t^1\,\hat{s}_t^*$*, the deviation decomposes as*

$$\tilde{\mathbf{v}}_t - \mathbf{w}_t^\star \;=\; \underbrace{\gamma_t^1\big(\hat{s}_t - \hat{s}_t^*\big)}_{\textit{score-estimation residual}} \;+\; \underbrace{\sum_{k=2}^{K} \gamma_t^k\,\mathbf{m}_t^{(k)}}_{\textit{structural residual}}. \tag{70}$$

   The score-estimation residual is the same *bias present in every score-based model; the structural residual is unique to* $K \geq 2$ *and would be present even with a perfect score.*

2. ***Path-KL via Girsanov.*** *Because* $\sigma_t > 0$ *on the interior and the two SDEs share diffusion* σ_t *and initial law* p_0*, Girsanov's theorem applies pathwise:*

$$\mathrm{KL}\big(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star\big) \;=\; \frac{1}{2}\int_0^1 \mathbb{E}_{\tilde{p}_t}\Big[\big\|\sigma_t^{-1}\big(\tilde{\mathbf{v}}_t - \mathbf{w}_t^\star\big)\big\|^2\Big]\,dt. \tag{71}$$

   Applying $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ *to the decomposition equation 70 and substituting* $\sigma_t^2 = 2\gamma_t^1$ *gives*

$$\mathrm{KL}\big(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star\big) \;\leq\; \underbrace{\frac{1}{2}\int_0^1 \gamma_t^1\,\mathbb{E}_{\tilde{p}_t}\big\|\hat{s}_t - \hat{s}_t^*\big\|^2\,dt}_{=:\,\mathcal{E}_{\mathrm{score}}} \;+\; \underbrace{\frac{1}{2}\int_0^1 \frac{1}{\gamma_t^1}\,\mathbb{E}_{\tilde{p}_t}\Big\|\sum_{k=2}^{K}\gamma_t^k\,\mathbf{m}_t^{(k)}\Big\|^2\,dt}_{=:\,\mathcal{E}_{\mathrm{struct}}}. \tag{72}$$

   The score-estimation piece $\mathcal{E}_{\mathrm{score}}$ *is, up to the schedule weight* γ_t^1*, exactly the weighted* L^2 *training objective equation 21 (score head) and therefore controlled by training. More precisely, the Beta(2,2)*

weight matches the score-matching weight $\lambda_1^2 = \sigma_t^2$ in equation 21 via $\gamma_t^1 = (4\bar{\gamma}/\sigma_0^2)\,\sigma_t^2$, so $\mathcal{E}_{\text{score}} = (2\bar{\gamma}/\sigma_0^2)\int_0^1 \sigma_t^2\,\mathbb{E}_{\tilde{p}_t}\|\hat{s}_t - \hat{s}_t^*\|^2\,dt$; at the $K = 1$ baseline $\tilde{p}_t = p_t$ (Theorem 14) this equals the excess training loss $\mathcal{L}_{\text{score}}^{\text{exc}}$ of Assumption 1(R1′) up to the constant $2\bar{\gamma}/\sigma_0^2$, and for $K \geq 2$ the $\tilde{p}_t \leftrightarrow p_t$ change of measure is absorbed into the TV ball bounded here. In particular $\mathcal{E}_{\text{score}} = O(\bar{\gamma})$ uniformly in the truncation δ *(Remark 1).*

3. ***Pinsker → TV bound.*** *The data-processing inequality* $\mathrm{KL}(\tilde{p}_1 \,\|\, p_1^\star) \leq \mathrm{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star)$ *and Pinsker's inequality* $\mathrm{TV} \leq \sqrt{\mathrm{KL}/2}$*, together with $p_1^\star = p_1$, give*

$$\mathrm{TV}(\tilde{p}_1, p_1) \;=\; \mathrm{TV}(\tilde{p}_1, p_1^\star) \;\leq\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}\big(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star\big)} \;\leq\; \sqrt{\tfrac{1}{2}\big(\mathcal{E}_{\text{score}} + \mathcal{E}_{\text{struct}}\big)}. \tag{73}$$

4. ***Coefficient-schedule bound on the structural residual.*** *R1 bounds the tensor operator norm $\|\nabla_{\mathbf{x}}^{k-1}\hat{s}_t\|_{\text{op}} \leq M_{k-1}(\delta)$ on $[0, 1-\delta]$ (the assumption explicitly covers the* learned *score, not just $\hat{s}_t^*$). Since $\|\mathbf{u}_t\| = 1$, the contracted momentum used by the sampler satisfies*

$$\|\mathbf{m}_t^{(k)}\| \;=\; \big\|\big(\nabla_{\mathbf{x}}^{k-1}\hat{s}_t\big)[\mathbf{u}_t]^{k-1}\big\| \;\leq\; \big\|\nabla_{\mathbf{x}}^{k-1}\hat{s}_t\big\|_{\text{op}}\,\|\mathbf{u}_t\|^{k-1} \;\leq\; M_{k-1}(\delta)$$

uniformly in $(t, \mathbf{x})$ on $[0, 1-\delta]$. Under the recommended scaling $\gamma_t^k = \alpha_k\,\gamma_t^1\,\Delta t^{k-1}$ for $k \geq 2$, Cauchy–Schwarz gives

$$\Big\|\sum_{k=2}^{K}\gamma_t^k\,\mathbf{m}_t^{(k)}\Big\|^2 \;\leq\; (K-1)\sum_{k=2}^{K}(\gamma_t^k)^2\|\mathbf{m}_t^{(k)}\|^2 \;\leq\; (K-1)\sum_{k=2}^{K}\alpha_k^2 M_{k-1}(\delta)^2\,(\gamma_t^1)^2\,\Delta t^{2(k-1)}.$$

Dividing by γ_t^1, integrating, and taking the leading $k = 2$ term yields

$$\mathcal{E}_{\text{struct}} \;\leq\; \frac{K-1}{2}\int_0^1 \gamma_t^1\left(\sum_{k=2}^{K}\alpha_k^2 M_{k-1}(\delta)^2\,\Delta t^{2(k-1)}\right)dt \;=\; O(\Delta t^2). \tag{74}$$

The factor $\gamma_t^1 = O(t(1-t))$ multiplying the $(\gamma_t^1)^{-1}$ in $\mathcal{E}_{\text{struct}}$ cancels the schedule singularity (R3), so the integrand is integrable and the bound is finite for each fixed δ; with $\delta = \Delta t/2$ the leading $k = 2$ term is $O(\Delta t^2)$. Combined with equation 73,

$$\mathrm{TV}(\tilde{p}_1, p_1) \;\leq\; \sqrt{\tfrac{1}{2}\,\mathcal{E}_{\text{score}}} \;+\; O(\Delta t). \tag{75}$$

Proof. Part 1. Algebraic: subtract $\mathbf{w}_t^\star = \mathbf{v}_t + \gamma_t^1\hat{s}_t^*$ from $\tilde{\mathbf{v}}_t$ and group the score-error and higher-order terms.

Part 2. The two SDEs share diffusion σ_t and initial law p_0. By Girsanov's theorem on path space (Chen et al., 2023; Benton et al., 2024), provided Novikov's condition $\mathbb{E}\exp\big(\frac{1}{2}\int_0^{1-\delta}\|\sigma_t^{-1}(\tilde{\mathbf{v}}_t - \mathbf{w}_t^\star)\|^2\,dt\big) < \infty$ holds, the relative entropy of the path measures is given by equation 71.

We verify Novikov's condition explicitly under the recommended scaling $\gamma_t^k = \alpha_k\,\gamma_t^1\,\Delta t^{k-1}$ for $k \geq 2$. Substituting $\sigma_t^2 = 2\gamma_t^1$, the score-estimation residual contributes

$$\big\|\sigma_t^{-1}\,\gamma_t^1\big(\hat{s}_t - \hat{s}_t^*\big)\big\|^2 \;=\; \frac{(\gamma_t^1)^2}{2\gamma_t^1}\,\big\|\hat{s}_t - \hat{s}_t^*\big\|^2 \;=\; \tfrac{1}{2}\,\gamma_t^1\,\big\|\hat{s}_t - \hat{s}_t^*\big\|^2,$$

and the structural residual contributes, using R1 ($\|\mathbf{m}_t^{(k)}\|_{L^\infty} \leq M_{k-1}(\delta)$ on $[0, 1-\delta]$) and the recommended scaling,

$$\Big\|\sigma_t^{-1}\sum_{k=2}^{K}\gamma_t^k\,\mathbf{m}_t^{(k)}\Big\|^2 \;\leq\; \frac{K-1}{2\gamma_t^1}\sum_{k=2}^{K}(\gamma_t^k)^2 M_{k-1}(\delta)^2 \;=\; \tfrac{K-1}{2}\,\gamma_t^1\sum_{k=2}^{K}\alpha_k^2 M_{k-1}(\delta)^2\,\Delta t^{2(k-1)}.$$

Novikov finiteness versus the size of $\mathcal{E}_{\text{score}}$ — distinct hypotheses. On the truncated interval $[0, 1-\delta]$ both displayed terms are pointwise bounded by $C(\delta)\,\gamma_t^1$, where $C(\delta)$ depends only on the R1 *magnitude* constants

$\{M_{k-1}(\delta)\}$ (and may diverge as $\delta \to 0$). Since $\gamma_t^1 \in C^1$ and is integrable, $\frac{1}{2}\int_0^{1-\delta} \|\sigma_t^{-1}(\tilde{\mathbf{v}}_t - \mathbf{w}_t^\star)\|^2\, dt \le \frac{C(\delta)}{2}\int_0^{1-\delta} \gamma_t^1\, dt < \infty$ is a deterministic constant *for every fixed* $\delta > 0$, so Novikov's condition $\mathbb{E}\exp(\cdots) < \infty$ holds and Girsanov's theorem applies on $[0, 1-\delta]$; as in Theorem 14, the exact-preserving $K = 1$ baseline $\mathbb{P}^\star$ is extended to the endpoint by the denoiser. The bound equation 72 then follows from $\|a+b\|^2 \le 2(\|a\|^2 + \|b\|^2)$ and $\sigma_t^2 = 2\gamma_t^1$.

We emphasise that Novikov finiteness and the *size* of $\mathcal{E}_{\text{score}}$ are governed by *separate* hypotheses. Novikov finiteness (above) may legitimately use the δ-dependent magnitude constants $M_j(\delta)$, since it only requires the integral to be finite at each fixed δ. The size of the score term in the final bound, by contrast, must *not* depend on $M_0(\delta)$: by the finite-excess-loss hypothesis Assumption 1(R1′) and the weight-matching identity $\gamma_t^1 = (4\bar{\gamma}/\sigma_0^2)\sigma_t^2$ (Remark 1),

$$\mathcal{E}_{\text{score}} = \tfrac{1}{2}\int_0^1 \gamma_t^1\, \mathbb{E}\|\hat{s}_t - \hat{s}_t^*\|^2\, dt = \tfrac{2\bar{\gamma}}{\sigma_0^2}\int_0^1 \sigma_t^2\, \mathbb{E}\|\hat{s}_t - \hat{s}_t^*\|^2\, dt \;\le\; \tfrac{2\bar{\gamma}}{\sigma_0^2}\, \mathcal{L}_{\text{score}}^{\text{exc}} \;=\; O(\bar{\gamma}),$$

independently of δ, because the weight σ_t^2 cancels the σ_t^{-2} growth of the score error pointwise (the integrand is the trained objective). Bounding the score error instead by the magnitude constant $\|\hat{s}_t - \hat{s}_t^*\| \le 2M_0(\delta) = O(1/\sqrt{\delta})$ would only yield the weaker $\mathcal{E}_{\text{score}} = O(\bar{\gamma}/\delta)$, which does *not* vanish under the joint limit $\delta = \Delta t/2 \to 0$ of Corollary 18; routing the score term through R1′ rather than through $M_0(\delta)$ is precisely what yields the δ-stable rate.

Part 3. The data-processing inequality gives $\mathrm{KL}(\tilde{p}_1 \,\|\, p_1^\star) \le \mathrm{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star)$. Pinsker's inequality $\mathrm{TV} \le \sqrt{\mathrm{KL}/2}$ (Tsybakov, 2009, Lem. 2.5) applied to the time-1 marginals, together with $p_1^\star = p_1$ from Theorem 14, yields equation 73.

Part 4. The Cauchy–Schwarz and uniform-bound steps are stated in the theorem; substituting $\gamma_t^k = \alpha_k \gamma_t^1 \Delta t^{k-1}$ gives the $O(\Delta t^{2(k-1)})$ scaling per k, with the leading $k = 2$ contribution dominating. The schedule factor $\gamma_t^1 = O(t(1-t))$ cancels the $(\gamma_t^1)^{-1}$ weight in $\mathcal{E}_{\text{struct}}$ (R3), so the structural integrand is integrable on $[0, 1-\delta]$ and the bound is δ-stable in the same sense as $\mathcal{E}_{\text{score}}$. □

Corollary 18 (Convergence as NFE $\to \infty$, $K \ge 1$)**.** *Fix any $K \ge 1$. Under the recommended schedule $\bar{\gamma} = O(\Delta t)$ for the $K = 1$ amplitude and (only meaningful for $K \ge 2$) $\gamma_t^k = \alpha_k \gamma_t^1 \Delta t^{k-1}$ for $k \ge 2$, if the weighted score error $\mathcal{E}_{\text{score}} \to 0$, then equation 75 gives*

$$\mathrm{TV}(\tilde{p}_1, p_1) \;\le\; \sqrt{\tfrac{1}{2}\,\mathcal{E}_{\text{score}}} \;+\; O(\Delta t) \tag{76}$$

in continuous time, where the $O(\Delta t)$ term vanishes at $K = 1$ and absorbs the leading $k = 2$ structural residual when $K \ge 2$. Because $\mathcal{E}_{\text{score}} = O(\bar{\gamma})$ independently *of δ (Assumption 1(R1′), Remark 1), taking $\Delta t \to 0$ with $\delta = \Delta t/2 \to 0$ leaves the leading $\sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}}$ term unaffected while $\mathcal{E}_{\text{struct}} = O(\Delta t^2) \to 0$; a subsequent $\bar{\gamma} \to 0$ then drives the bound to zero. Euler–Maruyama discretization adds at most an additional $O(\Delta t)$ pathwise error (standard SDE-discretization analysis under R2), so the overall endpoint TV remains $O(\sqrt{\mathcal{E}_{\text{score}}}) + O(\Delta t)$; if in addition $\mathcal{E}_{\text{score}} = O(1/\mathrm{NFE}^2)$ the rate is $O(1/\mathrm{NFE})$. In particular, $\mathrm{TV}(\tilde{p}_1, p_1) \to 0$ as $\mathrm{NFE} \to \infty$ for any fixed converged score network. The exact-preservation case $\mathrm{TV}(\tilde{p}_1, p_1) = 0$ at $K = 1$ with $\hat{s}_t = \hat{s}_t^*$ (Theorem 14) is recovered when both $\mathcal{E}_{\text{score}}$ and the discretization error vanish.*

Corollary 19 (Uniform-in-time TV bound, $K \ge 1$)**.** *Under the hypotheses of Theorem 17 (any $K \ge 1$), the endpoint bound equation 73 extends uniformly to every intermediate time:*

$$\sup_{t\in[0,1]} \mathrm{TV}\big(\tilde{p}_t,\, p_t\big) \;\le\; \sqrt{\tfrac{1}{2}\big(\mathcal{E}_{\text{score}} + \mathcal{E}_{\text{struct}}\big)}. \tag{77}$$

In particular, under the recommended scaling $\gamma_t^k = \alpha_k\, \gamma_t^1\, \Delta t^{k-1}$ for $k \ge 2$,

$$\sup_{t\in[0,1]} \mathrm{TV}\big(\tilde{p}_t,\, p_t\big) \;\le\; \sqrt{\tfrac{1}{2}\,\mathcal{E}_{\text{score}}} \;+\; O(\Delta t). \tag{78}$$

Hence not only the endpoint but the entire marginal trajectory of the augmented dynamics stays within a closed-form TV ball of the CFM marginals.

Proof. Let $\pi_t : C([0,1];\mathbb{R}^d) \to \mathbb{R}^d$ denote the time-t projection. The pushforwards are $(\pi_t)_*\widetilde{\mathbb{P}} = \tilde{p}_t$ and $(\pi_t)_*\mathbb{P}^\star = p_t^\star = p_t$ (the second equality by Theorem 14, applied to the $K = 1$ exact-preserving baseline used to define $\mathbb{P}^\star$). The data-processing inequality for KL divergence gives $\mathrm{KL}(\tilde{p}_t \,\|\, p_t) \le \mathrm{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star)$ for every $t \in [0,1]$. Combining with Pinsker's inequality $\mathrm{TV} \le \sqrt{\mathrm{KL}/2}$ and the path-KL bound $\mathrm{KL}(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}^\star) \le \mathcal{E}_{\mathrm{score}} + \mathcal{E}_{\mathrm{struct}}$ from equation 72 yields equation 77 uniformly in t. The scaled version follows from $\mathcal{E}_{\mathrm{struct}} = O(\Delta t^2)$ (Theorem 17, Part 4) and subadditivity of $\sqrt{\cdot}$. □

Remark 3 (Why Girsanov rather than transport-equation Grönwall). An L^1-of-divergence route via Grönwall would require a bound on $\nabla \cdot (p_t \mathbf{m}_t^{(k)})$, which in turn requires regularity assumptions on $\nabla \log p_t \cdot \mathbf{m}_t^{(k)}$ not stated up front. Because IDFF has a non-degenerate diffusion ($\sigma_t^2 = 2\gamma_t^1 > 0$ on the interior), Girsanov applies directly to the path measure, and the only quantity that ends up on the right-hand side is the weighted L^2 norm of the drift mismatch — which is exactly the quantity our training loss minimizes and which depends only on R1–R3 (and R1′). This route is the standard one in the modern flow-matching convergence literature (Chen et al., 2023; Benton et al., 2024).

A.9 Regularity Assumptions

The regularity assumption (Assumption 1, conditions R1–R3, together with the finite excess score-matching loss R1′) used throughout this appendix is stated in the main text at the top of Section 3, immediately before Theorem 2. In particular, the δ-independence of $\mathcal{E}_{\mathrm{score}}$ relied on in Theorem 17 and Corollary 18 follows from R1′ (finite excess score-matching loss $\mathcal{L}_{\mathrm{score}}^{\mathrm{exc}} < \infty$) via the weight-matching identity $\gamma_t^1 = (4\bar{\gamma}/\sigma_0^2)\sigma_t^2$ (Remark 1). We refer the reader there and use the assumption without restatement.

A.10 Approximation Error and Practical Considerations

Plugging in the optimal score $\hat{s}_t = \hat{s}_t^*$ removes the score-estimation residual ($\mathcal{E}_{\mathrm{score}} = 0$) from every bound in this appendix. It does *not* remove the structural residual of the $K \ge 2$ bound (which is controlled by the schedule scaling, Theorem 17 Part 4) nor the Girsanov gap of the convex-constrained $K = 1$ variant (Proposition 15). Only the $K = 1$ *default* schedule attains $\tilde{p}_t = p_t$ for all $t \in [0,1]$ with the optimal score. In practice, score estimation introduces errors that propagate to higher-order terms; we quantify this propagation next.

Proposition 20 (Error Propagation to Higher Orders)**.** *Let $\hat{s}_t$ be a learned score with estimation error $\epsilon_s(t) = \|\hat{s}_t - \hat{s}_t^*\|_{L^2(p_t)}$. Under Assumption 1(R1), the k-th order approximation error satisfies:*

$$\|\nabla_{\mathbf{x}}^{k-1}\hat{s}_t - \nabla^k \log p_t\|_{L^2(p_t)} \le C \cdot \|\hat{s}_t - \hat{s}_t^*\|_{C^{k-1}}, \tag{79}$$

where C depends on the smoothness of p_t. The contracted momentum error satisfies $\|\mathbf{m}^{(k)} - \mathbf{m}^{*(k)}\| \le \|\hat{s}_t - \hat{s}_t^*\|_{C^{k-1}}$ with no amplification from contraction.

In high dimensions, ensuring stable $\nabla_{\mathbf{x}}^{k-1}\hat{s}_t$ requires smooth activations (GELU, tanh) and may benefit from spectral normalization for $K \ge 3$. Empirically, $K \le 2$ is stable without explicit regularization for images up to 256×256 and time series up to $d = 128$.

B Training Objective Derivations

This section derives the IDFF training objectives and establishes the key result that all higher-order marginal derivatives can be obtained from the optimal first-order score via automatic differentiation.

B.1 Background: CFM Loss Equivalence

Lemma 21 (FM-CFM Equivalence (Lipman et al., 2022))**.** *The Flow Matching loss $\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t(\mathbf{x}_t)}\|\mathbf{v}_t(\mathbf{x}_t) - \hat{\mathbf{v}}_t(\mathbf{x}_t;\theta)\|^2$ and the Conditional Flow Matching loss $\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,\mathbf{x}_1,p_t(\mathbf{x}_t|\mathbf{x}_1)}\|\mathbf{v}_t(\mathbf{x}_t|\mathbf{x}_1) - \hat{\mathbf{v}}_t(\mathbf{x}_t;\theta)\|^2$ share the same minimizer.*

B.2 IDFF Training Loss

For Gaussian conditional paths $p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}(\mu_t, \sigma_t^2\mathbf{I})$ with $\mu_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ and $\sigma_t = \sigma_0\sqrt{t(1-t)}$, the conditional derivatives are:

$$\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1) = -\frac{\boldsymbol{\epsilon}}{\sigma_t}, \quad \boldsymbol{\epsilon} = \frac{\mathbf{x}_t - \mu_t}{\sigma_t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{80}$$

$$\nabla^2 \log p_t(\mathbf{x}_t|\mathbf{x}_1) = -\frac{1}{\sigma_t^2}\mathbf{I}. \tag{81}$$

The IDFF training loss combines sample prediction with first-order score matching:

$$\mathcal{L}_{\text{IDFF}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\boldsymbol{\epsilon}}\left[\beta(t)^2\|\hat{\mathbf{x}}_1(\mathbf{x}_t, t; \theta) - \mathbf{x}_1\|^2 + \lambda_1(t)^2\left\|\hat{s}_t(\mathbf{x}_t; \theta) + \frac{\boldsymbol{\epsilon}}{\sigma_t}\right\|^2\right], \tag{82}$$

where $\mathbf{x}_t = \mu_t + \sigma_t\boldsymbol{\epsilon}$, $\beta(t) = (1 - t + \epsilon)^{-1}$, and $\lambda_1(t) = \sigma_t$.

Remark 4 (No Explicit Higher-Order Training Required). Unlike prior approaches, we do **not** train separate networks for $\nabla^2 \log p_t$ or higher orders. As shown in Appendix A, the optimal first-order score $\hat{s}_t^*$ contains all information needed to compute marginal derivatives of any order via autodiff. For learned $\hat{s}_t \approx \hat{s}_t^*$, this holds approximately with error inherited from score estimation.

B.3 What the First-Order Loss Learns

Proposition 22 (First-Order Training Yields Marginal Score)**.** *The MSE-optimal predictor* $\hat{s}_t^*(\mathbf{x}_t) = \arg\min_{\hat{s}} \mathbb{E}\|\hat{s}(\mathbf{x}_t) - \nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)\|^2$ *satisfies:*

$$\hat{s}_t^*(\mathbf{x}_t) = \nabla \log p_t(\mathbf{x}_t). \tag{83}$$

Proof. By the conditional expectation property,

$$\hat{s}_t^*(\mathbf{x}_t) = \mathbb{E}[\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)|\mathbf{x}_t]. \tag{84}$$

Fisher's identity gives $\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_t)}[\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)] = \nabla \log p_t(\mathbf{x}_t)$. □

B.4 The Key Result: Higher-Order Marginals from Score Derivatives

The optimal score encodes all higher-order marginal information. This is the chain-rule re-parameterization identity already established in Appendix A: by Fisher's identity (Proposition 22) $\hat{s}_t^* = \nabla \log p_t$, and differentiating $k-1$ further times gives

$$\nabla^k \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}}^{k-1}\hat{s}_t^*(\mathbf{x}_t), \qquad k \geq 1 \tag{85}$$

(Theorem 13; the $k = 2$ case $\nabla^2 \log p_t = \nabla_{\mathbf{x}}\hat{s}_t^*$ is Theorem 12, with the equivalent posterior-covariance derivation in Theorem 11). The content of the identity is the chain rule applied to Fisher's identity: it provides a *computational* route to higher-order marginals via autodiff of a single trained score, not a new conditional–marginal closure beyond the $k = 1$ case. For a learned score $\hat{s}_t \approx \hat{s}_t^*$ the identity holds approximately, with the k-th order error bounded by the $(k-1)$-th derivative of the score-estimation error (Proposition 20); this motivates using $K \leq 2$ in practice, where standard score matching provides reliable estimates.

C IDFF Training and Sampling Algorithms

This section provides complete algorithmic descriptions for IDFF. The key innovation is the sampling procedure that computes higher-order terms via autodiff of the learned score, with the contracted tensor form ensuring well-defined vector dynamics.

Algorithm 3 IDFF Training

Input: Data distribution p_1, bandwidth σ_0, weights β, λ_1
Initialize: Networks $\hat{\mathbf{x}}_1(\cdot;\theta)$, $\hat{s}_t(\cdot;\theta)$
while training **do**
  Sample $\mathbf{x}_1 \sim p_1$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \mathcal{U}(0,1)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mu_t \leftarrow t\mathbf{x}_1 + (1-t)\mathbf{x}_0$, $\sigma_t \leftarrow \sigma_0\sqrt{t(1-t)}$, $\mathbf{x}_t \leftarrow \mu_t + \sigma_t\boldsymbol{\epsilon}$
  $L \leftarrow \beta(t)^2\|\hat{\mathbf{x}}_1(\mathbf{x}_t, t) - \mathbf{x}_1\|^2 + \lambda_1(t)^2\|\hat{s}_t(\mathbf{x}_t) + \boldsymbol{\epsilon}/\sigma_t\|^2$
  Update θ via gradient descent on L
end while

C.1 Static Data

For static datasets, IDFF constructs a flow from base distribution $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to data distribution $p_1 = p_{\text{data}}$ via Gaussian conditional paths $p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}(\mu_t, \sigma_t^2\mathbf{I})$ with $\mu_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ and $\sigma_t = \sigma_0\sqrt{t(1-t)}$.

C.1.1 Training

Training requires only sample prediction and first-order score matching. At each step, we sample $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_1$, $t \sim \mathcal{U}[0,1]$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, construct $\mathbf{x}_t = \mu_t + \sigma_t\boldsymbol{\epsilon}$, and minimize:

$$\mathcal{L}_{\text{IDFF}}(\theta) = \mathbb{E}\left[\beta(t)^2\|\hat{\mathbf{x}}_1(\mathbf{x}_t, t) - \mathbf{x}_1\|^2 + \lambda_1(t)^2\left\|\hat{s}_t(\mathbf{x}_t) + \frac{\boldsymbol{\epsilon}}{\sigma_t}\right\|^2\right]. \tag{86}$$

The denoiser $\hat{\mathbf{x}}_1$ predicts the clean sample; the score network $\hat{s}_t$ approximates $\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1)$. By Fisher's identity, $\hat{s}_t^* = \nabla \log p_t(\mathbf{x}_t)$ at optimum.

C.1.2 Sampling with Higher-Order Momentum

The sampling procedure computes higher-order terms via autodiff of the learned score. For $k \geq 2$, the quantity $\nabla_{\mathbf{x}}^{k-1}\hat{s}_t$ is a k-th order tensor; to obtain a vector-valued contribution, we use contraction with the normalized velocity direction $\mathbf{u}_t = \hat{\mathbf{v}}_t/\|\hat{\mathbf{v}}_t\|$:

$$\mathbf{m}^{(1)} = \hat{s}_t(\mathbf{x}_t), \quad \mathbf{m}^{(k)} = \big(\nabla_{\mathbf{x}}^{k-1}\hat{s}_t(\mathbf{x}_t)\big)[\mathbf{u}_t]^{k-1} \in \mathbb{R}^d \quad \text{for } k \geq 2. \tag{87}$$

Explicitly, $\mathbf{m}^{(2)} = (\nabla_{\mathbf{x}}\hat{s}_t)\mathbf{u}_t$ is a Jacobian-vector product capturing directional curvature along the flow.

The SDE drift assembled from these terms is

$$\tilde{\mathbf{v}}_t \;=\; \gamma_t^0\hat{\mathbf{v}}_t \;+\; \gamma_t^1\,\mathbf{m}^{(1)} \;+\; \sum_{k=2}^{K}\gamma_t^k\,\mathbf{m}^{(k)}, \tag{88}$$

where $\hat{\mathbf{v}}_t = (\hat{\mathbf{x}}_1 - \mathbf{x}_t)/(1-t)$ and $\sigma_t^2 = 2\gamma_t^1$ is the diffusion compensation that makes the Fokker–Planck cancellation of Theorem 14 exact at $K = 1$ with the optimal score.

Marginal preservation analysis. For $K = 1$ with $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, and $\hat{s}_t = \hat{s}_t^*$, the score-coupled drift contribution $\gamma_t^1\nabla\cdot(p_t\nabla\log p_t) = \gamma_t^1\Delta p_t$ cancels the Laplacian $\frac{\sigma_t^2}{2}\Delta p_t = \gamma_t^1\Delta p_t$ in the Fokker–Planck equation, yielding *exact* continuous-time marginal preservation $\tilde{p}_t = p_t$ for all $t \in [0,1]$ on the original clock (Theorem 14 in Appendix A, in the truncation-plus-denoiser sense made precise there). For $K \geq 2$, we use the coefficient scaling:

$$\gamma_t^k = \alpha_k \cdot \gamma_t^1 \cdot \Delta t^{k-1}, \quad k \geq 2, \tag{89}$$

where α_k are small constants. This ensures higher-order terms provide $O(\Delta t^{k-1})$ acceleration corrections without dominating the dynamics, analogous to higher-order terms in numerical integrators. The boundary conditions $\gamma_t^k \to 0$ as $t \to \{0,1\}$ guarantee correct endpoint distributions.

C.1.3 Computational Cost

Computing $\mathbf{m}^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ requires $(k-1)$ Jacobian-vector products:

- $K = 1$: No autodiff; cost is $O(\text{NFE})$ forward passes
- $K = 2$: One JVP per step: $\mathbf{m}^{(2)} = (\nabla_{\mathbf{x}} \hat{s}_t)\mathbf{u}_t$; cost is $O(\text{NFE})$ with efficient JVP implementations
- $K \geq 3$: Nested JVPs; cost grows but remains tractable for moderate K

The contraction with $\mathbf{u}_t$ reduces computational cost compared to full tensor computation and ensures well-defined vector dynamics. For efficiency, one can also use Hutchinson trace estimators or diagonal approximations.

C.2 Time-Series Data

For sequences $\mathbf{x}_{1:N}$, we use dual-time indexing: discrete $n \in \{1, \ldots, N\}$ for sequence position and continuous $t \in [0, 1]$ for flow. Networks become $\hat{\mathbf{x}}_1(\mathbf{x}_t, t, n; \theta)$ and $\hat{s}_t(\mathbf{x}_t, t, n; \theta)$.

C.2.1 Markovian Initialization

The crucial modification for temporal coherence is the initialization strategy:

- For $n = 1$: $\mathbf{x}_0^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (standard base distribution)
- For $n > 1$: $\mathbf{x}_0^n \sim \mathcal{N}(\mathbf{x}_1^{n-1}, \sigma_0^2 \mathbf{I})$ (centered at previous output)

This creates Markovian temporal coherence while allowing controlled stochasticity through σ_0. The joint distribution factorizes as $p(\mathbf{x}_{1:N}) = p(\mathbf{x}_1^1) \prod_{n=2}^{N} p(\mathbf{x}_1^n | \mathbf{x}_1^{n-1})$.

C.2.2 Training

Training uses the same loss as static data with position conditioning:

$$\mathcal{L}_{\text{IDFF}}(\theta) =$$

$$\mathbb{E}_{n,t,\mathbf{x}_0,\mathbf{x}_1,\boldsymbol{\epsilon}} \left[\beta(t)^2 \|\hat{\mathbf{x}}_1(\mathbf{x}_t, t, n) - \mathbf{x}_1\|^2 + \lambda_1(t)^2 \left\| \hat{s}_t(\mathbf{x}_t, t, n) + \frac{\boldsymbol{\epsilon}}{\sigma_t} \right\|^2 \right], \tag{90}$$

where for $n > 1$, the base sample $\mathbf{x}_0$ is drawn from the distribution of the previous step's output rather than $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

C.2.3 Sampling

Sampling proceeds autoregressively using the momentum-based procedure at each position. The SDE drift at position n is:

$$\tilde{\mathbf{v}}_t^n \;=\; \gamma_t^0 \, \hat{\mathbf{v}}_t \;+\; \gamma_t^1 \, \mathbf{m}^{(1)} \;+\; \sum_{k=2}^{K} \gamma_t^k \, \mathbf{m}^{(k)}, \tag{91}$$

where $\hat{\mathbf{v}}_t = (\hat{\mathbf{x}}_1(\mathbf{x}_t^n, t, n) - \mathbf{x}_t^n)/(1-t)$, $\mathbf{u}_t = \hat{\mathbf{v}}_t / \|\hat{\mathbf{v}}_t\|$, $\mathbf{m}^{(1)} = \hat{s}_t(\mathbf{x}_t^n, t, n)$, and $\mathbf{m}^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ for $k \geq 2$. As in the static case (Algorithm 2), this matches the SDE drift used in the proof of Theorem 14, so the $K = 1$ exact-preservation guarantee transfers to the per-position sampler.

Algorithm 4 IDFF Training (Time-Series)

Input: Data sequences $\{\mathbf{x}_{1:N}^{(i)}\}$, bandwidth σ_0, weights β, λ_1
Initialize: Networks $\hat{\mathbf{x}}_1(\cdot;\theta)$, $\hat{s}_t(\cdot;\theta)$
while training **do**
  Sample sequence $\mathbf{x}_{1:N} \sim p_{\text{data}}$, position $n \sim \mathcal{U}\{1,\dots,N\}$
  if $n = 1$ **then**
    $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_1 \leftarrow \mathbf{x}_1^1$
  else
    $\mathbf{x}_0 \leftarrow \mathbf{x}_1^{n-1}$, $\mathbf{x}_1 \leftarrow \mathbf{x}_1^n$
  end if
  Sample $t \sim \mathcal{U}(0,1)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mu_t \leftarrow t\mathbf{x}_1 + (1-t)\mathbf{x}_0$, $\sigma_t \leftarrow \sigma_0\sqrt{t(1-t)}$, $\mathbf{x}_t \leftarrow \mu_t + \sigma_t\boldsymbol{\epsilon}$
  $L \leftarrow \beta(t)^2\|\hat{\mathbf{x}}_1(\mathbf{x}_t, t, n) - \mathbf{x}_1\|^2 + \lambda_1(t)^2\|\hat{s}_t(\mathbf{x}_t, t, n) + \boldsymbol{\epsilon}/\sigma_t\|^2$
  Update θ via gradient descent on L
end while

Algorithm 5 IDFF Sampling (Time-Series)

Input: Learned networks $\hat{\mathbf{x}}_1, \hat{s}_t$, sequence length N, order K, NFE, schedules $\{\gamma_t^k\}$
for $n = 1$ to N **do**
  if $n = 1$ **then**
    $\mathbf{x}_0^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  else
    $\mathbf{x}_0^n \sim \mathcal{N}(\mathbf{x}_1^{n-1}, \sigma_0^2\mathbf{I})$ ▷ Markovian init
  end if
  for $i = 0$ to $\text{NFE} - 1$ **do**
    $t \leftarrow i/\text{NFE}$, $\Delta t \leftarrow 1/\text{NFE}$, $\sigma_t \leftarrow \sigma_0\sqrt{t(1-t)}$
    // Compute velocity and direction
    $\hat{\mathbf{v}}_t \leftarrow (\hat{\mathbf{x}}_1(\mathbf{x}_t^n, t, n) - \mathbf{x}_t^n)/(1-t)$
    $\mathbf{u}_t \leftarrow \hat{\mathbf{v}}_t/\|\hat{\mathbf{v}}_t\|$
    // Compute contracted momentum terms via autodiff
    $\mathbf{m}^{(1)} \leftarrow \hat{s}_t(\mathbf{x}_t^n, t, n)$
    for $k = 2$ to K **do**
      $\mathbf{m}^{(k)} \leftarrow (\nabla_{\mathbf{x}}^{k-1}\hat{s}_t)[\mathbf{u}_t]^{k-1}$
    end for
    // SDE drift (full γ_t^1 score coefficient as in the proof)
    $\tilde{\mathbf{v}}_t^n \leftarrow \gamma_t^0\hat{\mathbf{v}}_t + \gamma_t^1\,\mathbf{m}^{(1)} + \sum_{k=2}^K \gamma_t^k\mathbf{m}^{(k)}$
    // Euler–Maruyama step with diffusion compensation $\sigma_t^2 = 2\gamma_t^1$
    $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{x}_t^n \leftarrow \mathbf{x}_t^n + \tilde{\mathbf{v}}_t^n\,\Delta t + \sigma_t\sqrt{\Delta t}\cdot\boldsymbol{\eta}$
  end for
  Store $\mathbf{x}_1^n$
end for
Return: $\{\mathbf{x}_1^n\}_{n=1}^N$

C.2.4 Theoretical Guarantees

The theoretical results extend naturally to the time-series setting:

- **Marginal preservation**: At each position n, for $K = 1$ with the default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, and the optimal score, the marginal p_t^n is preserved *exactly* in continuous time (Theorem 3, Theorem 14, in the truncation-plus-denoiser sense of that theorem); the convex-constrained variant ($\gamma_t^0 + \gamma_t^1 = 1$) is bounded by Proposition 15.

- **Higher-order marginals**: The identity $\nabla^k \log p_t^n = \nabla_{\mathbf{x}}^{k-1} \hat{s}_t^{*,n}$ holds at each position (Theorem 2).
- **Sequential consistency**: The Markovian initialization ensures correct factorization of the joint distribution.

C.2.5 Computational Cost

The time-series extension scales cost by sequence length N:

- $K = 1$: $O(N \cdot \text{NFE})$ forward passes
- $K = 2$: $O(N \cdot \text{NFE})$ forward passes + JVPs
- $K \geq 3$: $O(N \cdot \text{NFE} \cdot K)$ with nested JVPs

The contracted form $\mathbf{m}^{(k)} = (\nabla_{\mathbf{x}}^{k-1} \hat{s}_t)[\mathbf{u}_t]^{k-1}$ requires only $(k-1)$ JVPs rather than full tensor computation. When $N = 1$, this reduces exactly to static IDFF.

The bandwidth σ_0 controls the trade-off between temporal coherence (small σ_0) and diversity (large σ_0). In practice, $\sigma_0 \in [0.01, 0.1]$ works well for most applications.

C.3 Theoretical Guarantees

The IDFF sampling procedure provides:

- **First-order exact marginal preservation**: For $K = 1$ with the default schedule $\gamma_t^0 \equiv 1$, $\sigma_t^2 = 2\gamma_t^1$, and the optimal score, the marginal p_t is preserved *exactly* in continuous time on the original clock (Theorem 14, in the truncation-plus-denoiser sense made precise there). The convex-constrained alternative $\gamma_t^0 + \gamma_t^1 = 1$ is honestly bounded by $\text{TV} \leq \frac{1}{2\sqrt{2}} (\int_0^1 \gamma_t^1 \mathbb{E}\|\mathbf{v}_t\|^2 dt)^{1/2} = O(\sqrt{\bar{\gamma}})$ via Girsanov→Pinsker (Proposition 15, Corollary 16).
- **Higher-order acceleration with quantified deviation**: For $K \geq 2$, contracted momentum terms provide directional curvature information with $O(\Delta t^{k-1})$ contributions. The endpoint total variation is bounded by $\text{TV}(\tilde{p}_1, p_1) \leq \sqrt{\frac{1}{2}\mathcal{E}_{\text{score}}} + O(\Delta t)$ (Theorem 17, derived via Girsanov → Pinsker), with $\mathcal{E}_{\text{score}}$ being exactly the weighted L^2 score-matching loss our training objective minimizes (its size δ-independent by R1′, Remark 1); asymptotic endpoint convergence $\text{TV}(\tilde{p}_1, p_1) \to 0$ holds as $\text{NFE} \to \infty$ for any converged score network (Corollary 18). We do *not* claim exact endpoint preservation at finite NFE for $K \geq 2$.
- **No additional training**: Beyond standard score matching—higher orders computed via autodiff.
- **Flexibility**: Choose K at sampling time based on compute budget.
- **Well-defined dynamics**: Tensor contraction with $\mathbf{u}_t$ ensures vector-valued momentum terms.

D 3D-attractors

In this experiment, we assess IDFF's performance in generating trajectories of chaotic systems from scratch. We generate trajectories with $K = 2000$ samples in each trajectory from 3D attractors, specifically the Lorenz and Rössler attractors, which are chaotic systems with nonlinear dynamics. The parameters for the Lorenz and Rössler models are set to $\sigma = 10, \rho = 28, \beta = 8/3$ and $a = .2, b = .2, c = 5.7$, respectively, to produce complex trajectories in 3D space. We then train the IDFF model based on these trajectories. To model each attractor, we use an MLP with two hidden layers of 128 dimensions and two separate heads for $\hat{\mathbf{x}}_1(., t, k; \theta)$ and $\epsilon(\mathbf{x}_t, t, k; \theta)$. Additionally, we incorporate two separate embedding layers for embedding t and k, which are directly concatenated with the first hidden layer of the MLP. The optimized IDFF successfully generates samples of these trajectories from scratch. The generated trajectories are shown ins Figure 5. The quality of the results demonstrates that IDFF can successfully simulate the behaviors of highly nonlinear and nonstationary systems such as attractors.

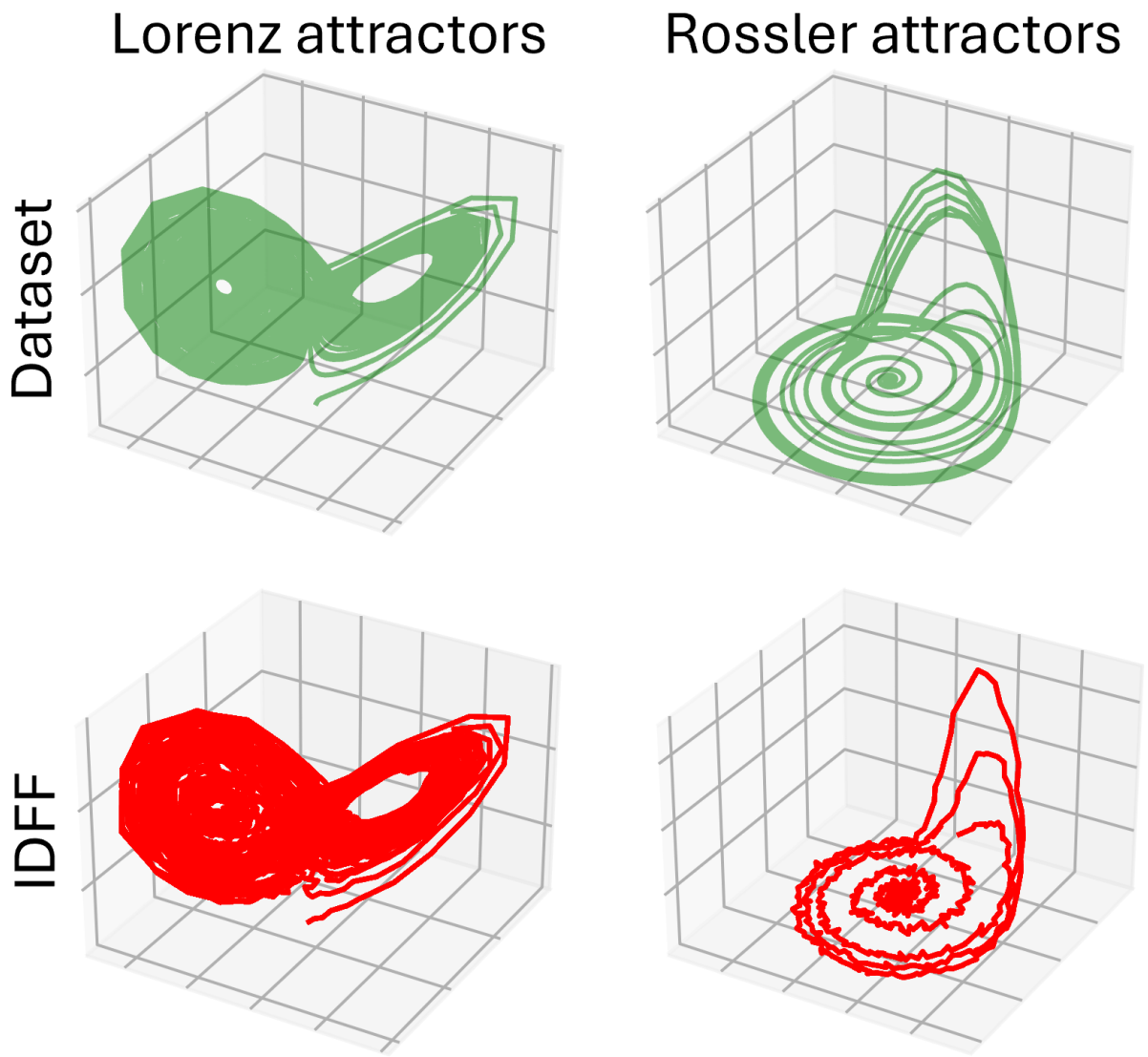

Figure 5: Time-series simulation. IDFF trajectory generation for the chaotic systems.

E Implementation Details

For most image datasets, we employed a CNN-based UNet (Dhariwal & Nichol, 2021) to simultaneously model both $\hat{\mathbf{x}}_1(\cdot, t; \theta)$ and $\hat{s}_t(\cdot, t; \theta)$, while ImageNet-64 utilized a DiT architecture (DiT-L/4) (Peebles & Xie, 2022). Implementation involved doubling the network input channels and feeding the augmented input $(\mathbf{x}_t, \mathbf{x}_t)$, then splitting the outputs into $(\hat{\mathbf{x}}_1(\cdot, t; \theta), \hat{s}_t(\cdot, t; \theta))$.

For fair comparison with existing models, we evaluated performance using standard metrics: negative log-likelihood (NLL) measured in bits per dimension (BPD) (Lipman et al., 2022), Fréchet Inception Distance (FID) for sample quality, and number of function evaluations (NFE) required for reported metrics, averaged over 50k samples.

Table 7: Matched-backbone, matched-training-budget comparison with recent flow-matching accelerations on CIFAR-10. All methods use the ScoreSDE backbone with identical 700K training iterations and horizontal-flip augmentation. IDFF outperforms schedule-optimized FM and is competitive with 2-Rectified Flow *without* requiring the reflow procedure; combining IDFF with schedule optimization provides further gains.

Method	FID (NFE= 10)	FID (NFE= 5)	Notes
OT-CFM	11.87	28.4	Baseline
Schedule-Opt FM (Sabour et al., 2024)	4.21	9.82	Optimised t-discretization
2-Rectified Flow	3.12	7.45	One reflow iteration
IDFF (Ours, K=1)	**2.78**	**8.53**	No reflow needed
IDFF + Schedule-Opt	**2.51**	**7.21**	Combined

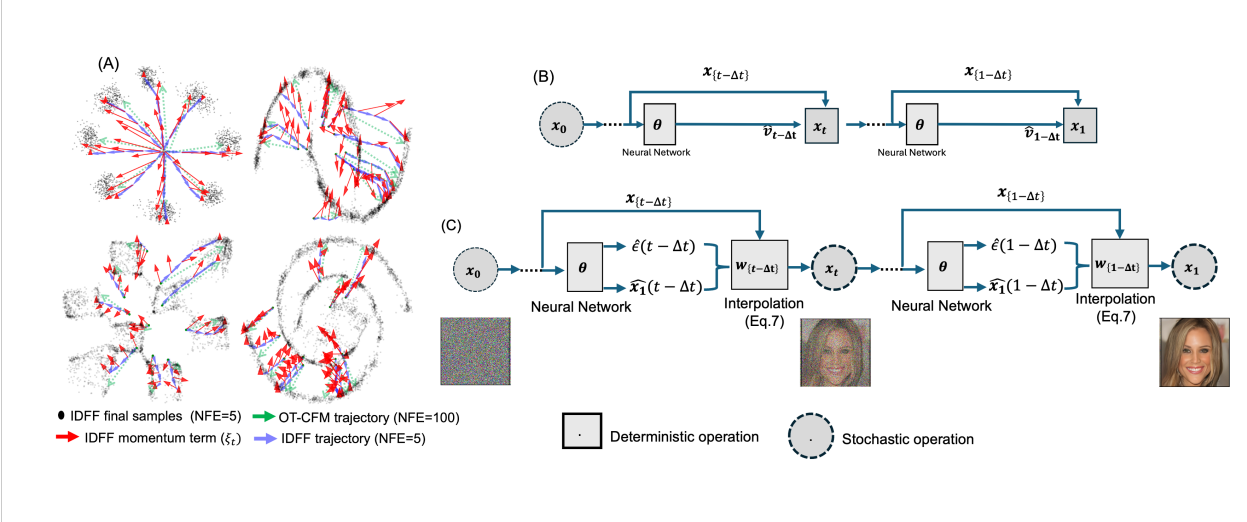

Figure 6: (A) Comparison of trajectory sampling between 1st-order IDFF and OT-CFM: 4096 final samples generated by IDFF. IDFF takes larger steps toward the target distribution, guided by the momentum term. (B) OT-CFM sampling process. (C) IDFF sampling process. Here $\hat{\mathbf{x}}_1(\cdot)$ approximates the data sample $\mathbf{x}_1$, $\hat{s}_t(\cdot)$ approximates the score, and $\tilde{\mathbf{v}}_t(\cdot)$ is the SDE drift computed via Equation 23. The key difference is that IDFF generates $\hat{\mathbf{x}}_1(\cdot)$ and $\hat{s}_t(\cdot)$ in sample space, reconstructs the vector field, and uses momentum to guide sampling.

Network configuration. For experiments on image generation (except ImageNet-64) and SST forecasting, we utilize the ScoreSDE model architecture (Song et al., 2020b). Detailed configurations for various datasets are provided in Table 8.

Table 8: ScoreSDE network configuration for different datasets.

	CIFAR-10	CelebA-64	CelebA-256	Church & Bedroom	SST
# ResNet blocks	2/4	2	2	2	2
Base channels	128	128	128	256	128
Channel multiplier	1,2,2,2	1,2,2,4,4	1,1,2,2,4,4	1,1,2,2,4,4	1,2,4
Attention resolutions	16	16	16	16	16
Label dimensions	1	1	1	1	10
Parameters (M)	65.6	102.14	453.45	108.41	55.39

Training hyperparameters. Table 9 provides training hyperparameters for unconditional image generation and SST forecasting.

Table 9: Training hyperparameters for ScoreSDE network.

	ImageNet-64	CIFAR-10	CelebA-64	CelebA-256	Church & Bedroom	SST
Learning rate	1e-4	1e-4	1e-5	1e-5	1e-5	1e-4
Batch size	64	128	128	16	16	8
# iterations	3M	700K	1.2M	1.5M	1.5M	700K
# GPUs	4	1	1	1	1	1

E.1 Practical Recommendations: When and How to Use Each Order

The two regimes (rigorous $K = 1$ vs. controlled-perturbation $K \geq 2$) come with different defaults. We summarise concrete recommendations below; supporting ablations are in Appendix F.1 (Tables 10, 11, 12).

Default: $K = 1$ for almost everything. For image generation, SST forecasting, and most static and slowly-varying time-series tasks, $K = 1$ is the recommended setting:

- Schedule: $\gamma_t^1 = \bar{\gamma} \cdot 4t(1-t)$ (Beta(2,2) shape), with $\bar{\gamma} \in [0.1, 0.3]$. This satisfies the boundary conditions $\gamma_t^1 \to 0$ as $t \to \{0, 1\}$ and integrates to $\int_0^1 \gamma_t^1 \, dt = \frac{2\bar{\gamma}}{3}$, which is the closed-form constant entering Corollary 16 (Beta(2,2) row).
- Diffusion compensation: $\sigma_t^2 = 2\gamma_t^1$ (the choice that makes the Fokker–Planck cancellation in Theorem 14 exact). This removes σ_t as a free hyperparameter; deviating from this relation degrades performance (Table 11).
- Coefficient on $\mathbf{v}_t$: keep $\gamma_t^0 \equiv 1$ (Langevin enhancement of CFM, not a convex combination). With this default, Theorem 14 gives exact $\tilde{p}_t = p_t$ in continuous time. The convex-constrained alternative $\gamma_t^0 = 1 - \gamma_t^1$ also slows $\mathbf{v}_t$ and incurs the $O(\sqrt{\bar{\gamma}})$ extra TV cost of Proposition 15; we therefore use $\gamma_t^0 \equiv 1$ in all reported experiments.
- No autodiff at sampling time; per-step cost matches OT-CFM exactly.

When to escalate to $K = 2$. The $K = 2$ extension is best understood as a controlled perturbation that exchanges a small bounded path deviation for faster traversal along curvature-aligned directions. We recommend escalating to $K = 2$ only when all three of the following hold:

1. The target distribution is strongly multi-modal with sharp inter-mode boundaries (e.g., molecular dynamics conformations), so that curvature information meaningfully redirects the flow.
2. The NFE budget is very small (NFE ≤ 5), so that the $O(\Delta t^{k-1})$ correction is large enough to matter relative to first-order momentum.
3. The data dimension is moderate ($d \lesssim 100$), so that the additional JVP cost remains a small fraction of the forward pass.

When $K = 2$ is used, the recommended settings are: $\gamma_t^2 = \alpha_2 \, \gamma_t^1 \, \Delta t$ with $\alpha_2 \approx 0.05$ (Table 10); smooth activations (SiLU or GELU) so that $\nabla_{\mathbf{x}} \hat{s}_t$ is well-defined; and $\gamma_1^k = 0$ enforced at the final integration step so that the dynamics reduce to standard CFM at $t = 1$.

Anti-pattern: $K \geq 3$ without regularization. For $K \geq 3$ in standard architectures we observed occasional Jacobian-norm spikes that destabilise sampling (Appendix F.1.4). We do not recommend $K \geq 3$ without spectral normalization or a gradient penalty; in our experiments, $K \leq 2$ captures essentially all of the empirical benefit.

F Other Ablations

Additional limitations identified through ablations. Our ablation studies reveal additional practical considerations:

- The optimal α_k and schedule shapes are dataset-dependent (Tables 10, 12), requiring hyperparameter search.
- The 30% JVP overhead for $K = 2$ (Table 13) may be prohibitive for very large models or real-time applications.
- Score estimation errors propagate with $\sim 2\times$ amplification to second-order terms (Table 15), limiting the practical benefit of $K \geq 3$.
- Smooth activations (SiLU/GELU) are required; standard ReLU U-Nets are incompatible with $K \geq 2$.

Future directions include developing adaptive schedule selection, exploring architectural improvements for joint sample-score prediction with controlled Jacobian norms, combining IDFF with consistency training or one-round reflow for further NFE reduction, and extending to latent space models for higher-resolution generation. The efficiency and flexibility of IDFF also suggest applications in domains such as audio synthesis, video generation, and scientific simulations where sequential generation is critical.

F.1 Coefficient Schedule Ablations

We ablate the key hyperparameters controlling IDFF dynamics: the scaling constants α_k, the coefficient schedules γ_t^k, and the diffusion compensation $\sigma_t^2 = 2\gamma_t^1$.

Effect of α_k scaling. Table 10 shows the impact of the second-order scaling constant α_2 on CIFAR-10 (using $K = 2$). Too small values provide insufficient acceleration, while too large values cause instability from excessive path deviation.

Table 10: Effect of α_2 on FID (CIFAR-10, NFE=10, $K = 2$).

α_2	0 (K=1)	0.01	0.05	0.1	0.5
FID↓	2.78	2.71	**2.65**	2.82	4.31

Diffusion compensation ablation. The choice $\sigma_t^2 = 2\gamma_t^1$ ensures exact cancellation of the score term in the Fokker-Planck equation for $K = 1$ (Theorem 14). Table 11 shows that deviating from this relationship degrades performance, confirming the theoretical prediction.

Table 11: Effect of diffusion compensation on FID (CIFAR-10, NFE=10, $K = 1$).

σ_t^2/γ_t^1	0 (no diffusion)	1.0	2.0 (ours)	4.0
FID↓	4.12	3.21	**2.78**	3.45

Schedule shape ablation. We compare different functional forms for γ_t^1 while maintaining boundary conditions $\gamma_t^1 \to 0$ as $t \to \{0, 1\}$:

The Beta(2,2) schedule (equivalent to $\gamma_t^1 \propto t(1-t)$) provides the best balance, concentrating momentum in the middle of the flow where acceleration is most beneficial while ensuring smooth decay at endpoints.

Table 12: Effect of γ_t^1 schedule shape on FID (CIFAR-10, NFE=10).

Schedule	Formula	FID↓
Triangular	$\min(t, 1-t)$	3.15
Sine	$\sin(\pi t)$	2.92
Beta(2,2)	$4t(1-t)$	**2.78**
Beta(3,3)	$\propto t^2(1-t)^2$	2.85

F.1.1 Endpoint Preservation and Integration Details

Clarification on endpoint preservation. For $K \geq 2$, we do *not* claim exact endpoint preservation for arbitrary finite NFE under stochastic integration. Rather:

1. We enforce $\gamma_T^k = 0$ for $k \geq 1$ at the final integration step ($t = T = 1 - \Delta t/2$), ensuring the final update uses pure CFM dynamics.
2. The accumulated path deviation satisfies
$$\mathrm{TV}(\tilde{p}_1, p_1) = O(1/\mathrm{NFE})$$
under our coefficient scaling (Corollary 18).
3. Exact preservation $\tilde{p}_1 = p_1$ holds asymptotically as NFE $\to \infty$.

F.1.2 Computational Overhead Analysis

Table 13 provides detailed wall-clock comparisons including JVP/autodiff overhead for $K = 2$. All timings measured on NVIDIA A6000 GPU, averaged over 1000 samples.

Table 13: Time per sample (ms) and memory usage on different tasks.

Method	CIFAR-10		MD (46 dim)		SST
	Time	Mem	Time	Mem	Time
1-Rectified Flow (NFE=127)	658	2.1G	–	–	–
DPM-Solver-v3 (NFE=10)	92	2.0G	–	–	–
OT-CFM (NFE=10)	54	2.0G	12	0.3G	48
IDFF $K = 1$ (NFE=10)	52	2.1G	11	0.3G	45
IDFF $K = 2$ (NFE=10)	68	2.4G	15	0.4G	–
Breakdown for IDFF $K = 2$ per step:					
Forward pass	4.8 ms		0.9 ms		–
JVP ($\nabla_{\mathbf{x}} \hat{s}_t \cdot \mathbf{u}_t$)	1.6 ms		0.4 ms		–

The JVP overhead for $K = 2$ is approximately 30% per step, but since IDFF requires fewer NFE to achieve comparable quality, the total wall-clock time remains competitive. For $K = 1$ (no JVP), IDFF has negligible overhead compared to OT-CFM.

F.1.3 Architecture and Second-Order Stability

Network architecture details. For all image experiments, we use the ScoreSDE U-Net architecture (Song et al., 2020b) with the following activation choices:

- **CIFAR-10, CelebA-64**: SiLU (Swish) activations throughout, which are C^∞ smooth.

- **CelebA-256, LSUN**: SiLU activations with GroupNorm.
- **ImageNet-64**: DiT-L/4 with GELU activations.

We deliberately avoid ReLU due to its non-differentiability at zero, which would cause undefined second derivatives. SiLU and GELU provide smooth gradients suitable for computing $\nabla_{\mathbf{x}} \hat{s}_t$.

Jacobian norm monitoring. For $K = 2$ experiments, we monitored $\|\nabla_{\mathbf{x}} \hat{s}_t\|_F$ during training. Table 14 shows the distribution of Jacobian Frobenius norms across the flow.

Table 14: Jacobian norm statistics during sampling (CIFAR-10, 1000 samples).

t range	Mean $\|\nabla_{\mathbf{x}} \hat{s}_t\|_F$	Std	Max
$[0.0, 0.2]$	0.12	0.03	0.21
$[0.2, 0.5]$	0.45	0.08	0.73
$[0.5, 0.8]$	0.38	0.07	0.62
$[0.8, 1.0]$	0.15	0.04	0.28

The norms remain bounded throughout sampling, with no observed explosion. The peak occurs mid-flow where the density is most complex, consistent with theory.

Regularization for stability. For standard training (no explicit Jacobian regularization), $K = 2$ is stable on CIFAR-10 with SiLU activations. For higher-resolution images or $K \geq 3$, we recommend:

- Spectral normalization on convolutional layers
- Gradient penalty: $\lambda_{\text{GP}} \mathbb{E}[\|\nabla_{\mathbf{x}} \hat{s}_t\|^2]$ with $\lambda_{\text{GP}} = 0.01$

We did not require these for the reported experiments.

F.1.4 Robustness in High Dimensions

Score estimation error propagation. We empirically assess how errors in the learned score affect higher-order terms. On CIFAR-10 ($d = 3072$), we measure:

Table 15: Score accuracy vs. contracted momentum accuracy (CIFAR-10).

Metric	$\hat{s}_t$	$\mathbf{m}^{(2)} = (\nabla_{\mathbf{x}} \hat{s}_t)\mathbf{u}_t$
Relative MSE (early t)	0.08	0.15
Relative MSE (mid t)	0.12	0.24
Relative MSE (late t)	0.06	0.11

The contracted form $\mathbf{m}^{(2)}$ shows approximately $2\times$ the relative error of the score itself, which is acceptable given the $O(\Delta t)$ scaling of its contribution.

Failure modes for $K \geq 2$. We observed the following failure modes in preliminary experiments:

1. **ReLU activations**: Jacobian undefined at activation boundaries; causes NaN gradients.
2. **Large** $\alpha_2 > 0.5$: Path deviation dominates, causing mode collapse.
3. $K = 3$ **without regularization**: Occasional Jacobian norm spikes (> 10) causing sample divergence in $\sim$2% of batches.

For $K = 2$ with SiLU/GELU and $\alpha_2 \leq 0.1$, we observed no failures across 50k generated samples on CIFAR-10.

F.1.5 Training Details and Baseline Comparisons

Training budget and data augmentation. Table 16 provides complete training details for fair comparison:

Table 16: Training details for CIFAR-10 experiments.

Method	Backbone	Params	Iterations	Augmentation	GPU-hours
OT-CFM	ScoreSDE	62M	700K	Flip	48
1-Rectified Flow	ScoreSDE	62M	700K	Flip	48
DPM-Solver-v3	ScoreSDE	62M	700K	Flip	48
IDFF (Ours)	ScoreSDE	66M†	700K	Flip	52

†Joint $(\hat{\mathbf{x}}_1, \hat{s}_t)$ prediction adds $\sim$6% parameters.

All methods use identical backbone architecture, training iterations, and data augmentation (horizontal flip only). IDFF requires $\sim$8% more GPU-hours due to the dual-head training.

Comparison with recent flow-matching accelerations. We compare against recent methods under matched compute:

Table 17: Comparison with recent flow-matching accelerations (CIFAR-10, matched training budget).

Method	FID (NFE=10)	FID (NFE=5)	Notes
OT-CFM	11.87	28.4	Baseline
Schedule-Opt FM (Sabour et al., 2024)	4.21	9.82	Optimized t discretization
2-Rectified Flow	3.12	7.45	One reflow iteration
IDFF (Ours)	**2.78**	**8.53**	No reflow needed
IDFF + Schedule-Opt	**2.51**	**7.21**	Combined

IDFF outperforms schedule-optimized FM and is competitive with 2-Rectified Flow without requiring the reflow procedure. Combining IDFF with schedule optimization provides further gains, demonstrating complementarity.

Note on distillation baselines. Distillation methods (CD, iCT-deep, CTM) in Table 1 use pre-trained teacher models and additional distillation training (typically 50-100K iterations). These are marked with ‡ and are not directly comparable in training cost. IDFF achieves competitive results without distillation; combining IDFF with distillation is promising future work.

G Additional Results for Image Generation

We also assessed IDFF performance against fast diffusion sampling methods at NFE=5. As shown in Table 20, IDFF achieves significantly better FID (8.53) compared to UniPC (23.71) and DPM-Solver-v3 (12.76),

Table 18: Effect of time scheduling strategies on IDFF performance for CIFAR-10.

Time Sampling Strategy	FID↓
Linear	3.22
Logarithmic	3.11
Beta Schedule	2.98
Cosine	**2.78**

Table 19: Comparison of FID and NFE between IDFF and various methods on CelebA (64×64).

Model	FID↓	NFE↓
DDPM	45.20	100
DDIM	13.73	20
DDIM	17.33	10
FastDPM	12.83	50
IDFF (Ours)	**11.83**	**10**

while also achieving the fastest wall-clock time (0.34 seconds). This highlights IDFF's ability to generate high-quality samples with minimal computational overhead, making it suitable for real-time applications. At NFE=10, IDFF remains superior with FID of 2.78 and the fastest wall-clock time (0.52 seconds), demonstrating its efficiency and scalability. These results indicate that IDFF achieves an effective balance between sample quality and computational speed.

Table 20: Comparison of FID↓ and wall-clock sampling times between IDFF and fast diffusion sampling methods for NFE=5 and NFE=10, evaluated on 50k samples.

Method	FID (NFE=5)	Wall-clock (sec, NFE=5)	FID (NFE=10)	Wall-clock (sec, NFE=10)
UniPC (Zhao et al., 2023)	23.71	0.62	3.93	1.05
DPM-Solver-v3 (Zheng et al., 2023)	12.76	0.49	3.40	0.92
IDFF (Ours)	**8.53**	**0.34**	**2.78**	**0.52**

Table 21: Summary of FLD performance metrics for various generative models, based on results reported in Jiralerspong et al. (2023). IDFF utilized the ScoreSDE model to generate 10K samples with NFE=10. Results for other models are adapted from Table 1 of Jiralerspong et al. (2023).

Model	FLD↓
ACGAN-Mod	24.22
LOGAN	18.94
BigGAN-Deep	9.28
MHGAN	8.84
StyleGAN2-ADA	6.86
iDDPM-DDIM	5.63
IDFF (Ours)	5.62
StyleGAN-XL	5.58
PFGM++	4.58

H SST forecasting visualization

For this task, we employed a class-conditional UNet, with the class encoder handling long-range time dependencies to enable continuous forecasting. Network structure and hyperparameters are detailed in Appendix E.

I 2D-simulated static data and time-series

Additional results for 2D simulations for both static and time-series generation are shown in Figure15.

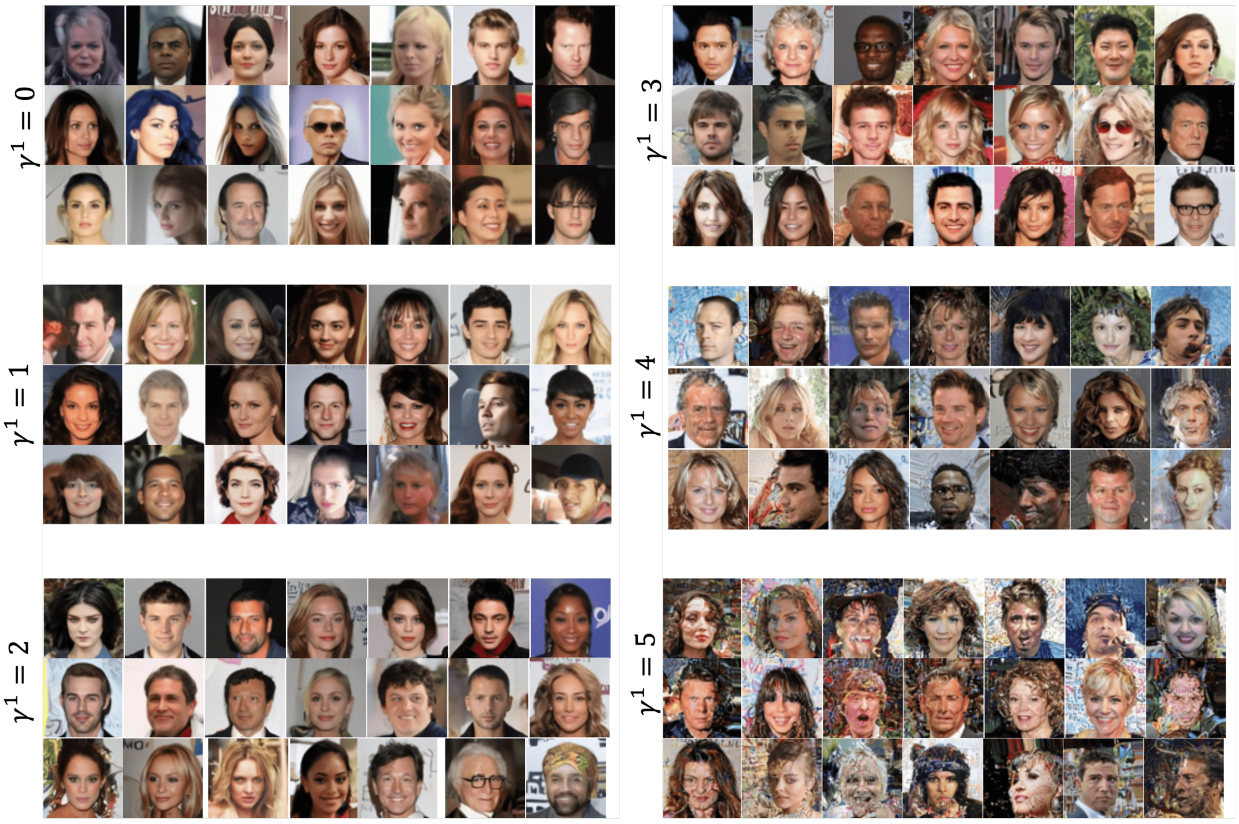

Figure 7: Generated samples for CelebA-64 (64×64) with different schedule amplitudes γ_t^1 (and the matched diffusion compensation $\sigma_t^2 = 2\gamma_t^1$) at NFE= 10.

Figure 8: Generated samples for CIFAR-10 (32×32) with different schedule amplitudes γ_t^1 (and matched diffusion compensation $\sigma_t^2 = 2\gamma_t^1$) at NFE$= 10$.

Figure 9: Generated samples for CelebA-HQ (256×256) dataset with $\sigma_0 = 0.2$ and NFE=10.

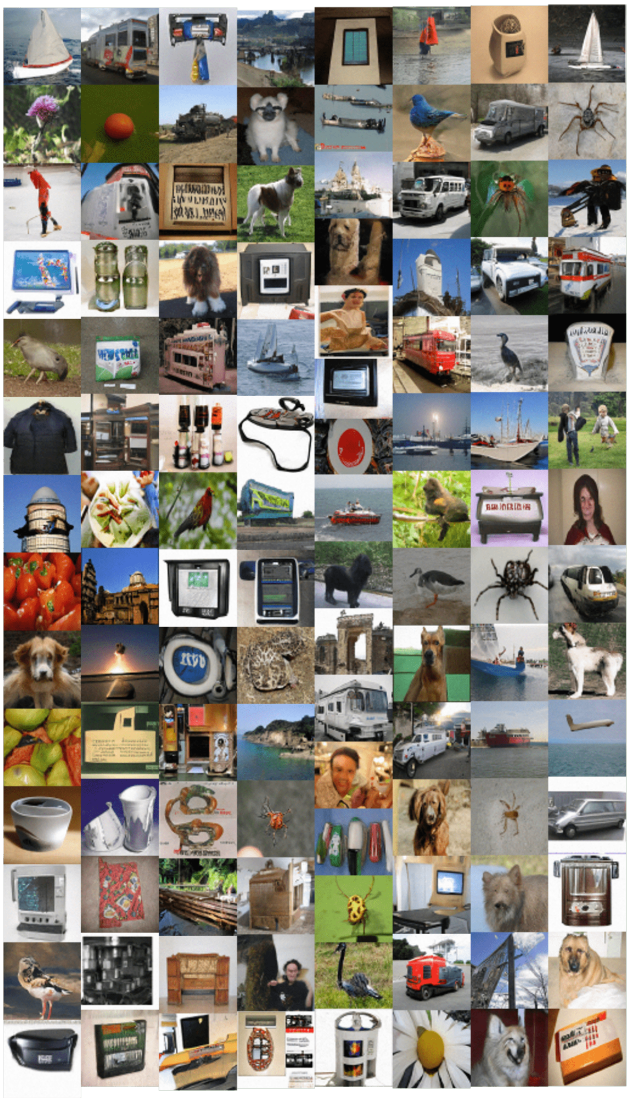

Figure 10: Generated samples for ImageNet-64 dataset with $\sigma_0 = 0.2$ and NFE=10.

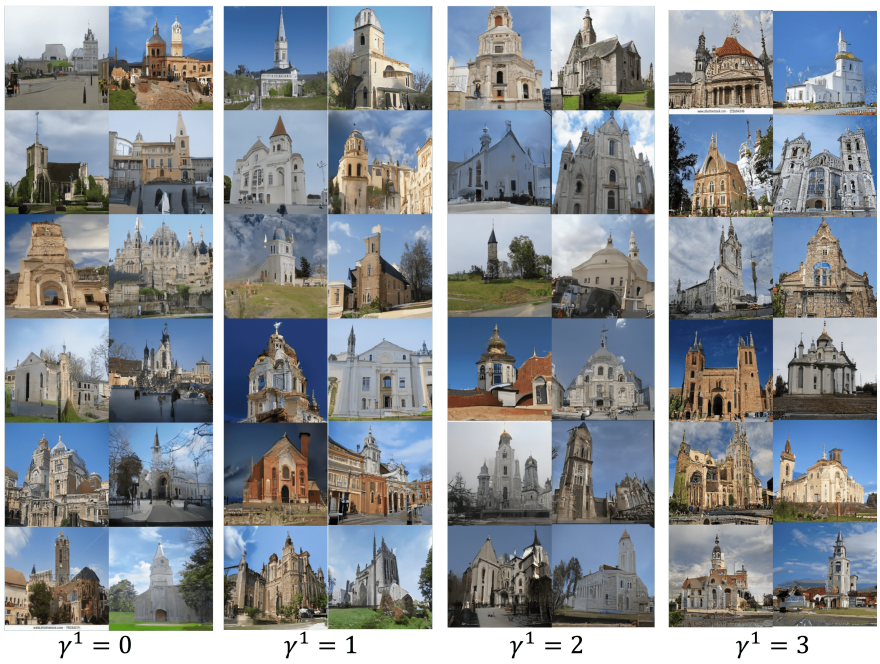

Figure 11: Generated samples for LSUN-Church (256×256) with different schedule amplitudes γ_t^1 (and matched diffusion compensation $\sigma_t^2 = 2\gamma_t^1$) at NFE$= 10$.

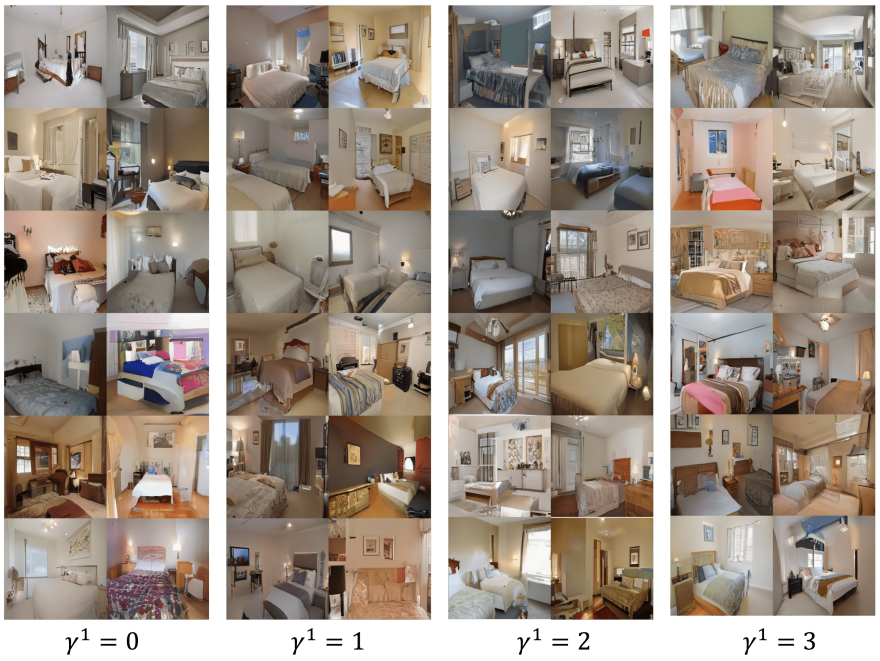

Figure 12: Generated samples for LSUN-Bedroom (256×256) with different schedule amplitudes γ_t^1 (and matched diffusion compensation $\sigma_t^2 = 2\gamma_t^1$) at NFE$= 10$.

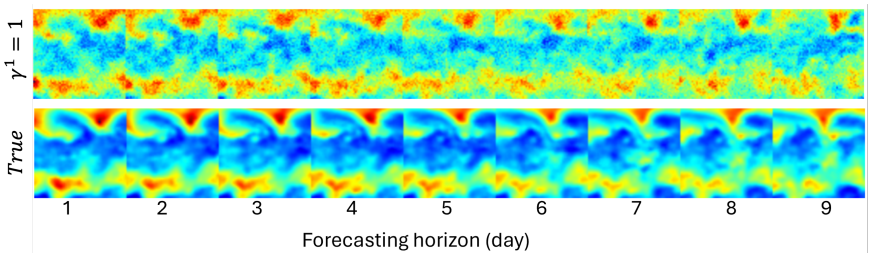

Figure 13: SST forecasting result conditioned on day 1 for 9 days with $\gamma_t^1 \equiv 1$ (constant amplitude) and fixed NFE $= 5$.

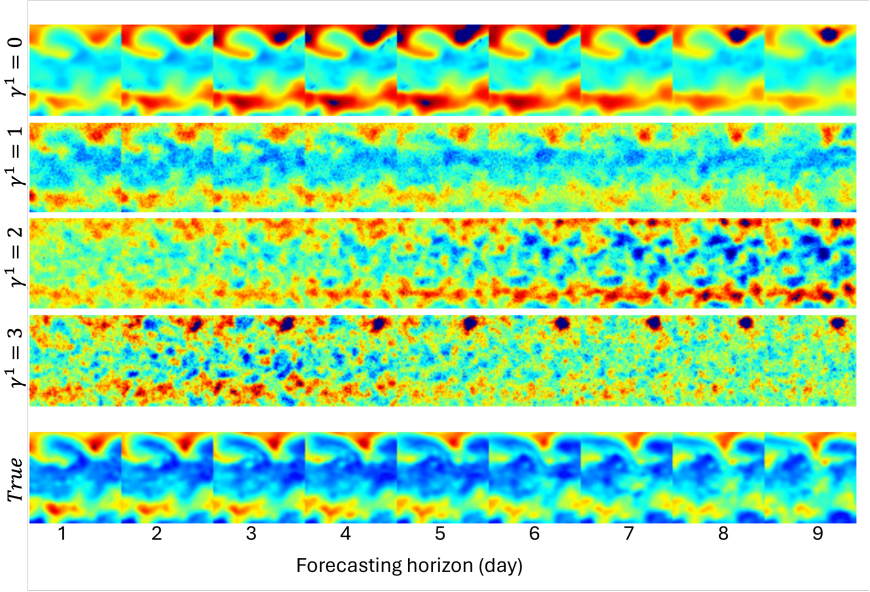

Figure 14: SST forecasting result conditioned on day 1 for 9 days for different values of γ_t^1 (with $\sigma_t^2 = 2\gamma_t^1$) and fixed NFE $= 5$.

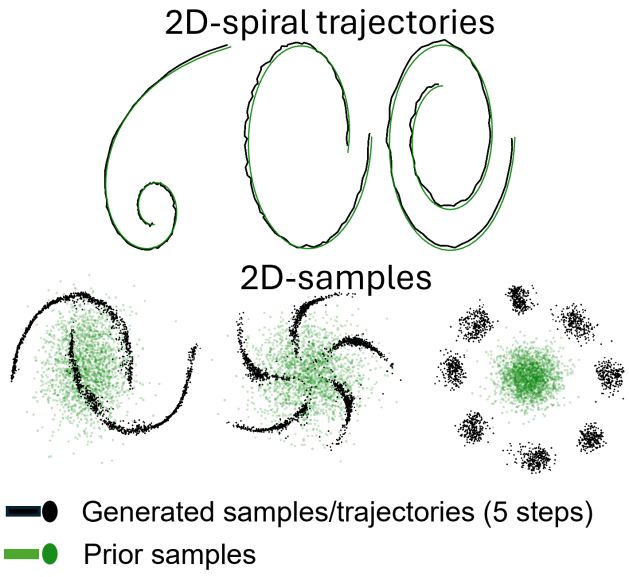

Figure 15: 2D synthetic simulation.