WHEN DOES UNROLLING HELP FLOW MATCHING? BACKPROPAGATION THROUGH TIME, OPTIMAL TRANSPORT, AND THE LINK TO CONSISTENCY AND SHORTCUT MODELS

PREPRINT

Quentin Fruytier
The University of Texas at Austin
qdf76@utexas.edu

July 4, 2026

ABSTRACT

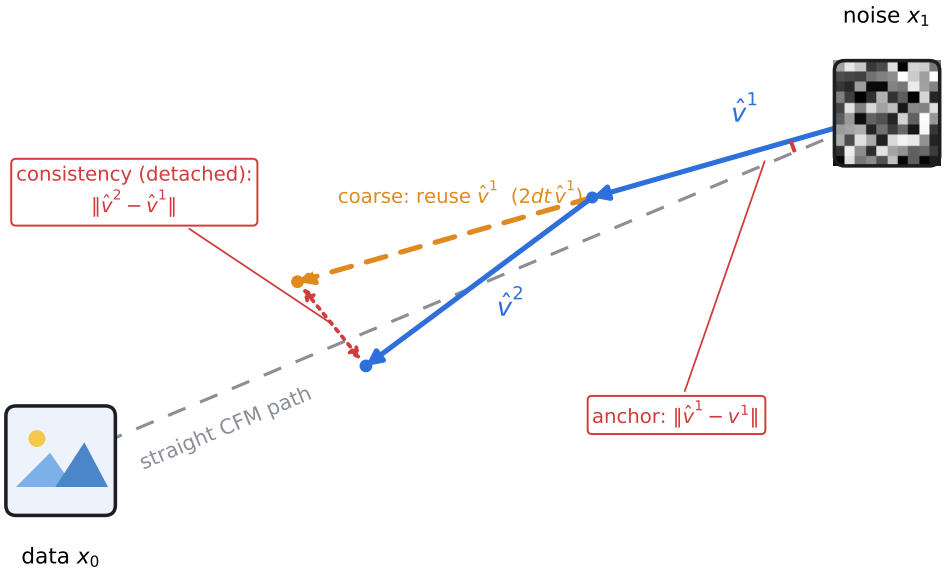
Flow-matching and diffusion models are trained with an *instantaneous* objective but are sampled by *integrating* an ODE, so training never sees the discretization error that inference compounds. The obvious fix—unroll the sampler for L steps and backpropagate through it (BPTT)—turns out to be the wrong default, but analyzing *why* yields our main result: **consistency and shortcut models fall out of unrolled flow matching once a gradient-variance argument is applied.** Expanding the k-step (negative) ELBO of the unrolled objective gives a weighted single-step flow-matching term *plus* trajectory cross-terms. The cross-terms are what let BPTT self-correct, but their gradient variance compounds as $(\mathrm{Var}(x_0 \mid x_t))^L$ and destabilizes training. Dropping them—equivalently, detaching the teacher between steps to *anchor* the objective—leaves exactly the single-step flow-matching loss plus a self-consistency loss, i.e. (up to weighting) consistency and shortcut models. We corroborate this empirically on 2D manifolds and CelebA-64^2: (i) the *field* parametrization is uniformly better than $\hat{x}_0$ under unrolling—$\hat{x}_0$'s $1/t$ loss and Jacobian scaling makes its gradient norm blow up near the data manifold; (ii) the anchored consistency objective with a geometric step schedule matches or beats both single-step training and full-BPTT NELBO unrolling; and (iii) minibatch optimal transport is a consistent variance-reducing stabilizer. Unrolled flow matching is thus best understood not as an end in itself, but as the derivation that anchors consistency- and shortcut-style training.

Keywords flow matching · diffusion models · consistency models · shortcut models · backpropagation through time · optimal transport

1 Introduction

Diffusion models (4, 8) and Continuous Normalizing Flows (2) have become the dominant paradigm for high-fidelity generative modeling. Flow Matching and Conditional Flow Matching (CFM) (5, 6, 10) recast this program as a simulation-free regression: rather than reversing a stochastic differential equation, one learns a deterministic velocity field whose ordinary differential equation (ODE) transports a simple source distribution to the data distribution along a prescribed probability path. Cold Diffusion (1) earlier established that Gaussian noise is not essential—arbitrary deterministic degradations can be reversed—and CFM makes the resulting deterministic transport explicit and easy to train.

The train/inference gap. Despite its elegance, CFM optimizes an *instantaneous* objective: it regresses the marginal velocity field at times t sampled independently and, crucially, at states that lie exactly on the ground-truth interpolation path. Inference does something categorically different—it *integrates* the learned field with a discrete solver over $[0, 1]$. A small directional error at step t displaces the sample off the data manifold and compounds at $t - \Delta t$; this



$$\mathcal{L} = 4\,\|\hat{v}^1 - v^1\|_2^2 + \|\hat{v}^2 - \hat{v}^1\|_2^2$$

Figure 1: **The WT-FM objective with cross-terms removed** (Section 4). From noise x_1, the ground-truth straight CFM path (gray) reaches the data x_0. The model's first step $\hat{v}^1$ (solid blue) is shared; the *coarse* step then continues along $\hat{v}^1$ for a doubled step (dashed orange, $2dt\,\hat{v}^1$), while the *fine* trajectory re-evaluates and turns with $\hat{v}^2$. Dropping the k-step ELBO cross-terms (Theorem 4.1) leaves two penalties: an *anchor* term $\|\hat{v}^1 - v^1\|$ matching the first step to the true velocity (single-step flow matching), and a *consistency* term $\|\hat{v}^2 - \hat{v}^1\|$ aligning the fine step with the coarse ($\hat{v}^1$-reusing) step with the teacher *detached*. Together they give $\mathcal{L} = 4\|\hat{v}^1 - v^1\|_2^2 + \|\hat{v}^2 - \hat{v}^1\|_2^2$—the consistency/shortcut-model structure.

discretization drift (a form of exposure bias) is never represented in the training signal, and the network is never asked to correct a state it did not itself produce.

Unrolling as the obvious fix. The natural remedy is to train the sampler the way it is used: unroll the discrete solver for a window of L steps, feed each step its own previous prediction, and backpropagate the terminal trajectory error through the intermediate *generated* states via Backpropagation Through Time (BPTT); we call this objective *Windowed Trajectory Flow Matching* (WT-FM). Run naively—as full BPTT on the terminal error, which we show equals the k-step negative ELBO—this does *not* beat well-tuned single-step training. Our contribution is not the objective itself, but what its analysis reveals.

Consistency and shortcut models are what unrolling reduces to (main result). Grounding the unrolled loss in the k-step (negative) ELBO and expanding it exposes a weighted single-step flow-matching term *plus* trajectory *cross-terms* (Section 3.3). The cross-terms are the only place BPTT couples steps: they enable self-correction, but their gradient variance compounds as $(\mathrm{Var}(x_0 \mid x_t))^L$ (Section 4.4) and is what makes full-BPTT unrolling fragile. A gradient-variance argument therefore *motivates dropping them*—which is exactly what detaching the teacher between steps does. What remains is the single-step FM loss plus a self-consistency loss between successive velocity estimates: up to weighting, *consistency models* (9) and *shortcut models* (3) (Section 4.2; illustrated in Figure 1). In this sense consistency and shortcut training are not separate heuristics but the *variance-anchored limit* of unrolled flow matching.

Field beats $\hat{x}_0$, and the reason is in the gradients. Empirically, the field parametrization is uniformly better than $\hat{x}_0$ under unrolling—on our 2D manifolds $\hat{x}_0$ is 130%–690% worse in Wasserstein-1 even with optimal-transport couplings, and on CelebA-64^2 its FID trails field. The mechanism (Sections C.3 and 4.4) is the $1/t$ factor in the $\hat{x}_0$ loss and transition Jacobian: at matched training loss it inflates the gradient norm near the data manifold ($t \to 0$) and makes it swing wildly across noise levels, whereas the field Jacobian $\mathbf{I} + dt\,\mathbf{J}_{\hat{v}}$ is numerically inert and stable.

Minibatch optimal transport (10, 7) further reduces gradient variance by collapsing $\mathrm{Var}(x_0 \mid x_t)$, and is a consistent stabilizer across all settings.

Contributions.

- **Consistency and shortcut models from unrolled BPTT plus variance minimization (main result).** We expand the k-step ELBO of the unrolled sampler into a single-step term plus cross-terms (Section 3.3), show that dropping the cross-terms leaves the single-step FM loss plus a self-consistency loss (Theorem 4.1), and show the *same* cross-terms drive an exponential-in-L gradient variance that dropping them removes (Theorem 4.2). Consistency and shortcut models are thus the variance-anchored limit of unrolled flow matching.
- **Why field beats $\hat{x}_0$ under unrolling** (Sections 3.2 and 4.4): the $\hat{x}_0$ loss/Jacobian $1/t$ scaling causes a near-manifold gradient blow-up, while the field Jacobian is a stable near-identity; we document this directly in gradient-norm measurements (Section C.3).
- **An empirical study** (Section 5) on 2D manifolds and CelebA-64^2: field $\gg \hat{x}_0$ under unrolling; the anchored consistency objective with a geometric step schedule matches or beats both single-step training and full-BPTT NELBO unrolling; and minibatch OT is a consistent stabilizer.

2 Related Work

Diffusion and flow-matching models. Diffusion models learn to reverse a noising SDE (4, 8); continuous normalizing flows (2) instead learn an ODE velocity field. Flow Matching and Conditional Flow Matching (CFM) (5, 6, 10) train that velocity by regression along a prescribed probability path, and Cold Diffusion (1) showed the corruption need not be Gaussian. All share the train/inference mismatch we study: the objective is instantaneous, but sampling integrates an ODE.

Optimal-transport couplings. Minibatch OT couplings straighten the CFM interpolant and reduce gradient variance by pairing source and target within a batch (10, 7). We adopt them and find they are the key stabilizer for unrolling.

Consistency and shortcut models. Consistency models (9) train a network whose outputs are self-consistent along a probability-flow trajectory—typically by distilling from a frozen teacher—so that one or few steps suffice at inference. Shortcut models (3) condition on the step size and add a self-consistency term tying a coarse jump to two fine steps. Both replace multi-step backpropagation with a *detached* self-consistency objective. Our contribution is to *derive* exactly this structure from the unrolled flow-matching ELBO: dropping the trajectory cross-terms (equivalently, detaching the teacher) reduces unrolled BPTT to a single-step FM loss plus a self-consistency loss (Section 4), a reduction we motivate by a gradient-variance argument.

Unrolling, exposure bias, and BPTT. Training the sampler the way it is used—unrolling the solver and backpropagating through it—is a natural remedy for the exposure bias created by teacher-forced training. We characterize when this helps for flow matching, and find that full BPTT is dominated by its detached (consistency/shortcut) limit for variance reasons.

3 Background

Let $p_1(x)$ denote the source distribution (e.g., standard Gaussian noise) and $p_0(x)$ denote the target data distribution. CFM defines a probability path $p_t(x)$ that smoothly interpolates between these distributions for $t \in [0, 1]$.

The most common formulation utilizes a linear interpolation path to define the ground-truth intermediate state x_t:

$$x_t = tx_1 + (1-t)x_0 \quad \text{for } t \in [0,1] \tag{1}$$

The optimal continuous vector field $v_t(x_t)$ that generates this deterministic flow is given by the derivative of the path w.r.t. t: $v_t(x_t) = \frac{x_0 - x_t}{t}$.

3.1 The Single-Step Training Objective

Writing $x_1 = \epsilon$ for the source (noise) sample and x_0 for the data sample, standard CFM regresses the network onto the marginal velocity. The two equivalent forms we use are the noise/velocity form and the target (x_0) form:

$$\min_\theta \mathbb{E}_{x_0,\epsilon,t;\, x_t=t\epsilon+(1-t)x_0} \left[\|(x_0 - \epsilon) - v_\theta(x_t, t)\|_2^2 \right], \tag{2}$$

$$\min_\theta \mathbb{E}_{x_0,\epsilon,t;\, x_t=t\epsilon+(1-t)x_0} \left[\left\| \tfrac{x_0 - x_t}{t} - v_\theta(x_t, t) \right\|_2^2 \right]. \tag{3}$$

In the target parametrization the network instead outputs a clean-data estimate $\hat{x}_0(x_t, t)$, and the velocity is defined implicitly through $v_\theta(x_t, t) = \frac{\hat{x}_0(x_t,t) - x_t}{t}$. Throughout, f_θ denotes the raw network output (v_θ or $\hat{x}_0$ depending on the parametrization).

The goal of CFM is to approximate this dynamic process. Consider a discrete Euler integration step of size dt pointing backward in time towards the data at $t = 0$:

$$x^*_{t-dt} = x_t + v_t(x_t)\, dt = x_t + \frac{dt}{t}(x_0 - x_t). \tag{4}$$

The network produces the corresponding state $\hat{x}_{t-dt}$ (through whichever parametrization is in use). The idealized single-step training objective minimizes the scaled L_2 norm of this step prediction error:

$$\mathcal{L}_{step} = \frac{1}{2} \left(\frac{1}{dt} \right)^2 \mathbb{E}_{t,x_0,x_1} \left[\|\hat{x}_{t-dt} - x^*_{t-dt}\|_2^2 \right]. \tag{5}$$

Expanding the terms shows this step-loss equals the continuous vector-field loss $\mathcal{L}_{VF} = \|\hat{v}(x_t, t) - v_t(x_t)\|^2$ scaled by $(dt)^2$.

3.2 Network parametrizations and single-step gradients

In practice, this dynamic is learned by training a network $f_\theta(x_t, t)$ on the single-step objective. The choice of what f_θ outputs alters the magnitude and geometry of the optimization landscape—and, as we show in Section 4, it is the single factor that determines whether unrolled BPTT is stable. We use the two parametrizations below; a third, direct-step prediction, behaves as an exploding RNN under unrolling and is deferred to Section A.1.

Let $e_{t-dt} = (\hat{x}_{t-dt} - x^*_{t-dt})$ denote the localized step prediction error. For a single-step window ($L = 1$), the parameter gradient is $\nabla_\theta \mathcal{L} = \left(\frac{1}{dt}\right)^2 e_{t-dt}^\top \nabla_\theta \hat{x}_{t-dt}$.

Velocity / field prediction ($\hat{x}_{t-dt} = x_t + dt\, f_\theta(x_t, t)$). The network outputs the vector field directly; the derivative w.r.t. θ extracts a factor of dt:

$$\nabla_\theta \mathcal{L}_{step} = (\hat{v}(x_t, t) - v_t(x_t))^\top \nabla_\theta f_\theta, \qquad \nabla_\theta \hat{x}_{t-dt} = dt\, \nabla_\theta f_\theta. \tag{6}$$

Target / $\hat{x}_0$ prediction ($\hat{x}_{t-dt} = x_t + \frac{dt}{t}(f_\theta(x_t, t) - x_t)$). The network outputs the clean-data estimate $\hat{x}_0(x_t, t)$ and the velocity is implicit; the derivative extracts a *dynamic* factor dt/t (the source of the near-manifold blow-up analyzed in Section 4):

$$\nabla_\theta \mathcal{L}_{step} = \frac{1}{t^2}(\hat{x}_0 - x_0)^\top \nabla_\theta f_\theta, \qquad \nabla_\theta \hat{x}_{t-dt} = \frac{dt}{t} \nabla_\theta f_\theta. \tag{7}$$

3.3 The k-step ELBO

3.3.1 Fundamental derivation of the diffusion ELBO

To establish the theoretical foundation of our unrolled training objective, we ground our loss in the Evidence Lower Bound (ELBO) of the data log-likelihood. Let x_0 be a data sample, and let $z_1, z_2, \ldots, z_N$ represent a standard forward diffusion process where noise is incrementally added until z_N becomes pure Gaussian noise.

Our goal is to maximize the marginal log-likelihood of the data, $\log p_\theta(x_0)$. By introducing the latent variables $z_{1:N}$ and the forward process $q(z_{1:N}|x_0)$, we can apply Jensen's Inequality to derive the variational lower bound:

$$\log p_\theta(x_0) = \log \int p_\theta(x_0, z_{1:N})\, dz_{1:N} \tag{8}$$

$$= \log \mathbb{E}_{q(z_{1:N}|x_0)} \left[\frac{p_\theta(x_0, z_{1:N})}{q(z_{1:N}|x_0)} \right] \tag{9}$$

$$\geq \mathbb{E}_{q(z_{1:N}|x_0)} \left[\log \frac{p_\theta(x_0, z_{1:K})}{q(z_{1:N}|x_0)} \right] \triangleq \text{ELBO} \tag{10}$$

We factorize the reverse generative process as $p_\theta(x_0, z_{1:N}) = p(z_N) \prod_{k=1}^{N} p_\theta(z_{k-1}|z_k)$, and the forward noising process as $q(z_{1:N}|x_0) = \prod_{k=1}^{N} q(z_k|z_{k-1})$. To make the objective tractable, we rewrite the forward process transitions using Bayes' theorem conditioned on x_0:

$$q(z_k|z_{k-1}) = \frac{q(z_{k-1}|z_k, x_0) q(z_k|x_0)}{q(z_{k-1}|x_0)} \tag{11}$$

Substituting this into the ELBO allows the fractional products to telescope perfectly. Isolating the conditioning terms yields the classical diffusion decomposition:

$$\text{ELBO} = \mathbb{E}_q \left[\log p(z_N) - \log q(z_N|x_0) + \sum_{k=2}^{N} \log \frac{p_\theta(z_{k-1}|z_k)}{q(z_{k-1}|z_k, x_0)} + \log p_\theta(x_0|z_1) \right] \tag{12}$$

Because the first term is a constant prior with no learnable parameters, the optimization objective simplifies to minimizing the sum of Kullback-Leibler (KL) divergences between the target denoising posterior and the model's reverse prediction.

To align this fundamental bound with the notation of our unrolled continuous ODE solver, we define x_i as the state of the model at integration step i. In this formulation, integration progresses backward in diffusion time: $i = 1$ corresponds to pure noise (z_N), and $i = N$ corresponds to the clean data (x_0). Mapping the indices accordingly ($z_k \to x_i$, $z_{k-1} \to x_{i+1}$), our discrete continuous-time objective becomes:

$$\mathcal{L}_{\text{ELBO}} = \sum_{i=1}^{N-1} \mathbb{E}_{q(x_{t_i}|x_{t_0})} \left[D_{\text{KL}} \Big(q(x_{t_{i+1}}|x_{t_i}, x_{t_0}) \parallel p_\theta(x_{t_{i+1}}|x_{t_i}) \Big) \right] \tag{13}$$

For continuous-time ODEs (such as Flow Matching and DDIM), these transitions can be modeled as Gaussians with an infinitesimally small variance $\sigma^2 \to 0$. The KL divergence between two Gaussians with identical variance reduces strictly to the scaled squared L_2 distance of their means: $D_{\text{KL}}(\mathcal{N}(\mu_q, \Sigma) \parallel \mathcal{N}(\mu_p, \Sigma)) = \frac{1}{2\sigma^2} \|\mu_q - \mu_p\|^2$.

3.3.2 The single-step instantaneous ELBO

In standard 1-step training, we evaluate the marginal transition from t_i to t_{i+1}. The target distribution $q(x_{i+1}|x_i, x_0)$ follows the true vector field $v^i = \frac{x_0 - x_i}{t_i}$, yielding a mean of $\mu_q = x_i + v^i dt_i$. The model distribution $p_\theta(x_{i+1}|x_i)$ follows the predicted vector field $\hat{v}^i$, yielding a mean of $\mu_p = x_i + \hat{v}^i dt_i$.

Substituting these means into the Gaussian KL divergence, the 1-step bound evaluates exactly to the scaled, instantaneous velocity error:

$$\mathcal{L}_{\text{1-step}} = \sum_{i=1}^{N-1} \mathbb{E}_{q(x_{t_i}|x_{t_0})} \left[\frac{1}{2\sigma^2} \|(x_{t_i} + v^i dt_i) - (x_{t_i} + \hat{v}^i dt_i)\|^2 \right] \tag{14}$$

$$\propto \sum_{i=1}^{N-1} \mathbb{E}_{q(x_{t_i}|x_{t_0})} \left[dt_i^2 \|v^i - \hat{v}^i\|^2 \right] \tag{15}$$

As shown in Eq. 15, the 1-step objective strictly penalizes local vector field deviations under teacher-forcing, remaining entirely agnostic to how these errors propagate through future integration steps. (Note: in practical implementations, the dt_i^2 term is often empirically discarded to stabilize optimization, prioritizing uniform weighting across time).

3.3.3 The k-step trajectory ELBO

When unrolling the solver for k steps, we no longer optimize the marginal 1-step conditional probabilities. Instead, we divide the N-step trajectory into sub-trajectories and optimize the joint transition probability over k-step blocks. For a block beginning at t_1 and ending at t_k, the objective evaluates the multi-step conditional divergence:

$$\mathcal{L}_{k\text{-step}} \propto \sum_{i=1}^{N/k-1} \mathbb{E}_{q(x_{t_N}|x_{t_0})} \left[D_{\mathrm{KL}}\Big(q(x_{t_{i+k}}|x_{t_i}, x_{t_0}) \parallel p_\theta(x_{t_{i+k}}|x_{t_i}) \Big) \right] \tag{16}$$

Here, the mean of the unrolled target distribution μ_q corresponds to the exact ground-truth linear interpolation x_k^*. The mean of the model's unrolled distribution μ_p corresponds to the accumulated approximated state $\hat{x}_k$. Applying the Gaussian KL reduction, the block-wise divergence perfectly collapses into the expected squared norm of the cumulative structural error E_k:

$$\mathcal{L}_{k\text{-step}} \propto \mathbb{E}_{q(x_{t_N}|x_{t_0})} \left[\|\hat{x}_{t_k} - x_{t_k}^*\|^2 \right] = \mathbb{E}_{q(x_{t_N}|x_{t_0})} \left[\|E_k\|^2 \right] \tag{17}$$

Substituting our previous exact derivation for E_k, the k-step ELBO is given by:

$$\mathcal{L}_{k\text{-step}} = \mathbb{E}_{x_{t_0}\sim P_x, x_{t_N}\sim q(x_{t_N}|x_{t_0}), t_0\sim P_t} \left[w_{t_k} \left\| \sum_{i=0}^{k-1} (\hat{v}^i - v^i) \frac{dt_i \cdot t_k}{t_{i+1}} \right\|^2 \right] \tag{18}$$

$$= \mathbb{E}_{x_{t_0}\sim P_x, x_{t_N}\sim q(x_{t_N}|x_{t_0}), t_0\sim P_t} \left[w_{t_k} \left\| \sum_{i=0}^{k-1} (\hat{x}_0^i - x_0) \left(\frac{t_k}{t_{i+1}} - \frac{t_k}{t_i} \right) \right\|^2 \right] \tag{19}$$

By algebraically expanding the squared sum, we isolate the fundamental difference between 1-step and k-step optimization:

$$\mathcal{L}_{k\text{-step}} = \mathbb{E}_{q(x_1|x_0)} \left[\underbrace{\sum_{i=1}^{k-1} \|\hat{v}^i - v^i\|^2 \left(\frac{dt_i \cdot t_k}{t_{i+1}} \right)^2}_{\text{Weighted 1-Step ELBO Terms}} + \underbrace{\sum_{i\neq j} (\hat{v}^i - v^i)^T (\hat{v}^j - v^j) \frac{dt_i dt_j t_k^2}{t_{i+1} t_{j+1}}}_{\text{Trajectory Self-Correction Cross-Terms}} \right] \tag{20}$$

This expansion reveals that the k-step ELBO is not merely a sum of local errors, but a joint trajectory bound. The presence of the dot-product cross-terms enables negative interference: an error introduced at step i can be explicitly offset by a compensatory error at step j. Consequently, unrolling inherently trains the model to anticipate and correct its own cumulative exposure bias. These same cross-terms drive the paper's central tension, and we return to them twice: Theorem 4.1 shows that *dropping* them collapses the objective to the single-step flow-matching loss plus a consistency loss (recovering consistency and shortcut models), while Theorem 4.2 shows that they are exactly what makes the unrolled gradient variance grow exponentially in the window. The cross-terms are thus simultaneously the source of self-correction and of instability—which is why, empirically, the detached (consistency/shortcut) form outperforms full BPTT.

4 Method: Windowed Trajectory Flow Matching

WT-FM trains the sampler the way it is used: from a sampled start time it unrolls the discrete solver for a window of L steps, feeds each step its own previous prediction, and penalizes the terminal trajectory error (Algorithm 1). The training signal therefore sees the same compounding discretization error as inference. Optimizing the objective naively requires backpropagation through time (BPTT) across the intermediate *generated* states. Below we (i) identify this objective as the k-step NELBO and decompose it (Section 4.1); (ii) reduce it, via a gradient-variance argument, to a single-step term plus a self-consistency term—recovering consistency and shortcut models (Section 4.2); and (iii) justify the two remaining design choices, the *field* parametrization (Section 4.3) and *minibatch-OT* couplings (Section 4.4). The resulting method is **minibatch OT + field parametrization + dropped cross-terms (anchored consistency)**.

Algorithm 1 Windowed Trajectory Flow Matching (WT-FM): one training step

Require: data x_0, source noise x_1; window L; step size dt; network f_θ; batch coupling π
$(x_0, x_1) \leftarrow \pi(x_0, x_1)$ ▷ minibatch-OT re-coupling
sample $t_0 \sim \mathcal{U}[L\,dt,\, 1]$; $\hat{x} \leftarrow t_0 x_1 + (1 - t_0)\, x_0$ ▷ teacher-forced window start
for $i = 1, \ldots, L$ **do**
  $t \leftarrow t_0 - (i{-}1)\, dt$
  $\hat{v} \leftarrow f_\theta(\hat{x}, t)$ ▷ field output, or $\hat{v} = (\hat{x}_0 - \hat{x})/t$ for the $\hat{x}_0$ parametrization
  $\hat{x} \leftarrow \hat{x} - dt\, \hat{v}$ ▷ Euler step on the model's own state (BPTT keeps the graph)
end for
$x^\star \leftarrow (t_0 - L\, dt)\, x_1 + (1 - t_0 + L\, dt)\, x_0$ ▷ true state after L steps
$\mathcal{L} \leftarrow \frac{1}{2}(1/dt)^2 \, \|\hat{x} - x^\star\|^2$ ▷ terminal error = k-step NELBO
(anchored variant: replace $\mathcal{L}$ by single-step FM + detached self-consistency, Section 4.2)
$\theta \leftarrow \theta - \eta\, \nabla_\theta \mathcal{L}$

4.1 The cumulative error and its anchor–consistency form

Let the integration process follow a monotonically decreasing time schedule $t_1, t_2, \ldots, t_k$, where t_1 is the initial time with partial noise and t_k is the current time. We define the dynamic step size at iteration i as $dt_i = t_{i+1} - t_i$. Note that because time flows backward in diffusion ($t_{i+1} < t_i$), dt_i is strictly negative.

Given the discrete Euler update for the approximated state $\hat{x}$ at iteration i, moving from t_i to t_{i+1}:

$$\hat{x}_{i+1} = \hat{x}_i + (\hat{x}_0^i - \hat{x}_i)\frac{t_i - t_{i+1}}{t_i} = \hat{x}_i - (\hat{x}_0^i - \hat{x}_i)\frac{dt_i}{t_i} \tag{21}$$

We define the true trajectory x_k^* as the case where the model perfectly predicts the ground truth x_0 at every step (i.e., $\hat{x}_0^i = x_0$). The cumulative error after reaching t_k is defined as $E_k = \hat{x}_k - x_k^*$.

Anchor–consistency decomposition. Unrolling the recurrence and telescoping (full derivation, backward-anchored variant, and $\hat{x}_0$-specific closed form in Section F) gives the exact cumulative error over any schedule:

$$E_k = -\sum_{i=1}^{k-1}(\hat{x}_0^i - x_0)\,\frac{dt_i\, t_k}{t_i t_{i+1}} = -t_k \sum_{i=1}^{k-1}(\hat{x}_0^i - x_0)\Big(\tfrac{1}{t_{i+1}} - \tfrac{1}{t_i}\Big). \tag{22}$$

Summation by parts (with $c_i \triangleq 1 - t_k/t_i$, decreasing in i, $c_k = 0$) rewrites this as an *anchor* term plus a *consistency* term; substituting the field relation $\hat{v}^i = (\hat{x}_0^i - \hat{x}_i)/t_i$ gives the equivalent velocity form:

$$E_k = -\Big[\underbrace{(\hat{x}_0^1 - x_0)\, c_1}_{\text{anchor}} + \underbrace{\textstyle\sum_{i=2}^{k-1}(\hat{x}_0^i - \hat{x}_0^{i-1})\, c_i}_{\text{consistency}}\Big] \tag{23}$$

$$= -\Big[(\hat{v}^1 - v^1)\, t_1 c_1 + \textstyle\sum_{i=2}^{k-1}(\hat{v}^i - \hat{v}^{i-1})\, t_i c_i\Big], \qquad t_i c_i = t_i - t_k. \tag{24}$$

The structural content is already visible: *the cumulative trajectory error is the single-step error at the window start* (anchor) *plus the step-to-step disagreement of successive clean-data / velocity estimates* (consistency)—exactly the quantity consistency models penalize. In the fully-denoised limit $t_k = 0$ it collapses to the last-step error $(\hat{x}_0^{k-1} - x_0)$.

4.2 Dropping the Cross-Terms: Reduction to Consistency and Shortcut Models

We now make precise the paper's central structural claim. Recall from Section 3.3 that expanding $\|E_k\|^2$ produces *weighted single-step terms plus cross-terms* $\sum_{i \neq j}(\hat{v}^i - v^i)^\top(\hat{v}^j - v^j)(\cdot)$. The cross-terms are the only place where BPTT couples different steps—they are what lets an error at step i be cancelled by a compensating error at step j—and, as we show in Section 4.4, they are also the dominant source of gradient-variance blow-up. Suppose we simply *drop* them (equivalently: detach the teacher between steps, so no gradient flows across the trajectory). What remains?

Starting from the field-error form and bounding the squared sum with Cauchy–Schwarz (which is exactly what discarding cross-terms achieves up to a constant):

$$\|E_k\|_2^2/(K^2 \cdot dt^2) = \frac{1}{K^2}\left\|K(\hat{v}^1 - v^1) + \sum_{i=2}^{k}\left(\hat{v}^i - \hat{v}^{i-1}\right)(K-i)\right\|_2^2 \tag{25}$$

$$\leq \frac{2}{K^2}\left[\left\|K(\hat{v}^1 - v^1)\right\|_2^2 + \left\|\sum_{i=2}^{k}\left(\hat{v}^i - \hat{v}^{i-1}\right)(K-i)\right\|_2^2\right] \tag{26}$$

$$= \frac{2}{K^2}\left[\left\|K(\hat{v}^1 - v^1)\right\|_2^2 + \left\|\sum_{i=2}^{k}\left(\hat{v}^i - \hat{v}^{1}\right)\right\|_2^2\right] \tag{27}$$

$$\leq \frac{2}{K^2}\left[\left\|K(\hat{v}^1 - v^1)\right\|_2^2 + (K-1)\sum_{i=2}^{k}\left\|\left(\hat{v}^i - \hat{v}^1\right)\right\|_2^2\right] \tag{28}$$

$$\approx \mathbb{E}_{i\sim Unif\{2,\dots,k-1\}}\left[\left\|\hat{v}^1 - v^1\right\|_2^2 + \left(\frac{K-1}{K}\right)^2\left\|\left(\hat{v}^i - \hat{v}^1\right)\right\|_2^2\right] \tag{29}$$

$$= \left\|\hat{v}^1 - v^1\right\|_2^2 + \left(\frac{K-1}{K}\right)^2 \mathbb{E}_{i\sim Unif\{2,\dots,k-1\}}\left[\left\|(K-i)\left(\hat{v}^i - \hat{v}^{i-1}\right)\right\|_2^2\right] \tag{30}$$

The right-hand side is a sum of two familiar pieces: the *classical single-step flow-matching loss* $\|\hat{v}^1 - v^1\|^2$, plus a *self-consistency loss* $\|\hat{v}^i - \hat{v}^{i-1}\|^2$ that penalizes disagreement between successive velocity estimates. We state this as our main structural result.

Proposition 4.1 (Cross-term-free unrolling is single-step FM plus a consistency loss)**.** *Let $\hat{v}^i$ be the velocity estimates along an L-step unrolled trajectory with v^1 the ground-truth velocity at the window start. Discarding the trajectory cross-terms of the k-step objective (equivalently, detaching the teacher between steps) upper-bounds the terminal error by*

$$\tfrac{1}{K^2 dt^2}\|E_k\|_2^2 \;\lesssim\; \underbrace{\left\|\hat{v}^1 - v^1\right\|_2^2}_{\text{single-step FM loss}} \;+\; \left(\tfrac{K-1}{K}\right)^2 \underbrace{\mathbb{E}_{i\sim\mathrm{Unif}\{2,\dots,k-1\}}\left[\left\|(K-i)\left(\hat{v}^i - \hat{v}^{i-1}\right)\right\|_2^2\right]}_{\text{self-consistency / straightening loss}}. \tag{31}$$

Approximating the intermediate states by the tangent line $\hat{x}^{i-1} \approx x_{t_0} + (i-1)\,dt\,\hat{v}^1$ (a two-step trajectory) collapses the consistency term to a comparison against a single reference velocity:

$$\|\hat{v}^1 - v^1\|_2^2 + \left(\tfrac{K-1}{K}\right)^2 \mathbb{E}_{i\sim\mathrm{Unif}\{2,\dots,k-1\}}\left[\left\|(K-i)\left(\hat{v}^i - \hat{v}^{i-1}\right)\right\|_2^2\right] \tag{32}$$

$$\approx \|\hat{v}^1 - v^1\|_2^2 + \tfrac{1}{4}\,\mathbb{E}_{i\sim\mathrm{Unif}\{2,\dots,k-1\}}\left[\left\|\hat{v}^{i\cdot dt} - \hat{v}^1\right\|_2^2\right]. \tag{33}$$

This is, up to weighting, the shortcut-model objective (3): a flow-matching term plus a self-consistency term between a fine and a coarse step. If shortcut models did *not* condition on the step size dt, their loss would read

$$\left\|\hat{v}^0 - v^0\right\|_2^2 + \mathbb{E}_{i\sim\mathrm{Unif}\{0,\dots,7\}}\left[\left(\tfrac{1}{2}\right)^2\left\|\hat{v}^{2^i\cdot dt}_{\mathrm{detach}} - \hat{v}^1\right\|_2^2\right]. \tag{34}$$

One discrepancy remains: the shortcut sampling strategy weights discretizations near dt much more heavily (geometric growth of the jump), whereas our derivation suggests a more uniform weighting. We attribute the gap to training stability or to the approximation error incurred when several model passes are collapsed into one large jump. Theorem 4.1 is the sense in which *unrolled flow matching relaxes to consistency and shortcut models once the cross-terms are removed*—and, empirically (Section 5), this detached form is what actually helps, whereas full BPTT with the cross-terms is fragile.

4.3 Why the field parametrization: transition Jacobians

BPTT stability is governed by the transition Jacobian $\mathbf{T}_t \triangleq \nabla_{x_t}\hat{x}_{t-dt}$, whose product $\prod_t \mathbf{T}_t$ carries the unrolled gradient. For the two parametrizations,

$$\mathbf{T}_t^{\text{field}} = \mathbf{I} + dt\,\mathbf{J}_{\hat{v}_\theta}, \qquad \mathbf{T}_t^{\hat{x}_0} = \left(1 - \tfrac{dt}{t}\right)\mathbf{I} + \tfrac{dt}{t}\,\mathbf{J}_{\hat{x}_0}. \tag{35}$$

The field Jacobian is a near-identity ($dt \ll 1$): its cross-step products are $\mathcal{O}(dt^2)$ and vanish, so BPTT behaves almost like a per-step stop-gradient—numerically inert but *stable*. The $\hat{x}_0$ Jacobian is adaptive—near the data manifold ($t \to dt$) it becomes $\mathbf{J}_{\hat{x}_0}$ and passes full cross-step gradient—but the $1/t$ factor it shares with the $\hat{x}_0$ loss (Section 3.2) makes the gradient blow up there. Empirically this is decisive: at matched single-step train loss, $\hat{x}_0$ carries $\sim 3\times$ the mean gradient norm and a far larger swing *across* noise levels (not higher per-level variance), with frequent clipping, whereas field gradients are stable (Section C.3). **We therefore adopt the field parametrization.** The full three-parametrization analysis (including the direct-step "exploding-RNN" case) and the gradient-stability measurements are deferred to Sections A, A.1 and C.3.

4.4 Why minibatch OT: collapsing the conditional variance

The remaining instability is statistical. With independent couplings, the conditional target variance $\mathrm{Var}(x_0 \mid x_t)$ grows exponentially as $t \to 1$ (strictly linear on a log axis across our benchmarks; Section B, Figures 3 and 4), and by the law of total variance the single-step gradient variance is proportional to it (full derivation in Section B). Under unrolling the L transition Jacobians multiply, so the variance compounds multiplicatively:

$$\mathrm{Var}\left(\prod_{j=1}^{L} \mathbf{T}^{target}_{t_0 - j\,dt}\right) \approx \left(\mathrm{Var}(x_0 | x_{t_0})\right)^L . \tag{36}$$

In the high-noise regime, multiplying the large pairing uncertainty L times causes the gradient to explode, feeding the optimizer statistical noise. This exponential-in-L dependence is a direct consequence of *keeping the trajectory cross-terms*; removing them changes the scaling qualitatively.

Lemma 4.2 (Dropping the cross-terms cuts loss and gradient variance)**.** *Write the terminal residual as* $E_k = \sum_{i=1}^{L} w_i\, \varepsilon_i$ *with per-step residuals* $\varepsilon_i = \hat{v}^i - v^i$ *and schedule weights* w_i*, and let* $\sigma^2 = \max_i \mathrm{Var}(\nabla_\theta \|\varepsilon_i\|^2)$*. In* full BPTT *the gradient of the terminal loss* $\|E_k\|^2$ *propagates through the product* $\prod_j \mathbf{T}_{t_0 - j\,dt}$*, so by Equation* (36) *its variance is multiplicative in the window,* $\mathrm{Var}(\nabla_\theta \mathcal{L}_k) \sim (\mathrm{Var}(x_0 \mid x_t))^L$*. If the cross-terms are dropped—i.e. the teacher is detached between steps so that gradients do not flow across the trajectory—the objective reduces to the additive form* $\sum_i w_i^2 \|\varepsilon_i\|^2$ *(Theorem 4.1), and its gradient is a sum of* L *single-step gradients with no shared random dependence, giving*

$$\mathrm{Var}\Big(\nabla_\theta \textstyle\sum_{i=1}^{L} w_i^2 \|\varepsilon_i\|^2\Big) \;=\; \sum_{i=1}^{L} w_i^4\, \mathrm{Var}\big(\nabla_\theta \|\varepsilon_i\|^2\big) \;=\; \mathcal{O}\big(L\,\sigma^2\big). \tag{37}$$

Hence dropping the cross-terms reduces the window-dependence of the gradient variance from exponential $(\mathrm{Var})^L$ *to* linear *in* L*.*

Theorem 4.2 is the variance-side counterpart of Theorem 4.1: the cross-terms that make unrolling expressive (self-correction) are exactly the terms whose variance compounds. Detaching the teacher—i.e. moving to the consistency/shortcut form—is therefore not only a modeling choice but a variance-reduction mechanism, which is consistent with our empirical finding that the detached objectives are what actually help. Minibatch OT attacks the same quantity from the other side, by shrinking the base factor $\mathrm{Var}(x_0 \mid x_t)$ itself.

Concretely, a non-crossing (bijective) batch assignment makes x_1 determine x_0, so $\mathrm{Var}_{\mathrm{OT}}(x_0 \mid x_t) \approx 0$ for all t and the base factor of Equation (36) collapses. We use minibatch-OT couplings throughout; the total-variance derivation and the conditional-variance measurements are given in Section B.

5 Experiments

We evaluate along two axes. On **2D manifolds** (Moons, S-Curve, Swiss Roll, Circle) we isolate the effect of parametrization $\times$ window $\times$ coupling under a controlled harness (batch 512, minibatch-OT coupling unless noted, constant solver, uniform-t, 15k steps, 100 NFE at evaluation), reporting Wasserstein-$1/2/3$ (density) and Chamfer (support). On **images** (CelebA 64^2 and CIFAR-10, 35M-parameter U-Net) we report FID. The questions are: does unrolling help, does parametrization matter, and does minibatch OT deliver on the variance theory of Section 4.4?

5.1 Synthetic Manifolds: Field $\gg \hat{x}_0$ Under Unrolling

Table 1 compares the two parametrizations under the terminal (NELBO) objective with minibatch-OT couplings across four 2D manifolds. The verdict is unambiguous: **field prediction dominates** $\hat{x}_0$ **at every dataset and window size**, by 130%–690% in Wasserstein-1 and by up to orders of magnitude in Chamfer. The $\hat{x}_0$ objective is also far less

stable—at $W \geq 3$ it can diverge outright (e.g. Moons $\hat{x}_0$, $W=3$ reaches Wass1 3.8 and Chamfer $\sim 2\times 10^3$; Table 3)—whereas field stays well-behaved. This is the empirical face of the $1/t$ gradient pathology of Sections C.3 and 4.4: the $\hat{x}_0$ loss and transition Jacobian carry a $1/t$ factor that blows the gradient up near the data manifold, while the field Jacobian $\mathbf{I} + dt\,\mathbf{J}_{\hat{v}}$ is a stable near-identity.

Table 1: Synthetic 2D manifolds (15k steps, 100 NFE, minibatch-OT couplings): final Wasserstein-1 and Chamfer (↓) for the field vs. $\hat{x}_0$ parametrization under the terminal (NELBO) objective. Field is uniformly better across all datasets and windows. Full ablations (Wass-2/3, $W \leq 5$, weighted variants) in Table 3.

Dataset	Param. (W)	Wass1 (↓)	Chamfer (↓)
Circle	field ($W=1$)	$\mathbf{0.0080 \pm 0.0008}$	$\mathbf{4.4\text{e-}5 \pm 2.3\text{e-}6}$
	field ($W=2$)	0.0076 ± 0.0005	$4.2\text{e-}5 \pm 1.1\text{e-}6$
	$\hat{x}_0$ ($W=1$)	0.0384 ± 0.0046	$3.0\text{e-}4 \pm 1.0\text{e-}4$
	$\hat{x}_0$ ($W=2$)	0.0332 ± 0.0102	$2.0\text{e-}4 \pm 1.0\text{e-}4$
Moons	field ($W=1$)	0.0182 ± 0.0133	0.0127 ± 0.0270
	field ($W=2$)	$\mathbf{0.0150 \pm 0.0038}$	$\mathbf{0.0006 \pm 0.0003}$
	$\hat{x}_0$ ($W=1$)	0.1433 ± 0.2325	1.5249 ± 3.4081
	$\hat{x}_0$ ($W=2$)	0.0350 ± 0.0124	0.0003 ± 0.0001
S-Curve	field ($W=1$)	0.3101 ± 0.0413	0.0170 ± 0.0054
	field ($W=2$)	$\mathbf{0.2776 \pm 0.0698}$	$\mathbf{0.0142 \pm 0.0031}$
	$\hat{x}_0$ ($W=1$)	0.7349 ± 0.0253	0.1310 ± 0.0435
	$\hat{x}_0$ ($W=2$)	0.4765 ± 0.0609	0.0472 ± 0.0089
Swiss Roll	field ($W=1$)	0.2379 ± 0.0575	$\mathbf{0.0089 \pm 0.0014}$
	field ($W=2$)	$\mathbf{0.2038 \pm 0.0589}$	0.0087 ± 0.0026
	$\hat{x}_0$ ($W=1$)	0.5479 ± 0.0690	0.3367 ± 0.0771
	$\hat{x}_0$ ($W=2$)	0.3245 ± 0.0671	0.0758 ± 0.0283

Two further observations sharpen the story (full ablations in Section C). First, *single-step ($W=1$) field training is already competitive*: full-BPTT NELBO unrolling gives only small gains on some datasets (Moons, Swiss Roll) and *degrades* on others (S-Curve at $W=5$), consistent with the cross-terms contributing more variance than signal. Second, and most useful, the *anchored consistency objective*—the cross-term-free reduction of Theorem 4.1, i.e. single-step flow matching plus a self-consistency term—*with a geometric step schedule* is the best unrolled variant: on S-Curve and Swiss Roll it improves Wasserstein-1 and Chamfer over both single-step and full-BPTT NELBO by 10–25% (Table 5). The configuration our theory predicts should work—drop the cross-terms, anchor the teacher—is exactly the one that does.

Minibatch OT is the consistent stabilizer underlying these runs: it collapses the conditional-pairing variance $\mathrm{Var}(x_0 \mid x_t)$ that unrolling would otherwise amplify as $(\mathrm{Var})^L$ (Section 4.4), and it is the single most reliable intervention across datasets and windows. Figure 2 visualizes the field-vs-$\hat{x}_0$ gap on the recovered manifolds.

5.2 Images: CelebA-64^2

Table 2 reports FID on CelebA-64^2 (35M U-Net); the ordering mirrors the synthetic story. *Field* prediction is the front-runner (best FID $\mathbf{5.28}$, field+EMA), $\hat{x}_0$ variants trail (7.78 with OT, 10–15 otherwise), and full-BPTT unrolling does not beat well-tuned single-step. The diagnostic is again the $1/t$ signature: single-step train loss is nearly identical for field and $\hat{x}_0$ at every noise level, but $\hat{x}_0$ carries $\sim 3\times$ the mean gradient norm and a far larger spread of that norm *across* noise levels (its per-noise-level coefficient of variation is *not* higher; Section C.3), and unrolling to $W=2$ makes the norm take off. Min-SNR clamping of the $1/t$ factor and a smooth surrogate weight (Section E) cut gradient-clip counts but do not overturn the ordering. FID—not validation loss, which plateaus at ≈ 0.21 for all runs—must select checkpoints. CelebA-64^2 is not comparable to CelebA-HQ-256^2 baselines; treat 5.28 as an internal progress marker, not a SOTA claim.

6 Conclusion

We studied unrolled flow matching—training the sampler the way it is used, via BPTT through the discrete solver—and found that its value is not as a standalone objective but as a *derivation*. Expanding the k-step ELBO of the unrolled loss yields a single-step flow-matching term plus trajectory cross-terms; the cross-terms enable self-correction but

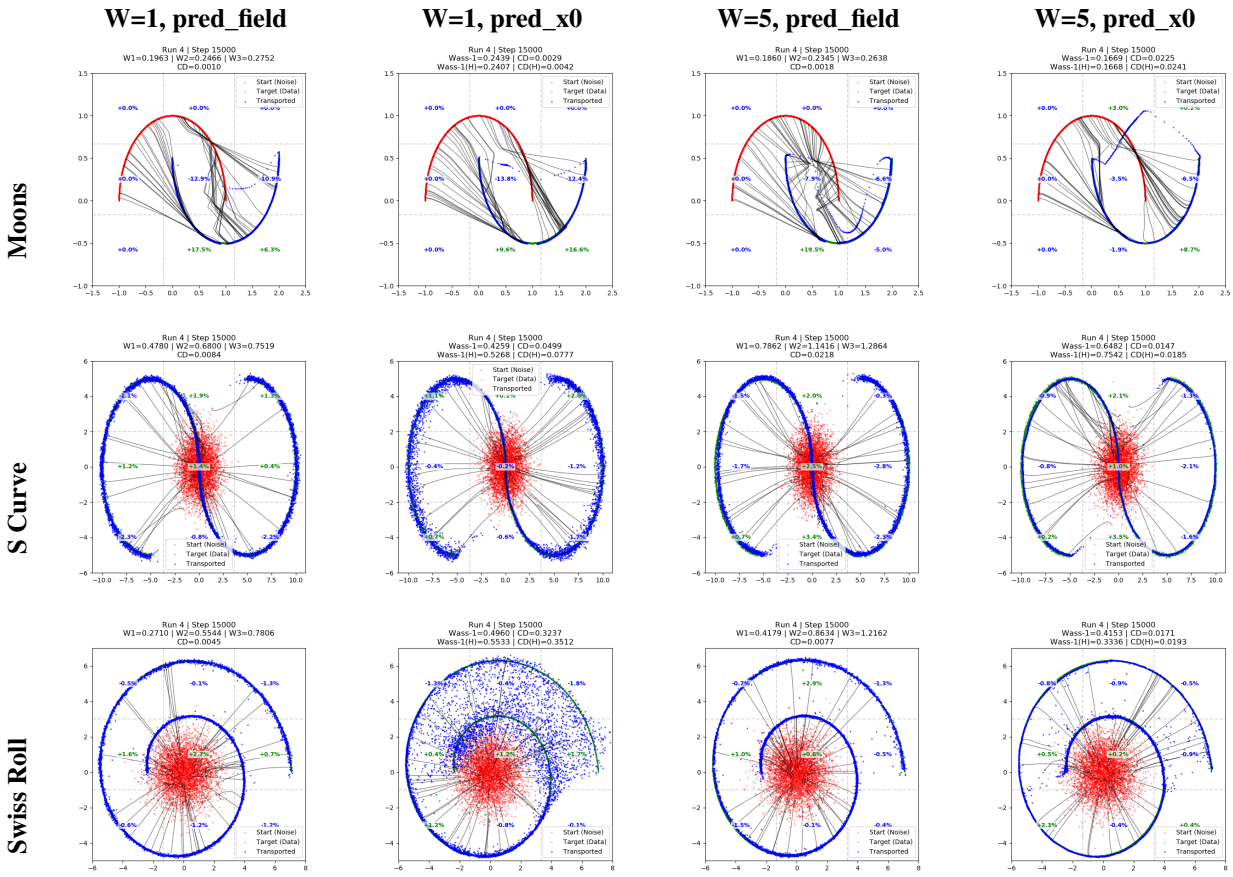


Figure 2: Visual comparison of final trained states (NELBO objective) across Moons, S Curve, and Swiss Roll datasets. The grid contrasts field prediction against $\hat{x}_0$ prediction for window sizes $W{=}1$ and $W{=}5$; field recovers the manifolds cleanly while $\hat{x}_0$ smears, especially at $W{=}5$.

Table 2: CelebA 64^2 (35M U-Net): best FID (↓) per run. The best result is *single-step* (field + EMA, $W{=}1$); unrolled BPTT does not beat it. CelebA 64^2 is not directly comparable to CelebA-HQ-256^2 baselines—treat 5.28 as an internal progress marker, not a SOTA claim.

Parametrization	**Stabilizer**	W	**Steps**	**Best FID**
Field	EMA	1	500k	**5.28**
$\hat{x}_0$	OT	2	500k	7.78
$\hat{x}_0$	OT + EMA	2	500k	8.63
Field	—	2	350k	9.14
$\hat{x}_0$	—	1	300k	9.99
$\hat{x}_0$	—	2	200k	14.60
$\hat{x}_0$ (no stop-grad)	—	5	200k	15.59

drive an exponential-in-L gradient variance (Theorem 4.2), and dropping them—detaching the teacher to anchor the objective—leaves exactly the single-step flow-matching loss plus a self-consistency loss, i.e. consistency and shortcut models (Theorem 4.1). Empirically this is borne out: on 2D manifolds and CelebA-64^2 the field parametrization dominates $\hat{x}_0$ (whose $1/t$ scaling destabilizes the gradient), full-BPTT NELBO unrolling does not beat single-step training, and the anchored consistency objective with a geometric schedule is the best unrolled variant, while minibatch OT is a consistent variance-reducing stabilizer throughout. The practical recipe is therefore: use the field parametrization, add OT, and if you unroll, detach the teacher and penalize consistency—i.e. prefer the consistency/shortcut form, now understood as the variance-anchored limit of unrolled flow matching.

A Transition-Jacobian Analysis (All Parametrizations)

BPTT stability is governed by the transition Jacobian $\mathbf{T}_t \triangleq \nabla_{x_t} \hat{x}_{t-dt}$; the unrolled gradient scales with the product $\prod_t \mathbf{T}_t$, and whether cross-step interaction survives determines whether the model can learn "what led to the error."

Velocity / field.

$$\mathbf{T}_t^{\text{field}} = \mathbf{I} + dt\, \mathbf{J}_{\hat{v}_\theta}. \tag{38}$$

Expanding $\prod(\mathbf{I} + dt\, \mathbf{J}_{\hat{v}_\theta})$, the linear sum dominates because cross-terms $dt^2 \mathbf{J}_m \mathbf{J}_n$ vanish ($dt \ll 1$). BPTT thus behaves almost like a per-step `stop_grad`: it cannot learn multi-step correction, but it is numerically stable—which, empirically, is what matters.

Target / $\hat{x}_0$.

$$\mathbf{T}_t^{\hat{x}_0} = \left(1 - \tfrac{dt}{t}\right)\mathbf{I} + \tfrac{dt}{t}\, \mathbf{J}_{\hat{x}_0}. \tag{39}$$

This is adaptive: at high noise ($t \approx 1$), dt/t is small and $\mathbf{T}_t \approx \mathbf{I}$ (safe propagation); near the data manifold ($t \approx dt$), $\mathbf{T}_t \approx \mathbf{J}_{\hat{x}_0}$ and full cross-step gradient flows. The same dt/t factor, however, amplifies gradient variance exactly where it passes gradient (Section B).

A.1 Direct-step prediction (exploding RNN)

A third parametrization outputs the next state directly, $\hat{x}_{t-dt} = f_\theta(x_t, t)$, with single-step gradient $\nabla_\theta \mathcal{L}_{\text{direct}} = \frac{1}{dt^2} e_{t-dt}^\top \nabla_\theta f_\theta$ and transition Jacobian

$$\mathbf{T}_t^{\text{direct}} = \mathbf{J}_{f_\theta}. \tag{40}$$

Lacking any residual identity, the unrolled gradient multiplies raw network Jacobians, guaranteeing vanishing/exploding gradients (especially at high noise). We do not use it.

Theory vs. practice. The Jacobian analysis identifies $\hat{x}_0$ as the parametrization BPTT *can* exploit, yet adaptivity cuts both ways: the steps where $\mathbf{T}_t \approx \mathbf{J}_{\hat{x}_0}$ passes gradient are exactly where multiplied Jacobians amplify variance (Section B). The field parametrization trades cross-step learning for a stable near-identity Jacobian, and empirically this wins (Sections C.3 and 5): the gradient-stability gap—not a difference in fit quality—is why we prefer field.

B Conditional Variance and Minibatch OT (Full Analysis)

Conditional pairing noise. Standard CFM samples x_1, x_0 independently, so the network models $\mathbb{E}[x_0 \mid x_t]$ with ambiguity $\mathrm{Var}(x_0 \mid x_t)$. As $t \to 1$ (near noise), knowing x_t says almost nothing about the randomly assigned x_0 and $\mathrm{Var}(x_0 \mid x_t)$ approaches the dataset variance; as $t \to 0$ (near data) it collapses to ≈ 0 (Figures 3 and 4).

From conditional variance to gradient variance. By the law of total variance, the single-step gradient variance is proportional to the pairing noise:

$$\mathrm{Var}(\nabla_\theta \mathcal{L}_{step}) = \mathbb{E}\big[\mathrm{Var}[\nabla_\theta \mathcal{L}_{step} \mid x_t]\big] + \mathrm{Var}\big[\mathbb{E}[\nabla_\theta \mathcal{L}_{step} \mid x_t]\big] \tag{41}$$

$$\approx \mathbb{E}_{x_t}\big[(\nabla_\theta \hat{x}_0)^\top \mathbf{Var}(x_0 \mid x_t)(\nabla_\theta \hat{x}_0)\big]. \tag{42}$$

Unrolling multiplies L target Jacobians, giving the multiplicative growth of Equation (36), $\mathrm{Var}(\prod_j \mathbf{T}^{\hat{x}_0}) \approx (\mathrm{Var}(x_0 \mid x_{t_0}))^L$.

Minibatch OT collapses the base factor. Re-coupling x_1, x_0 within the batch by minimizing Euclidean transport cost yields a non-crossing assignment in which x_1 determines x_0; this maximizes covariance and collapses $\mathrm{Var}_{\text{OT}}(x_0 \mid x_t) \approx 0$ for all t, so the $(\mathrm{Var})^L$ penalty of Equation (36) is neutralized and multi-step gradients become stable. This is the empirical linchpin of Section 5.

C Synthetic Ablations for Multi-Step Unrolling Implementation Choices

Performance is assessed by Wasserstein-$\{1, 2, 3\}$ distance, Chamfer distance, visual reconstruction of the target manifold, and straightness of the sampling paths.

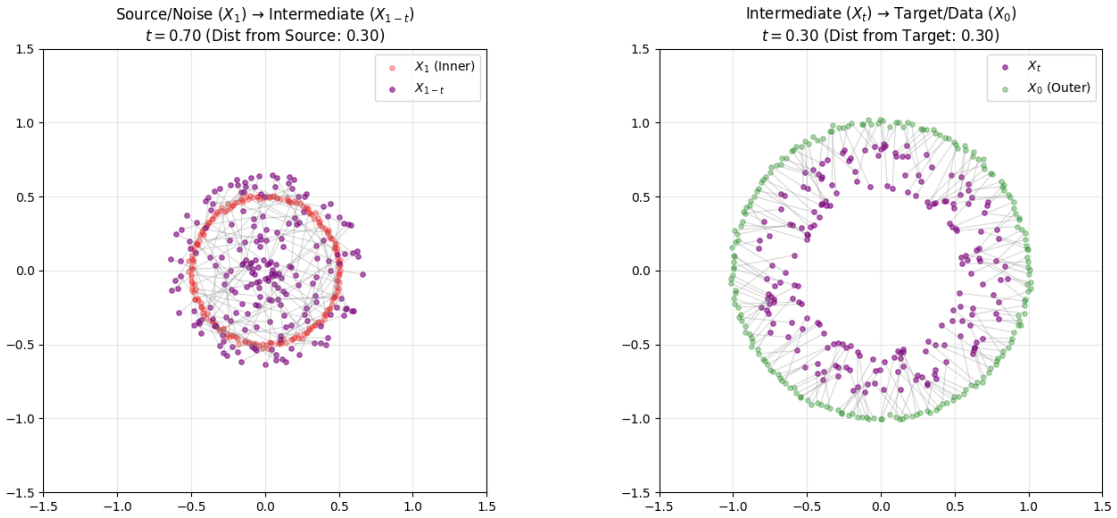
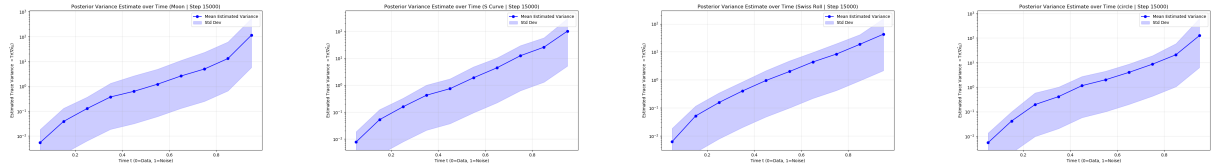


Figure 3: Conditional target variance. **Left:** at high noise ($t = 0.70$) the state-to-target pairings are highly stochastic, i.e. large $\mathrm{Var}(x_0 \mid x_t)$. **Right:** at low noise ($t = 0.30$) they are near-deterministic, $\mathrm{Var}(x_0 \mid x_t) \approx 0$.



(a) Moons (b) S Curve (c) Swiss Roll (d) Circle

Figure 4: Hutchinson-trace estimate of the conditional target variance $\mathrm{Tr}(\mathrm{Var}(x_0 \mid x_t))$ vs. t (log scale, step 15k). The strictly linear trend shows exponential growth as $t \to 1$.

Objective terminology. The tables below use the following objective names, which map to the analysis in Sections 3.3 and 4.2:

- **NELBO** (`Final_Error`) and **NELBO-field** (`Final_Error_field`): the terminal squared error $\|\sum_i \varepsilon_i\|^2$ over the window, trained with *full* BPTT. This keeps the trajectory cross-terms and is (up to weighting) the k-step negative ELBO of Section 3.3. **NELBO-weighted** adds the smooth x_0 weight of Section E.
- **Consistency** (`consistency_loss`): the cross-term-free reduction of Theorem 4.1—single-step flow matching plus a self-consistency term $\|\hat{v}^1 - \overline{\hat{v}^{2:W}}\|^2$ with the $0.8/0.2$ weighting—trained with gradient through both terms. **Consistency-detached** (`consistency_loss_detached`) freezes the later-step teacher, making the self-consistency term a distillation loss; this is the form closest to consistency and shortcut models.
- **Straighten / Straighten-detached**: an add-on straightening regularizer $\lambda\|\mathrm{op}_i - \mathrm{op}_{i-1}\|^2$ on successive network outputs (detached teacher in the latter).

C.1 Coarse Grained Ablations

Unless noted, the default harness is: batch size 512, minibatch-OT coupling, constant ODE solver, uniform-t sampling. The most robust BPTT setup we found is *field prediction, final-error loss, L=2–3 loops*.

Parametrization (field vs. $\hat{x}_0$) and number of loops (L=1–5). Table 3 reports final validation metrics for both parametrizations across window sizes, with percentages relative to the field, W=1 default. The field parametrization is essentially flat in W, whereas $\hat{x}_0$ (without OT) is far worse at every window—e.g., Circle $\hat{x}_0$ at W=1 is $+380\%$ over field.

Table 3: Final Validation Metrics (100 NFE, 15k Steps) Comparing Parametrizations (NELBO vs. NELBO-weighted) across different BPTT window sizes during training. Percentages show change relative to the default setting: NELBO | field | W=1 for each dataset.

Data Type	Setting	Wass1 (↓)	Wass2 (↓)	Wass3 (↓)	Chamfer (↓)	Velocity First (↓)
Circle	NELBO | field | W=1	0.0080 ± 0.0008	0.0094 ± 0.0009	0.0103 ± 0.0010	4.38e-5 ± 2.30e-6	0.0001 ± 0.0000
	NELBO | field | W=2	0.0076 ± 0.0005 (-5.0%)	0.0089 ± 0.0005 (-5.3%)	0.0098 ± 0.0005 (-4.9%)	4.17e-5 ± 1.11e-6 (-4.8%)	0.0001 ± 0.0000 (0.0%)
	NELBO | field | W=3	0.0081 ± 0.0013 (+1.2%)	0.0095 ± 0.0015 (+1.1%)	0.0104 ± 0.0016 (+1.0%)	4.36e-5 ± 2.30e-6 (-0.5%)	0.0001 ± 0.0000 (0.0%)
	NELBO | field | W=5	0.0079 ± 0.0005 (-1.2%)	0.0092 ± 0.0006 (-2.1%)	0.0101 ± 0.0006 (-1.9%)	4.29e-5 ± 1.39e-6 (-2.1%)	0.0001 ± 0.0000 (0.0%)
	NELBO | x0 | W=1	0.0384 ± 0.0046 (+380.0%)	0.0455 ± 0.0063 (+384.0%)	0.0505 ± 0.0070 (+390.3%)	0.0003 ± 0.0001 (+584.9%)	0.0058 ± 0.0053 (+5700.0%)
	NELBO | x0 | W=2	0.0332 ± 0.0102 (+315.0%)	0.0382 ± 0.0119 (+306.4%)	0.0418 ± 0.0128 (+305.8%)	0.0002 ± 0.0001 (+356.6%)	0.0012 ± 0.0005 (+1100.0%)
	NELBO | x0 | W=3	0.0227 ± 0.0069 (+183.8%)	0.0258 ± 0.0075 (+174.5%)	0.0282 ± 0.0081 (+173.8%)	0.0001 ± 0.0001 (+128.3%)	0.0007 ± 0.0002 (+600.0%)
	NELBO | x0 | W=5	0.0267 ± 0.0080 (+233.8%)	0.0312 ± 0.0096 (+231.9%)	0.0348 ± 0.0110 (+237.9%)	0.0003 ± 0.0002 (+584.9%)	0.0007 ± 0.0005 (+600.0%)
	NELBO-weighted | field | W=1	0.0091 ± 0.0004 (+13.8%)	0.0108 ± 0.0005 (+14.9%)	0.0119 ± 0.0004 (+15.5%)	5.56e-5 ± 1.66e-6 (+26.9%)	0.0001 ± 0.0000 (0.0%)
	NELBO-weighted | field | W=2	0.0091 ± 0.0011 (+13.8%)	0.0107 ± 0.0012 (+13.8%)	0.0118 ± 0.0012 (+14.6%)	5.68e-5 ± 1.59e-6 (+29.7%)	0.0001 ± 0.0000 (0.0%)
	NELBO-weighted | field | W=3	0.0092 ± 0.0010 (+15.0%)	0.0107 ± 0.0010 (+13.8%)	0.0118 ± 0.0011 (+14.6%)	5.64e-5 ± 3.39e-6 (+28.8%)	0.0001 ± 0.0000 (0.0%)
	NELBO-weighted | field | W=5	0.0082 ± 0.0004 (+2.5%)	0.0096 ± 0.0004 (+2.1%)	0.0106 ± 0.0004 (+2.9%)	5.03e-5 ± 2.24e-6 (+14.8%)	0.0001 ± 0.0000 (0.0%)
	NELBO-weighted | x0 | W=1	0.0134 ± 0.0026 (+67.5%)	0.0151 ± 0.0027 (+60.6%)	0.0164 ± 0.0028 (+59.2%)	6.28e-5 ± 1.04e-5 (+43.4%)	0.0004 ± 0.0002 (+300.0%)
	NELBO-weighted | x0 | W=2	0.0175 ± 0.0072 (+118.8%)	0.0196 ± 0.0074 (+108.5%)	0.0211 ± 0.0079 (+104.9%)	9.28e-5 ± 4.89e-5 (+111.9%)	0.0004 ± 0.0002 (+300.0%)
	NELBO-weighted | x0 | W=3	0.0161 ± 0.0071 (+101.2%)	0.0181 ± 0.0076 (+92.6%)	0.0197 ± 0.0081 (+91.3%)	7.20e-5 ± 3.24e-5 (+64.4%)	0.0002 ± 0.0001 (+100.0%)
	NELBO-weighted | x0 | W=5	0.0171 ± 0.0026 (+113.8%)	0.0190 ± 0.0025 (+102.1%)	0.0205 ± 0.0026 (+99.0%)	8.92e-5 ± 2.74e-5 (+103.7%)	0.0002 ± 0.0001 (+100.0%)
Moon	NELBO | field | W=1	0.0182 ± 0.0133	0.0717 ± 0.0994	0.1386 ± 0.2052	0.0127 ± 0.0270	6.78e-5 ± 6.84e-5
	NELBO | field | W=2	0.0150 ± 0.0038 (-17.6%)	0.0273 ± 0.0050 (-61.9%)	0.0436 ± 0.0063 (-68.5%)	0.0006 ± 0.0003 (-95.3%)	6.64e-5 ± 3.42e-5 (-2.1%)
	NELBO | field | W=3	0.0210 ± 0.0189 (+15.4%)	0.0592 ± 0.0659 (-17.4%)	0.1069 ± 0.1190 (-22.9%)	0.0064 ± 0.0118 (-49.6%)	5.92e-5 ± 3.91e-5 (-12.7%)
	NELBO | field | W=5	0.0110 ± 0.0033 (-39.6%)	0.0219 ± 0.0025 (-69.5%)	0.0395 ± 0.0033 (-71.5%)	0.0004 ± 0.0001 (-96.9%)	7.88e-5 ± 6.87e-5 (+16.2%)
	NELBO | x0 | W=1	0.1433 ± 0.2325 (+687.4%)	0.5910 ± 1.2175 (+724.3%)	1.0523 ± 2.2371 (+659.2%)	1.5249 ± 3.4081 (1.2e+04%)	0.0057 ± 0.0046 (+8307.1%)
	NELBO | x0 | W=2	0.0350 ± 0.0124 (+92.3%)	0.0448 ± 0.0156 (-37.5%)	0.0513 ± 0.0167 (-63.0%)	0.0003 ± 0.0001 (-97.6%)	0.0029 ± 0.0019 (+4177.3%)
	NELBO | x0 | W=3	3.8117 ± 8.3619 (2.1e+04%)	20.3334 ± 44.2812 (2.8e+04%)	36.6467 ± 79.2092 (2.6e+04%)	1981.84 ± 4429.24 (1.6e+07%)	0.0026 ± 0.0027 (+3734.8%)
	NELBO | x0 | W=5	0.0364 ± 0.0044 (+100.0%)	0.0476 ± 0.0086 (-33.6%)	0.0544 ± 0.0140 (-60.8%)	0.0005 ± 0.0007 (-96.1%)	0.0006 ± 0.0002 (+785.0%)
	NELBO-weighted | field | W=1	0.0348 ± 0.0078 (+91.2%)	0.0595 ± 0.0113 (-17.0%)	0.0780 ± 0.0118 (-43.7%)	0.0027 ± 0.0009 (-78.7%)	0.0006 ± 0.0003 (+785.0%)
	NELBO-weighted | field | W=2	0.0326 ± 0.0067 (+79.1%)	0.0579 ± 0.0088 (-19.2%)	0.0758 ± 0.0101 (-45.3%)	0.0025 ± 0.0007 (-80.3%)	0.0006 ± 0.0002 (+785.0%)
	NELBO-weighted | field | W=3	0.0342 ± 0.0115 (+87.9%)	0.0551 ± 0.0140 (-23.2%)	0.0697 ± 0.0150 (-49.7%)	0.0020 ± 0.0010 (-84.3%)	0.0004 ± 0.0003 (+490.0%)
	NELBO-weighted | field | W=5	0.0279 ± 0.0024 (+53.3%)	0.0522 ± 0.0086 (-27.2%)	0.0718 ± 0.0159 (-48.2%)	0.0020 ± 0.0009 (-84.3%)	0.0001 ± 0.0001 (+47.5%)
	NELBO-weighted | x0 | W=1	0.0578 ± 0.0370 (+217.6%)	0.4265 ± 0.4560 (+494.8%)	0.9857 ± 1.0871 (+611.2%)	0.3468 ± 0.5661 (+2630.7%)	0.0006 ± 0.0003 (+785.0%)
	NELBO-weighted | x0 | W=2	8.43e6 ± 1.88e7 (4.6e+10%)	6.80e7 ± 1.52e8 (9.5e+10%)	1.43e8 ± 3.19e8 (1.0e+11%)	2.31e16 ± 5.16e16 (1.8e+20%)	0.0008 ± 0.0009 (+1079.9%)
	NELBO-weighted | x0 | W=3	1.43e6 ± 3.19e6 (7.9e+09%)	1.04e7 ± 2.33e7 (1.5e+10%)	2.08e7 ± 4.65e7 (1.5e+10%)	5.44e14 ± 1.22e15 (4.3e+18%)	0.0003 ± 0.0004 (+342.5%)
	NELBO-weighted | x0 | W=5	1954.71 ± 4362.84 (1.1e+07%)	21327.91 ± 47608.38 (3.0e+07%)	49192.98 ± 109813.60 (3.5e+07%)	2.27e9 ± 5.07e9 (1.8e+13%)	0.0001 ± 0.0001 (+47.5%)
S Curve	NELBO | field | W=1	0.3101 ± 0.0413	0.5442 ± 0.0925	0.6348 ± 0.1048	0.0170 ± 0.0054	0.0759 ± 0.0164
	NELBO | field | W=2	0.2776 ± 0.0698 (-10.5%)	0.4681 ± 0.1339 (-14.0%)	0.5375 ± 0.1647 (-15.3%)	0.0142 ± 0.0031 (-16.5%)	0.1176 ± 0.0548 (+54.9%)
	NELBO | field | W=3	0.2798 ± 0.0280 (-9.8%)	0.4532 ± 0.0844 (-16.7%)	0.5350 ± 0.0888 (-15.7%)	0.0151 ± 0.0038 (-11.2%)	0.0685 ± 0.0204 (-9.7%)
	NELBO | field | W=5	0.3483 ± 0.0487 (+12.3%)	0.6165 ± 0.0608 (+13.3%)	0.7081 ± 0.0618 (+11.5%)	0.0183 ± 0.0043 (+7.6%)	0.0773 ± 0.0318 (+1.8%)
	NELBO | x0 | W=1	0.7349 ± 0.0253 (+137.0%)	1.2613 ± 0.0590 (+131.8%)	1.5152 ± 0.0553 (+138.7%)	0.1310 ± 0.0435 (+670.6%)	0.6160 ± 0.2812 (+711.6%)
	NELBO | x0 | W=2	0.4765 ± 0.0609 (+53.7%)	0.7755 ± 0.1013 (+42.5%)	0.9185 ± 0.1203 (+44.7%)	0.0472 ± 0.0089 (+177.6%)	0.3198 ± 0.0869 (+321.3%)
	NELBO | x0 | W=3	0.4320 ± 0.1185 (+39.3%)	0.7305 ± 0.2438 (+34.2%)	0.8945 ± 0.3439 (+40.9%)	0.0339 ± 0.0113 (+99.4%)	0.2263 ± 0.1236 (+198.2%)
	NELBO | x0 | W=5	0.3679 ± 0.0715 (+18.6%)	0.5954 ± 0.1195 (+9.4%)	0.6551 ± 0.1273 (+3.2%)	0.0118 ± 0.0039 (-30.6%)	0.1208 ± 0.0158 (+59.2%)
	NELBO-weighted | field | W=1	0.3954 ± 0.1069 (+27.5%)	0.6720 ± 0.1681 (+23.5%)	0.8699 ± 0.1904 (+37.0%)	0.1183 ± 0.0153 (+595.9%)	0.2002 ± 0.0890 (+163.8%)
	NELBO-weighted | field | W=2	0.3553 ± 0.0794 (+14.6%)	0.5993 ± 0.1491 (+10.1%)	0.7968 ± 0.1467 (+25.5%)	0.1096 ± 0.0140 (+544.7%)	0.1382 ± 0.0354 (+82.1%)
	NELBO-weighted | field | W=3	0.3635 ± 0.0911 (+17.2%)	0.6481 ± 0.1457 (+19.1%)	0.8297 ± 0.1458 (+30.7%)	0.0923 ± 0.0221 (+442.9%)	0.1335 ± 0.0285 (+75.9%)
	NELBO-weighted | field | W=5	0.2964 ± 0.0649 (-4.4%)	0.5341 ± 0.0844 (-1.9%)	0.6902 ± 0.0804 (+8.7%)	0.0760 ± 0.0104 (+347.1%)	0.1248 ± 0.0208 (+64.4%)
	NELBO-weighted | x0 | W=1	0.4602 ± 0.0147 (+48.4%)	0.7940 ± 0.1490 (+45.9%)	0.9749 ± 0.2424 (+53.6%)	0.0342 ± 0.0083 (+101.2%)	0.5215 ± 0.2715 (+587.1%)
	NELBO-weighted | x0 | W=2	0.3609 ± 0.0539 (+16.4%)	0.5985 ± 0.1457 (+10.0%)	0.6954 ± 0.1743 (+9.5%)	0.0275 ± 0.0119 (+61.8%)	0.3665 ± 0.1445 (+382.9%)
	NELBO-weighted | x0 | W=3	0.4734 ± 0.0965 (+52.7%)	0.7322 ± 0.1455 (+34.5%)	0.8329 ± 0.1804 (+31.2%)	0.0430 ± 0.0305 (+152.9%)	0.1766 ± 0.0463 (+132.7%)
	NELBO-weighted | x0 | W=5	0.3225 ± 0.0996 (+4.0%)	0.5254 ± 0.1740 (-3.5%)	0.5885 ± 0.2017 (-7.3%)	0.0172 ± 0.0129 (+1.2%)	0.1470 ± 0.0235 (+93.7%)
Swiss Roll	NELBO | field | W=1	0.2379 ± 0.0575	0.4265 ± 0.0841	0.6277 ± 0.1558	0.0089 ± 0.0014	0.0385 ± 0.0090
	NELBO | field | W=2	0.2038 ± 0.0589 (-14.3%)	0.3720 ± 0.1231 (-12.8%)	0.5460 ± 0.1928 (-13.0%)	0.0087 ± 0.0026 (-2.2%)	0.0557 ± 0.0196 (+44.7%)
	NELBO | field | W=3	0.1895 ± 0.0434 (-20.3%)	0.3704 ± 0.0637 (-13.2%)	0.5313 ± 0.1075 (-15.4%)	0.0116 ± 0.0020 (+30.3%)	0.0385 ± 0.0061 (0.0%)
	NELBO | field | W=5	0.2076 ± 0.0438 (-12.7%)	0.3507 ± 0.0942 (-17.8%)	0.4841 ± 0.1667 (-22.9%)	0.0134 ± 0.0026 (+50.6%)	0.0474 ± 0.0118 (+23.1%)
	NELBO | x0 | W=1	0.5479 ± 0.0690 (+130.3%)	0.7827 ± 0.0853 (+83.5%)	0.9443 ± 0.0961 (+50.4%)	0.3367 ± 0.0771 (+3683.1%)	0.3422 ± 0.1931 (+788.8%)
	NELBO | x0 | W=2	0.3245 ± 0.0671 (+36.4%)	0.6045 ± 0.1249 (+41.7%)	0.8415 ± 0.1744 (+34.1%)	0.0758 ± 0.0283 (+751.7%)	0.1540 ± 0.0237 (+300.0%)
	NELBO | x0 | W=3	0.3641 ± 0.0685 (+53.0%)	0.6874 ± 0.1054 (+61.2%)	0.9737 ± 0.1236 (+55.1%)	0.0573 ± 0.0291 (+543.8%)	0.1285 ± 0.0566 (+233.8%)
	NELBO | x0 | W=5	0.3009 ± 0.0704 (+26.5%)	0.6483 ± 0.1138 (+52.0%)	0.9718 ± 0.1617 (+54.8%)	0.0172 ± 0.0086 (+93.3%)	0.0776 ± 0.0217 (+101.6%)
	NELBO-weighted | field | W=1	0.2032 ± 0.0458 (-14.6%)	0.3513 ± 0.0700 (-17.6%)	0.4757 ± 0.0598 (-24.2%)	0.0626 ± 0.0085 (+603.4%)	0.0432 ± 0.0085 (+12.2%)
	NELBO-weighted | field | W=2	0.2227 ± 0.0648 (-6.4%)	0.3742 ± 0.0955 (-12.3%)	0.4979 ± 0.0941 (-20.7%)	0.0508 ± 0.0096 (+470.8%)	0.0623 ± 0.0151 (+61.8%)
	NELBO-weighted | field | W=3	0.2080 ± 0.0403 (-12.6%)	0.3579 ± 0.0469 (-16.1%)	0.4892 ± 0.0506 (-22.1%)	0.0552 ± 0.0092 (+520.2%)	0.0496 ± 0.0117 (+28.8%)
	NELBO-weighted | field | W=5	0.1936 ± 0.0454 (-18.6%)	0.3358 ± 0.0501 (-21.3%)	0.4586 ± 0.0386 (-26.9%)	0.0435 ± 0.0145 (+388.8%)	0.0597 ± 0.0118 (+55.1%)
	NELBO-weighted | x0 | W=1	0.2618 ± 0.0277 (+10.0%)	0.5736 ± 0.0958 (+34.5%)	0.8696 ± 0.2140 (+38.5%)	0.0071 ± 0.0023 (-20.2%)	0.2495 ± 0.0927 (+548.1%)
	NELBO-weighted | x0 | W=2	0.2840 ± 0.0557 (+19.4%)	0.5986 ± 0.0975 (+40.4%)	0.9060 ± 0.1374 (+44.3%)	0.0067 ± 0.0036 (-24.7%)	0.1277 ± 0.0268 (+231.7%)
	NELBO-weighted | x0 | W=3	0.2940 ± 0.0824 (+23.6%)	0.5915 ± 0.1264 (+38.7%)	0.8720 ± 0.1887 (+38.9%)	0.0087 ± 0.0031 (-2.2%)	0.0930 ± 0.0200 (+141.6%)
	NELBO-weighted | x0 | W=5	0.3021 ± 0.0522 (+27.0%)	0.5846 ± 0.0917 (+37.1%)	0.8623 ± 0.2192 (+37.4%)	0.0107 ± 0.0038 (+20.2%)	0.1134 ± 0.0855 (+194.5%)

C.2 Fine Grained Ablations

C.2.1 Re-weighted Loss:

C.2.2 Batch size (512 - 8192)

Table 4: Final Validation Metrics (100 NFE, 15k Steps) Comparing Parametrizations (NELBO vs. NELBO-weighted).

Data Type	Setting	Wass1	Wass2	Wass3	Chamfer	Velocity First
Circle	NELBO | field | W=1	0.0080 ± 0.0008	0.0094 ± 0.0009	0.0103 ± 0.0010	4.38e-5 ± 2.30e-6	0.0001 ± 0.0000
	NELBO | field | W=2	0.0076 ± 0.0005	0.0089 ± 0.0005	0.0098 ± 0.0005	4.17e-5 ± 1.11e-6	0.0001 ± 0.0000
	NELBO | field | W=3	0.0081 ± 0.0013	0.0095 ± 0.0015	0.0104 ± 0.0016	4.36e-5 ± 2.30e-6	0.0001 ± 0.0000
	NELBO | field | W=5	0.0079 ± 0.0005	0.0092 ± 0.0006	0.0101 ± 0.0006	4.29e-5 ± 1.39e-6	0.0001 ± 0.0000
	NELBO | x0 | W=1	0.0384 ± 0.0046	0.0455 ± 0.0063	0.0505 ± 0.0070	0.0003 ± 0.0001	0.0058 ± 0.0053
	NELBO | x0 | W=2	0.0332 ± 0.0102	0.0382 ± 0.0119	0.0418 ± 0.0128	0.0002 ± 0.0001	0.0012 ± 0.0005
	NELBO | x0 | W=3	0.0227 ± 0.0069	0.0258 ± 0.0075	0.0282 ± 0.0081	0.0001 ± 0.0001	0.0007 ± 0.0002
	NELBO | x0 | W=5	0.0267 ± 0.0080	0.0312 ± 0.0096	0.0348 ± 0.0110	0.0003 ± 0.0002	0.0007 ± 0.0005
	NELBO-weighted | field | W=1	0.0091 ± 0.0004	0.0108 ± 0.0005	0.0119 ± 0.0004	5.56e-5 ± 1.66e-6	0.0001 ± 0.0000
	NELBO-weighted | field | W=2	0.0091 ± 0.0011	0.0107 ± 0.0012	0.0118 ± 0.0012	5.68e-5 ± 1.59e-6	0.0001 ± 0.0000
	NELBO-weighted | field | W=3	0.0092 ± 0.0010	0.0107 ± 0.0010	0.0118 ± 0.0011	5.64e-5 ± 3.39e-6	0.0001 ± 0.0000
	NELBO-weighted | field | W=5	0.0082 ± 0.0004	0.0096 ± 0.0004	0.0106 ± 0.0004	5.03e-5 ± 2.24e-6	0.0001 ± 0.0000
	NELBO-weighted | x0 | W=1	0.0134 ± 0.0026	0.0151 ± 0.0027	0.0164 ± 0.0028	6.28e-5 ± 1.04e-5	0.0004 ± 0.0002
	NELBO-weighted | x0 | W=2	0.0175 ± 0.0072	0.0196 ± 0.0074	0.0211 ± 0.0079	9.28e-5 ± 4.89e-5	0.0004 ± 0.0002
	NELBO-weighted | x0 | W=3	0.0161 ± 0.0071	0.0181 ± 0.0076	0.0197 ± 0.0081	7.20e-5 ± 3.24e-5	0.0002 ± 0.0001
	NELBO-weighted | x0 | W=5	0.0171 ± 0.0026	0.0190 ± 0.0025	0.0205 ± 0.0026	8.92e-5 ± 2.74e-5	0.0002 ± 0.0001
Moon	NELBO | field | W=1	0.0182 ± 0.0133	0.0717 ± 0.0994	0.1386 ± 0.2052	0.0127 ± 0.0270	6.78e-5 ± 6.84e-5
	NELBO | field | W=2	0.0150 ± 0.0038	0.0273 ± 0.0050	0.0436 ± 0.0063	0.0006 ± 0.0003	6.64e-5 ± 3.42e-5
	NELBO | field | W=3	0.0210 ± 0.0189	0.0592 ± 0.0659	0.1069 ± 0.1190	0.0064 ± 0.0118	5.92e-5 ± 3.91e-5
	NELBO | field | W=5	0.0110 ± 0.0033	0.0219 ± 0.0025	0.0395 ± 0.0033	0.0004 ± 0.0001	7.88e-5 ± 6.87e-5
	NELBO | x0 | W=1	0.1433 ± 0.2325	0.5910 ± 1.2175	1.0523 ± 2.2371	1.5249 ± 3.4081	0.0057 ± 0.0046
	NELBO | x0 | W=2	0.0350 ± 0.0124	0.0448 ± 0.0156	0.0513 ± 0.0167	0.0003 ± 0.0001	0.0029 ± 0.0019
	NELBO | x0 | W=3	3.8117 ± 8.3619	20.3334 ± 44.2812	36.6467 ± 79.2092	1981.84 ± 4429.24	0.0026 ± 0.0027
	NELBO | x0 | W=5	0.0364 ± 0.0044	0.0476 ± 0.0086	0.0544 ± 0.0140	0.0005 ± 0.0007	0.0006 ± 0.0002
	NELBO-weighted | field | W=1	0.0348 ± 0.0078	0.0595 ± 0.0113	0.0780 ± 0.0118	0.0027 ± 0.0009	0.0006 ± 0.0003
	NELBO-weighted | field | W=2	0.0326 ± 0.0067	0.0579 ± 0.0088	0.0758 ± 0.0101	0.0025 ± 0.0007	0.0006 ± 0.0002
	NELBO-weighted | field | W=3	0.0342 ± 0.0115	0.0551 ± 0.0140	0.0697 ± 0.0150	0.0020 ± 0.0010	0.0004 ± 0.0003
	NELBO-weighted | field | W=5	0.0279 ± 0.0024	0.0522 ± 0.0086	0.0718 ± 0.0159	0.0020 ± 0.0009	0.0001 ± 0.0001
	NELBO-weighted | x0 | W=1	0.0578 ± 0.0370	0.4265 ± 0.4560	0.9857 ± 1.0871	0.3468 ± 0.5661	0.0006 ± 0.0003
	NELBO-weighted | x0 | W=2	8.43e6 ± 1.88e7	6.80e7 ± 1.52e8	1.43e8 ± 3.19e8	2.31e16 ± 5.16e16	0.0008 ± 0.0009
	NELBO-weighted | x0 | W=3	1.43e6 ± 3.19e6	1.04e7 ± 2.33e7	2.08e7 ± 4.65e7	5.44e14 ± 1.22e15	0.0003 ± 0.0004
	NELBO-weighted | x0 | W=5	1954.71 ± 4362.84	21327.91 ± 47608.38	49192.98 ± 109813.60	2.27e9 ± 5.07e9	0.0001 ± 0.0001
S Curve	NELBO | field | W=1	0.3101 ± 0.0413	0.5442 ± 0.0925	0.6348 ± 0.1048	0.0170 ± 0.0054	0.0759 ± 0.0164
	NELBO | field | W=2	0.2776 ± 0.0698	0.4681 ± 0.1339	0.5375 ± 0.1647	0.0142 ± 0.0031	0.1176 ± 0.0548
	NELBO | field | W=3	0.2798 ± 0.0280	0.4532 ± 0.0844	0.5350 ± 0.0888	0.0151 ± 0.0038	0.0685 ± 0.0204
	NELBO | field | W=5	0.3483 ± 0.0487	0.6165 ± 0.0608	0.7081 ± 0.0618	0.0183 ± 0.0043	0.0773 ± 0.0318
	NELBO | x0 | W=1	0.7349 ± 0.0253	1.2613 ± 0.0590	1.5152 ± 0.0553	0.1310 ± 0.0435	0.6160 ± 0.2812
	NELBO | x0 | W=2	0.4765 ± 0.0609	0.7755 ± 0.1013	0.9185 ± 0.1203	0.0472 ± 0.0089	0.3198 ± 0.0869
	NELBO | x0 | W=3	0.4320 ± 0.1185	0.7305 ± 0.2438	0.8945 ± 0.3439	0.0339 ± 0.0113	0.2263 ± 0.1236
	NELBO | x0 | W=5	0.3679 ± 0.0715	0.5954 ± 0.1195	0.6551 ± 0.1273	0.0118 ± 0.0039	0.1208 ± 0.0158
	NELBO-weighted | field | W=1	0.3954 ± 0.1069	0.6720 ± 0.1681	0.8699 ± 0.1904	0.1183 ± 0.0153	0.2002 ± 0.0890
	NELBO-weighted | field | W=2	0.3553 ± 0.0794	0.5993 ± 0.1491	0.7968 ± 0.1467	0.1096 ± 0.0140	0.1382 ± 0.0354
	NELBO-weighted | field | W=3	0.3635 ± 0.0911	0.6481 ± 0.1457	0.8297 ± 0.1458	0.0923 ± 0.0221	0.1335 ± 0.0285
	NELBO-weighted | field | W=5	0.2964 ± 0.0649	0.5341 ± 0.0844	0.6902 ± 0.0804	0.0760 ± 0.0104	0.1248 ± 0.0208
	NELBO-weighted | x0 | W=1	0.4602 ± 0.0147	0.7940 ± 0.1490	0.9749 ± 0.2424	0.0342 ± 0.0083	0.5215 ± 0.2715
	NELBO-weighted | x0 | W=2	0.3609 ± 0.0539	0.5985 ± 0.1457	0.6954 ± 0.1743	0.0275 ± 0.0119	0.3665 ± 0.1445
	NELBO-weighted | x0 | W=3	0.4734 ± 0.0965	0.7322 ± 0.1455	0.8329 ± 0.1804	0.0430 ± 0.0305	0.1766 ± 0.0463
	NELBO-weighted | x0 | W=5	0.3225 ± 0.0996	0.5254 ± 0.1740	0.5885 ± 0.2017	0.0172 ± 0.0129	0.1470 ± 0.0235
Swiss Roll	NELBO | field | W=1	0.2379 ± 0.0575	0.4265 ± 0.0841	0.6277 ± 0.1558	0.0089 ± 0.0014	0.0385 ± 0.0090
	NELBO | field | W=2	0.2038 ± 0.0589	0.3720 ± 0.1231	0.5460 ± 0.1928	0.0087 ± 0.0026	0.0557 ± 0.0196
	NELBO | field | W=3	0.1895 ± 0.0434	0.3704 ± 0.0637	0.5313 ± 0.1075	0.0116 ± 0.0020	0.0385 ± 0.0061
	NELBO | field | W=5	0.2076 ± 0.0438	0.3507 ± 0.0942	0.4841 ± 0.1667	0.0134 ± 0.0026	0.0474 ± 0.0118
	NELBO | x0 | W=1	0.5479 ± 0.0690	0.7827 ± 0.0853	0.9443 ± 0.0961	0.3367 ± 0.0771	0.3422 ± 0.1931
	NELBO | x0 | W=2	0.3245 ± 0.0671	0.6045 ± 0.1249	0.8415 ± 0.1744	0.0758 ± 0.0283	0.1540 ± 0.0237
	NELBO | x0 | W=3	0.3641 ± 0.0685	0.6874 ± 0.1054	0.9737 ± 0.1236	0.0573 ± 0.0291	0.1285 ± 0.0566
	NELBO | x0 | W=5	0.3009 ± 0.0704	0.6483 ± 0.1138	0.9718 ± 0.1617	0.0172 ± 0.0086	0.0776 ± 0.0217
	NELBO-weighted | field | W=1	0.2032 ± 0.0458	0.3513 ± 0.0700	0.4757 ± 0.0598	0.0626 ± 0.0085	0.0432 ± 0.0085
	NELBO-weighted | field | W=2	0.2227 ± 0.0648	0.3742 ± 0.0955	0.4979 ± 0.0941	0.0508 ± 0.0096	0.0623 ± 0.0151
	NELBO-weighted | field | W=3	0.2080 ± 0.0403	0.3579 ± 0.0469	0.4892 ± 0.0506	0.0552 ± 0.0092	0.0496 ± 0.0117
	NELBO-weighted | field | W=5	0.1936 ± 0.0454	0.3358 ± 0.0501	0.4586 ± 0.0386	0.0435 ± 0.0145	0.0597 ± 0.0118
	NELBO-weighted | x0 | W=1	0.2618 ± 0.0277	0.5736 ± 0.0958	0.8696 ± 0.2140	0.0071 ± 0.0023	0.2495 ± 0.0927
	NELBO-weighted | x0 | W=2	0.2840 ± 0.0557	0.5986 ± 0.0975	0.9060 ± 0.1374	0.0067 ± 0.0036	0.1277 ± 0.0268
	NELBO-weighted | x0 | W=3	0.2940 ± 0.0824	0.5915 ± 0.1264	0.8720 ± 0.1887	0.0087 ± 0.0031	0.0930 ± 0.0200
	NELBO-weighted | x0 | W=5	0.3021 ± 0.0522	0.5846 ± 0.0917	0.8623 ± 0.2192	0.0107 ± 0.0038	0.1134 ± 0.0855

C.2.3 Discretization Schedule: [5, 100], [10,100], [25,100], [50,100], None

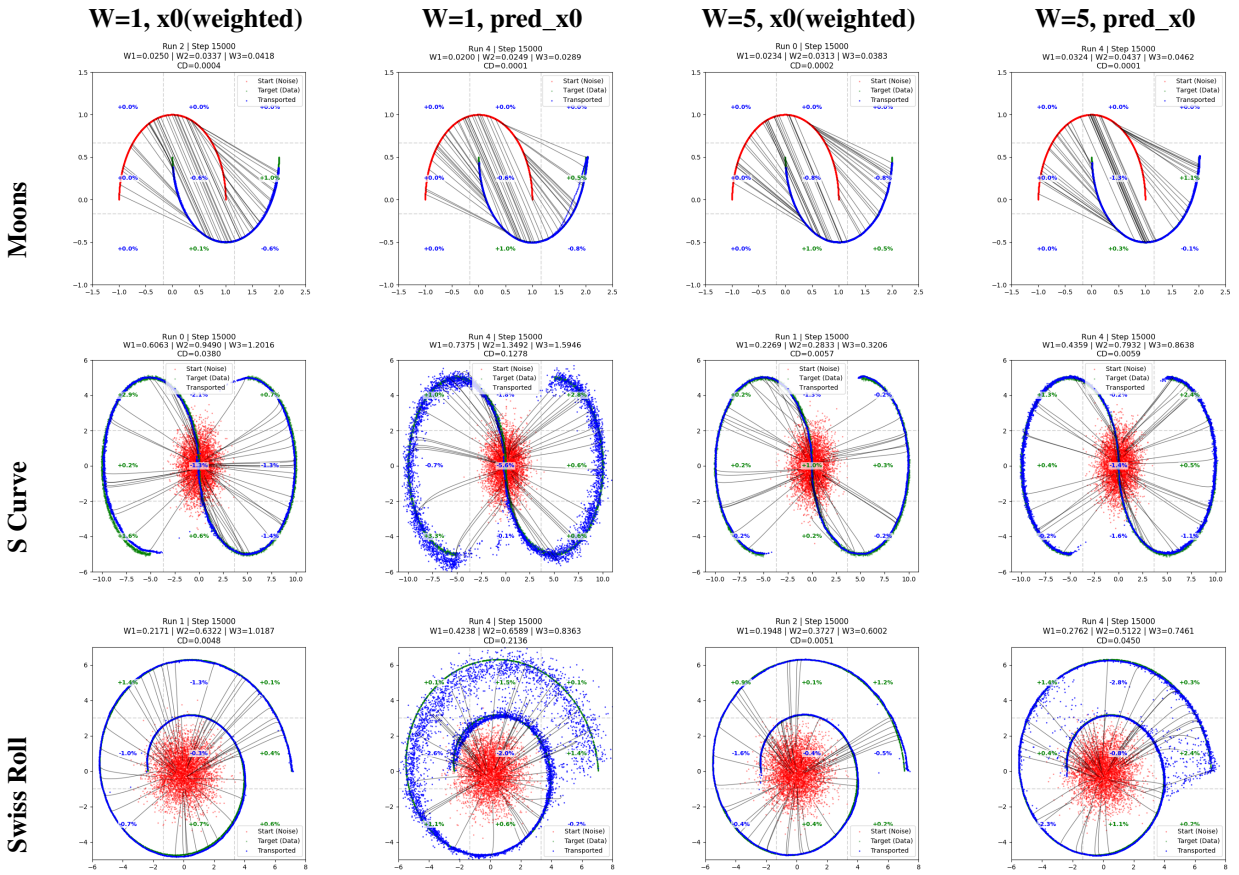


Figure 5: Visual comparison of final trained states (Final-Error loss vs. reweighted Final-Error loss) across Moons, S Curve, and Swiss Roll datasets. The grid assumes pred_x0 parameterization with BPTT Window Sizes W=1 and W=5.

C.3 Gradient Stability: Field vs. x_0 (CelebA)

We compare training stability between the field and x_0 parametrizations under the single- and double-step field-error loss. The single-step *train loss* is essentially identical across the two parametrizations at every noise level, yet their average gradient norm and gradient-norm spread (measured over 5,000-step intervals) differ sharply (Figures 6 and 7). This is attributable to the $1/t$ scaling in the x_0 loss, under which large but perceptually inconsequential single-pixel errors near $t \to 0$ produce very large gradients; the field parametrization has no such factor. Moving to $W{=}2$ with x_0 makes the gradient norm "take off," inflating gradient-clip counts. Two mitigations reduce the effect: clamping the t used in the field computation to a floor (a min-SNR-style clamp), and increasing batch size to average down the added variance in the unrolled setting.

An important nuance (Figures 8 and 9): the discrepancy is *not* a higher per-noise-level variance for x_0. On the contrary, the *normalized* standard deviation (coefficient of variation) of the gradient norm is slightly *lower* for x_0; the larger raw spread comes from the model experiencing very different mean gradient norms *across* noise levels, driven by the $1/t$ blow-up near the data manifold.

We also observe that the single-step error is largest in the "near pure noise" ($t \in [0.75, 1]$) and "near data manifold" ($t \in [0, 0.25]$) regimes for both parametrizations, and that the objective is hardest to improve near the data manifold (its log-loss curve stays flattest during training). This argues for placing *more* weight there, somewhat against a plain logit-normal schedule, which under-weights $t \approx 0.25$.

To remedy this gradient explosion, we apply a min-SNR reweighting that clamps the $1/t$ scaling in the loss to $t \geq \gamma$ (here $\gamma = 0.1$).

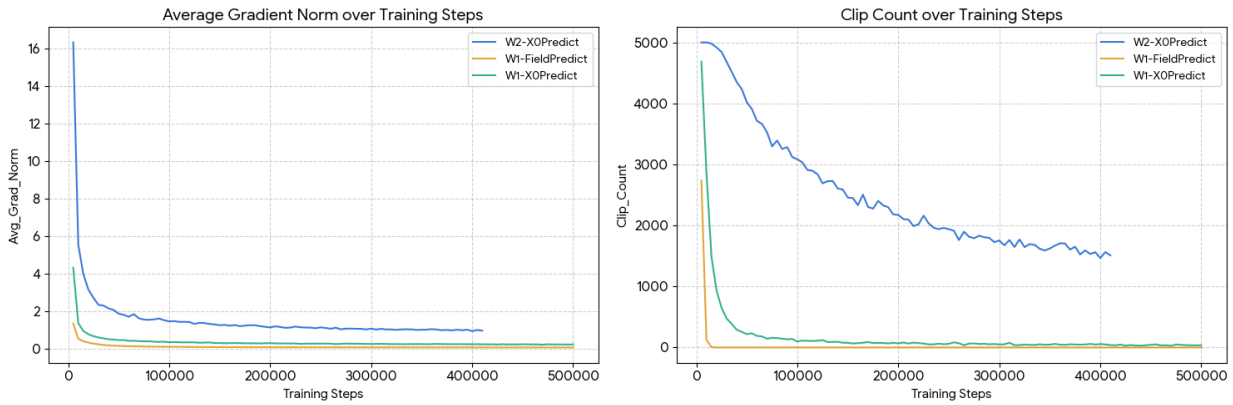


Figure 6: Comparison of the Average Gradient Norm (left) and Gradient Clip Count (right) over training steps for W2-X0Predict, W1-FieldPredict, and W1-X0Predict.

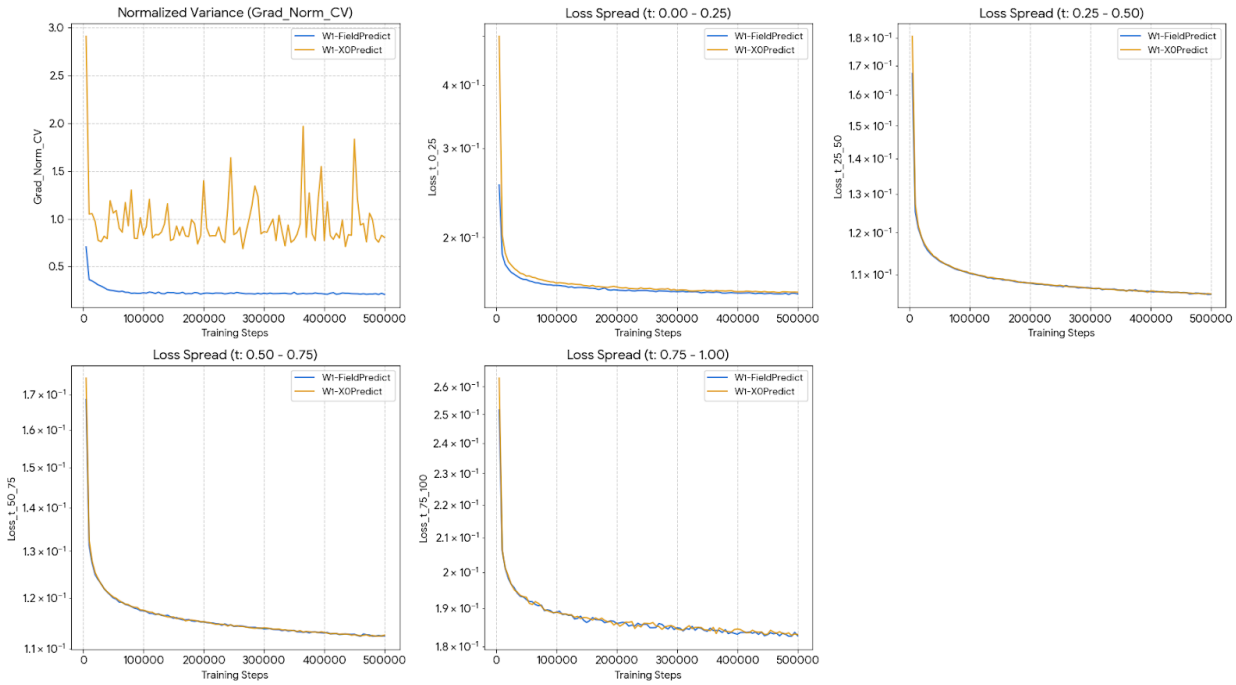


Figure 7: Comparison of the Normalized Variance (Grad_Norm_CV) and the Loss Spread across different noising factor bins (t) over training steps for W1-FieldPredict and W1-X0Predict. The loss spreads are displayed on a logarithmic scale.

Table 5: Final Validation Metrics (15k Steps) compiling NELBO-field, Consistency, and Consistency-detached. Percentages show change relative to the baseline: NELBO-field | field | Sched=None | W=1.

Data Type	Setting	NFE	Wass1 (↓)	Wass2 (↓)	Wass3 (↓)	Chamfer (↓)	Velocity First (↓)
S Curve	NELBO-field | field | Sched=None | W=1	1	0.4932 ± 0.0671 (0.0%)	0.8885 ± 0.0398 (0.0%)	1.1947 ± 0.0381 (0.0%)	0.5164 ± 0.0132 (0.0%)	0.0813 ± 0.0280 (0.0%)
		2	0.3851 ± 0.0666 (0.0%)	0.7464 ± 0.0601 (0.0%)	1.0413 ± 0.0434 (0.0%)	0.2964 ± 0.0430 (0.0%)	0.0813 ± 0.0280 (0.0%)
		5	0.3554 ± 0.0721 (0.0%)	0.6842 ± 0.0886 (0.0%)	0.9016 ± 0.0986 (0.0%)	0.1043 ± 0.0237 (0.0%)	0.0813 ± 0.0280 (0.0%)
	NELBO-field | field | Sched=None | W=2	1	0.5979 ± 0.0530 (+21.2%)	1.0189 ± 0.0699 (+14.7%)	1.3122 ± 0.0642 (+9.8%)	0.5851 ± 0.0243 (+13.3%)	0.0658 ± 0.0113 (-19.1%)
		2	0.4750 ± 0.0606 (+23.3%)	0.8269 ± 0.0774 (+10.8%)	1.0917 ± 0.0843 (+4.8%)	0.3213 ± 0.0872 (+8.4%)	0.0658 ± 0.0113 (-19.1%)
		5	0.4338 ± 0.0642 (+22.1%)	0.7306 ± 0.1085 (+6.8%)	0.9105 ± 0.1285 (+1.0%)	0.1072 ± 0.0332 (+2.8%)	0.0658 ± 0.0113 (-19.1%)
	NELBO-field | field | Sched=None | W=3	1	0.5830 ± 0.1051 (+18.2%)	1.0011 ± 0.1134 (+12.7%)	1.3152 ± 0.1057 (+10.1%)	0.6222 ± 0.0242 (+20.5%)	0.0673 ± 0.0214 (-17.2%)
		2	0.4225 ± 0.0858 (+9.7%)	0.7818 ± 0.1221 (+4.7%)	1.0864 ± 0.1018 (+4.3%)	0.3446 ± 0.0316 (+16.3%)	0.0673 ± 0.0214 (-17.2%)
		5	0.3674 ± 0.1024 (+3.4%)	0.6607 ± 0.1653 (-3.4%)	0.8625 ± 0.1624 (-4.3%)	0.1115 ± 0.0101 (+6.9%)	0.0673 ± 0.0214 (-17.2%)
	NELBO-field | field | Sched=geometric | W=1	1	0.4374 ± 0.0665 (-11.3%)	0.8252 ± 0.0972 (-7.1%)	1.1402 ± 0.0840 (-4.6%)	0.4545 ± 0.0330 (-12.0%)	0.2216 ± 0.0887 (+172.6%)
		2	0.3831 ± 0.0857 (-0.5%)	0.7189 ± 0.1495 (-3.7%)	0.9895 ± 0.1498 (-5.0%)	0.2472 ± 0.0398 (-16.6%)	0.2216 ± 0.0887 (+172.6%)
		5	0.3899 ± 0.1128 (+9.7%)	0.7326 ± 0.2257 (+7.1%)	0.9384 ± 0.2833 (+4.1%)	0.1044 ± 0.0321 (0.0%)	0.2216 ± 0.0887 (+172.6%)
	NELBO-field | field | Sched=geometric | W=2	1	0.6014 ± 0.0996 (+21.9%)	0.9971 ± 0.0929 (+12.2%)	1.3006 ± 0.0801 (+8.9%)	0.6952 ± 0.0878 (+34.6%)	0.4140 ± 0.0554 (+409.2%)
		2	0.4047 ± 0.0747 (+5.1%)	0.7593 ± 0.1106 (+1.7%)	1.0500 ± 0.1075 (+0.8%)	0.3243 ± 0.0345 (+9.4%)	0.4140 ± 0.0554 (+409.2%)
		5	0.3720 ± 0.0591 (+4.7%)	0.6969 ± 0.1521 (+1.9%)	0.9293 ± 0.1878 (+3.1%)	0.1418 ± 0.0192 (+36.0%)	0.4140 ± 0.0554 (+409.2%)
	NELBO-field | field | Sched=geometric | W=3	1	0.6653 ± 0.0721 (+34.9%)	1.0695 ± 0.0712 (+20.4%)	1.3726 ± 0.0660 (+14.9%)	0.8387 ± 0.0893 (+62.4%)	0.2560 ± 0.0800 (+214.9%)
		2	0.4255 ± 0.0788 (+10.5%)	0.7812 ± 0.0910 (+4.7%)	1.0780 ± 0.0672 (+3.5%)	0.4044 ± 0.0256 (+36.4%)	0.2560 ± 0.0800 (+214.9%)
		5	0.3730 ± 0.1090 (+5.0%)	0.6536 ± 0.1549 (-4.5%)	0.8962 ± 0.1480 (-0.6%)	0.1821 ± 0.0023 (+74.6%)	0.2560 ± 0.0800 (+214.9%)
	Consistency | field | Sched=geometric | W=2	1	0.4179 ± 0.0340 (-15.3%)	0.7380 ± 0.0247 (-16.9%)	1.0470 ± 0.0176 (-12.4%)	0.4120 ± 0.0268 (-20.2%)	0.5155 ± 0.1164 (+534.1%)
		2	0.3824 ± 0.0254 (-0.7%)	0.6930 ± 0.0541 (-7.2%)	0.9475 ± 0.0643 (-9.0%)	0.2244 ± 0.0248 (-24.3%)	0.5155 ± 0.1164 (+534.1%)
		5	0.4238 ± 0.0256 (+19.2%)	0.7758 ± 0.1026 (+13.4%)	0.9609 ± 0.1215 (+6.6%)	0.0994 ± 0.0194 (-4.7%)	0.5155 ± 0.1164 (+534.1%)
	Consistency | field | Sched=geometric | W=3	1	0.4648 ± 0.0786 (-5.8%)	0.8403 ± 0.0959 (-5.4%)	1.1491 ± 0.0815 (-3.8%)	0.4467 ± 0.0437 (-13.5%)	0.2528 ± 0.0992 (+210.9%)
		2	0.4416 ± 0.0851 (+14.7%)	0.8040 ± 0.1191 (+7.7%)	1.0818 ± 0.1262 (+3.9%)	0.2342 ± 0.0138 (-21.0%)	0.2528 ± 0.0992 (+210.9%)
		5	0.4762 ± 0.0979 (+34.0%)	0.8764 ± 0.1527 (+28.1%)	1.1303 ± 0.1999 (+25.4%)	0.0909 ± 0.0177 (-12.8%)	0.2528 ± 0.0992 (+210.9%)
	Consistency-detached | field | Sched=geometric | W=2	1	0.4973 ± 0.1007 (+0.8%)	0.8304 ± 0.0664 (-6.5%)	1.1201 ± 0.0470 (-6.2%)	0.4239 ± 0.0341 (-17.9%)	0.3298 ± 0.1064 (+305.7%)
		2	0.4653 ± 0.1203 (+20.8%)	0.7908 ± 0.0891 (+5.9%)	1.0655 ± 0.0570 (+2.3%)	0.3263 ± 0.0199 (+10.1%)	0.3298 ± 0.1064 (+305.7%)
		5	0.4824 ± 0.1475 (+35.7%)	0.7829 ± 0.1345 (+14.4%)	1.0352 ± 0.0975 (+14.8%)	0.2690 ± 0.0199 (+157.9%)	0.3298 ± 0.1064 (+305.7%)
	Consistency-detached | field | Sched=geometric | W=3	1	0.4343 ± 0.0904 (-11.9%)	0.8006 ± 0.0679 (-9.9%)	1.1037 ± 0.0495 (-7.6%)	0.4224 ± 0.0244 (-18.2%)	0.2023 ± 0.0888 (+148.8%)
		2	0.3825 ± 0.0804 (-0.7%)	0.7101 ± 0.0645 (-4.9%)	0.9867 ± 0.0573 (-5.2%)	0.2741 ± 0.0396 (-7.5%)	0.2023 ± 0.0888 (+148.8%)
		5	0.3791 ± 0.0758 (+6.7%)	0.6893 ± 0.0733 (+0.7%)	0.9332 ± 0.0806 (+3.5%)	0.1943 ± 0.0417 (+86.3%)	0.2023 ± 0.0888 (+148.8%)
Swiss Roll	NELBO-field | field | Sched=None | W=1	1	0.2909 ± 0.0281 (0.0%)	0.4962 ± 0.0303 (0.0%)	0.6676 ± 0.0306 (0.0%)	0.1858 ± 0.0173 (0.0%)	0.0645 ± 0.0208 (0.0%)
		2	0.2271 ± 0.0198 (0.0%)	0.4052 ± 0.0386 (0.0%)	0.5550 ± 0.0420 (0.0%)	0.0813 ± 0.0105 (0.0%)	0.0645 ± 0.0208 (0.0%)
		5	0.2184 ± 0.0230 (0.0%)	0.3983 ± 0.0618 (0.0%)	0.5562 ± 0.1088 (0.0%)	0.0316 ± 0.0058 (0.0%)	0.0645 ± 0.0208 (0.0%)
	NELBO-field | field | Sched=None | W=2	1	0.3122 ± 0.0249 (+7.3%)	0.5188 ± 0.0283 (+4.6%)	0.6890 ± 0.0277 (+3.2%)	0.1952 ± 0.0125 (+5.1%)	0.0464 ± 0.0194 (-28.1%)
		2	0.2378 ± 0.0486 (+4.7%)	0.4271 ± 0.0671 (+5.4%)	0.5827 ± 0.0622 (+5.0%)	0.0878 ± 0.0040 (+8.0%)	0.0464 ± 0.0194 (-28.1%)
		5	0.2246 ± 0.0592 (+2.8%)	0.4052 ± 0.1013 (+1.7%)	0.5612 ± 0.1227 (+1.5%)	0.0356 ± 0.0052 (+12.7%)	0.0464 ± 0.0194 (-28.1%)
	NELBO-field | field | Sched=None | W=3	1	0.3358 ± 0.0307 (+15.4%)	0.5370 ± 0.0386 (+8.2%)	0.6888 ± 0.0272 (+3.2%)	0.1967 ± 0.0123 (+5.9%)	0.0468 ± 0.0167 (-27.4%)
		2	0.2533 ± 0.0475 (+11.5%)	0.4114 ± 0.0604 (+1.5%)	0.5366 ± 0.0513 (-3.3%)	0.0880 ± 0.0140 (+8.2%)	0.0468 ± 0.0167 (-27.4%)
		5	0.2247 ± 0.0477 (+2.9%)	0.3512 ± 0.0626 (-11.8%)	0.4560 ± 0.0568 (-18.0%)	0.0352 ± 0.0102 (+11.4%)	0.0468 ± 0.0167 (-27.4%)
	NELBO-field | field | Sched=geometric | W=1	1	0.2735 ± 0.0274 (-6.0%)	0.4747 ± 0.0262 (-4.3%)	0.6514 ± 0.0237 (-2.4%)	0.1859 ± 0.0113 (0.0%)	0.0548 ± 0.0136 (-15.0%)
		2	0.2112 ± 0.0327 (-7.0%)	0.3960 ± 0.0496 (-2.3%)	0.5672 ± 0.0605 (+2.2%)	0.0835 ± 0.0120 (+2.7%)	0.0548 ± 0.0136 (-15.0%)
		5	0.2110 ± 0.0344 (-3.4%)	0.4292 ± 0.0607 (+7.8%)	0.6372 ± 0.0976 (+14.6%)	0.0346 ± 0.0094 (+9.5%)	0.0548 ± 0.0136 (-15.0%)
	NELBO-field | field | Sched=geometric | W=2	1	0.3471 ± 0.0153 (+19.3%)	0.5791 ± 0.0253 (+16.7%)	0.7648 ± 0.0297 (+14.6%)	0.2747 ± 0.0110 (+47.8%)	0.1340 ± 0.0242 (+107.8%)
		2	0.2469 ± 0.0202 (+8.7%)	0.4475 ± 0.0303 (+10.4%)	0.6172 ± 0.0306 (+11.2%)	0.1418 ± 0.0069 (+74.4%)	0.1340 ± 0.0242 (+107.8%)
		5	0.2193 ± 0.0327 (+0.4%)	0.3925 ± 0.0784 (-1.5%)	0.5529 ± 0.1167 (-0.6%)	0.0611 ± 0.0034 (+93.4%)	0.1340 ± 0.0242 (+107.8%)
	NELBO-field | field | Sched=geometric | W=3	1	0.3813 ± 0.0268 (+31.1%)	0.6092 ± 0.0318 (+22.8%)	0.7943 ± 0.0372 (+19.0%)	0.3050 ± 0.0178 (+64.2%)	0.1114 ± 0.0550 (+72.7%)
		2	0.2609 ± 0.0292 (+14.9%)	0.4803 ± 0.0399 (+18.5%)	0.6637 ± 0.0468 (+19.6%)	0.1549 ± 0.0127 (+90.5%)	0.1114 ± 0.0550 (+72.7%)
		5	0.2334 ± 0.0316 (+6.9%)	0.4352 ± 0.0502 (+9.3%)	0.6258 ± 0.0800 (+12.5%)	0.0711 ± 0.0054 (+125.0%)	0.1114 ± 0.0550 (+72.7%)
	Consistency | field | Sched=geometric | W=2	1	0.2532 ± 0.0244 (-13.0%)	0.4489 ± 0.0409 (-9.5%)	0.6204 ± 0.0522 (-7.1%)	0.1474 ± 0.0081 (-20.7%)	0.1421 ± 0.0411 (+120.3%)
		2	0.2041 ± 0.0357 (-10.1%)	0.3861 ± 0.0881 (-4.7%)	0.5407 ± 0.1330 (-2.6%)	0.0634 ± 0.0057 (-22.0%)	0.1421 ± 0.0411 (+120.3%)
		5	0.2065 ± 0.0395 (-5.4%)	0.3905 ± 0.1231 (-2.0%)	0.5341 ± 0.2031 (-4.0%)	0.0274 ± 0.0034 (-13.3%)	0.1421 ± 0.0411 (+120.3%)
	Consistency | field | Sched=geometric | W=3	1	0.2623 ± 0.0154 (-9.8%)	0.4765 ± 0.0320 (-4.0%)	0.6530 ± 0.0509 (-2.2%)	0.1691 ± 0.0107 (-9.0%)	0.1161 ± 0.0418 (+80.0%)
		2	0.2114 ± 0.0294 (-6.9%)	0.4373 ± 0.0846 (+7.9%)	0.6419 ± 0.1445 (+15.7%)	0.0645 ± 0.0023 (-20.7%)	0.1161 ± 0.0418 (+80.0%)
		5	0.2113 ± 0.0343 (-3.3%)	0.4718 ± 0.1063 (+18.5%)	0.7345 ± 0.1594 (+32.1%)	0.0237 ± 0.0021 (-25.0%)	0.1161 ± 0.0418 (+80.0%)
	Consistency-detached | field | Sched=geometric | W=2	1	0.2776 ± 0.0124 (-4.6%)	0.4826 ± 0.0125 (-2.7%)	0.6607 ± 0.0205 (-1.0%)	0.1762 ± 0.0153 (-5.2%)	0.1133 ± 0.0157 (+75.7%)
		2	0.2363 ± 0.0055 (+4.1%)	0.4393 ± 0.0200 (+8.4%)	0.6164 ± 0.0339 (+11.1%)	0.1229 ± 0.0118 (+51.2%)	0.1133 ± 0.0157 (+75.7%)
		5	0.2264 ± 0.0080 (+3.7%)	0.4003 ± 0.0279 (+0.5%)	0.5585 ± 0.0515 (+0.4%)	0.0777 ± 0.0070 (+145.9%)	0.1133 ± 0.0157 (+75.7%)
	Consistency-detached | field | Sched=geometric | W=3	1	0.2772 ± 0.0316 (-4.7%)	0.4784 ± 0.0222 (-3.6%)	0.6488 ± 0.0173 (-2.8%)	0.1735 ± 0.0133 (-6.6%)	0.0831 ± 0.0249 (+28.8%)
		2	0.2195 ± 0.0303 (-3.3%)	0.3998 ± 0.0189 (-1.3%)	0.5567 ± 0.0167 (+0.3%)	0.0910 ± 0.0060 (+11.9%)	0.0831 ± 0.0249 (+28.8%)
		5	0.2079 ± 0.0371 (-4.8%)	0.3739 ± 0.0322 (-6.1%)	0.5223 ± 0.0329 (-6.1%)	0.0427 ± 0.0027 (+35.1%)	0.0831 ± 0.0249 (+28.8%)

Table 6: Final Validation Metrics (15k Steps) comparing the NELBO objective and the anchored Consistency loss (the 2-NFE row matches the initial step, i.e. backward self-consistency).

Data Type	Setting	NFE	Wass1 (↓)	Wass2 (↓)	Wass3 (↓)	Chamfer (↓)	Velocity First (↓)
S Curve	NELBO | field | Sched=geometric | W=2	1	0.6054 ± 0.0372	1.0009 ± 0.0599	1.2992 ± 0.0597	0.6818 ± 0.0396	0.3552 ± 0.1514
		2	0.4125 ± 0.0430	0.7496 ± 0.0534	1.0301 ± 0.0588	0.3199 ± 0.0274	0.3552 ± 0.1514
		5	0.3871 ± 0.0431	0.6793 ± 0.0578	0.8900 ± 0.0613	0.1397 ± 0.0125	0.3552 ± 0.1514
	Consistency | field | Sched=geometric | W=2	1	0.4190 ± 0.0507	0.7561 ± 0.0377	1.0756 ± 0.0284	0.4378 ± 0.0258	0.2248 ± 0.0910
		2	0.3628 ± 0.0571	0.6925 ± 0.0464	1.0086 ± 0.0327	0.3445 ± 0.0207	0.2248 ± 0.0910
		5	0.3571 ± 0.0649	0.6525 ± 0.0569	0.9394 ± 0.0516	0.2649 ± 0.0222	0.2248 ± 0.0910

In our continuous-time Flow Matching framework utilizing an x_0-parameterization, the geometrically optimal loss weight for the unrolled prediction evaluates to $1/t^2$. However, this theoretically pure formulation presents severe optimization challenges at the trajectory boundaries. As $t \to 0$, the $1/t^2$ coefficient asymptotically explodes, causing microscopic high-frequency errors to generate infinite gradients. Conversely, as $t \to 1$, the raw mean squared error $||\hat{x}_0 - x_0||^2$ becomes astronomically large, as the network is forced to regress the clean data manifold from nearly pure Gaussian noise. These instabilities are significantly exacerbated by our use of a logit-normal time sampling distribution ($\sigma = 1.5$). While this distribution is highly advantageous for exposing the network to the critical boundary conditions of the ODE trajectory, the heavy-tailed sampling concentrates batch density exactly where the $1/t^2$ gradient variance is most destructive, leading to catastrophic batch hijacking in finite-batch training regimes.

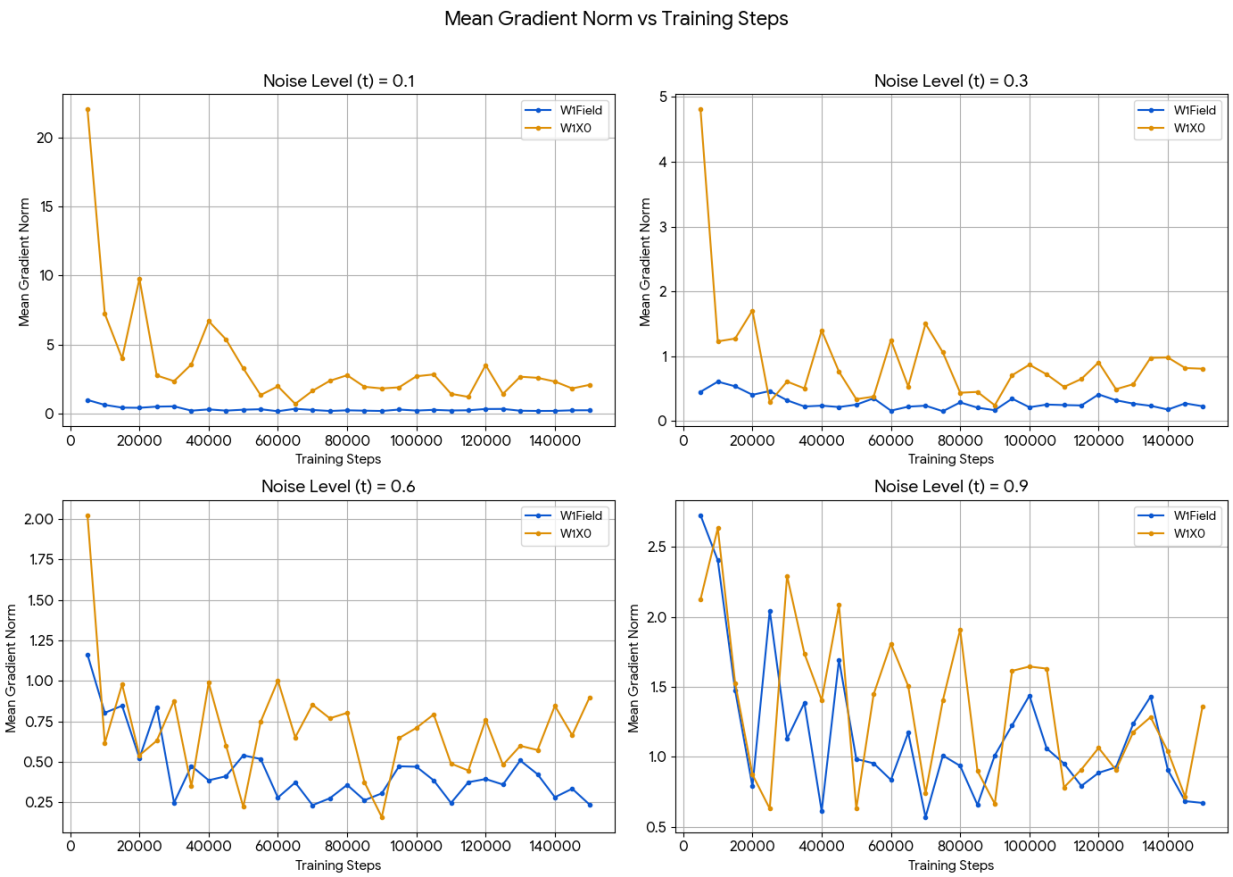


Figure 8: Mean gradient norm over training steps for W1Field and W1X0, evaluated at discrete noise levels $t \in \{0.1, 0.3, 0.6, 0.9\}$.

To reconcile the boundary-focused logit-normal sampling with gradient stability, we introduce a smooth, C^∞ continuous surrogate weighting scheme conceptually aligned with the Min-SNR strategy. Rather than applying the raw $1/t$ scaling or relying on piecewise hard-clipping, we use $w_t = \frac{(1-t_0)t_0}{dt(-0.5\cdot(t_0-0.5)^2+0.25)}$.

D Window Size Ablations

The optimal window size L represents a tradeoff between inference-horizon context and BPTT instability. As evidenced by our experiments, increasing the window size to $L = 10$ consistently degraded performance across all parameterizations unless strict variance mitigation was enforced. With OT couplings active, $L = 5$ emerged as the optimal threshold, providing sufficient temporal context to learn trajectory curvature correction without accumulating destructive numerical precision errors in the Transition Jacobians.

E Variance-Aware Optimal Transport Regularization

When Minibatch OT cannot be applied (e.g., massive datasets where global couplings are required), we must dynamically regularize the network in high-variance regimes.

We introduce a velocity regularization term $\lambda(x_t, t)\|v_\theta(x_t)\|^2$, where λ scales with the pairing uncertainty. Utilizing Tweedie's formula, the conditional variance is proportional to the trace of the Transition Jacobian:

$$\text{Var}(x_0|x_t) \approx \frac{t^2}{1-t}\nabla_{x_t}\hat{x}_0(x_t, t) \tag{43}$$

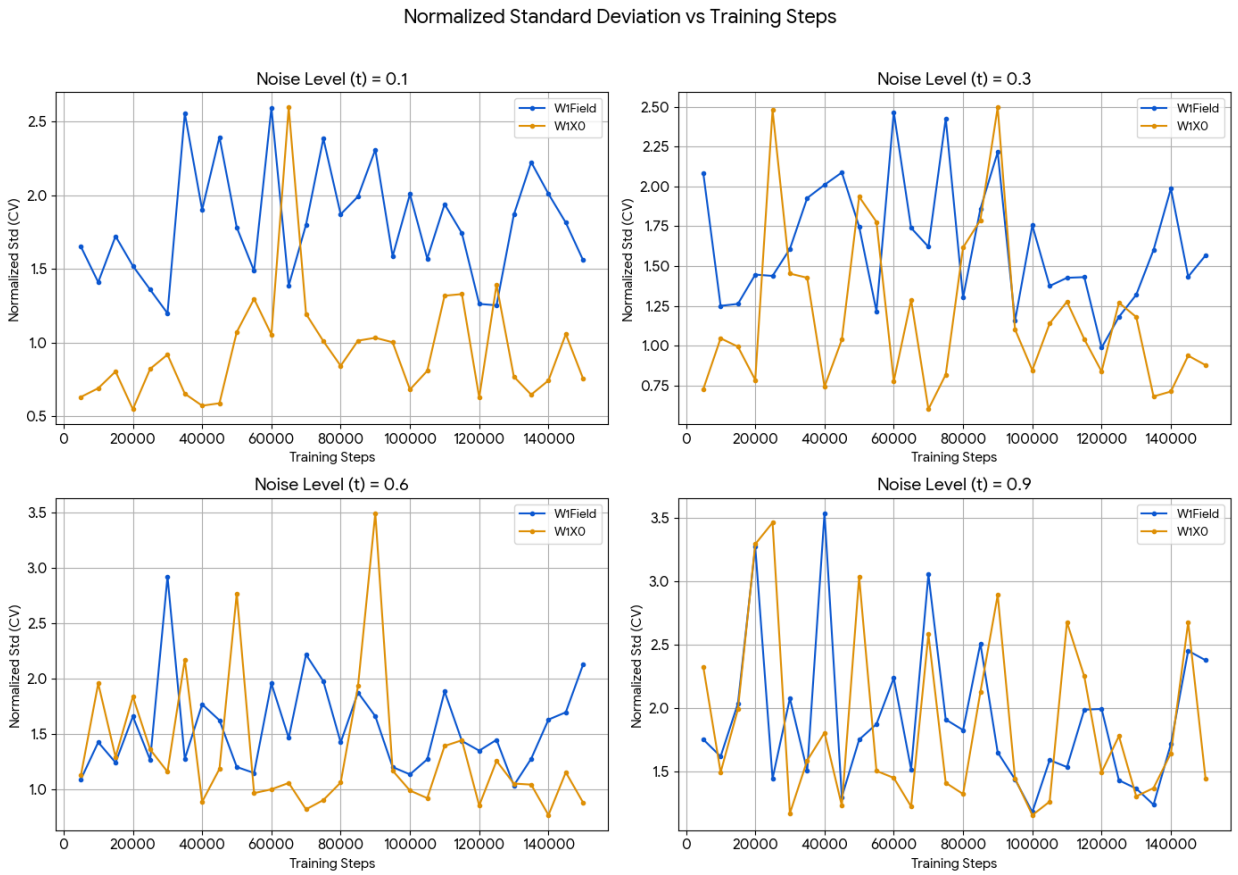


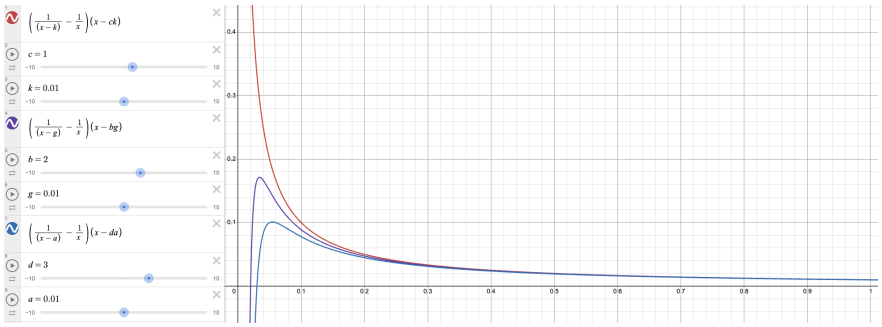
Figure 9: Normalized standard deviation (Coefficient of Variation) of the gradient norm over training steps for W1Field and W1X0 across different noise levels t.

Hutchinson Trace Estimator: To scale λ dynamically per-batch, we estimate this trace using a standard normal probe vector $z \sim \mathcal{N}(0, I)$:

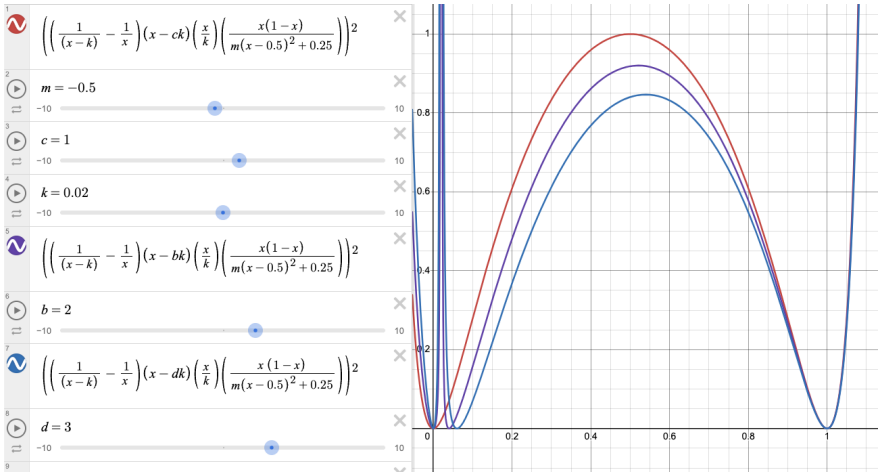
$$\mathrm{Tr}(\mathrm{Var}(x_0|x_t)) \approx \frac{t^2}{1-t}\mathbb{E}_z\left[z^\top \nabla_{x_t}\hat{x}_0(x_t, t)z\right] \tag{44}$$

While highly adaptive, this requires an expensive Vector-Jacobian Product (VJP) during every forward pass.

Heuristic Schedule: Empirical measurement of the trace estimator reveals that pairing variance grows strictly exponentially as $t \to 1$. Consequently, we can bypass the expensive VJP computation by employing a fixed exponential schedule $\lambda_{heur}(t) = \lambda_0 \exp(m \cdot t)$, successfully shielding the unrolled optimizer from catastrophic variance at a fraction of the computational cost.



(a) $w_t = 1$ (uniform weighting).



(b) $w_t = \frac{(1-t_0)t_0}{dt(-0.5(t_0-0.5)^2+0.25)}$ (smooth Min-SNR surrogate).

Figure 10: Expected loss curves over the time schedule under two loss weightings w_{t_k}: uniform (top) vs. the smooth C^∞ surrogate used to tame the $1/t$ blow-up of the $\hat{x}_0$ parametrization (bottom).

F Full Derivation of the Cumulative Error E_k

This appendix gives the step-by-step derivation of Equations (22) and (23), including the backward-anchored variant and the field form. The setup is that of Section 4.1: a monotonically decreasing schedule $t_1, \dots, t_k$ with $dt_i = t_{i+1} - t_i < 0$, Euler update $\hat{x}_{i+1} = \hat{x}_i - (\hat{x}_0^i - \hat{x}_i)\, dt_i/t_i$, ground-truth trajectory x_k^* (set $\hat{x}_0^i = x_0$), and $E_k = \hat{x}_k - x_k^*$.

F.1 Unrolling the Recurrence

Grouping $\hat{x}_i$ and using $1 + dt_i/t_i = t_{i+1}/t_i$,

$$\hat{x}_{t_{i+1}} = \hat{x}_{t_i}\Big(\tfrac{t_{i+1}}{t_i}\Big) - \hat{x}_0^i\Big(\tfrac{dt_i}{t_i}\Big), \qquad \Longrightarrow \qquad \frac{\hat{x}_{t_{i+1}}}{t_{i+1}} = \frac{\hat{x}_{t_i}}{t_i} - \hat{x}_0^i \frac{dt_i}{t_i t_{i+1}}. \tag{45}$$

Summing from $i = 1$ to $k-1$ and multiplying by t_k gives the exact unrolled state:

$$\hat{x}_{t_k} = \frac{t_k}{t_1}\hat{x}_{t_0} - \sum_{i=1}^{k-1} \hat{x}_0^i \frac{dt_i \cdot t_k}{t_i t_{i+1}}. \tag{46}$$

F.2 Error Formulation

Using $\frac{-dt_i}{t_i t_{i+1}} = \frac{1}{t_{i+1}} - \frac{1}{t_i}$, the ground-truth trajectory telescopes:

$$x_{t_k}^* = \frac{t_k}{t_1}x_{t_0} + x_0\, t_k \sum_{i=1}^{k-1}\Big(\tfrac{1}{t_{i+1}} - \tfrac{1}{t_i}\Big) = \frac{t_k}{t_1}x_{t_0} + x_0\Big(1 - \tfrac{t_k}{t_1}\Big), \tag{47}$$

and subtracting yields Equation (22): $E_k = -\sum_{i=1}^{k-1}(\hat{x}_0^i - x_0)\, dt_i t_k/(t_i t_{i+1})$.

F.3 Anchor–Consistency Decomposition (summation by parts)

Let $e_i = \hat{x}_0^i - x_0$ and $S_i = 1 - t_k/t_i$, so the coefficient equals $S_i - S_{i+1}$. Summation by parts gives

$$\begin{aligned} E_k &= -\sum_{i=1}^{k-1} e_i(S_i - S_{i+1}) = -\left[e_1 S_1 + \sum_{i=2}^{k-1}(e_i - e_{i-1})S_i - e_{k-1}S_k\right] \\ &= -\left[\underbrace{(\hat{x}_0^1 - x_0)\big(1 - \tfrac{t_k}{t_1}\big)}_{\text{anchor}} + \underbrace{\sum_{i=2}^{k-1}(\hat{x}_0^i - \hat{x}_0^{i-1})\big(1 - \tfrac{t_k}{t_i}\big)}_{\text{consistency}}\right], \end{aligned} \tag{48}$$

since $S_k = 0$; the coefficients satisfy $c_1 \geq \cdots \geq c_{k-1} \geq c_k = 0$.

F.4 Backward-Anchored Variant

Writing $e_i = e_{k-1} - \sum_{j=i}^{k-2}(e_{j+1} - e_j)$, substituting, and swapping summation order (Fubini) telescopes the inner sum to $\sum_{i=2}^{k-1}(e_i - e_{i-1})(S_1 - S_i)$, giving the end-anchored form

$$E_k = -\left[\underbrace{(\hat{x}_0^{k-1} - x_0)\big(1 - \tfrac{t_k}{t_1}\big)}_{\text{final anchor}} - \underbrace{\sum_{i=2}^{k-1}(\hat{x}_0^i - \hat{x}_0^{i-1})\big(\tfrac{t_k}{t_i} - \tfrac{t_k}{t_1}\big)}_{\text{retroactive consistency}}\right]. \tag{49}$$

F.5 Equivalent Field-Error Form

With $\hat{v}^i = (\hat{x}_0^i - \hat{x}_i)/t_i$ (so $\hat{v}^i - v^i = (\hat{x}_0^i - x_0)/t_i$),

$$E_k = -\sum_{i=1}^{k-1}(\hat{v}^i - v^i)\, \frac{dt_i \cdot t_k}{t_{i+1}}. \tag{50}$$

Applying the same substitution to the anchor–consistency decomposition and expanding the intermediate state via $\hat{x}_i = \frac{t_i}{t_{i-1}}\hat{x}_{i-1} - \frac{t_i - t_{i-1}}{t_{i-1}}\hat{x}_0^{i-1}$, the clean-data increments become velocity increments:

$$E_k = -\left[(\hat{v}^1 - v^1)\, t_1 c_1 + \sum_{i=2}^{k-1} (\hat{v}^i - \hat{v}^{i-1})\, t_i c_i\right], \qquad t_i c_i = t_i - t_k. \tag{51}$$

G Lower-Bound Hierarchy and Variance Reduction

The windowed objective can be motivated from a different angle than the ELBO of Section 3.3: as a tighter, lower-variance *estimator* of the true inference error. We make this precise below. The results are worth stating because they sharpen the paper's central tension—the windowed loss is provably a better proxy for what we care about, and yet, empirically (Section 5), full BPTT does not win. The resolution is that the bounds here assume benign (linear) error propagation and account only for time-sampling variance; they do *not* model the cross-term gradient-variance blow-up of Section 4.4, which is what dominates at image scale.

Definition G.1 (Terminal-loss estimators)**.** Let T be the number of inference steps and $\hat{x}_{t\to t+k}$ the state after unrolling k steps from a teacher-forced start at t. Define

$$\mathcal{J}^*(\theta) = \mathbb{E}_{x_0,x_1}\big[\,\|\hat{x}_{T\to 0} - x_0^*\|\,\big] \qquad \text{(true terminal loss)}, \tag{52}$$

$$\mathcal{J}_1(\theta) = T\cdot\mathbb{E}_{t,x}\big[\,\|\hat{x}_{t\to t+1} - x_{t+1}^*\|\,\big] \qquad \text{(single-step estimator)}, \tag{53}$$

$$\mathcal{J}_L(\theta) = \tfrac{T}{L}\cdot\mathbb{E}_{t,x}\big[\,\|\hat{x}_{t\to t+L} - x_{t+L}^*\|\,\big] \qquad (L\text{-window estimator}). \tag{54}$$

Assumption G.1 (Accumulated error)**.** Let $\epsilon = \mathbb{E}[\|\hat{x}_{t\to t+1} - x_{t+1}^*\|]$ be the local truncation error given a perfect input, and $e(k) = \mathbb{E}[\|\hat{x}_{t\to t+k} - x_{t+k}^*\|]$ the accumulated error after k steps. Distribution shift makes error accumulate at least linearly: $e(k) \geq k\,\epsilon$, with strict inequality once the trajectory drifts off the training manifold.

Theorem G.1 (Estimator hierarchy)**.** *Under Assumption G.1 with strict accumulation ($e(k) > k\epsilon$ for $k > 1$), for $1 < L < T$,*

$$\mathcal{J}_1(\theta) \;<\; \mathcal{J}_L(\theta) \;<\; \mathcal{J}^*(\theta), \tag{55}$$

and the gap $\mathcal{J}^* - \mathcal{J}_L$ shrinks as $L \to T$.

Proof. By definition $\mathcal{J}_1 = T\,e(1) = T\epsilon$. For the window estimator, $\mathcal{J}_L = \frac{T}{L}e(L) > \frac{T}{L}(L\epsilon) = T\epsilon = \mathcal{J}_1$. For the true loss, $\mathcal{J}^* = e(T)$; since drift compounds so that the average per-step error grows with horizon, $\frac{e(T)}{T} > \frac{e(L)}{L}$, hence $\mathcal{J}^* > \frac{T}{L}e(L) = \mathcal{J}_L$. As $L \to T$, $\mathcal{J}_L \to \mathcal{J}^*$, so the window estimator is a strictly tighter proxy than single-step FM, which underestimates the true error by a factor of T versus only T/L for the window. □

Lemma G.2 (Variance scaling)**.** *If the estimator variance is dominated by stochastic time sampling and the per-step errors are approximately uncorrelated within a window, then*

$$\mathrm{Var}(\mathcal{J}_L) \approx \tfrac{1}{L}\,\mathrm{Var}(\mathcal{J}_1), \tag{56}$$

and in the full-unrolling limit the time-sampling variance vanishes (the integral over $[0,T]$ is deterministic in t; only the data variance in (x_0, x_1) remains).

Proof. $\mathcal{J}_1$ evaluates a single sampled time and inherits its across-time (heteroscedastic) variance σ^2. $\mathcal{J}_L$ averages the error over L steps; treating the segment error as a mean of L terms, $\mathrm{Var}(\frac{1}{L}\sum_{i=1}^{L}\mathrm{error}_i) \propto \sigma^2/L$. Full unrolling sums all time steps deterministically, removing time-sampling stochasticity entirely. □

Remark G.1 (Why the tighter, lower-variance estimator can still lose). Theorem G.1 and Theorem G.2 treat the *loss* as a scalar estimator and assume linear error propagation. They do not model the *gradient* of the unrolled loss, whose variance scales as $(\mathrm{Var}(x_0 \mid x_t))^L$ once the transition Jacobians are multiplied (Section 4.4). At image scale this gradient-variance term, together with the $1/t$ blow-up of the $\hat{x}_0$ parametrization, overwhelms the estimator-level benefits—which is exactly why minibatch OT (variance removal) and the detached consistency/shortcut form (Theorem 4.1) are what actually help in practice, rather than full BPTT.

References

[1] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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

[7] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning (ICML)*, 2023.

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

[9] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

[10] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.