Inference-Time Self-Aligned Drift for Few-Step Flow Matching

Shigui Li[1,2*] **JIAN XU**[3] **Wei Chen**[1] **Junmei Yang**[1]
Fei Wang[2] **Daiguo Zhou**[2] **Delu Zeng**[1†]
[1]South China University of Technology [2]MiLM Plus, Xiaomi Inc. [3]AIP, RIKEN

Abstract

Flow Matching (FM) provides an efficient formulation for continuous-time generative modeling through deterministic transport. However, coarse ordinary differential equation (ODE) sampling can suppress variation in the generated distribution when generation is reduced to only a few steps. Stochastic differential equation (SDE) sampling offers a source of random fluctuations, but simply adding isotropic noise to a coarse update cannot generally restore the lost covariance. To improve few-step generation, we propose Self-Aligned Drift (SADrift), an inference-time framework that combines a corrective drift estimated from velocity history with local stochastic refinement. We estimate the drift from the difference between the current velocity and a moving average of past predictions, then integrate it together with Gaussian fluctuations through an exact local Ornstein–Uhlenbeck transition. The noise scale is calibrated from the model's response to preceding sampling perturbations. For linear FM, we quantify the covariance omitted by a population Euler step and relate it to velocity changes along the flow. By accounting for stochastic history and calibration bias, we derive conditions for reducing the leading local covariance error. The refinement requires no retraining, external guidance models, or additional neural function evaluations (NFE). On ImageNet-256 with pretrained SiT-XL/2, SADrift reduces Euler FID from 8.35 to 3.82 at 10 NFE and achieves FID 2.00 with Heun2 at 20 NFE. We also obtain improved VBench temporal consistency scores with Wan2.1-1.3B and TurboDiffusion at 4 NFE.

1 Introduction

Flow matching (FM) learns a velocity field that transports noise into data [1, 2], providing an efficient framework for continuous-time generative modeling. Together with diffusion models, it underlies modern image and video generators [3–5]. In linear FM, training uses straight paths between paired noise and data samples, while generation follows the learned field through an ordinary differential equation (ODE). Reducing the number of neural function evaluations (NFE) accelerates generation, but each step must then approximate the transport over a longer interval. This coarse approximation can suppress variation present in the continuous-time flow.

In linear FM, the population velocity correctly transports the prescribed marginals in continuous time. A finite Euler step, however, returns the conditional mean of the next state on the paired interpolation, omitting variation around that mean. For an input drawn from the FM marginal, the law of total covariance gives the exact covariance deficit. Independent source–target pairing provides an extreme example: a single population Euler step from noise to data maps every initial sample to the data mean. We refer to this finite-step averaging effect as *conditional-mean concentration*. Since the missing

*Work done during an internship at Xiaomi Inc., Shigui Li (sgl.shiguili@gmail.com).
†Corresponding author: Delu Zeng (dlzeng@scut.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Qualitative comparison at few-step settings (4 NFE). Both sequences are generated using the prompt: "*A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.*" Results are obtained using *TurboDiffusion* [19] on the T2V foundation model of *Wan2.1* [20].

covariance need not be isotropic, adding an independent isotropic perturbation alone cannot generally restore it.

Existing approaches improve few-step generation through changes to training or sampling. Distillation and consistency training learn predictors for short trajectories [6–8], while path design modifies the transport learned during training [9–11]. For pretrained models, high-order solvers improve integration accuracy by combining velocity or score predictions [12–14]. Stochastic refinement is provided by predictor–corrector methods [4] and restart sampling, which alternates noise injection with backward ODE integration [15]. Previous evaluations can also supply information for subsequent updates: adaptive momentum samplers reuse past predictions to stabilize or accelerate diffusion sampling [16, 17], while Free Hunch estimates denoiser covariance without additional denoiser evaluations [18]. We investigate how velocity history and the model's response to previous sampling perturbations can jointly guide a few-step FM update.

To improve few-step generation, we introduce *Self-Aligned Drift* (SADrift), an inference-time framework for pretrained FM models. We construct a corrective drift by comparing the current velocity with a moving average of past predictions. The resulting drift is integrated with Gaussian fluctuations through a local Ornstein–Uhlenbeck (OU) process centered on the base proposal. Its exact transition damps the local correction and requires no additional model evaluations. At the next step, we use the model's response to the preceding perturbation to calibrate the new noise scale. This feedback couples velocity memory, local damping, and random forcing in non-equilibrium memory dynamics, which we interpret through the generalized Langevin perspective [21, 22].

In summary, our contributions are threefold. First, we develop an inference-time correction that combines a history-based drift with Gaussian fluctuations through a local OU transition, without retraining or additional model evaluations. Second, we quantify the covariance deficit of a population Euler step in linear FM and relate it to material acceleration. Accounting for stochastic history and calibration bias, we derive conditions for reducing the leading local covariance error. Third, we demonstrate improvements with pretrained image and video models. On ImageNet-256, SADrift reduces Euler FID from 8.35 to 3.82 at 10 NFE (54.3%) and Heun2 FID from 2.20 to 2.00 at 20 NFE (9.1%) (Table 1). The gains persist with a SiT-adapted DPM-Solver++ at matched NFE. With Wan2.1-1.3B and TurboDiffusion, SADrift improves the VBench temporal flickering, subject consistency, and background consistency scores at 4 NFE (Figure 2).

2 Preliminaries

We begin by reviewing score-based diffusion and its connection to probability-flow ODEs. We then introduce linear flow matching and use Doob's h-transform to describe how conditioning on a target endpoint modifies the generative dynamics.

2.1 Score-based diffusion and stochastic exploration

Consider a generative diffusion $\mathbf{x}_t \in \mathbb{R}^D$ with smooth marginal density p_t, evolving from noise at $t = 0$ to data at $t = 1$:

$$\mathrm{d}\mathbf{x}_t = \mathbf{b}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t, \qquad \mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}). \tag{1}$$

Here $\mathbf{w}_t$ is standard Brownian motion, $g(t)$ is a scalar diffusion coefficient, and $\mathbf{s}$ is the score. The probability-flow velocity associated with this SDE is [4]

$$\mathbf{v}(\mathbf{x}, t) = \mathbf{b}(\mathbf{x}, t) - \frac{g(t)^2}{2}\mathbf{s}(\mathbf{x}, t). \tag{2}$$

The Fokker–Planck equation then becomes

$$\partial_t p_t + \nabla \cdot (p_t \mathbf{v}) = 0, \qquad \dot{\mathbf{x}}(t) = \mathbf{v}(\mathbf{x}(t), t). \tag{3}$$

The SDE and ODE thus share the same marginal densities. Conversely, given a probability-flow velocity $\mathbf{v}$, we can introduce Brownian motion while preserving these marginals by choosing $\mathbf{b} = \mathbf{v} + g^2\mathbf{s}/2$. The score correction compensates for the diffusion introduced by the noise.

2.2 Flow matching and conditional-mean concentration

Flow matching learns a probability-flow velocity directly [1]. Let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{x}_{\mathrm{data}} \sim p_{\mathrm{data}}$ be independent, with finite second moments. Here $\mathbf{I}$ is the identity matrix in $\mathbb{R}^D$. For linear interpolation, define

$$\mathbf{x}_t = (1 - t)\boldsymbol{\epsilon} + t\mathbf{x}_{\mathrm{data}}, \qquad \mathbf{u} = \mathbf{x}_{\mathrm{data}} - \boldsymbol{\epsilon}. \tag{4}$$

The population minimizer of the squared-error objective is the conditional mean velocity $\mathbf{v}$. We denote the conditional covariance around this mean by $\mathbf{C}$:

$$\mathbf{v}(\mathbf{x}, t) = \mathbb{E}[\mathbf{u} \mid \mathbf{x}_t = \mathbf{x}], \qquad \mathbf{C}(\mathbf{x}, t) = \mathrm{Cov}(\mathbf{u} \mid \mathbf{x}_t = \mathbf{x}). \tag{5}$$

Here $\mathbf{C}(\mathbf{x}, t) \in \mathbb{R}^{D\times D}$ is symmetric and positive semidefinite. The population velocity $\mathbf{v}$ satisfies Eq. (3). The paired paths in Eq. (4) are straight; the marginal ODE paths can curve as the conditional velocity changes. Gaussian conditioning also gives

$$\mathbf{s}(\mathbf{x}, t) = \frac{t\mathbf{v}(\mathbf{x}, t) - \mathbf{x}}{1 - t}, \qquad 0 < t < 1. \tag{6}$$

Under the usual regularity conditions, this score can be substituted into Eq. (2) to construct a diffusion with the same continuous-time marginals as linear FM. An Euler approximation introduces a further averaging effect: the conditional mean velocity advances the sample to the conditional mean of the next paired state. We quantify the resulting loss of variation in Section 3.1.

2.3 Conditional bridges and the Doob h-transform

The score–flow relation also describes how endpoint conditioning changes the probability current. Let $q_{1|t}(\mathbf{y} \mid \mathbf{x})$ be the transition density of the diffusion in Eq. (1) to a terminal point $\mathbf{y}$ with $p_1(\mathbf{y}) > 0$. For $h(\mathbf{x}, t) = q_{1|t}(\mathbf{y} \mid \mathbf{x}) > 0$, the Doob h-transform gives the conditioned drift [23]

$$\mathbf{b}^h(\mathbf{x}, t) = \mathbf{b}(\mathbf{x}, t) + g(t)^2\nabla_{\mathbf{x}} \log h(\mathbf{x}, t). \tag{7}$$

The diffusion coefficient is unchanged. On interior times where these quantities are smooth, the conditioned density is $p_t^h(\mathbf{x}) = p_t(\mathbf{x})h(\mathbf{x}, t)/p_1(\mathbf{y})$. Consequently, its score and probability-flow velocity satisfy

$$\mathbf{s}^h = \mathbf{s} + \nabla \log h, \qquad \mathbf{v}^h = \mathbf{v} + \frac{g(t)^2}{2}\nabla \log h. \tag{8}$$

Endpoint conditioning therefore changes both the stochastic drift and the probability-flow velocity through the posterior score. In linear FM, each endpoint pair specifies a conditional velocity, and averaging over compatible pairs gives the marginal field. Their spread around this mean is measured by the covariance in Eq. (5), which enters the finite-step analysis below.

3 Method

We first quantify the covariance deficit of a population Euler step in linear FM and relate it to material acceleration. We then use velocity history to construct a correction and integrate it with Gaussian fluctuations through a local OU process. Finally, we analyze the resulting sampler's leading local distributional error.

3.1 Conditional-mean concentration and its corrective current

For $h > 0$ with $t + h \leq 1$ and an input $\mathbf{x}_t \sim p_t$, the population Euler proposal satisfies

$$\mathbf{y} = \mathbf{x}_t + h\mathbf{v}(\mathbf{x}_t, t) = \mathbb{E}[\mathbf{x}_{t+h} \mid \mathbf{x}_t]. \tag{9}$$

This follows directly from the paired interpolation $\mathbf{x}_{t+h} = \mathbf{x}_t + h\mathbf{u}$. Taking the conditional mean preserves the overall mean but removes the conditional variation, as formalized below.

Proposition 3.1 (Conditional-mean concentration)**.** *The proposal in Eq.* (9) *has the same mean as* $\mathbf{x}_{t+h}$*. For every convex* φ *with finite expectations,* $\mathbb{E}[\varphi(\mathbf{y})] \leq \mathbb{E}[\varphi(\mathbf{x}_{t+h})]$*, and*

$$\mathrm{Cov}(\mathbf{x}_{t+h}) - \mathrm{Cov}(\mathbf{y}) = h^2\mathbb{E}[\mathbf{C}(\mathbf{x}_t, t)]. \tag{10}$$

The covariance deficit is positive semidefinite and vanishes exactly when $\mathbf{u}$ *is determined by* $\mathbf{x}_t$ *almost surely.*

At $t = 0$ and $h = 1$, independence in Eq. (4) gives $\mathbf{y} = \mathbb{E}[\mathbf{x}_{\mathrm{data}}]$: a single Euler step collapses all initial samples to the data mean. The proposition quantifies the same averaging effect over shorter intervals. Continuous integration of the population field reproduces the prescribed marginals; the covariance loss arises from replacing that transport with a finite Euler update.

To recover the omitted covariance, we relate it to how the population velocity changes along an ODE trajectory. Writing the velocity Jacobian as $\mathbf{J}_\mathbf{v} = (\partial_{x_j} v_k)_{k,j=1}^D$, we define the material acceleration by

$$\mathbf{a}(\mathbf{x}, t) = \partial_t \mathbf{v}(\mathbf{x}, t) + \mathbf{J}_\mathbf{v}(\mathbf{x}, t)\mathbf{v}(\mathbf{x}, t). \tag{11}$$

For a matrix field $\mathbf{B}$, we use the row-wise divergence $(\nabla \cdot \mathbf{B})_k = \sum_j \partial_{x_j} B_{kj}$. Since $\mathbf{u}$ is constant along each paired path, its velocity moment equation is

$$\partial_t(p_t\mathbf{v}) + \nabla \cdot \left(p_t\mathbf{v}\mathbf{v}^\top + p_t\mathbf{C}\right) = \mathbf{0}. \tag{12}$$

Combining this equation with continuity gives

$$\boxed{p_t\mathbf{a} = -\nabla \cdot (p_t\mathbf{C})}. \tag{13}$$

For a smooth observable φ, integration by parts expresses the effect of this acceleration in terms of the conditional covariance. With $\mathbf{B} : \mathbf{H} = \mathrm{tr}(\mathbf{B}^\top\mathbf{H})$ and vanishing boundary terms, we obtain

$$\mathbb{E}[\nabla\varphi(\mathbf{x}_t)^\top\mathbf{a}(\mathbf{x}_t, t)] = \mathbb{E}[\mathbf{C}(\mathbf{x}_t, t) : \nabla^2\varphi(\mathbf{x}_t)]. \tag{14}$$

Taylor expansion of the next paired state and the Euler proposal, under the corresponding smoothness and moment conditions, gives

$$\mathbb{E}[\varphi(\mathbf{x}_{t+h})] - \mathbb{E}[\varphi(\mathbf{y})] = \frac{h^2}{2}\mathbb{E}[\mathbf{C}(\mathbf{x}_t, t) : \nabla^2\varphi(\mathbf{x}_t)] + \mathcal{O}(h^3). \tag{15}$$

Comparing Eqs. (14) and (15) shows that a displacement $h^2\mathbf{a}/2$ matches the leading weak contribution omitted by Euler. This gives the corrected proposal

$$\mathbf{y}^\star = \mathbf{x}_t + h\mathbf{v}(\mathbf{x}_t, t) + \frac{h^2}{2}\mathbf{a}(\mathbf{x}_t, t). \tag{16}$$

In particular, the current identity gives

$$\mathbb{E}[\mathbf{a}(\mathbf{x}_t, t)] = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{x}_t, \mathbf{a}(\mathbf{x}_t, t)) = \mathbb{E}[\mathbf{C}(\mathbf{x}_t, t)]. \tag{17}$$

Through its correlation with the current state, the acceleration correction recovers the omitted covariance at leading order. We therefore estimate the acceleration from velocity history to construct the drift without explicit velocity derivatives. Appendix A gives the derivations of the moment identities.

3.2 Self-aligned memory and stochastic refinement

At inference time, we have access to the frozen field $\mathbf{v}_\theta$, including its conditioning inputs and any guidance rule. Let $\mathbf{a}_\theta$ denote its material acceleration. We approximate this acceleration using the difference between the current velocity and an exponential moving average of past predictions. For a grid $0 = t_0 < \cdots < t_N = 1$, write $h_i = t_{i+1} - t_i$ and $\mathbf{v}_i = \mathbf{v}_\theta(\mathbf{x}_i, t_i)$, where $\mathbf{x}_i$ is the numerical state. The history uses this starting velocity and is updated once per base-solver step. We compare it with the incoming moving average:

$$\mathbf{d}_i = \mathbf{v}_i - \mathbf{m}_i, \qquad \mathbf{m}_{i+1} = \beta \mathbf{m}_i + (1-\beta)\mathbf{v}_i, \qquad 0 \le \beta < 1. \tag{18}$$

With $\mathbf{m}_0 = \mathbf{v}_0$, unrolling the recursion expresses the residual as a weighted sum of velocity changes:

$$\mathbf{d}_i = \sum_{j=0}^{i-1} \beta^{i-1-j}(\mathbf{v}_{j+1} - \mathbf{v}_j), \qquad \mathbf{d}_0 = \mathbf{0}. \tag{19}$$

Because the residual accumulates velocity changes across several intervals, we divide by their effective time lag to obtain an acceleration estimate:

$$\ell_0 = 0, \qquad \ell_{i+1} = h_i + \beta \ell_i, \qquad \hat{\mathbf{a}}_i = \frac{\mathbf{d}_i}{\ell_i} \quad (i \ge 1). \tag{20}$$

We set $\hat{\mathbf{a}}_0 = \mathbf{0}$. The lag equals $\sum_{j<i} \beta^{i-1-j} h_j$ and is positive for $i \ge 1$. On a uniform grid, $\ell_i = h(1-\beta^i)/(1-\beta)$. Along a smooth ODE trajectory, Taylor expansion shows that this normalized difference approximates material acceleration. With stochastic refinement, past predictions also contain responses to the injected noise. We account for these responses in the history analysis of Section 3.3. Feature modulation within this history-drift branch is described in Appendix E.

To complement the directed history correction with isotropic noise, we use the closest isotropic approximation to the conditional velocity covariance as the scalar calibration target:

$$c(\mathbf{x}, t) = \frac{\operatorname{tr} \mathbf{C}(\mathbf{x}, t)}{D} = \arg\min_{s \ge 0} \|\mathbf{C}(\mathbf{x}, t) - s\mathbf{I}\|_F^2. \tag{21}$$

Here $\|\cdot\|_F$ is the Frobenius norm, and $c(\mathbf{x}, t)$ is the mean conditional velocity variance per coordinate. For the population field, Gaussian conditioning in Eq. (4) gives, for $0 < t < 1$,

$$t\mathbf{C}(\mathbf{x}, t) = \mathbf{I} + (1-t)\mathbf{J}_{\mathbf{v}}(\mathbf{x}, t). \tag{22}$$

Taking the trace expresses $c(\mathbf{x}, t)$ through the velocity divergence. We estimate this divergence from the model's response to the preceding Gaussian perturbation, reusing information already available during sampling. We write the perturbation as $\nu_{i-1}\boldsymbol{\xi}_{i-1}$, with coordinate standard deviation $\nu_{i-1} > 0$. For $1 \le i < N$, Gaussian integration by parts leads us to the estimator

$$\tilde{c}_i = \frac{1}{t_i}\left(1 + \frac{1-t_i}{D\nu_{i-1}} \mathbf{d}_i^\top \boldsymbol{\xi}_{i-1}\right), \qquad \hat{c}_i = \operatorname{clip}(\tilde{c}_i). \tag{23}$$

Here clip projects onto a fixed positive interval. The clipping rule and fixed initial scale $\hat{c}_0$ are specified in Appendix B. The inner product measures the velocity response in the direction of the known perturbation. The incoming memory is independent of this draw, so subtracting it leaves the expected probe value unchanged. Clipping provides a positive, bounded noise scale, with the resulting bias quantified in the analysis.

We combine the history correction and calibrated noise in a local OU process around the base solver's proposal. Let $\mathbf{y}_i$ denote this proposal, computed using the starting velocity $\mathbf{v}_i$; for Euler, $\mathbf{y}_i = \mathbf{x}_i + h_i \mathbf{v}_i$. Mean reversion controls the accumulated drift, while the exact transition determines both the directed displacement and its Gaussian fluctuations. The refinement displacement follows

$$\mathrm{d}\mathbf{z}_i(\tau) = (-\kappa \mathbf{z}_i(\tau) + \mathbf{r}_i)\,\mathrm{d}\tau + \sqrt{h_i}\sigma_i\,\mathrm{d}\mathbf{w}_i(\tau), \quad \mathbf{z}_i(0) = \mathbf{0}, \quad 0 \le \tau \le h_i. \tag{24}$$

The Brownian motion is fresh at each step. The factor $\sqrt{h_i}$ makes the random displacement of order h_i for bounded σ_i, so its covariance has the same h_i^2 order as the Euler deficit. Choose a fixed history weight $0 < \lambda < 1$ and damping rate $\kappa > 0$, and define

$$b_i = \frac{1 - \mathrm{e}^{-\kappa h_i}}{\kappa}, \qquad \alpha_i = \frac{\lambda b_i}{h_i}, \qquad \nu_i^2 = h_i^2(1-\alpha_i)\hat{c}_i. \tag{25}$$

Algorithm 1 SADrift with memory and noise-response calibration

Require: Frozen field $\mathbf{v}_\theta$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, grid $\{t_i\}_{i=0}^N$
Require: $0 \le \beta < 1$, $0 < \lambda < 1$, $\kappa > 0$
Require: Base ODE solver and fixed initial scale $\hat{c}_0$ (Appendix B)
$\ell_0 \leftarrow 0$
for $i = 0, \dots, N-1$ **do**
  $h_i \leftarrow t_{i+1} - t_i, \quad \mathbf{v}_i \leftarrow \mathbf{v}_\theta(\mathbf{x}_i, t_i)$
  if $i = 0$ **then**
    $\mathbf{m}_0 \leftarrow \mathbf{v}_0, \quad \hat{\mathbf{a}}_0 \leftarrow \mathbf{0}$
  else
    $\mathbf{d}_i \leftarrow \mathbf{v}_i - \mathbf{m}_i, \quad \hat{\mathbf{a}}_i \leftarrow \mathbf{d}_i / \ell_i$
    Compute $\tilde{c}_i, \hat{c}_i$ from Eq. (23)
  end if
  Compute b_i, α_i, ν_i from Eq. (25)
  $\mathbf{y}_i \leftarrow$ base solver proposal using $\mathbf{x}_i, \mathbf{v}_i, h_i$
  Draw fresh $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_{i+1} \leftarrow \mathbf{y}_i + (\alpha_i h_i^2/2)\hat{\mathbf{a}}_i + \nu_i \boldsymbol{\xi}_i$
  $\mathbf{m}_{i+1} \leftarrow \beta \mathbf{m}_i + (1-\beta)\mathbf{v}_i, \quad \ell_{i+1} \leftarrow h_i + \beta \ell_i$
  Retain $\boldsymbol{\xi}_i, \nu_i$ for the next probe
end for
return $\mathbf{x}_N$

To obtain these coefficients from the OU transition, set

$$\mathbf{r}_i = \frac{\lambda h_i}{2}\hat{\mathbf{a}}_i, \qquad \sigma_i^2 = \frac{2\kappa\nu_i^2}{h_i(1 - \mathrm{e}^{-2\kappa h_i})}. \tag{26}$$

Here α_i determines the strength of the integrated history correction. Since $0 < \alpha_i < \lambda < 1$, the noise variance is positive. The two components satisfy

$$b_i \mathbf{r}_i = \frac{\alpha_i h_i^2}{2}\hat{\mathbf{a}}_i, \qquad \alpha_i + \frac{\nu_i^2}{h_i^2 \hat{c}_i} = 1. \tag{27}$$

We use Euler's leading covariance deficit as the reference for this allocation. By Eq. (17), for an input from p_{t_i}, the ideal history displacement $\alpha_i h_i^2 \mathbf{a}/2$ contributes $\alpha_i h_i^2 \mathbb{E}[\mathbf{C}]$ to the covariance at leading order. We assign the remaining fraction $1 - \alpha_i$ of the estimated isotropic component to the noise. Since $\alpha_i = \lambda b_i / h_i$ includes the damping factor $b_i < h_i$, the allocation uses the drift actually accumulated during the OU refinement.

The exact transition gives

$$\boxed{\mathbf{x}_{i+1} = \mathbf{y}_i + \frac{\alpha_i h_i^2}{2}\hat{\mathbf{a}}_i + \nu_i \boldsymbol{\xi}_i, \qquad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}. \tag{28}$$

Let $\mathcal{H}_i$ contain all states, model evaluations, and previous Gaussian draws available before drawing $\boldsymbol{\xi}_i$. Conditional on $\mathcal{H}_i$, every coefficient in the update is fixed and $\boldsymbol{\xi}_i$ is independent. In particular,

$$\mathbb{E}[\|\mathbf{x}_{i+1} - \mathbf{y}_i\|^2 \mid \mathcal{H}_i] = \frac{\alpha_i^2 h_i^4}{4}\|\hat{\mathbf{a}}_i\|^2 + D\nu_i^2. \tag{29}$$

At the next step of Algorithm 1, the field evaluation at the refined state supplies the starting velocity for the base solver and updates both the history and noise estimate. The base solver may use further evaluations to form its proposal; the OU refinement adds none. For the next update, we retain the velocity memory, the preceding Gaussian draw, the effective time lag, and the preceding noise scale. Appendix D gives the Markov representation obtained by including this history in the sampling state.

3.3 Distributional effect of stochastic self-correction

Because random perturbations affect both the sample and the velocities stored in memory, we analyze the history error jointly with the numerical state. We consider Algorithm 1 with Euler base proposals.

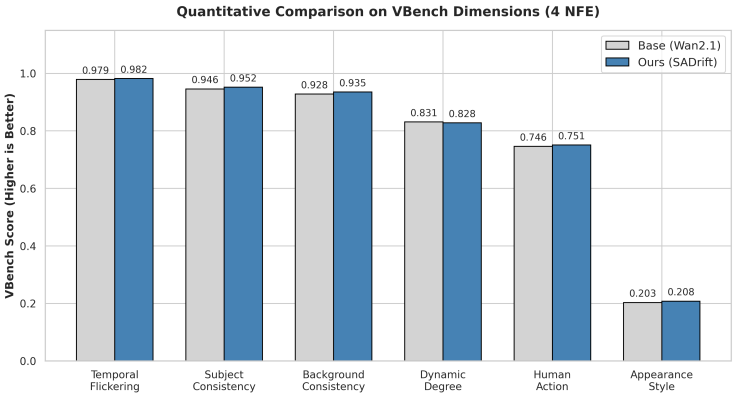


Figure 2: VBench consistency and action scores for the Wan2.1 baseline and SADrift at 4 NFE.

Let $h = \max_i h_i$ and consider quasiuniform grids, $\rho h \le h_i \le h$, for a fixed $\rho > 0$. The regularity conditions in Appendix B ensure uniform moments and justify Taylor expansion along the stochastic updates.

Proposition B.1 in Appendix B bounds the history error in the relevant joint moments at order h. For the stochastic term, Proposition C.1 in Appendix C identifies the mean unclipped covariance estimate through Gaussian integration by parts.

We now combine the history bound and calibration identity to compare one corrected step with exact transport from the same numerical input. Let $\mathbf{x}_i^+ = \Phi_{t_i,t_{i+1}}(\mathbf{x}_i)$, where Φ is the population ODE flow. Write $\mathbf{C}_i = \mathbf{C}(\mathbf{x}_i, t_i)$ and define

$$\bar{\mathbf{C}}_i = \mathbb{E}[\mathbf{C}_i], \qquad \bar{c}_i = \frac{\operatorname{tr} \bar{\mathbf{C}}_i}{D}, \qquad \varepsilon_i = \mathbb{E}[\hat{c}_i - \tilde{c}_i]. \tag{30}$$

Here $\bar{c}_i = \mathbb{E}[c(\mathbf{x}_i, t_i)]$, and ε_i is the mean bias introduced by clipping.

Theorem 3.2 (Local correction of conditional-mean concentration)**.** *Suppose the sampler uses Euler base proposals under the population field and starts from the prescribed Gaussian prior. Under the conditions of Appendix C, for $1 \le i < N$ and $\varphi \in C_b^4(\mathbb{R}^D)$,*

$$\mathbb{E}[\varphi(\mathbf{x}_i^+)] - \mathbb{E}[\varphi(\mathbf{x}_{i+1})] = \frac{h_i^2(1-\alpha_i)}{2}\mathbb{E}[(\mathbf{C}_i - \hat{c}_i\mathbf{I}) : \nabla^2\varphi(\mathbf{x}_i)] + \mathcal{O}(h^3). \tag{31}$$

The means differ by $\mathcal{O}(h^3)$, and

$$\operatorname{Cov}(\mathbf{x}_i^+) - \operatorname{Cov}(\mathbf{x}_{i+1}) = h_i^2(1-\alpha_i)[\bar{\mathbf{C}}_i - (\bar{c}_i + \varepsilon_i)\mathbf{I}] + \mathcal{O}(h^3). \tag{32}$$

The covariance remainder is bounded in Frobenius norm.

The history displacement supplies an α_i fraction of the leading covariance deficit, while the Gaussian term compensates for the isotropic component of the remainder. The residual covariance error depends on the anisotropy of $\bar{\mathbf{C}}_i$ and the clipping bias ε_i. Compared with the Euler discrepancy $h_i^2\bar{\mathbf{C}}_i + \mathcal{O}(h^3)$, the leading corrected error has Frobenius norm at most $1 - \alpha_i$ times as large when $|\varepsilon_i| \le \bar{c}_i$. If $\varepsilon_i = 0$, the leading discrepancy has zero trace, so total variance is recovered to this order; the remaining covariance error is proportional to $\bar{\mathbf{C}}_i - \bar{c}_i\mathbf{I}$. For other smooth observables, Eq. (31) retains the dependence on the state-dependent covariance, calibrated scale, and Hessian of the observable. These estimates describe the sampler through a local expansion with a finite-step remainder; the analytical exactness of Eq. (28) concerns the locally frozen OU subproblem.

For a learned field, additional errors enter through the velocity, its material acceleration, and its divergence. Appendix C traces these terms through the base proposal, history correction, and noise estimate.

4 Experiments

We evaluate SADrift on ImageNet image synthesis and TurboDiffusion video generation to examine its effect on generation quality at fixed sampling budgets. Each paired comparison uses the same

pretrained model and NFE budget for the base and corrected samplers. We also examine the contributions of history and noise, with further controls and evaluation details in Appendix E.

4.1 Experimental setup

We conduct our image-generation experiments on ImageNet 256×256 using the pretrained SiT-XL/2 model (675M parameters) and a classifier-free guidance scale of 1.5 [5]. For each configuration, we generate 50,000 samples and compute FID, sFID, Inception Score (IS), precision, and recall. Paired runs share the checkpoint, guidance, timestep schedule, initial latent and class draws, sample count, and evaluation pipeline. Our NFE counts include all model evaluations, including the predictor and corrector calls in Heun2. We reuse the starting velocity from each integration interval for the history correction and compute the OU update without additional model calls.

Unless otherwise stated, we use a memory coefficient of $\beta = 0.9$ and an OU damping rate of $\kappa = 1.0$. We report correction strength as the overall scale $1 + \lambda$: the values 1.25, 1.50, 1.70, and 1.75 correspond to $\lambda \in \{0.25, 0.50, 0.70, 0.75\}$. In Algorithm 1, the history weight λ determines the integrated correction through $\alpha_i = \lambda b_i / h_i$, while Eq. (26) relates the Gaussian displacement's standard deviation ν_i to the OU diffusion amplitude σ_i. For the baseline, we disable both correction branches and use the underlying solver. Appendix E details the scale convention and component definitions.

4.2 Few-step image generation

Across the budgets and correction scales in Table 1, we obtain lower FID with SADrift for both Euler and Heun2. At 10 NFE with Euler, the best result occurs at $1 + \lambda = 1.75$: FID decreases from 8.35 to 3.82, while IS increases from 184.11 to 224.74. At 20 NFE with Heun2, both $1 + \lambda = 1.70$ and $1 + \lambda = 1.75$ achieve FID 2.00, compared with 2.20 for the base solver. The table also includes 100-NFE ODE and SDE runs using the same SiT backbone.

We extend the comparison to DPM-Solver++ (2M) [13] using a SiT-specific adapter for the noise schedule and prediction parameterization. At a fixed history weight of $\lambda = 0.50$, FID decreases from 12.3051 to 7.3772 at 5 NFE, from 3.5653 to 2.7336 at 10 NFE, and from 2.4592 to 2.1464 at 20 NFE. The relative reductions are 40.05%, 23.33%, and 12.72%, respectively, with the largest gain at 5 NFE. At the same reported scale $1 + \lambda = 1.50$, Euler and Heun2 also improve at every tested budget (Table 5 in Appendix E). Table 6 gives the full DPM-Solver++ scale sweep. The role of each base solver's local error in selecting the refinement coefficients is discussed in Appendix D.

SADrift achieves FID 2.00 at 20 NFE with the frozen SiT backbone, compared with the published SiT result of 2.06 at 250 NFE. Table 2 places this result alongside other generative models with their published training and sampling setups. The paired results in Table 1 measure the sampler's effect with the backbone held fixed.

4.3 Component analysis and sensitivity

To examine the role of history and noise, we compare the full method with controls that remove either the drift $\mathbf{r}_i$ or the Gaussian displacement $\nu_i \boldsymbol{\xi}_i$ at 10 NFE (Table 3). Without history drift, FID remains close to the native solver: 3.2407 versus 3.2524 for Heun2 and 8.5657 versus 8.4539 for Euler. Retaining history while removing random perturbations lowers FID to 2.2315 and 4.7189, respectively. The full method further lowers FID to 2.2039 for Heun2 and 4.7117 for Euler. These comparisons indicate that the history correction accounts for most of the observed FID improvement, with Gaussian fluctuations providing a smaller additional reduction.

The uniform-feature-weight control and additional metrics are reported in Appendix E (Table 8).

The gains with adapted DPM-Solver++ persist across all four positive correction scales at 5, 10, and 20 NFE; we use $\lambda = 0.50$ for the fixed-scale comparison. We also examine memory decay in a separate controlled Euler experiment. As β increases from 0.50 to 0.99, FID changes from 6.939 to 6.780. Our default $\beta = 0.90$ gives FID 6.830, within 0.050 of the best observed value. These results indicate modest sensitivity to the memory decay in Eq. (18) over the tested range. Full sensitivity results appear in Appendix E.

Table 1: Performance of SiT-XL/2 on ImageNet 256×256. Flow matching baselines and **SADrift** are compared within each solver and NFE group. Parentheses report the overall scale $1 + \lambda$ defined in Section 4.1. SADrift rows are ordered by this scale, and baseline rows are gray. Best and second best results within each group are highlighted.

NFE	Method $(1 + \lambda)$	IS ↑	(Δ IS)	FID ↓	(Δ FID)	sFID ↓	Prec. ↑	Rec. ↑
	Baseline (Euler)	58.02	-	52.64	-	44.10	0.3692	0.4353
	SADrift (1.25)	79.37	(+36.8%)	36.33	(-31.0%)	29.53	0.4707	0.4417
5	SADrift (1.5)	98.94	(+70.5%)	25.66	(-51.3%)	19.21	0.5554	0.4501
	SADrift (1.7)	112.93	(+94.6%)	20.14	(-61.7%)	13.61	0.6068	0.4647
	SADrift (1.75)	**116.03**	**(+100.0%)**	**19.06**	**(-63.8%)**	**12.52**	**0.6172**	**0.4660**
	Baseline (Euler)	184.11	-	8.35	-	8.83	0.7303	0.5234
	SADrift (1.25)	205.38	(+11.6%)	5.44	(-34.8%)	5.86	0.7717	0.5297
10	SADrift (1.5)	217.89	(+18.3%)	4.23	(-49.3%)	**5.09**	0.7939	0.5326
	SADrift (1.7)	223.84	(+21.6%)	3.87	(-53.7%)	5.54	0.8065	0.5328
	SADrift (1.75)	**224.74**	**(+22.1%)**	**3.82**	**(-54.3%)**	5.75	**0.8082**	**0.5346**
	Baseline (Euler)	231.60	-	3.33	-	5.33	0.7902	0.5736
	SADrift (1.25)	242.07	(+4.5%)	2.70	(-18.9%)	4.44	0.8040	**0.5769**
20	SADrift (1.5)	248.75	(+7.4%)	2.46	(-26.1%)	**4.40**	0.8132	0.5742
	SADrift (1.7)	251.91	(+8.8%)	**2.44**	**(-26.7%)**	4.80	0.8151	0.5683
	SADrift (1.75)	**252.60**	**(+9.1%)**	2.45	(-26.4%)	4.94	**0.8162**	0.5682
	Baseline (Heun2)	148.39	-	14.64	-	22.11	0.6590	0.5155
	SADrift (1.25)	165.57	(+11.6%)	11.04	(-24.6%)	17.61	0.6899	0.5334
5	SADrift (1.5)	175.83	(+18.5%)	9.07	(-38.0%)	15.08	0.7096	0.5442
	SADrift (1.7)	**179.47**	**(+20.9%)**	8.46	(-42.2%)	**14.55**	**0.7143**	**0.5472**
	SADrift (1.75)	179.37	(+20.9%)	**8.45**	**(-42.3%)**	14.64	0.7137	0.5458
	Baseline (Heun2)	233.24	-	3.15	-	6.59	0.7882	0.5852
	SADrift (1.25)	241.66	(+3.6%)	2.65	(-15.9%)	5.88	0.7977	0.5860
10	SADrift (1.5)	247.84	(+6.3%)	2.41	(-23.5%)	**5.78**	0.8027	0.5910
	SADrift (1.7)	250.00	(+7.2%)	**2.36**	**(-25.1%)**	6.01	0.8056	0.5906
	SADrift (1.75)	**250.27**	**(+7.3%)**	**2.36**	**(-25.1%)**	6.11	**0.8059**	**0.5937**
	Baseline (Heun2)	250.65	-	2.20	-	4.73	0.8009	0.6000
	SADrift (1.25)	253.43	(+1.1%)	2.10	(-4.5%)	4.61	0.8043	0.5996
20	SADrift (1.5)	256.40	(+2.3%)	2.04	(-7.3%)	**4.57**	0.8071	0.5992
	SADrift (1.7)	258.97	(+3.3%)	**2.00**	**(-9.1%)**	4.60	**0.8094**	**0.6005**
	SADrift (1.75)	**259.38**	**(+3.5%)**	**2.00**	**(-9.1%)**	4.62	0.8091	0.5986
100	ODE (euler)	253.16	-	2.12	-	4.47	0.8059	0.5981
100	SDE (Heun)	270.29	-	2.08	-	4.63	0.8260	0.5867

4.4 Video generation

We apply SADrift to Wan2.1-1.3B with TurboDiffusion at 4 NFE [20, 19] and compare videos generated from the same prompt in Figure 1. The VBench evaluation [28] shows improvements in temporal flickering from 0.979 to 0.982, subject consistency from 0.946 to 0.952, and background consistency from 0.928 to 0.935 (Figure 2). Human action and appearance style also increase, from 0.746 to 0.751 and from 0.203 to 0.208, respectively, alongside a small decrease in dynamic degree from 0.831 to 0.828.

Conclusions

We introduced SADrift, an inference-time framework for improving few-step flow matching with pretrained models. By reusing velocity history, we construct a corrective drift and combine it with Gaussian fluctuations through an exact local Ornstein–Uhlenbeck transition. Our analysis of linear FM relates this correction to the covariance omitted by Euler and characterizes its leading local error

Table 2: Comparison of generative models on ImageNet 256×256. Epochs report published model training for the reference methods and additional training for SADrift; the frozen SiT backbone retains its pretraining. The SADrift row reports the best observed result in Table 1.

Setting	Method	#Params	Epochs	NFE	IS ↑	FID ↓	Prec. ↑	Rec. ↑
Published result	DiT-XL/2 [24]	675M	1400	250	278.2	2.27	0.83	0.57
	SiT-XL/2 [5]	675M	1400	250	270.3	2.06	0.82	0.59
Published result	LightningDiT [25]	675M	800	125	295.3	1.35	0.79	0.65
Published result	MDTv2 [26]	675M	1080	250	314.7	1.58	0.79	0.65
Published result	REPA [27]	675M	800	250	305.7	1.42	0.80	0.65
Inference-time (Ours)	SADrift (w. SiT)	675M	0	20	259.4	2.00	0.81	0.60

Table 3: History-drift and Gaussian-noise controls at 10 NFE. Each solver is evaluated in a matched 50K-sample comparison. The noise-free control retains deterministic refinement. Additional metrics and the feature-weight control appear in Table 8.

	Heun2		Euler	
Variant	**FID ↓**	**sFID ↓**	**FID ↓**	**sFID ↓**
Base solver	3.2524	6.5678	8.4539	8.8328
Without history drift	3.2407	6.6205	8.5657	8.8273
Without stochastic noise	2.2315	4.7507	4.7189	5.6873
Full (SADrift)	2.2039	4.7874	4.7117	5.6980

under the stated conditions. Without retraining or additional model evaluations, SADrift achieves FID 2.00 with Heun2 on ImageNet-256 at 20 NFE and improves VBench temporal consistency scores at 4 NFE.

Limitations and broader impact The quality of SADrift samples depends on the accuracy of the supplied velocity field. The method also requires storage for velocity memory and the preceding random perturbation. By making image and video generation more efficient, it can increase access to these models, including their use to create misleading synthetic media.

Acknowledgments and Disclosure of Funding

This work was supported in part by grants from the National Natural Science Foundation of China (52539005), the China Scholarship Council (202306150167), the fundamental research program of Guangdong, China (2023A1515011281), the Guangdong Basic and Applied Basic Research Foundation (24202107190000687), and the Foshan Science and Technology Research Project (2220001018608). We also thank Xiaomi Technology Group for providing resources for this research.

References

[1] Lipman, Y., R. T. Q. Chen, H. Ben-Hamu, et al. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*. 2023.

[2] Liu, X., C. Gong, qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*. 2023.

[3] Ho, J., A. Jain, P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, vol. 33, pages 6840–6851. 2020.

[4] Song, Y., J. Sohl-Dickstein, D. P. Kingma, et al. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*. 2021.

[5] Ma, N., M. Goldstein, M. S. Albergo, et al. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[6] Salimans, T., J. Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*. 2022.

[7] Song, Y., P. Dhariwal, M. Chen, et al. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

[8] Liu, X., X. Zhang, J. Ma, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*. 2024.

[9] Albergo, M. S., E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*. 2023.

[10] Lee, S., B. Kim, J. C. Ye. Minimizing trajectory curvature of ode-based generative models. In *International Conference on Machine Learning*, pages 18957–18973. PMLR, 2023.

[11] Tong, A., K. FATRAS, N. Malkin, et al. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

[12] Lu, C., Y. Zhou, F. Bao, et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, vol. 35, pages 5775–5787. 2022.

[13] —. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025.

[14] Zhao, W., L. Bai, Y. Rao, et al. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[15] Xu, Y., M. Deng, X. Cheng, et al. Restart sampling for improving generative processes. *Advances in Neural Information Processing Systems*, 36:76806–76838, 2023.

[16] Wang, X., A.-D. Dinh, D. Liu, et al. Boosting diffusion models with an adaptive momentum sampler. *arXiv preprint arXiv:2308.11941*, 2023.

[17] Wen, Z., X. Deng, P. Luo, et al. Score-based generative models with adaptive momentum. *arXiv preprint arXiv:2405.13726*, 2024.

[18] Rissanen, S., M. Heinonen, A. Solin. Free Hunch: Denoiser covariance estimation for diffusion models without extra costs. In *International Conference on Learning Representations*. 2025.

[19] Zhang, J., K. Zheng, K. Jiang, et al. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025.

[20] Wan, T., A. Wang, B. Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[21] Zwanzig, R. Memory effects in irreversible thermodynamics. *Physical Review*, 124:983–992, 1961.

[22] Mori, H. Transport, collective motion, and brownian motion. *Progress of Theoretical Physics*, 33(3):423–455, 1965.

[23] Doob, J. *Stochastic Processes*. Probability and Statistics Series. Wiley, 1953.

[24] Peebles, W., S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205. 2023.

[25] Yao, J., B. Yang, X. Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712. 2025.

[26] Gao, S., P. Zhou, M.-M. Cheng, et al. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23164–23173. 2023.

[27] Singh, J., X. Leng, Z. Wu, et al. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025.

[28] Huang, Z., Y. He, J. Yu, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818. 2024.

[29] Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[30] Batifol, S., A. Blattmann, F. Boesel, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

[31] Batzolis, G., J. Stanczuk, C.-B. Schönlieb, et al. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[32] Blattmann, A., R. Rombach, H. Ling, et al. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575. 2023.

[33] Chefer, H., P. Esser, D. Lorenz, et al. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

[34] Deng, M., H. Li, T. Li, et al. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

[35] Dhariwal, P., A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, vol. 34, pages 8780–8794. 2021.

[36] Du, Y., M. Plainer, R. Brekelmans, et al. Doob's lagrangian: A sample-efficient variational approach to transition path sampling. In *Advances in Neural Information Processing Systems*. 2024.

[37] Ho, J., T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 2021.

[38] Karras, T., M. Aittala, T. Aila, et al. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*. 2022.

[39] Kattumannil, S. K. On stein's identity and its applications. *Statistics & probability letters*, 79(12):1444–1449, 2009.

[40] Lee, K., H. Liu, M. Ryu, et al. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[41] Luo, S., Y. Tan, L. Huang, et al. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[42] Peluchetti, S. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

[43] Podell, D., Z. English, K. Lacey, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*. 2024.

[44] Rombach, R., A. Blattmann, D. Lorenz, et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695. 2022.

[45] Sauer, A., D. Lorenz, A. Blattmann, et al. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

[46] Shi, Y., V. De Bortoli, A. Campbell, et al. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

[47] Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[48] Wallace, B., M. Dang, R. Rafailov, et al. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238. 2024.

[49] Xu, J., X. Liu, Y. Wu, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[50] Yin, T., M. Gharbi, R. Zhang, et al. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623. 2024.

[51] Yu, S., S. Kwak, H. Jang, et al. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*. 2025.

[52] Zheng, H., W. Nie, A. Vahdat, et al. Fast sampling of diffusion models via operator learning. In *International conference on machine learning*, pages 42390–42402. PMLR, 2023.

A Conditional moments and the corrective current

Proposition 3.1 follows from finite second moments, with the convex inequality valid whenever both expectations exist. For the differential identities below, we work on intervals where the densities and conditional moments are smooth and assume sufficient integrability to exchange differentiation and integration. Boundary terms are assumed to vanish in each integration by parts.

Proof of Proposition 3.1. The paired interpolation gives $\mathbf{x}_{t+h} = \mathbf{x}_t + h\mathbf{u}$ and hence $\mathbf{y} = \mathbb{E}[\mathbf{x}_{t+h} \mid \mathbf{x}_t]$. Conditional Jensen's inequality proves the convex inequality, and taking expectations proves equality of the means. Moreover,

$$\mathrm{Cov}(\mathbf{x}_{t+h} \mid \mathbf{x}_t) = h^2\mathbf{C}(\mathbf{x}_t, t).$$

The law of total covariance then gives

$$\mathrm{Cov}(\mathbf{x}_{t+h}) = \mathrm{Cov}(\mathbf{y}) + h^2\mathbb{E}[\mathbf{C}(\mathbf{x}_t, t)].$$

The last term is positive semidefinite and vanishes exactly when $\mathbb{E}\|\mathbf{u} - \mathbb{E}[\mathbf{u} \mid \mathbf{x}_t]\|^2 = 0$. At $t = 0$, the population velocity is $\mathbf{v}(\boldsymbol{\epsilon}, 0) = \mathbb{E}[\mathbf{x}_{\mathrm{data}}] - \boldsymbol{\epsilon}$, which proves the one-step example. □

Derivation of Eqs. (12)–(14). Let $p_t(\mathbf{x}, \mathbf{u})$ be the joint density of the paired state and velocity. Constant paired velocity gives

$$\partial_t p_t(\mathbf{x}, \mathbf{u}) + \nabla_{\mathbf{x}} \cdot [\mathbf{u} p_t(\mathbf{x}, \mathbf{u})] = 0.$$

Integration over $\mathbf{u}$ gives continuity. Multiplication by $\mathbf{u}$ before integration gives

$$\partial_t(p_t\mathbf{v}) + \nabla_{\mathbf{x}} \cdot \big(p_t\mathbb{E}[\mathbf{u}\mathbf{u}^\top \mid \mathbf{x}_t = \mathbf{x}]\big) = \mathbf{0}.$$

Substitute $\mathbb{E}[\mathbf{u}\mathbf{u}^\top \mid \mathbf{x}_t = \mathbf{x}] = \mathbf{v}\mathbf{v}^\top + \mathbf{C}$. Using continuity to cancel the terms multiplying $\mathbf{v}$ yields $p_t\mathbf{a} = -\nabla \cdot (p_t\mathbf{C})$. Componentwise integration by parts gives

$$\begin{aligned}\int p_t \nabla\varphi^\top \mathbf{a}\,\mathrm{d}\mathbf{x} &= -\sum_{k,j}\int \partial_{x_k}\varphi\,\partial_{x_j}(p_t C_{kj})\,\mathrm{d}\mathbf{x} \\ &= \sum_{k,j}\int p_t C_{kj}\partial_{x_j x_k}\varphi\,\mathrm{d}\mathbf{x}.\end{aligned}$$

This is Eq. (14). □

Linear and quadratic test functions give Eq. (17). We use $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{y} - \mathbb{E}\mathbf{y})^\top]$. Integration of $x_k p_t a_j$ by parts gives $\mathbb{E}[C_{jk}]$. Symmetry of the conditional covariance completes the matrix identity.

Derivation of Eq. (15). Assume $\varphi \in C_b^3(\mathbb{R}^D)$ and $\mathbb{E}\|\mathbf{u}\|^3 < \infty$. Expand $\varphi(\mathbf{x}_t + h\mathbf{u})$ and $\varphi(\mathbf{x}_t + h\mathbf{v}(\mathbf{x}_t, t))$ about $\mathbf{x}_t$. The first-order terms have the same expectation by conditional averaging. The difference of the quadratic terms is

$$\frac{h^2}{2}\mathbb{E}\big[\big(\mathbb{E}[\mathbf{u}\mathbf{u}^\top \mid \mathbf{x}_t] - \mathbf{v}\mathbf{v}^\top\big) : \nabla^2\varphi(\mathbf{x}_t)\big] = \frac{h^2}{2}\mathbb{E}[\mathbf{C} : \nabla^2\varphi(\mathbf{x}_t)].$$

Bounded third derivatives and conditional Jensen's inequality $\mathbb{E}\|\mathbf{v}\|^3 \le \mathbb{E}\|\mathbf{u}\|^3$ bound the remaining terms by a constant times h^3. □

Score, flow, and bridge identities. For a scalar diffusion coefficient depending only on time, the Fokker–Planck equation is $\partial_t p_t = -\nabla \cdot (p_t\mathbf{b}) + g^2\Delta p_t/2$, where $\Delta = \nabla \cdot \nabla$ is the spatial Laplacian. Since $p_t\mathbf{s} = \nabla p_t$, substitution of Eq. (2) gives Eq. (3). For the independent Gaussian interpolation, let $\boldsymbol{\mu}(\mathbf{x}, t) = \mathbb{E}[\mathbf{x}_{\mathrm{data}} \mid \mathbf{x}_t = \mathbf{x}]$. Differentiating its Gaussian conditional density yields

$$\mathbf{s}(\mathbf{x}, t) = -\frac{\mathbf{x} - t\boldsymbol{\mu}(\mathbf{x}, t)}{(1-t)^2}, \qquad \boldsymbol{\mu}(\mathbf{x}, t) = \mathbf{x} + (1-t)\mathbf{v}(\mathbf{x}, t).$$

This proves Eq. (6).

For $t < 1$ and a positive smooth transition density of the reference diffusion, the bridge function satisfies $(\partial_t + \mathcal{L}_t)h = 0$, where $\mathcal{L}_t f = \mathbf{b}^\top \nabla f + g^2 \Delta f/2$. The transformed generator is $\mathcal{L}_t^h f = h^{-1}[\mathcal{L}_t(hf) - f\mathcal{L}_t h]$. The product rule gives drift $\mathbf{b} + g^2 \nabla \log h$. The conditioned initial law is reweighted by $h(\mathbf{x}, 0)$; the resulting marginal density is $p_t h / p_1(\mathbf{y})$. Taking its logarithmic gradient and subtracting $g^2 \mathbf{s}^h/2$ from the transformed drift proves Eq. (8).

Proof of Eq. (22). For $0 < t < 1$, let $\boldsymbol{\mu}(\mathbf{x}, t) = \mathbb{E}[\mathbf{x}_{\text{data}} \mid \mathbf{x}_t = \mathbf{x}]$. The Gaussian conditional density has mean $t\mathbf{x}_{\text{data}}$ and covariance $(1-t)^2\mathbf{I}$. Differentiating the posterior mean gives

$$\nabla_{\mathbf{x}} \boldsymbol{\mu}(\mathbf{x}, t) = \frac{t}{(1-t)^2} \operatorname{Cov}(\mathbf{x}_{\text{data}} \mid \mathbf{x}_t = \mathbf{x}).$$

The paired identities $\mathbf{x}_{\text{data}} = \mathbf{x}_t + (1-t)\mathbf{u}$, $\boldsymbol{\mu} = \mathbf{x} + (1-t)\mathbf{v}$, and $\operatorname{Cov}(\mathbf{x}_{\text{data}} \mid \mathbf{x}_t = \mathbf{x}) = (1-t)^2\mathbf{C}$ then imply $\mathbf{I} + (1-t)\mathbf{J}_{\mathbf{v}} = t\mathbf{C}$. □

B History along the stochastic sampling trajectory

To analyze the history estimator on the actual sampling trajectory, we separate the velocity increments into acceleration and noise responses. Their different contributions to the weak error lead to Proposition B.1.

Noise-scale stabilization. We use fixed positive clipping bounds:

$$\operatorname{clip}(s) = \min\{c_{\max}, \max\{c_{\min}, s\}\}, \qquad 0 < c_{\min} < c_{\max} < \infty.$$

The fixed initial scale satisfies $\hat{c}_0 \in [c_{\min}, c_{\max}]$. The lower bound keeps the Gaussian perturbation nonzero for the next probe, and the upper bound controls its moments uniformly over the grid.

Regularity and moment conditions. We consider Euler base proposals. For the error estimates as $h \to 0$, assume that the evaluated field is jointly C^3 in state and time and globally Lipschitz in the state, uniformly in time. Its time derivative has at most linear growth, and its derivatives through order three have polynomial growth, uniformly on the time interval. Initial states have moments of every finite order. The grid obeys $\rho h \le h_i \le h$ with fixed $\rho > 0$, and $\beta, \lambda, \kappa, c_{\min}, c_{\max}$ are fixed as $h \to 0$. The linear-growth conditions give the uniform moment bounds used below.

The arguments below apply to either the population or learned field. Within this section, write $\mathbf{v}$ and $\mathbf{a}$ for the evaluated field and its material acceleration. For a random vector, $\|\mathbf{z}\|_{L^p} = (\mathbb{E}\|\mathbf{z}\|^p)^{1/p}$. The constant K may change between estimates but is independent of the grid. All bounds are for a fixed dimension and time horizon.

Proposition B.1 (History accuracy under stochastic feedback)**.** *For the supplied field, set* $\mathbf{e}_i = \hat{\mathbf{a}}_i - \mathbf{a}(\mathbf{x}_i, t_i)$*. Under the stated regularity conditions, for every globally Lipschitz vector test function* $\boldsymbol{\psi}$ *and* $1 \le i < N$*,*

$$\left|\mathbb{E}[\mathbf{e}_i^\top \boldsymbol{\psi}(\mathbf{x}_i)]\right| \le K_\psi h, \qquad \|\operatorname{sym} \operatorname{Cov}(\mathbf{x}_i, \mathbf{e}_i)\|_F \le Kh. \tag{33}$$

Here $\operatorname{sym} \mathbf{B} = (\mathbf{B} + \mathbf{B}^\top)/2$ *is the symmetric part of* $\mathbf{B}$*. The constants are independent of the grid size.*

The noise responses in the velocity increments form a weighted sum of martingale differences. Although they remain in the history estimate, their expected inner product with a Lipschitz function of the current state is of order h. This is the error that enters the local weak expansion when the estimated acceleration replaces the exact acceleration.

Moment bounds for the sampling update. The calibration bounds and Eq. (25) give

$$h_i \sqrt{(1-\lambda)c_{\min}} \le \nu_i \le h_i \sqrt{c_{\max}}.$$

Thus every probe denominator is positive. Write $X_i = \|\mathbf{x}_i\|_{L^p}$ and $D_i = \|\mathbf{d}_i\|_{L^p}$. The lag satisfies $\ell_i \ge \rho h$ for $i \ge 1$. Linear growth of the field, the update, and the history recursion imply

$$X_{i+1} \le (1 + Kh)X_i + Kh + KhD_i, \qquad D_{i+1} \le (\beta + Kh)D_i + Kh(1 + X_i).$$

For sufficiently small h, $\beta + Kh < 1$. Discrete Gronwall applied to $X_i + D_i$, followed by the geometric recursion for D_i, gives

$$\sup_i X_i \le K_p, \qquad \sup_i D_i \le K_p h, \qquad \sup_i \|\hat{\mathbf{a}}_i\|_{L^p} \le K_p.$$

In the recursion for $X_i + D_i$, the coefficient of D_i is $\beta + 2Kh \le 1 + K'h$, and the other coefficients also have this form. Once X_i is bounded, the D_i recursion has forcing of order h. The same estimates give $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_{L^p} \le Kh_i$. Clipping bounds the adaptive noise scale on every trajectory, and each new Gaussian draw is independent of $\mathcal{H}_i$; these properties justify the same moment estimates under feedback.

Proof of Proposition B.1. Taylor expansion of the evaluated velocity along one sampling step gives

$$\mathbf{v}_{j+1} - \mathbf{v}_j = h_j \mathbf{a}(\mathbf{x}_j, t_j) + \nu_j \mathbf{J}_{\mathbf{v}}(\mathbf{x}_j, t_j)\boldsymbol{\xi}_j + \boldsymbol{\rho}_j, \qquad \|\boldsymbol{\rho}_j\|_{L^2} \le Kh_j^2. \tag{34}$$

The directed OU displacement is of order h_j^2 in every required moment. Quadratic Taylor terms are also of this order because the state increment has L^p norm of order h_j. Polynomial derivative bounds and the moment estimates above justify the displayed remainder estimate.

Using $\ell_i = \sum_{j<i} \beta^{i-1-j} h_j$ in the exact history kernel yields

$$\hat{\mathbf{a}}_i - \mathbf{a}(\mathbf{x}_i, t_i) = \frac{1}{\ell_i} \sum_{j<i} \beta^{i-1-j} \nu_j \mathbf{J}_{\mathbf{v}}(\mathbf{x}_j, t_j)\boldsymbol{\xi}_j + \boldsymbol{\rho}_i^{\mathrm{hist}}.$$

Smoothness, polynomial growth, and the increment bounds give $\|\mathbf{a}(\mathbf{x}_j, t_j) - \mathbf{a}(\mathbf{x}_i, t_i)\|_{L^2} \le K(i-j)h$. Consequently, summing the Taylor remainders and the change in acceleration over the exponentially weighted history gives $\|\boldsymbol{\rho}_i^{\mathrm{hist}}\|_{L^2} \le Kh$. Here $\sum_{k\ge 0}(k+1)\beta^k < \infty$ and $\ell_i \ge \rho h$.

For each $j < i$, the coefficient $\nu_j \mathbf{J}_{\mathbf{v}}(\mathbf{x}_j, t_j)$ is $\mathcal{H}_j$-measurable and $\boldsymbol{\xi}_j$ is fresh. Its inner product with $\boldsymbol{\psi}(\mathbf{x}_j)$ therefore has zero expectation. Subtracting this term from the martingale sum replaces $\boldsymbol{\psi}(\mathbf{x}_i)$ by $\boldsymbol{\psi}(\mathbf{x}_i) - \boldsymbol{\psi}(\mathbf{x}_j)$ within the expectation. Cauchy–Schwarz and Lipschitz continuity give

$$\frac{1}{\ell_i} \sum_{j<i} \beta^{i-1-j} \|\nu_j \mathbf{J}_{\mathbf{v}}(\mathbf{x}_j, t_j)\boldsymbol{\xi}_j\|_{L^2} \|\boldsymbol{\psi}(\mathbf{x}_i) - \boldsymbol{\psi}(\mathbf{x}_j)\|_{L^2} \le K_\psi h.$$

The history remainder has the same bound when paired with $\boldsymbol{\psi}(\mathbf{x}_i)$, whose second moment is bounded. This proves the first claim. Constant and coordinate-linear vector test functions control the error mean and all entries of its cross-covariance with $\mathbf{x}_i$, proving the second claim. □

Relation to smooth-history approximation. On a smooth velocity trace, the normalized history admits a deterministic Taylor bound.

Lemma B.2 (Approximation by normalized smooth history)**.** *Suppose* $\mathbf{v}_j = \mathbf{g}(t_j)$ *and* $\|\mathbf{g}''(s)\| \le K$ *over* $[t_0, t_i]$. *With the incoming EMA weights* ω_{ij},

$$\|\hat{\mathbf{a}}_i - \mathbf{g}'(t_i)\| \le \frac{K}{2\ell_i} \sum_{j<i} \omega_{ij}(t_i - t_j)^2. \tag{35}$$

On a uniform grid, the bound is at most $Kh(1+\beta)/(2(1-\beta))$.

Proof. For $i \ge 1$, $\omega_{i0} = \beta^{i-1}$ and $\omega_{ij} = (1-\beta)\beta^{i-1-j}$ for $1 \le j < i$. The weights sum to one, and $\ell_i = \sum_{j<i} \omega_{ij}(t_i - t_j)$. Taylor expansion of $\mathbf{g}(t_j)$ about t_i proves the first bound. On a uniform grid, the tail weight at ages of at least k steps is β^{k-1}, so

$$\frac{\sum_{j<i} \omega_{ij}(t_i - t_j)^2}{\ell_i} = h \frac{\sum_{k=1}^{i} (2k-1)\beta^{k-1}}{\sum_{k=1}^{i} \beta^{k-1}} \le h \frac{1+\beta}{1-\beta}.$$

The finite weighted average is no larger than the infinite geometric average of the increasing sequence $2k-1$. □

C Noise calibration and local distributional correction

We analyze one corrected step from the state produced by the preceding numerical updates. To apply the population moment identities to this state, we bound the difference between its distribution and the prescribed FM marginal, then use this bound in the proof of Theorem 3.2.

Exact OU transition. Conditional on $\mathcal{H}_i$, the local coefficients are fixed. The integrating-factor solution is

$$\mathbf{z}_i(h_i) = b_i \mathbf{r}_i + \sqrt{h_i}\sigma_i \int_0^{h_i} \mathrm{e}^{-\kappa(h_i - s)}\, \mathrm{d}\mathbf{w}_i(s).$$

The stochastic term is centered Gaussian with covariance

$$h_i \sigma_i^2 \int_0^{h_i} \mathrm{e}^{-2\kappa s}\, \mathrm{d}s\, \mathbf{I} = \frac{h_i \sigma_i^2}{2\kappa}(1 - \mathrm{e}^{-2\kappa h_i})\mathbf{I} = \nu_i^2 \mathbf{I}.$$

Substituting Eq. (26) gives the transition in Eq. (28) and its conditional second moment. Moreover, $b_i \le h_i$ and $\alpha_i = \lambda + \mathcal{O}(h_i)$. Thus ν_i^2 specifies the covariance per coordinate of the integrated OU displacement.

The Gaussian term contributes covariance according to the mean calibrated scale. To characterize that mean under the population field, we condition on $\mathcal{H}_{i-1}$. This fixes $\mathbf{m}_i$ and ν_{i-1}, while $\mathbf{x}_i$ remains affine in the fresh draw $\boldsymbol{\xi}_{i-1}$. Gaussian integration by parts then expresses the probe expectation through $\operatorname{tr} \mathbf{J}_{\mathbf{v}}$. Together with Eq. (22), it yields the following calibration identity.

Proposition C.1 (Noise-response calibration)**.** *With $\mathbf{v}_\theta = \mathbf{v}$ and $1 \le i < N$, the unclipped probe in Eq.* (23) *satisfies*

$$\mathbb{E}[\tilde{c}_i \mid \mathcal{H}_{i-1}] = \frac{1}{D}\mathbb{E}[\operatorname{tr} \mathbf{C}(\mathbf{x}_i, t_i) \mid \mathcal{H}_{i-1}]. \tag{36}$$

The expectations are taken under the incoming law generated by Algorithm 1. The probe therefore targets the expected covariance scale of the states reached by the sampler.

Proof of Proposition C.1. Conditional on $\mathcal{H}_{i-1}$, the state $\mathbf{x}_i$ is affine in $\boldsymbol{\xi}_{i-1}$, with derivative $\nu_{i-1}\mathbf{I}$. The incoming memory $\mathbf{m}_i$ is fixed under this conditioning. Gaussian integration by parts gives

$$\mathbb{E}\left[\frac{(\mathbf{v}(\mathbf{x}_i, t_i) - \mathbf{m}_i)^\top \boldsymbol{\xi}_{i-1}}{\nu_{i-1}} \,\middle|\, \mathcal{H}_{i-1}\right] = \mathbb{E}[\operatorname{tr} \mathbf{J}_{\mathbf{v}}(\mathbf{x}_i, t_i) \mid \mathcal{H}_{i-1}].$$

Combining this with Eq. (22) proves the result. The conditioning fixes the incoming memory and noise amplitude while averaging over the perturbation that produces the current state. □

The same argument gives a weighted identity. For a smooth scalar function f with integrable terms, use the covariance scale $c(\mathbf{x}, t)$ from Eq. (21). Then

$$\mathbb{E}[(\tilde{c}_i - c(\mathbf{x}_i, t_i))f(\mathbf{x}_i)] = \frac{1 - t_i}{D t_i}\mathbb{E}[\mathbf{d}_i^\top \nabla f(\mathbf{x}_i)]. \tag{37}$$

For bounded ∇f, the right-hand side is $\mathcal{O}(h/t_i)$ by the history moment bound. For $f = 1$, the right-hand side vanishes and recovers the unbiased calibration identity. Clipping changes this expectation by ε_i, as defined in Eq. (30).

Distribution along the sampling trajectory. In addition to the conditions in Appendix B, assume the population field has a global ODE flow and that $\mathbf{C}$, $\mathbf{a}$, and their state derivatives through order two have polynomial growth uniformly in time. We use test functions whose state derivatives through order two have polynomial growth; their compositions with the backward flow have the same bounds uniformly on the time interval. The moment identities of Appendix A are assumed to hold, with integrable terms and vanishing boundary terms. These sufficient conditions include smooth linear population fields with nondegenerate Gaussian data.

For every such test function f, the numerical sampler satisfies

$$\left|\mathbb{E}[f(\mathbf{x}_i)] - \mathbb{E}_{p_{t_i}}[f]\right| = \mathcal{O}(h). \tag{38}$$

Indeed, fix the final time of this comparison and compose f with the exact backward ODE flow. Its time derivative plus its spatial gradient paired with $\mathbf{v}$ is zero. Conditional Taylor expansion along the numerical step cancels the first-order transport term. The history correction has first moment of order h_i^2, the fresh noise has zero conditional mean and second moment of order h_i^2, and the remaining terms are integrable by the moment bounds. The expected local defect is therefore $\mathcal{O}(h_i^2)$. Summing over the grid gives $\sum_i h_i^2 \le h$ and proves Eq. (38). Each local expansion conditions on the sampling history and averages over the fresh Gaussian draw.

Applying this estimate to the identities under p_{t_i} gives

$$\mathbb{E}[\mathbf{a}(\mathbf{x}_i, t_i)] = \mathcal{O}(h), \qquad \operatorname{sym}\operatorname{Cov}(\mathbf{x}_i, \mathbf{a}(\mathbf{x}_i, t_i)) = \bar{\mathbf{C}}_i + \mathcal{O}(h).$$

Thus the population moment identities hold under the numerical input distribution up to an error of order h. The same argument applies to Eq. (14) for the smooth test functions in the theorem.

Proof of Theorem 3.2. Write $\mathbf{x} = \mathbf{x}_i$, $\mathbf{v} = \mathbf{v}(\mathbf{x}_i, t_i)$, $\mathbf{a} = \mathbf{a}(\mathbf{x}_i, t_i)$, and $\mathbf{y} = \mathbf{x} + h_i\mathbf{v}$. The exact ODE step has the expansion

$$\mathbf{x}_i^+ = \mathbf{y} + \frac{h_i^2}{2}\mathbf{a} + \mathcal{O}_{L^2}(h_i^3).$$

The remainder has L^2 norm of order h_i^3. The SADrift step is $\mathbf{x}_{i+1} = \mathbf{y} + (\alpha_i h_i^2/2)\hat{\mathbf{a}}_i + \nu_i\boldsymbol{\xi}_i$. Conditional Taylor expansion over the fresh Gaussian gives

$$\mathbb{E}[\varphi(\mathbf{x}_{i+1})] - \mathbb{E}[\varphi(\mathbf{y})] = \frac{h_i^2}{2}\mathbb{E}[\alpha_i\hat{\mathbf{a}}_i^\top\nabla\varphi(\mathbf{x}) + (1-\alpha_i)\hat{c}_i\Delta\varphi(\mathbf{x})] + \mathcal{O}(h^3).$$

Evaluating the derivatives at $\mathbf{x}$ rather than $\mathbf{y}$ changes the expansion by $\mathcal{O}(h^3)$. Proposition B.1, applied to $\boldsymbol{\psi} = \nabla\varphi$, replaces $\hat{\mathbf{a}}_i$ by $\mathbf{a}$ in this expectation with the same remainder. Subtract from the exact-flow expansion and use the transferred weak-current identity to obtain Eq. (31). Constant vector tests, the history error mean, and $\mathbb{E}\mathbf{a} = \mathcal{O}(h)$ give the stated mean discrepancy.

For covariance, $\nu_i\boldsymbol{\xi}_i$ is centered conditional on $\mathcal{H}_i$, so all cross terms with $\mathcal{H}_i$-measurable variables vanish. Thus

$$\operatorname{Cov}(\mathbf{x}_{i+1}) = \operatorname{Cov}(\mathbf{y}) + \alpha_i h_i^2 \operatorname{sym}\operatorname{Cov}(\mathbf{x}, \hat{\mathbf{a}}_i) + \mathbb{E}[\nu_i^2]\mathbf{I} + \mathcal{O}(h^3).$$

The exact flow has covariance $\operatorname{Cov}(\mathbf{y}) + h_i^2 \operatorname{sym}\operatorname{Cov}(\mathbf{x}, \mathbf{a}) + \mathcal{O}(h^3)$. The stochastic-history bound and the incoming weak-error estimate replace both cross-covariances by $\bar{\mathbf{C}}_i$ up to order h. Finally, Proposition C.1 gives $\mathbb{E}[\hat{c}_i] = \bar{c}_i + \varepsilon_i$. Substituting $\mathbb{E}[\nu_i^2] = h_i^2(1-\alpha_i)(\bar{c}_i + \varepsilon_i)$ proves Eq. (32). □

Covariance allocation and its calibration condition. For a positive semidefinite matrix $\mathbf{C}$, write $c = \operatorname{tr}(\mathbf{C})/D$. The identity $\|\mathbf{C} - s\mathbf{I}\|_F^2 = \|\mathbf{C} - c\mathbf{I}\|_F^2 + D(s-c)^2$ proves Eq. (21), since $c \ge 0$. The traceless part of $\bar{\mathbf{C}}_i$ is orthogonal to $\mathbf{I}$ in the Frobenius inner product. Therefore

$$\|\bar{\mathbf{C}}_i - (\bar{c}_i + \varepsilon_i)\mathbf{I}\|_F^2 = \|\bar{\mathbf{C}}_i - \bar{c}_i\mathbf{I}\|_F^2 + D\varepsilon_i^2.$$

If $|\varepsilon_i| \le \bar{c}_i$, this is at most $\|\bar{\mathbf{C}}_i\|_F^2$. The factor $1 - \alpha_i < 1$ gives the stated contraction of the leading local covariance discrepancy whenever $\bar{\mathbf{C}}_i \neq \mathbf{0}$. Taking the trace of the theorem gives

$$\operatorname{tr}\big[\operatorname{Cov}(\mathbf{x}_i^+) - \operatorname{Cov}(\mathbf{x}_{i+1})\big] = -Dh_i^2(1-\alpha_i)\varepsilon_i + \mathcal{O}(h^3).$$

When $\varepsilon_i = 0$, the leading total-variance deficit vanishes. This comparison uses exact transport from the same numerical input, whose accumulated distributional error is bounded by Eq. (38).

Learned-field terms. Let $\boldsymbol{\delta} = \mathbf{v}_\theta - \mathbf{v}$ and denote its Jacobian by $\mathbf{J}_{\boldsymbol{\delta}}$, with $\boldsymbol{\delta}_i = \boldsymbol{\delta}(\mathbf{x}_i, t_i)$. To isolate the base-proposal error, compare corrected outputs $\mathbf{y}_\theta^+$ and $\mathbf{y}^+$ from a common input, keeping the history correction, noise amplitude ν_i, and Gaussian vector $\boldsymbol{\xi}_i$ fixed. Changing only the base-proposal velocity gives $\mathbf{y}_\theta^+ - \mathbf{y}^+ = h_i\boldsymbol{\delta}_i$. The Lipschitz inequality for φ yields

$$\big|\mathbb{E}[\varphi(\mathbf{y}_\theta^+)] - \mathbb{E}[\varphi(\mathbf{y}^+)]\big| \le h_i\|\nabla\varphi\|_\infty\,\mathbb{E}\|\boldsymbol{\delta}_i\|. \tag{39}$$

The history proposition estimates the supplied field's acceleration, whose difference from the population acceleration is

$$\mathbf{a}_\theta - \mathbf{a} = \partial_t\boldsymbol{\delta} + \mathbf{J}_\mathbf{v}\boldsymbol{\delta} + \mathbf{J}_{\boldsymbol{\delta}}\mathbf{v} + \mathbf{J}_{\boldsymbol{\delta}}\boldsymbol{\delta}.$$

Gaussian integration by parts remains valid for the learned field. Along the resulting sampling trajectory, the unclipped probe satisfies

$$\mathbb{E}[\tilde{c}_i] - \frac{1}{D}\mathbb{E}[\operatorname{tr} \mathbf{C}(\mathbf{x}_i, t_i)] = \frac{1 - t_i}{D t_i}\mathbb{E}[\operatorname{tr} \mathbf{J}_{\boldsymbol{\delta}}(\mathbf{x}_i, t_i)].$$

Clipping adds $\mathbb{E}[\hat{c}_i - \tilde{c}_i]$ to the calibration bias. The velocity error affects the base proposal, the velocity and derivative errors affect the history correction, and the divergence error affects the expected noise scale.

D Memory representation and implementation conventions

Non-equilibrium memory. Eliminating the moving average gives the exact representation

$$\mathbf{x}_{i+1} = \mathbf{y}_i + \sum_{j=0}^{i-1} \frac{\lambda h_i b_i}{2\ell_i} \beta^{i-1-j} (\mathbf{v}_{j+1} - \mathbf{v}_j) + \nu_i \boldsymbol{\xi}_i, \qquad i \geq 1. \tag{40}$$

The drift depends on past velocity increments through exponentially decaying weights, while the noise amplitude depends on the response to the preceding perturbation. For $i \geq 1$ the augmented state $(\mathbf{x}_i, \mathbf{m}_i, \ell_i, \boldsymbol{\xi}_{i-1}, \nu_{i-1})$ is Markovian on the prescribed grid for a one-step base solver. A multistep base solver additionally retains its own solver history. The position process alone generally is not.

This dependence on past velocity changes, together with OU damping and random forcing, motivates the generalized Langevin interpretation [21, 22]. As the sampler advances, the refinement center moves with the base proposal and the corrections adapt to the trajectory. The resulting dynamics are driven and non-equilibrium, with drift and noise chosen to compensate for the covariance deficit of the numerical update.

Continuous memory interpretation. For a smooth velocity trace $\mathbf{g}(t)$ and relaxation time $\tau_{\mathrm{m}} > 0$, the auxiliary equation $\dot{\mathbf{m}} = (\mathbf{g} - \mathbf{m})/\tau_{\mathrm{m}}$ has solution

$$\mathbf{m}(t) = \mathrm{e}^{-(t-t_0)/\tau_{\mathrm{m}}} \mathbf{m}(t_0) + \frac{1}{\tau_{\mathrm{m}}} \int_{t_0}^{t} \mathrm{e}^{-(t-s)/\tau_{\mathrm{m}}} \mathbf{g}(s)\, \mathrm{d}s.$$

With $\mathbf{m}(t_0) = \mathbf{g}(t_0)$, its residual equals the exponentially weighted integral of $\mathrm{d}\mathbf{g}$. A fixed continuous relaxation time corresponds to $\beta_i = \mathrm{e}^{-h_i/\tau_{\mathrm{m}}}$. Algorithm 1 uses a fixed β per integration interval.

Time, initialization, and numerical evaluation. We use increasing generation time throughout, so models parameterized by decreasing noise time require the corresponding velocity sign conversion. At initialization, the history correction is zero and the covariance scale is the fixed input $\hat{c}_0$. Subsequent probes use $t_i > 0$ and $\nu_{i-1} > 0$, with the residual computed from the incoming memory before the EMA update in Algorithm 1. For small κh_i, computing $b_i = -\operatorname{expm1}(-\kappa h_i)/\kappa$ avoids numerical cancellation. The integrated OU noise is sampled directly as $\nu_i \boldsymbol{\xi}_i$.

Other base solvers. In Algorithm 1, the base solver constructs the proposal before OU refinement. Heun2 combines the starting velocity with a velocity evaluated at its predicted endpoint; a multistep solver also uses its own stored predictions. The SADrift memory is updated once per integration interval using the starting velocity. All internal solver evaluations count toward NFE.

The velocity moment identity in Eq. (13) depends on the FM interpolation. Noise-response calibration also holds for any base proposal that is fixed before the fresh Gaussian draw, since the next state remains conditionally affine in that draw. The finite-step covariance error depends on the base solver. Euler's deficit is given by Eq. (10) and analyzed in Theorem 3.2; a second-order proposal already contains the leading material-acceleration term.

Deterministic limit. As $\kappa \to 0$ and $\lambda \to 1$, $\alpha_i \to 1$ and $\nu_i^2 \to 0$ for bounded $\hat{c}_i$. The state update approaches deterministic history extrapolation. For Euler base proposals and $\beta = 0$, this limit is the variable-step two-step Adams–Bashforth update after its Euler initialization. The covariance probe is used only at positive noise levels.

Relation to stochastic and momentum samplers. Restart sampling revisits a noisier state and follows a backward ODE trajectory [15]. Adaptive momentum samplers also reuse sampling history: **(author?)** [16] use Adam-inspired moment estimates to smooth reverse sampling, and **(author?)** [17] incorporate adaptive momentum into Langevin sampling and stochastic correctors. In our method, the current-minus-EMA residual measures velocity changes and supplies an additive correction to the base FM proposal. The effective-lag normalization connects this residual to material acceleration, while the local OU transition integrates it with Gaussian fluctuations. This construction uses the stored history and existing model evaluations within each integration interval.

Table 4: Some of the datasets, models, benchmarks, and codebases used.

Name	URL	License
SiT	https://github.com/willisma/SiT	MIT License
TurboDiffusion	https://github.com/thu-ml/TurboDiffusion	Apache License 2.0
Self-Forcing	https://github.com/guandeh17/Self-Forcing	Apache License 2.0
Wan2.1-T2V-1.3B	https://huggingface.co/Wan-AI/Wan2.1-T2V-1.3B	Apache License 2.0
VBench	https://github.com/Vchitect/VBench	Apache License 2.0
ImageNet-256×256	https://www.image-net.org/	\

E Additional experimental results

Evaluation protocol. For the SiT-XL/2 experiments on ImageNet 256×256, each configuration uses one fixed-seed set of 50,000 generated samples. We compute all metrics over that complete set, using the ImageNet reference statistics for FID and sFID. In paired runs, we hold the checkpoint, CFG scale, timestep schedule, evaluation budget, latent and class draws, and metric implementation fixed.

We conduct the component study at an overall scale of $1+\lambda = 1.70$, with $\beta = 0.9$ and $\kappa = 1.0$. Each variant is compared with its solver's native baseline evaluated within the same study. These baseline runs are separate from the main scale sweep, and each reported value comes from one generated sample set.

Parameter and component correspondence. We denote the reported overall scale by $1 + \lambda$ and its increment above one by λ. To describe the component controls, we use the integrated history displacement and coordinate noise variance from Algorithm 1:

$$b_i \mathbf{r}_i = \frac{\alpha_i h_i^2}{2\ell_i} \mathbf{d}_i \quad (i \geq 1), \qquad \nu_i^2 = h_i^2 (1 - \alpha_i) \hat{c}_i. \tag{41}$$

Here β controls the moving average, ℓ_i is its effective lag, and κ determines the OU factor b_i. The algorithm's coefficient rules determine the integrated weight α_i and perturbation scale ν_i at each step. The full transition also depends on the history estimator, OU initial condition, and noise rule.

For the control without history drift, we set $\mathbf{r}_i = \mathbf{0}$ while retaining the velocity memory and noise-response calibration. For the noise-free control, we omit the Gaussian perturbation and skip noise-response calibration, while retaining the history correction with the same coefficient α_i. For the base solver, we disable both correction branches. Setting $\lambda = 0$ alone leaves the Gaussian term active. Effective-lag normalization and noise-response calibration were not evaluated separately in these component comparisons.

Feature modulation within the history drift. In the weighted configuration, we modulate individual entries of the history residual before OU integration. With the elementwise weights represented by a diagonal matrix $\mathbf{W}_i \in \mathbb{R}^{D \times D}$, the drift becomes

$$\mathbf{r}_i = \frac{\lambda h_i}{2\ell_i} \mathbf{W}_i \mathbf{d}_i \quad (i \geq 1), \qquad \mathbf{r}_0 = \mathbf{0}. \tag{42}$$

The weights act on the residual inside the drift branch, before integration by the local OU process. The uniform-feature-weight control sets $\mathbf{W}_i = \mathbf{I}$, recovering the drift in Eq. (26). The residual $\mathbf{d}_i$ in the EMA and noise-response probe is unchanged; $\hat{c}_i$ determines the Gaussian perturbation

scale through Eq. (25). Algorithm 1 and Theorem 3.2 use this identity-weighted formulation. The component results in Table 8 include the uniform-weight control within the history branch.

Fixed-scale solver comparison. At the common reported scale $1 + \lambda = 1.50$ ($\lambda = 0.50$), SADrift lowers FID with all three solvers at 5, 10, and 20 NFE (Table 5). The main-text results in Table 1 provide the full Euler and Heun2 scale sweeps, together with the 100-NFE baseline runs.

Table 5: Few-step sampling on ImageNet 256×256 with frozen SiT-XL/2. SADrift results use the reported scale $1 + \lambda = 1.50$ at every budget. DPM-Solver++ (2M) uses the SiT-specific adapter. Relative reductions use the full reported FID precision in Tables 1 and 6, before the rounding shown here.

Base solver	NFE	Base FID ↓	+ SADrift FID ↓	Reduction
Euler	5	52.64	25.66	51.25%
Euler	10	8.35	4.23	49.34%
Euler	20	3.33	2.46	26.13%
Heun2	5	14.64	9.07	38.05%
Heun2	10	3.15	2.41	23.49%
Heun2	20	2.20	2.04	7.27%
DPM-Solver++ (2M)	5	12.31	7.38	40.05%
DPM-Solver++ (2M)	10	3.57	2.73	23.33%
DPM-Solver++ (2M)	20	2.46	2.15	12.72%

Correction strength with DPM-Solver++. Varying the correction strength for adapted DPM-Solver++ (2M), we find that every tested positive scale improves FID at each budget (Table 6). The best scale depends on the budget: $\lambda = 0.75$ at 5 NFE and $\lambda = 0.50$ at 10 and 20 NFE. We use $\lambda = 0.50$ throughout the fixed-scale comparison in Table 5.

Table 6: Correction-scale sweep with SiT-adapted DPM-Solver++ (2M). The columns report the overall scale $1 + \lambda$ and its increment λ. Gray rows use the base solver with both SADrift correction branches disabled. Results use CFG scale 1.5 and 50,000 samples. The lowest FID at each budget is bold.

NFE	$1 + \lambda$	λ	IS ↑	FID ↓	sFID ↓	Prec. ↑	Rec. ↑
5	1.00	0.00	152.4358	12.3051	9.6981	0.68344	0.5073
5	1.25	0.25	167.9261	9.0751	6.7397	0.72440	0.5034
5	1.50	0.50	178.6653	7.3772	5.9250	0.74560	0.5083
5	1.70	0.70	181.9859	6.7369	6.3181	0.75896	0.5107
5	1.75	0.75	182.3149	**6.6502**	6.5235	0.76048	0.5074
10	1.00	0.00	223.3429	3.5653	6.7718	0.79908	0.5657
10	1.25	0.25	231.7928	2.9571	5.9010	0.80936	0.5676
10	1.50	0.50	238.1568	**2.7336**	6.2863	0.81344	0.5695
10	1.70	0.70	240.9438	2.7912	7.1398	0.81462	0.5671
10	1.75	0.75	240.9095	2.8377	7.4114	0.81482	0.5670
20	1.00	0.00	243.5873	2.4592	5.5046	0.80190	0.5904
20	1.25	0.25	250.1663	2.2153	5.0111	0.80650	0.5895
20	1.50	0.50	255.1077	**2.1464**	5.0518	0.81296	0.5892
20	1.70	0.70	258.3707	2.1724	5.3969	0.81454	0.5899
20	1.75	0.75	258.9626	2.1931	5.5223	0.81452	0.5894

Memory decay. In the Euler experiment of Table 7, we vary only β to examine sensitivity to the length of velocity memory. FID spans 0.159 across the tested range, with the default coefficient within 0.050 of the best value; precision and recall each vary by less than 0.003. A larger β retains past predictions over more integration intervals in Eq. (18), but the small changes in these metrics indicate modest sensitivity over the tested range.

Table 7: Sensitivity to the memory coefficient β in a controlled Euler evaluation at 10 NFE. The correction scale and all other settings are fixed within the sweep. Each row uses 50,000 ImageNet samples with matched initial latent and class draws; the default coefficient is $\beta = 0.90$.

β	IS ↑	FID ↓	sFID ↓	Prec. ↑	Rec. ↑
0.50	188.419	6.939	9.539	0.7418	0.5763
0.70	188.580	6.902	9.411	0.7429	0.5788
0.90	188.921	6.830	9.159	0.7435	0.5777
0.95	189.110	6.805	9.064	0.7442	0.5777
0.99	189.253	6.780	8.979	0.7444	0.5791

Additional component controls. We examine feature modulation by setting $\mathbf{W}_i = \mathbf{I}$ in Eq. (42). This uniform-weight configuration achieves FIDs of 2.2229 for Heun2 and 4.7775 for Euler, compared with 3.2524 and 8.4539 for their native solvers (Table 8).

The full method achieves FIDs of 2.2039 for Heun2 and 4.7117 for Euler, improving on the uniform-weight configuration. Compared with the noise-free controls, whose FIDs are 2.2315 and 4.7189, respectively, Gaussian perturbations provide a small additional reduction in FID. The larger improvement over the native solvers is retained by the history correction without stochastic noise.

Table 8: Component ablations at 10 NFE. SADrift configurations use an overall scale of $1 + \lambda = 1.70$. Comparisons within each solver use matched evaluation settings and 50,000 samples.

Variant	IS ↑	FID ↓	sFID ↓	Prec. ↑	Rec. ↑
Heun2					
Base solver	231.0284	3.2524	6.5678	0.79086	0.5818
Without history drift	230.4675	3.2407	6.6205	0.78984	0.5855
Without stochastic noise	252.2355	2.2315	4.7507	0.80862	0.5977
Uniform feature weights	253.6233	2.2229	4.8788	0.80766	0.6013
Full (SADrift)	249.6061	2.2039	4.7874	0.80866	0.5940
Euler					
Base solver	182.8326	8.4539	8.8328	0.73176	0.5293
Without history drift	183.6585	8.5657	8.8273	0.72894	0.5275
Without stochastic noise	213.6536	4.7189	5.6873	0.78248	0.5398
Uniform feature weights	212.5640	4.7775	6.0290	0.78038	0.5373
Full (SADrift)	215.6647	4.7117	5.6980	0.78374	0.5382

Datasets and software resources. Table 4 lists project links and license information for the datasets, models, benchmarks, and codebases.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately summarize our main contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Limitations are explicitly addressed in a dedicated Limitations and Future Impact.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions and proofs are provided in the theory section.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We repeated the experiments and all settings are described in the experimental section.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and data will be publicly released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See the experimental section for detailed settings, hyperparameters, and evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We are repeated experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were conducted on a single NVIDIA H200 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed broader impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not involve high-risk models or datasets.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credited all existing assets and cited licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: No crowdsourcing or human subjects were involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing assistance, not as a core method component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.