STOCHASTIC SAMPLING WITH 1D BROWNIAN MOTION FOR FLOW MATCHING

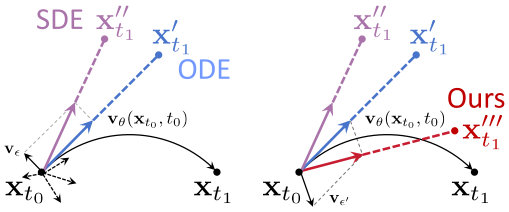
Anonymous authors
Paper under double-blind review

ABSTRACT

Flow Matching (FM) models are commonly sampled with deterministic ordinary differential equations. Stochastic sampling can improve trajectory diversity and sample quality, but conventional stochastic differential equation (SDE) samplers often require a large number of function evaluations (NFE). Building on the classical marginal-preservation identity for the Fokker–Planck equation, we develop an adapted low-rank step-process construction that converts low-dimensional Brownian motion into a practical, training-free sampler. The **S**implified **1**D-**N**oise-**G**uided **S**ampling (SONGS) framework, in its rank-one form, injects a scalar Brownian increment along a history-dependent estimate of the local discretization-error direction. Freezing this direction on each integration interval eliminates the Jacobian and divergence evaluations required by a general state-dependent diffusion and supports an improved small-noise stochastic Runge–Kutta implementation. Experiments on CIFAR-10, FFHQ, and ImageNet demonstrate that SONGS achieves superior sampling performance at low NFE.

1 INTRODUCTION

Flow matching (FM) learns a velocity field that transports a source distribution to the data distribution and supports efficient deterministic ordinary differential equation (ODE) sampling (Lipman et al., 2023; Liu et al., 2023). Stochastic sampling is attractive because it yields non-deterministic trajectories and can improve peak generation quality and downstream exploration (Albergo et al., 2025; Ma et al., 2024; Liu et al., 2025). In practice, however, standard stochastic differential equation (SDE) samplers often require many neural-network evaluations, and high-dimensional multiplicative noise complicates higher-order discretization through coupled iterated Itô integrals (Särkkä & Solin, 2019; Higham & Kloeden, 2021).



(a) ODE/SDE Euler (b) SONGS

Figure 1: At a large step size, isotropic SDE noise may perturb the sample in poorly oriented directions. SONGS instead injects scalar noise along an estimated local-error direction.

Conventional diffusion samplers use a d-dimensional Brownian motion for a d-dimensional state. The Fokker–Planck equation, however, constrains the diffusion covariance rather than the dimension of the driving Brownian motion. This classical observation permits low-rank diffusion matrices but does not by itself yield a scalable sampler: a general state-dependent direction introduces Jacobian, divergence, and score terms, while a history-dependent direction must remain nonanticipative. Our goal is to turn this identity into an implementable stochastic correction for FM.

We propose **SONGS**, a training-free sampling framework whose rank-one form injects a scalar Brownian increment along an estimated local-error direction. The direction is computed from completed solver stages, stored in a history buffer, and frozen over the next time interval. This adapted step-process construction preserves the target marginal path while making the direction's spatial derivatives vanish analytically. It also reduces the iterated stochastic integrals to the scalar-noise case. Figure 1 illustrates the resulting correction.

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

Our contributions are:

- We formulate a rank-m adapted step-process family that operationalizes the classical marginal-preservation identity with Brownian dimension independent of data dimension.
- We instantiate the framework with a general two-stage improved small-noise stochastic Runge–Kutta (SRK) method. At zero noise, the implementation reduces to a second-order Runge–Kutta (RK2) family, with Heun obtained when $\xi = 1$.
- We evaluate quality, diversity, and runtime across pixel- and latent-space models. Our single-step implementation performs strongly with a very small number of function evaluations (NFE), while comparisons at larger budgets identify a remaining gap to advanced multistep solvers.

2 Background and Numerical Challenge

Let p_0 be a source distribution and p_1 the data distribution. A flow-matching model learns a velocity field $\mathbf{v}_\theta$ for the ODE

$$d\mathbf{x}_t = \mathbf{v}(\mathbf{x}_t, t)\, dt, \qquad \mathbf{x}_0 \sim p_0, \tag{1}$$

whose solution has marginals p_t and terminal law p_1 (Lipman et al., 2023). Writing $\mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, the same marginal path is induced by

$$d\mathbf{x}_t = \left[\mathbf{v}(\mathbf{x}_t, t) + \frac{\omega_t^2}{2}\mathbf{s}(\mathbf{x}_t, t)\right] dt + \omega_t\, d\mathbf{w}_t, \tag{2}$$

where $\mathbf{w}_t$ is a d-dimensional Brownian motion. This follows from the continuity and Fokker–Planck equations and includes the probability-flow/reverse-SDE correspondence of diffusion models up to time reparameterization (Song et al., 2021b); details are in Appendix A.1. For reference, the standard forward diffusion SDE and its probability-flow ODE are denoted

$$d\mathbf{y}_t = f(t)\mathbf{y}_t dt + g(t) d\mathbf{w}_t, \tag{3}$$

$$d\mathbf{y}_t = \left[f(t)\mathbf{y}_t - \tfrac{1}{2} g(t)^2 \nabla_{\mathbf{y}} \log q_t(\mathbf{y}_t)\right] dt. \tag{4}$$

Euler–Maruyama discretizes Eq. (2) as

$$\hat{\mathbf{x}}_{t_{k+1}} = \hat{\mathbf{x}}_{t_k} + \left[\mathbf{v}(\hat{\mathbf{x}}_{t_k}, t_k) + \tfrac{1}{2}\omega_{t_k}^2 \mathbf{s}(\hat{\mathbf{x}}_{t_k}, t_k)\right] h_k + \omega_{t_k}\sqrt{h_k}\, \boldsymbol{\epsilon}_k. \tag{5}$$

Higher-order simulation becomes difficult when the diffusion is state-dependent. To make this dependence explicit, consider

$$d\mathbf{x}_t = \tilde{\mathbf{v}}(\mathbf{x}_t, t) dt + \mathbf{V}(\mathbf{x}_t, t) d\mathbf{w}_t. \tag{6}$$

Definition 2.1 (Strong convergence order)**.** A discretization has strong order p if $\mathbb{E}\|\mathbf{x}_T - \hat{\mathbf{x}}_T\| \le Ch^p$ under the standard regularity conditions.

Lemma 2.2. *For one Euler–Maruyama step of Eq.* (6)*, the local error admits an Itô–Taylor expansion containing drift integrals, mixed drift–diffusion integrals, and iterated diffusion integrals. In particular, the latter include*

$$\sum_j \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathcal{A}_{2,j}(\mathbf{V})_s\, dw_j(s)\, d\mathbf{w}(t), \tag{7}$$

which couple Brownian components and are costly to approximate when the noise dimension is large.

The full expansion, operators, and proof are given in Appendices A.2 and B.2. This motivates reducing the Brownian dimension while retaining a useful state-aware stochastic correction.

Why the noise dimension matters. For general noncommutative d-dimensional noise, higher-order strong methods may require $O(d^2)$ coupled iterated integrals. Additive or commutative noise can simplify these terms; diagonality alone does not determine the cost. With a scalar driver, repeated Itô integrals instead reduce to functions of a single Wiener increment. For example,

$$\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} dW(s)\, dW(t) = \tfrac{1}{2}\big((\Delta W_k)^2 - h_k\big). \tag{8}$$

This simplification concerns the dimension of the random driver, not the dimension of the generated state: a scalar increment can still produce a d-dimensional update when multiplied by a direction in $\mathbb{R}^d$. The remaining challenge is to choose that direction without violating adaptedness or introducing costly spatial derivatives.

3 METHOD

3.1 MARGINAL-PRESERVING LOW-RANK NOISE

The following proposition is a direct low-rank specialization of the classical Fokker–Planck marginal-preservation identity. We use it as a starting point rather than claiming the identity itself as the methodological novelty.

Proposition 3.1. *Let W be scalar Brownian motion, ω_t a bounded deterministic schedule, and $\mathbf{v}^e(\mathbf{x}, t)$ a differentiable direction field. Under Assumption B.1, the SDE*

$$d\mathbf{x}_t = \left[\mathbf{v}(\mathbf{x}_t, t) + \omega_t^2 \mathbf{g}_t \mathbf{v}^e(\mathbf{x}_t, t)\right] dt + \omega_t \mathbf{v}^e(\mathbf{x}_t, t) dW_t \tag{9}$$

has the same marginals p_t as Eq. (1)*, where*

$$\mathbf{g}_t = \tfrac{1}{2}\left[\nabla_{\mathbf{x}} \mathbf{v}^e + \left(\nabla_{\mathbf{x}} \cdot \mathbf{v}^e + (\mathbf{v}^e)^\top \mathbf{s}\right) I_d\right]. \tag{10}$$

For an affine Gaussian path $\mathbf{x}_t \overset{d}{=} \alpha_t \mathbf{x}_1 + \sigma_t \mathbf{x}_0$, the score is obtained from the FM velocity by

$$\mathbf{s}(\mathbf{x}, t) = \frac{\alpha_t \mathbf{v}(\mathbf{x}, t) - \dot{\alpha}_t \mathbf{x}}{\dot{\alpha}_t \sigma_t^2 - \alpha_t \dot{\sigma}_t \sigma_t}. \tag{11}$$

The proof and endpoint conditions are in Appendix B.3. Equation (9) reduces the Brownian dimension, but a neural direction field still requires expensive Jacobian and divergence terms. Our main construction removes this obstacle.

To see the source of the correction, let $\mathbf{a}(\mathbf{x}, t) = \omega_t \mathbf{v}^e(\mathbf{x}, t)$. The drift correction that cancels the Fokker–Planck diffusion term is

$$\tfrac{1}{2} \nabla_{\mathbf{x}} \cdot \left(\mathbf{a}\mathbf{a}^\top\right) + \tfrac{1}{2} \mathbf{a}\mathbf{a}^\top \mathbf{s}. \tag{12}$$

Substituting $\mathbf{a} = \omega_t \mathbf{v}^e$ yields exactly $\omega_t^2 \mathbf{g}_t \mathbf{v}^e$ in Eq. (9). Thus the proposition is an identity at the level of marginal evolution. It does not assert that every direction is numerically useful; the practical question is how to obtain a cheap, nonanticipative direction that targets finite-step error.

3.2 ADAPTED STEP-PROCESS CONSTRUCTION

Definition 3.2. A process $\mathbf{S}(t)$ is a step process on a partition $\{t_k\}_{k=0}^K$ if $\mathbf{S}(t) = \mathbf{S}_k$ for $t \in [t_k, t_{k+1})$, where $\mathbf{S}_k$ is $\mathcal{F}(t_k)$-measurable and square-integrable.

Remark 3.3. Each interval uses a direction computed only from completed steps. It is then frozen until the next endpoint, so it is nonanticipative and has zero spatial Jacobian and divergence on that interval.

Proposition 3.4. *For m adapted step processes $\{\mathbf{S}_r\}_{r=1}^m$, define $\tilde{\mathbf{S}}_r(t) = \omega_r(t)\mathbf{S}_r(t)$ and*

$$d\mathbf{x}_t = \left[\mathbf{v}(\mathbf{x}_t, t) + \tfrac{1}{2} \sum_{r=1}^m \tilde{\mathbf{S}}_r \tilde{\mathbf{S}}_r^\top \mathbf{s}(\mathbf{x}_t, t)\right] dt + \sum_{r=1}^m \tilde{\mathbf{S}}_r \, dW_r. \tag{13}$$

Under the regularity conditions in Appendix B.1, Eq. (13) *has a unique solution with the same marginal path as Eq.* (1)*.*

This follows by applying Proposition 3.1 interval by interval. Because $\mathbf{S}_{r,k}$ is fixed with respect to the current state on $[t_k, t_{k+1})$, the Jacobian and divergence terms vanish analytically, leaving only the projected score term in Eq. (13). The complete proof is in Appendix B.5. The cases $m = d$ with standard basis directions and $m = 1$ recover conventional isotropic noise and the rank-one construction, respectively.

This construction separates two roles that are coupled in a conventional SDE. The integer m controls the stochastic rank, while the vectors $\mathbf{S}_{r,k} \in \mathbb{R}^d$ control where the random correction acts. Choosing $m \ll d$ reduces random-number generation and the number of cross-component Itô integrals, yet the update remains full-dimensional. The rank-one case is especially attractive for few-step sampling because it requires only one scalar Brownian driver per sample.

Algorithm 1 SONGS–improved SRK with a general two-stage parameter

Require: $\mathbf{x}_0$, schedule $\{t_k\}_{k=0}^{K}$, model $\mathbf{v}_\theta$, scales $\{\omega_k, \xi_k\}$ with $0 < \xi_k \le 1$, $\varepsilon_{\mathrm{dir}} > 0$

$\hat{\mathbf{x}}_{t_0} \leftarrow \mathbf{x}_0$; $\mathbf{v}^e_{t_0} \leftarrow 0$
for $k = 0$ to $K-1$ **do**
  $h_k \leftarrow t_{k+1} - t_k$; sample independent $\epsilon_k, \eta_k \sim \mathcal{N}(0,1)$
  $\Delta W_k \leftarrow \sqrt{h_k}\epsilon_k$; $\bar{J}_k \leftarrow \sqrt{h_k}(\epsilon_k + \eta_k/\sqrt{3})/2$
  $\mathbf{g}_k \leftarrow \sqrt{2}\,\omega_k \mathbf{v}^e_{t_k}$ **frozen over both stages**
  $\mathbf{v}_1 \leftarrow \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k)$; $\mathbf{s}_1 \leftarrow \textsc{ScoreFromVelocity}(\mathbf{v}_1, \hat{\mathbf{x}}_{t_k}, t_k)$
  $\tilde{\mathbf{v}}_1 \leftarrow \mathbf{v}_1 + \omega_k^2 \mathbf{v}^e_{t_k} (\mathbf{v}^e_{t_k})^\top \mathbf{s}_1$
  $\mathbf{z}_k \leftarrow \hat{\mathbf{x}}_{t_k} + \xi_k h_k \tilde{\mathbf{v}}_1 + 2\xi_k \mathbf{g}_k \bar{J}_k$
  $\mathbf{v}_2 \leftarrow \mathbf{v}_\theta(\mathbf{z}_k, t_k + \xi_k h_k)$; $\mathbf{s}_2 \leftarrow \textsc{ScoreFromVelocity}(\mathbf{v}_2, \mathbf{z}_k, t_k + \xi_k h_k)$
  $\tilde{\mathbf{v}}_2 \leftarrow \mathbf{v}_2 + \omega_k^2 \mathbf{v}^e_{t_k} (\mathbf{v}^e_{t_k})^\top \mathbf{s}_2$
  $\tilde{\mathbf{v}} \leftarrow (1 - 1/(2\xi_k))\tilde{\mathbf{v}}_1 + \tilde{\mathbf{v}}_2/(2\xi_k)$
  $\hat{\mathbf{x}}_{t_{k+1}} \leftarrow \hat{\mathbf{x}}_{t_k} + h_k \tilde{\mathbf{v}} + \mathbf{g}_k \Delta W_k$
  $E_k \leftarrow h_k(\tilde{\mathbf{v}}_2 - \tilde{\mathbf{v}}_1)/(2\xi_k)$; $\mathbf{v}^e_{t_{k+1}} \leftarrow E_k / \max(\|E_k\|_2, \varepsilon_{\mathrm{dir}})$
end for
Return: $\hat{\mathbf{x}}_{t_K}$

3.3 SONGS WITH IMPROVED STOCHASTIC RUNGE–KUTTA

SONGS denotes the low-rank sampling framework in Eq. (13), not a particular numerical solver. Our main implementation, ***SONGS–improved SRK***, uses the two-stage small-noise SRK family of Buckwar et al. (2010) with a free stage parameter ξ_k. Setting $\xi_k = 1$ gives the improved Heun scheme (RIe4WM); setting $\xi_k = 1/2$ gives the improved midpoint scheme (RIe3WM). A second implementation, ***SONGS–PL1W1***, uses the tableau of Rößler (2006) and is described separately in Appendix A.3.2.

For the main implementation, set $m = 1$ and freeze both ω_k and $\mathbf{v}^e_{t_k}$ on interval k. The direction is computed from a completed step and used only on the next interval. Comparing the improved SRK update with Euler–Maruyama from the same state and Brownian increment gives

$$E_k = \hat{\mathbf{x}}_{t_{k+1}} - \mathbf{x}^{\mathrm{EM}}_{t_{k+1}} = \frac{h_k}{2\xi_k}(\tilde{\mathbf{v}}_2 - \tilde{\mathbf{v}}_1), \qquad \mathbf{v}^e_{t_{k+1}} = \frac{E_k}{\max(\|E_k\|_2, \varepsilon_{\mathrm{dir}})}. \tag{14}$$

The positive constant $\varepsilon_{\mathrm{dir}}$ prevents division by zero. We initialize $\mathbf{v}^e_{t_0} = 0$, so the first interval is deterministic two-stage RK, coinciding with Heun when $\xi_0 = 1$. Score evaluation is skipped whenever $\omega_k = 0$. The new direction is $\mathcal{F}(t_{k+1})$-measurable and cannot affect the completed interval.

The mixed integral $J_k = \int_{t_k}^{t_{k+1}} (W_t - W_{t_k})\, dt$ is correlated with ΔW_k, and $\bar{J}_k = J_k/h_k$. The factor $2\xi_k$ in the predictor follows from the small-noise SRK order condition for the mixed time–Brownian integral. Replacing this mixed-integral perturbation by ΔW_k instead gives the frozen-diffusion PL1W1 update. Each implementation uses two velocity evaluations per full step; improved SRK draws two independent scalar Gaussians, whereas PL1W1 draws one.

Appendix A.3.1 derives the coefficients and states the scope of the original small-noise error bounds. In particular, the $O(h^2 + \epsilon^2 h)$ root-mean-square bound applies to additive noise under the source's smoothness and stability assumptions; it is not an unconditional convergence theorem for history-adaptive SONGS. Numerical convergence to an SDE and preservation of the target marginal path are separate questions. Our finite-NFE experiments test the combined effect of discretization, learned fields, and the stochastic correction.

4 EXPERIMENTS

4.1 EXPERIMENTAL PROTOCOL

We evaluate training-free sampling on unconditional CIFAR-10 32×32, unconditional FFHQ 64×64, conditional ImageNet 64×64, and SiT-XL/2 on ImageNet 256×256 (Krizhevsky, 2009; Karras et al., 2019; Russakovsky et al., 2015; Ma et al., 2024). Unless stated otherwise, we compute Fréchet inception distance (FID) from 50K generated samples. The main comparisons use matched model-evaluation budgets within each dataset. We distinguish default and log signal-to-noise ratio

Table 1: FID (↓) for training-free samplers at matched NFE. Results use unconditional CIFAR-10 and FFHQ, and conditional ImageNet. SONGS is most effective in the very-low-NFE regime.

Method	CIFAR-10		FFHQ		ImageNet	
	3	5	3	5	3	5
ODE samplers						
Heun (Karras et al., 2022)	468.28	319.80	466.77	344.78	249.40	89.63
DDIM (Song et al., 2021a)	93.36	49.66	215.70	74.68	82.96	43.81
DPM-Solver-2 (Lu et al., 2022)	153.60	43.27	78.21	43.93	140.20	42.41
DPM-Solver++(3M) (Lu et al., 2025)	110.00	24.97	86.45	22.51	91.52	25.49
UniPC (Zhao et al., 2023)	109.60	23.98	86.43	21.40	91.38	24.36
iPNDM (Zhang & Chen, 2023)	47.98	13.59	45.98	17.17	58.53	18.99
SDE samplers						
Improved SDE (Karras et al., 2022)	468.28	358.47	466.76	344.77	437.45	248.39
SA-Solver (default) (Xue et al., 2023)	305.22	264.04	193.90	164.02	245.47	193.97
SA-Solver (log-SNR) (Xue et al., 2023)	155.31	23.89	126.07	20.03	104.49	30.47
SONGS (ours)	**15.00**	**6.09**	**34.21**	**15.35**	**29.52**	**14.26**

Table 2: CIFAR-10 FID (↓) across NFE.

Method	3	5	7	9
DDIM	93.36	49.66	27.93	18.43
Heun	468.28	319.80	99.42	37.80
DPM-Solver-2	153.60	43.27	16.69	8.65
DPM-Solver++(3M)	110.00	24.97	6.74	3.42
iPNDM	47.98	13.59	5.08	**3.17**
SONGS	**15.00**	**6.09**	**4.43**	3.83

Table 3: Runtime at matched NFE on one RTX 6000 Ada.

Data	Method	s/sample
CIFAR-10	Heun2	0.0266
CIFAR-10	SONGS	0.0282
CIFAR-10	Euler–Maruyama	0.0283
ImageNet 256	Heun2	0.4352
ImageNet 256	SONGS	0.4411
ImageNet 256	Euler–Maruyama	0.4588

(log-SNR) schedules where shown; the literature peak-quality table in the appendix also includes different checkpoints and budgets. We compare against representative ODE solvers (DDIM, Heun, DPM-Solver-2, DPM-Solver++(3M), UniPC, and iPNDM) and SDE solvers (Euler–Maruyama, Improved SDE, and SA-Solver).

For pixel-space experiments, NFE counts neural-network evaluations directly. For SiT with classifier-free guidance (CFG), we report CFG-NFE and use guidance scale 1.5; an ODE step with two guided model calls therefore has a different step count from a one-stage SDE update at the same CFG-NFE. We compare methods at matched reported CFG-NFE rather than matched step count. For latent-space evaluation, we also report spatial FID (sFID) and inception score (IS). Additional quantitative results and schedule ablations are provided in Appendix E.

4.2 Very-Low-NFE Generation

Table 1 shows that SONGS consistently improves FID at 3–5 NFE. On CIFAR-10 at 5 NFE, for example, SONGS obtains 6.09 FID, compared with 13.59 for the strongest listed ODE baseline and 23.89 for the strongest listed SDE baseline. The gains are largest when the integration interval is long, which is the regime in which both deterministic truncation error and learned-field error are most consequential.

4.3 Behavior Across Sampling Budgets

Table 2 compares our single-step improved-SRK implementation of SONGS with advanced multistep ODE samplers as NFE increases. Our implementation achieves the lowest FID among the compared methods at 3–7 NFE, but at 9 NFE its FID of 3.83 is higher than the 3.42 of DPM-Solver++(3M) and the 3.17 of iPNDM. These results reveal a remaining performance gap for the current single-step implementation at the larger tested budget; they do not establish a limitation of the SONGS framework. Developing efficient SONGS implementations based on stochastic multistep methods is a promising direction for future work.

SONGS adds 6.0% wall-clock time over Heun2 on CIFAR-10 and 1.4% on ImageNet 256, while remaining comparable to Euler–Maruyama. The implementation constructs the direction from

Table 4: SiT-XL/2 on ImageNet 256×256, evaluated with 50K samples, guidance scale 1.5, and matched CFG-NFE. Lower FID/sFID and higher IS are better.

Solver	Type	FID ↓			sFID ↓			IS ↑		
		10	14	18	10	14	18	10	14	18
Heun2	ODE	14.84	5.66	3.64	22.50	10.86	7.48	143.80	205.19	226.64
Euler–Maruyama	SDE	56.60	23.13	12.78	68.63	36.97	24.29	51.45	117.84	163.47
SONGS	SDE	**4.23**	**2.88**	**2.46**	**6.64**	**5.31**	**4.87**	**212.33**	**231.32**	**239.73**

Table 5: CIFAR-10 quality and diversity at matched NFE; higher is better.

Method (NFE)	Prec.	Recall	Density	Coverage
SONGS (3)	**.5747**	**.5215**	**.7150**	**.5833**
DPM++(3M) (3)	.5394	.1687	.4822	.1201
SONGS (5)	**.6189**	**.5787**	**.8395**	**.7243**
DPM++(3M) (5)	.5664	.4375	.6330	.5114
SONGS (7)	**.6371**	**.5961**	**.9147**	**.7740**
DPM++(3M) (7)	.6329	.5655	.8884	.7438

Table 6: Mean 2-Wasserstein distance (↓) over five seeds.

Dataset	Euler	Heun2	SONGS
8Gaussians	0.5645	0.5709	**0.5631**
Moons	0.2649	0.2607	**0.2601**
S-curve	0.3019	0.3064	**0.2992**

quantities already computed in the two solver stages and requires no additional Jacobian–vector product (JVP).

4.4 Latent-Space Generation and Diversity

Table 4 shows that the improvement transfers from pixel-space diffusion models to a latent flow model. To test whether stochastic correction improves distributional support rather than FID alone, Table 5 compares SONGS with DPM-Solver++(3M) using precision, recall, density, and coverage. The largest gains occur in recall and coverage at 3–5 NFE, supporting improved exploration rather than merely sharper samples. The gap narrows at 7 NFE, consistent with the FID crossover in Table 2.

4.5 Marginal Recovery in Two Dimensions

Following the protocol of Tong et al. (2024), we sample three two-dimensional target distributions from a standard Gaussian source. Every method uses the same trained FM model, target samples, and 100 model evaluations; Table 6 reports the mean over five seeds. SONGS recovers all three targets with the lowest mean 2-Wasserstein distance. The differences are small, so these results support marginal recovery but do not establish a statistically significant advantage over the deterministic baselines.

4.6 Qualitative Results and Ablations

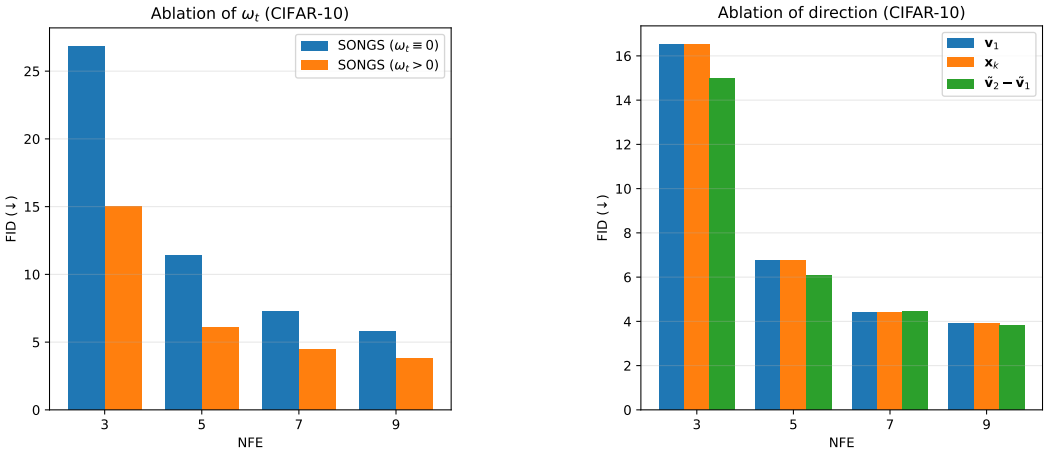
Figure 2 compares samples at the same model-evaluation budget. At 3 and 5 NFE, SONGS produces more coherent object structure than DPM-Solver++(3M), matching the quantitative trend in Table 1. Additional FFHQ, AFHQv2, and ImageNet samples are provided in Appendix D.

Figure 3 isolates the stochastic coefficient and direction estimator. Setting $\omega = 0$ recovers the deterministic RK backbone used in the ablation (Heun when $\xi = 1$); a small nonzero coefficient improves FID throughout the low-NFE range. The two-stage discrepancy in Eq. (14) is the most reliable direction overall, although the alternatives are close at 7–9 NFE. Complete coefficient sweeps and full-simulation comparisons are in Appendix E.



DPM-Solver++(3M), 3 NFE | SONGS, 3 NFE | DPM-Solver++(3M), 5 NFE | SONGS, 5 NFE

Figure 2: Qualitative comparison on CIFAR-10 under matched NFE. All panels use the same pretrained model.



(a) $\omega = 0$ vs. $\omega > 0$.

(b) Direction choice.

Figure 3: Ablations on CIFAR-10.

5 Analysis and Interpretation

5.1 Marginal Preservation Versus Numerical Accuracy

The marginal-preservation result and the finite-step algorithm make different claims. Proposition 3.4 is a continuous-time statement: if the velocity and score are exact and the adapted step process satisfies the stated regularity conditions, then Eq. (13) has the same one-time marginals as the FM ODE. The proposition does not imply that every discretization, direction, or stochastic schedule has the same finite-step error.

For a fixed interval and a rank-one direction $\mathbf{S}_k$, the diffusion covariance is $\omega_t^2 \mathbf{S}_k \mathbf{S}_k^\top$. Because $\mathbf{S}_k$ is constant with respect to the current state, the Fokker–Planck correction reduces to

$$\tfrac{1}{2}\omega_t^2 \mathbf{S}_k \mathbf{S}_k^\top \mathbf{s}(\mathbf{x}, t). \tag{15}$$

Adding Eq. (15) to the drift exactly cancels the diffusion contribution to the marginal evolution. This cancellation explains why the Brownian dimension can be smaller than the state dimension. It also explains why adaptedness is essential: if the direction used information from the current interval's future Brownian increment, the interval-wise Fokker–Planck argument would no longer apply. Algorithm 1 therefore uses the discrepancy computed at step k only from step $k+1$ onward.

The implementation introduces two approximations. First, it replaces the exact velocity and score with learned estimates. Second, it discretizes the resulting SDE with a finite number of stages. The source SRK error bounds concern discretization under regularity assumptions; their scope for frozen, history-adaptive coefficients is discussed in Appendix A.3.1. Our empirical claims concern the combined error under a fixed, pretrained generative model.

Table 7: Per-sample, per-step computational structure. "Extra model eval." is measured relative to the corresponding two-stage backbone.

Sampler	Brownian dim.	Gaussian draws	Coupled Itô terms	Extra model eval.
Full SDE	d	d	potentially $O(d^2)$	method dependent
General rank-one SDE	1	method dependent	none across dimensions	JVP/divergence
Improved SRK (Alg. 1)	1	2	none across dimensions	0

5.2 Why an Adaptive Stochastic Correction Can Help

The direction estimator in Eq. (14) is motivated by embedded error estimation. Both stages start from the same state, and their discrepancy identifies a direction in which the local approximation changes most strongly. Injecting noise along this direction is more targeted than perturbing every ambient coordinate independently. The estimator is inexpensive because both stage velocities are already required by the two-stage RK backbone.

The role of a nonzero stochastic coefficient can be illustrated by separating model error from discretization error. Let the learned quantities be $\hat{\mathbf{v}} = \mathbf{v} + \mathbf{e}_v$ and $\hat{\mathbf{s}} = \mathbf{s} + \mathbf{e}_s$, and let $\|\mathbf{S}_k\|_2 = 1$. The drift perturbation projected onto the active direction is

$$\mathbf{S}_k^\top \mathbf{e}_{\text{drift}} = \mathbf{S}_k^\top \mathbf{e}_v + \tfrac{1}{2}\omega_t^2 \mathbf{S}_k^\top \mathbf{e}_s. \tag{16}$$

A projected error diagnostic for $\omega_t > 0$ is

$$\mathcal{E}_\parallel(\omega) := \tfrac{1}{2}\int_0^1 \omega_t^{-2}\mathbb{E}_{p_t}\left[\left(\mathbf{S}_k^\top \mathbf{e}_v + \tfrac{1}{2}\omega_t^2 \mathbf{S}_k^\top \mathbf{e}_s\right)^2\right] dt. \tag{17}$$

Equation (17) is a directional diagnostic, not a bound on full-distribution KL. A rank-one diffusion cannot control arbitrary drift errors orthogonal to its noise direction. It captures the observed bias–variance trade-off: an extremely small ω_t suppresses the correction, whereas an excessively large value amplifies score error and stochastic discretization error. The coefficient sweep in Appendix E shows that the result is sensitive to this scale and therefore requires tuning. The reported improvement over $\omega = 0$ in Figure 3 provides the direct empirical test.

5.3 Cost and Intended Regime

Table 7 separates model-evaluation cost from stochastic-integration overhead. A full-dimensional state-dependent SDE may require d Gaussian variables per sample and coupled iterated integrals; evaluating a general learned direction may additionally require Jacobian or divergence estimates. Algorithm 1 uses two scalar Gaussian draws for one Brownian driver and no extra model evaluation or Jacobian–vector product beyond its two-stage backbone. Its tensor operations remain $O(d)$ because the scalar draw is multiplied by a d-dimensional direction.

The single-step improved-SRK implementation is particularly effective at very low NFE. Table 2 shows that advanced multistep ODE solvers outperform this implementation at 9 NFE. This empirical gap motivates combining the SONGS framework with stochastic multistep methods; whether such implementations improve performance at larger budgets remains to be evaluated.

The current realization also has clear failure modes. A poorly estimated direction can place noise along an irrelevant axis; an overly aggressive ω_t can magnify score error; and the rank-one update cannot simultaneously correct several independent error directions. These limitations motivate rank-m extensions, but increasing m increases random-number generation and tensor work; general state-dependent extensions may also introduce cross-component stochastic integrals. The appropriate rank should therefore depend on both the error geometry and the available sampling budget.

6 Related Work

Training-free integration. Fast diffusion and FM sampling is dominated by specialized ODE solvers and predictor–corrector or multistep schemes (Lu et al., 2022; Zhang & Chen, 2023; Zhao et al.,

2023; Zheng et al., 2023; Lu et al., 2025). Efficient Integrators develops conjugate and splitting integrators for diffusion dynamics (Pandey et al., 2024), whereas Bespoke Solvers optimizes model-specific ODE-solver parameters (Shaul et al., 2024). TADA augments deterministic dynamics with momentum and higher-dimensional initial noise (Chen et al., 2025). SONGS instead changes the rank and direction of the stochastic correction during integration. It requires no additional training and targets the very-low-NFE regime.

Stochastic sampling. EDM and SA-Solver improve stochastic sampling for diffusion models (Karras et al., 2022; Xue et al., 2023), while SEEDS constructs exponential SDE solvers with strong-convergence guarantees (Gonzalez et al., 2023). GLASS Flows constructs an inner flow model from a pretrained model for efficient stochastic transition sampling in reward alignment (Holderrieth et al., 2026). SONGS addresses a different question: it constructs a marginal-preserving rank-one stochastic correction for base FM generation using a history-adapted step-process direction. Its two implementations use existing improved small-noise SRK (Buckwar et al., 2010) and PL1W1 (Rößler, 2006) schemes. Our contribution lies in the low-rank FM construction and practical direction estimator, rather than in generic SRK theory.

7 Conclusion

We introduced the SONGS low-rank sampling framework with improved SRK and PL1W1 implementations. By freezing a history-dependent direction over each integration interval, SONGS preserves the target marginal path without evaluating the Jacobian or divergence of a neural direction field. Experiments show favorable quality, diversity, and runtime trade-offs in the very-low-NFE regime, while the current single-step implementation still trails advanced multistep ODE solvers at the larger budget tested in Table 2.

Limitations. The direction estimator and stochastic schedule require tuning. The two-dimensional distribution results have overlapping uncertainty, and our high-resolution analysis isolates noise-generation cost rather than end-to-end text-to-image latency. One promising direction is to develop efficient SONGS implementations using stochastic multistep methods. Learned rank-m directions and exponential integrators are additional directions for future work.

Reproducibility Statement

The main algorithm and its information flow are given in Algorithm 1. Appendix B.1 states the technical assumptions and Appendix B.3 provides the complete marginal-preservation proof. Appendix A.3 derives both numerical implementations and their zero-noise limits, while Appendix E reports additional quantitative results and schedule ablations.

AI Use Statement

Generative AI tools were used to assist with drafting and revising portions of the manuscript, language editing, document formatting, figure-label adjustments, organizing reviewer feedback and related literature, checking reference formatting, and suggesting manuscript structure. They were not used to generate experimental measurements or to independently establish mathematical claims. The authors reviewed all AI-assisted text, calculations, citations, and edits and take responsibility for the final content of the manuscript.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://jmlr.org/papers/v26/23-1605.html.

Evelyn Buckwar, Andreas Rößler, and Renate Winkler. Stochastic Runge–Kutta methods for Itô SODEs with small noise. *SIAM Journal on Scientific Computing*, 32(4):1789–1808, 2010. doi: 10.1137/090763275. URL https://doi.org/10.1137/090763275.

Tianrong Chen, Huangjie Zheng, David Berthelot, Jiatao Gu, Joshua Susskind, and Shuangfei Zhai. TADA: Improved diffusion sampling with training-free augmented dynamics. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/aa1d8cf866c4a684ef2e066dd200f8e4-Abstract-Conference.html.

Lawrence C Evans. *An introduction to stochastic differential equations*, volume 82. American Mathematical Society, 2013. URL https://www.ams.org/books/mbk/082/.

James Foster and Karen Habermann. Brownian bridge expansions for Lévy area approximations and particular values of the Riemann zeta function. *Combinatorics, Probability and Computing*, 32(3): 370–397, 2023.

Martin Gonzalez, Nelson Fernandez Pinto, Thuy Tran, Elies Gherbi, Hatem Hajri, and Nader Masmoudi. SEEDS: Exponential SDE solvers for fast high-quality sampling from diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d6f764aae383d9ff28a0f89f71defbd9-Abstract-Conference.html.

Desmond Higham and Peter Kloeden. *An Introduction to the Numerical Simulation of Stochastic Differential Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021. doi: 10.1137/1.9781611976434. URL https://doi.org/10.1137/1.9781611976434.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Peter Holderrieth, Uriel Singer, Tommi Jaakkola, Ricky T. Q. Chen, Yaron Lipman, and Brian Karrer. GLASS flows: Efficient inference for reward alignment of flow and diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ad6ac52e02e14e0f2ddc0d4587ab47f1-Abstract-Conference.html.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b578f2a52a0229873fefc2a4b06377fa-Abstract.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=2uAaGwlP_V.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025. ISSN 2731-5398. doi: 10.1007/s11633-025-1562-4. URL https://doi.org/10.1007/s11633-025-1562-4.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Jan Mrongowius and Andreas Rößler. On the approximation and simulation of iterated stochastic integrals and the corresponding Lévy areas in terms of a multidimensional Brownian motion. *Stochastic Analysis and Applications*, 40(3):397–425, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6 edition, 2003. doi: 10.1007/978-3-642-14394-6. URL https://doi.org/10.1007/978-3-642-14394-6.

Kushagra Pandey, Maja Rudolph, and Stephan Mandt. Efficient integrators for diffusion generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bfa20cc46a90fff34af616b738c10f9b-Abstract-Conference.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Andreas Rößler. Runge–Kutta methods for Itô stochastic differential equations with scalar noise. *BIT Numerical Mathematics*, 46(1):97–110, 03 2006. doi: 10.1007/s10543-005-0039-7. URL https://doi.org/10.1007/s10543-005-0039-7.

Neta Shaul, Juan Perez, Ricky T. Q. Chen, Ali Thabet, Albert Pumarola, and Yaron Lipman. Bespoke solvers for generative flow models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/35c1d69d23bb5dd6b9abcd68be005d5c-Abstract-Conference.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Fangyikang Wang, Hubery Yin, Lei Qian, Yinan Li, Shaobin Zhuang, Huminhao Zhu, Yilin Zhang, Yanlong Tang, Chao Zhang, Hanbin Zhao, Hui Qian, and Chen Li. Unleashing high-quality image generation in diffusion sampling using second-order Levenberg–Marquardt–Langevin. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10453–10464, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Wang_Unleashing_High-Quality_Image_Generation_in_Diffusion_Sampling_Using_Second-Order_Levenberg-Marquardt-Langevin_ICCV_2025_paper.pdf.

Magnus Wiktorsson. Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions. *The Annals of Applied Probability*, 11(2):470–487, 2001.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16782–16795. Curran Associates, Inc., 2022.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. SA-Solver: Stochastic Adams solver for fast sampling of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=f6a9XVFYIo.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Loek7hfb46P.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=hrkmlPhp1u.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. DPM-Solver-v3: Improved diffusion ODE solver with empirical model statistics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9fWKExmKa0.

A SUPPLEMENTARY THEORY

A.1 AFFINE GAUSSIAN CONDITIONAL PATH

Given the affine conditional flow

$$\psi_t(\mathbf{x}_0|\mathbf{x}_1) := \alpha_t \mathbf{x}_1 + \sigma_t \mathbf{x}_0 \tag{18}$$

where $\mathbf{x}_0 \sim p_0$ and $\alpha, \sigma : [0,1] \to [0,1]$ are differentiable with $\alpha_0 = 0 = \sigma_1$ and $\alpha_1 = 1 = \sigma_0$. Assume that $\psi_t(\mathbf{x}_0|\mathbf{x}_1)$ is the solution of

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}^e(\mathbf{x}_t, t|\mathbf{x}_1).$$

If p_0 is a d-dimensional standard Gaussian distribution, that is, $p_0 = \mathcal{N}(0, I)$, then the probability path

$$p_t(\cdot|\mathbf{x}_1) := \mathcal{N}(\alpha_t \mathbf{x}_1, \sigma_t^2 I), t \in (0,1)$$

generated by the above $\mathbf{v}^e(\cdot, t|\mathbf{x}_1)$ is called the (affine) Gaussian conditional path.

The score of the Gaussian conditional path is

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{x}_1) = -\frac{\mathbf{x} - \alpha_t \mathbf{x}_1}{\sigma_t^2}.$$

Then the score of the corresponding marginal probability path is

$$\begin{aligned}
\nabla \log p_t(\mathbf{x}) &= \int \frac{\nabla p_t(\mathbf{x}|\mathbf{x}_1) p_1(\mathbf{x}_1)}{p_t(\mathbf{x})} \, d\mathbf{x}_1 \\
&= \int \frac{\nabla \log p_t(\mathbf{x}|\mathbf{x}_1) p_t(\mathbf{x}|\mathbf{x}_1) p_1(\mathbf{x}_1)}{p_t(\mathbf{x})} \, d\mathbf{x}_1 \\
&= \mathbb{E}\left[\nabla \log p_t(\mathbf{x}_t|\mathbf{x}_1) \mid \mathbf{x}_t = \mathbf{x}\right] \\
&= -\mathbb{E}\left[\frac{1}{\sigma_t^2}(\mathbf{x}_t - \alpha_t \mathbf{x}_1) \mid \mathbf{x}_t = \mathbf{x}\right] \\
&= -\frac{1}{\sigma_t}\mathbb{E}\left[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}\right]
\end{aligned} \tag{19}$$

The literature on diffusion refers to $\mathbb{E}\left[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}\right]$ as *noise-prediction*, or ϵ*-prediction*. According to the form in (Liu et al., 2023),

$$\begin{aligned}
\mathbf{v}_t(\mathbf{x}) &= \mathbb{E}\left[\dot{\alpha}_t \mathbf{x}_1 + \dot{\sigma}_t \mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}\right] \\
&= \dot{\alpha}_t \mathbb{E}\left[\frac{\mathbf{x}_t - \sigma_t \mathbf{x}_0}{\alpha_t} \mid \mathbf{x}_t = \mathbf{x}\right] + \dot{\sigma}_t \mathbb{E}\left[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}\right] \\
&= \frac{\dot{\alpha}_t}{\alpha_t}\mathbf{x} + [\dot{\sigma}_t - \sigma_t \frac{\dot{\alpha}_t}{\alpha_t}]\mathbb{E}\left[\mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}\right]
\end{aligned}$$

Then

$$\mathbf{v}_t(\mathbf{x}) = \frac{d \log \alpha_t}{dt}\mathbf{x} - (\frac{1}{2}\frac{d\sigma_t^2}{dt} - \sigma_t^2 \frac{d \log \alpha_t}{dt})\mathbf{s}(\mathbf{x}, t). \tag{20}$$

where

$$\mathbf{s}(\mathbf{x}, t) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

Let $r_f : [0,1] \to [0,T]$ be a strictly monotonically decreasing function with $r_f(0) = T$ and $r_f(1) = 0$. Eq. (20) coincides with the probability flow ODE formulation of diffusion models with diffusion time $r = r_f(t)$ for some r_f. In diffusion models, the Gaussian conditional path is constructed by the forward process. For example, choosing the scheduler

$$\alpha_t = e^{-\frac{\beta_t}{2}}, \sigma_t = \sqrt{1 - e^{-\beta_t}}, \beta_0 \gg 0, \beta_1 = 0$$

defines a conditional probability path $p_t(\mathbf{x}|\mathbf{x}_1), t \in [0,1]$ the same as that of Variance Preserving (VP) diffusion SDE. However, such $p_t(\mathbf{x}|\mathbf{x}_1), t \in [0,1]$ cannot satisfy $p_0 = \mathcal{N}(0, I)$ exactly. In this case, flow matching (FM) is equivalent to diffusion modeling with different time and model output parameterization. Therefore, FM subsumes the ODE-based sampling of the classic diffusion models.

A.2 DIFFICULTIES IN HIGHER CONVERGENCE ORDER

We can expand some terms of (40) to increase the convergence order. For example, recall (39), $f = \mathcal{A}_{2,j}(\mathbf{V})$ and $t = s$ lead to

$$\begin{aligned}\mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_s, s) = \mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_{t_k}, t_k) \\ + \int_{t_k}^{s} \mathcal{A}_1\mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_\tau, \tau)\, d\tau \\ + \sum_{\ell=1}^{m} \int_{t_k}^{s} \mathcal{A}_{2,\ell}\mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_\tau, \tau)\, dw_\ell(\tau).\end{aligned}$$

Then

$$\begin{aligned}\mathbf{x}_{t_{k+1}} = \mathbf{x}_{t_k} + \tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k)\, h_k + \mathbf{V}(\mathbf{x}_{t_k}, t_k)\, \Delta\mathbf{w}_k \\ + \sum_j \mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_{t_k}, t_k)\chi_{k,j} + \mathbf{r}\end{aligned}$$

where $\mathbf{r}$ is the remainder term and

$$\chi_{k,j} = \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} d\mathbf{w}_j(\tau)\, d\mathbf{w}(s).$$

To make the notation explicit, one may write the full matrix of second-order iterated Itô integrals on $[t_k, t_{k+1}]$ as

$$I_{ij}^{(k)} := \int_{t_k}^{t_{k+1}} \int_{t_k}^{s} d\mathbf{w}_i(\tau)\, d\mathbf{w}_j(s), \qquad i, j = 1, \ldots, m.$$

The vector $\chi_{k,j}$ collects the entries $I_{j\ell}^{(k)}$ over $\ell = 1, \ldots, m$. In particular, the diagonal terms satisfy the exact identity

$$I_{ii}^{(k)} = \frac{1}{2}\big((\Delta\mathbf{w}_{k,i})^2 - h_k\big),$$

whereas the off-diagonal terms $I_{ij}^{(k)}$ for $i \neq j$ are Lévy-area-type quantities and generally do not admit a similarly simple closed form in terms of the Brownian increments alone (Wiktorsson, 2001; Mrongowius & Rößler, 2022; Foster & Habermann, 2023). Hence, once $m \geq 2$, a second-order simulation typically has to deal with a whole family of coupled cross terms $\{I_{ij}^{(k)}\}_{i,j=1}^{m}$, whose cardinality is $O(m^2)$ in general. This is precisely the source of the extra approximation burden absent in the scalar-noise case.

In practice, classical high-order SDE solvers therefore require a separate simulation or approximation procedure for these iterated integrals (Särkkä & Solin, 2019; Wiktorsson, 2001; Mrongowius & Rößler, 2022). Such procedures introduce an additional error on top of the discretization error of the numerical scheme itself. Roughly speaking, if one replaces $I_{ij}^{(k)}$ by approximations $\widehat{I}_{ij}^{(k)}$, then the local error contains an extra contribution proportional to the collection $\{I_{ij}^{(k)} - \widehat{I}_{ij}^{(k)}\}_{i,j}$; since there are $O(m^2)$ such terms, the aggregate mean-square error budget and the associated implementation cost both grow quickly with the Brownian dimension unless additional structure is exploited. This is one reason why directly applying generic higher-order multidimensional SDE solvers becomes unattractive in the high-dimensional generative setting.

Similarly, notice that by (39), for $j = 1, 2, \ldots, m$ the terms

$$\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathcal{A}_{2,j}(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\, d\mathbf{w}_j(s)\, dt$$

in Eq. (40) satisfy

$$\begin{aligned}&\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathcal{A}_{2,j}(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\, d\mathbf{w}_j(s)\, dt \\ &= \mathcal{A}_{2,j}(\tilde{\mathbf{v}})(\mathbf{x}_{t_k}, t_k) \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} d\mathbf{w}_j(s)\, dt + \mathbf{r}'\end{aligned}$$

where $\mathbf{r}'$ is a remainder term. The cross-term Itô integral

$$\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} d\mathbf{w}_j(s)\, dt = \int_{t_k}^{t_{k+1}} (\mathbf{w}_j(t) - \mathbf{w}_j(t_k))\, dt \tag{21}$$

is correlated with $\Delta\mathbf{w}_k$. For a full-dimensional driver, $m = d$ can be very large (up to millions), (21) for $j = 1, 2, \ldots, m$ are not easy to handle. In particular, for higher-order SDE simulation, one may even need to take into account the combinations of iterated and cross terms. Therefore, the design of high-order SDE algorithms is non-trivial.

A.3 Two numerical implementations of the SONGS framework

SONGS specifies the low-rank stochastic dynamics and the history-based direction; a numerical integrator specifies how to advance those dynamics. We distinguish *SONGS–improved SRK* (Algorithm 1) from *SONGS–PL1W1* (Algorithm 2). The former follows the improved small-noise methods of Buckwar et al. (2010); the latter follows the PL1W1 tableau in Rößler (2006, Fig. 4.7, p. 106). The numerical schemes and their classical order results are prior work. Below we summarize the relevant derivations in our notation and then specialize them to a frozen SONGS interval.

Throughout this subsection, write $\mathbf{u}_k = \mathbf{v}^e_{t_k}$ and define

$$\mathbf{g}_k = \sqrt{2}\,\omega_k \mathbf{u}_k, \qquad \mathbf{f}_k(\mathbf{x}, t) = \mathbf{v}_\theta(\mathbf{x}, t) + \omega_k^2 \mathbf{u}_k \mathbf{u}_k^\top \mathbf{s}_\theta(\mathbf{x}, t). \tag{22}$$

Both $\mathbf{u}_k$ and ω_k are fixed throughout $[t_k, t_{k+1}]$, including its endpoint stage; the next interval's coefficients are installed only after the update. When $\omega_k = 0$, we set the correction to zero directly and skip score evaluation, including in both algorithms. Conditional on the information available at t_k, the diffusion $\mathbf{g}_k$ is constant on this interval. This is an interval-wise numerical statement, not by itself a proof of unconditional target-marginal preservation.

A.3.1 Improved SRK: source derivation and SONGS specialization

Mixed time–Brownian integral. The improvement in Buckwar et al. (2010, Sec. 3.3, Eqs. (3.17)–(3.18)) is to include mixed time–Brownian integrals rather than only the terminal Brownian increment. For a scalar driver, let

$$\Delta W_k = W_{t_{k+1}} - W_{t_k}, \qquad J_k = \int_{t_k}^{t_{k+1}} (W_t - W_{t_k})\, dt = \int_{t_k}^{t_{k+1}} (t_{k+1} - s)\, dW_s. \tag{23}$$

Itô isometry gives

$$\mathbb{E}[(\Delta W_k)^2] = h_k, \qquad \mathbb{E}[J_k^2] = h_k^3/3, \qquad \mathbb{E}[\Delta W_k J_k] = h_k^2/2. \tag{24}$$

Since these integrals are jointly Gaussian, their joint law is sampled exactly by

$$\Delta W_k = \sqrt{h_k}\epsilon_k, \qquad \bar{J}_k := J_k/h_k = \frac{\sqrt{h_k}}{2}(\epsilon_k + \eta_k/\sqrt{3}), \qquad \epsilon_k, \eta_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{25}$$

This is the construction in Buckwar et al. (2010, Sec. 5.1, p. 1803). The auxiliary draw η_k represents the Brownian-bridge contribution to J_k; it does not introduce a second Brownian driver into the SDE.

Why the stage coefficient is 2. For fixed additive diffusion $\mathbf{g}$, the source's Eqs. (4.5)–(4.6) take the form

$$\mathbf{H}_i = \mathbf{x} + h\sum_j a_{ij}\mathbf{f}(\mathbf{H}_j, t + c_j h) + \sum_j b_{ij}\mathbf{g}\, J/h, \qquad \mathbf{x}^+ = \mathbf{x} + h\sum_i \beta_i \mathbf{f}(\mathbf{H}_i, t + c_i h) + \mathbf{g}\Delta W. \tag{26}$$

The relevant order conditions, extracted from the source's Eq. (3.24), are

$$\boldsymbol{\beta}^\top \mathbf{1} = 1, \qquad \boldsymbol{\beta}^\top A\mathbf{1} = \tfrac{1}{2}, \qquad \boldsymbol{\beta}^\top B\mathbf{1} = 1. \tag{27}$$

Indeed, the additive-noise Itô expansion contains

$$\mathbf{x}^+ = \mathbf{x} + h\mathbf{f} + \mathbf{g}\Delta W + \tfrac{1}{2}h^2(\partial_t \mathbf{f} + D\mathbf{f}\,\mathbf{f}) + D\mathbf{f}\,\mathbf{g}J + \text{higher-order terms}.$$

Expanding the stages in Eq. (26) gives coefficients $\boldsymbol{\beta}^\top A\mathbf{1}$ and $\boldsymbol{\beta}^\top B\mathbf{1}$ for the deterministic and mixed-integral terms, respectively. The latter must equal one to match $D\mathbf{f}\,\mathbf{g}J$.

The explicit two-stage family in Buckwar et al. (2010, Fig. 3.4, p. 1800) has $a_{21} = 1/(2\theta)$, $b_{21} = 1/\theta$, and $\boldsymbol{\beta} = (1-\theta, \theta)$. The choice $\theta = 1/2$ gives the improved Heun scheme RIe4WM:

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{28}$$

For completeness, the source's general multiplicative-noise tableau with $\zeta = 1$ also has $\widehat{A}_{21} = 1$, $\boldsymbol{\gamma} = (0, 1)$, and $\boldsymbol{\eta} = (1, -1)$. With constant additive diffusion, the extra diffusion-stage evaluations collapse to $\mathbf{g}\Delta W$ and the $\boldsymbol{\eta}$ contribution vanishes.

Substituting Eq. (22) yields the general form used in Algorithm 1:

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}_k(\hat{\mathbf{x}}_{t_k}, t_k), \\ \mathbf{H}_2 &= \hat{\mathbf{x}}_{t_k} + \xi h_k \mathbf{k}_1 + 2\xi \mathbf{g}_k \bar{J}_k, \\ \mathbf{k}_2 &= \mathbf{f}_k(\mathbf{H}_2, t_k + \xi h_k), \\ \hat{\mathbf{x}}_{t_{k+1}} &= \hat{\mathbf{x}}_{t_k} + h_k \left(1 - \frac{1}{2\xi}\right) \mathbf{k}_1 + \frac{h_k}{2\xi} \mathbf{k}_2 + \mathbf{g}_k \Delta W_k. \end{aligned} \tag{29}$$

In a general two-stage parameterization $a_{21} = \xi$, the corresponding coefficients must be $b_{21} = 2\xi$ and $\boldsymbol{\beta} = (1 - 1/(2\xi), 1/(2\xi))$. Thus the predictor perturbation is $2\xi\mathbf{g}_k\bar{J}_k$, not $\mathbf{g}_k\bar{J}_k$ for arbitrary ξ. The choice $\xi = 1$ gives the improved Heun scheme RIe4WM, while $\xi = 1/2$ gives the improved midpoint scheme RIe3WM.

Local error and global error: scope of the source result. Let the source SDE have diffusion $G = \epsilon\widehat{G}$, with $0 \le \epsilon \le 1$ denoting noise magnitude, and let h be the largest step. Here ϵ is a deterministic small parameter, distinct from the Gaussian draw ϵ_k. For sufficiently smooth coefficients with the required moment bounds and stability, Buckwar et al. (2010, Sec. 3.3) obtain root-mean-square error $O(h^2 + \epsilon^2 h^{1/2})$ for general multiplicative noise. The improved additive-noise estimate is sharper.

In the additive case, Eq. (4.2) of that paper bounds the local truncation remainder ρ by

$$\|\mathbb{E}\rho\| = O(h^3 + \epsilon^2 h^2), \qquad \|\rho\|_{L^2} = O(h^3 + \epsilon^2 h^2 + \epsilon h^{5/2}). \tag{30}$$

The method's remainder has the same bounds. The source's Theorem 2.1 converts local mean and L^2 error bounds c_1, c_2 into a global bound of the form $C(c_1/h + c_2/\sqrt{h})$, under its Lipschitz, growth, moment, and stability assumptions. Consequently, as stated in Sec. 4 (p. 1802),

$$\max_n \|\mathbf{X}(t_n) - \mathbf{Y}_n\|_{L^2} = O(h^2 + \epsilon^2 h + \epsilon h^2 + \epsilon^2 h^{3/2}) = O(h^2 + \epsilon^2 h). \tag{31}$$

This is a two-parameter error estimate. At fixed nonzero ϵ, it does not assert strong order two; if $\epsilon^2 = O(h)$ and the constants remain uniform, the bound scales as $O(h^2)$.

For SONGS, conditioning on past information makes $\mathbf{g}_k$ additive on each interval, which justifies the coefficient specialization above. A global bound for directions recomputed from numerical history additionally requires uniform conditional local-error estimates, moment bounds, and stability of the direction/history update as the grid changes. These requirements are not established by quoting the source theorem, nor do the numerical order conditions prove target-marginal preservation. We therefore report the original error results with their assumptions rather than claim an unconditional strong-order theorem for adaptive SONGS.

A.3.2 PL1W1: SOURCE TABLEAU AND THE SECOND IMPLEMENTATION

Rößler (2006, Fig. 4.7, p. 106) explicitly names the scheme *PL1W1*, with deterministic order $p_D = 2$ and stochastic *weak* order $p_S = 2$. The accompanying text attributes it to an extension of Platen's autonomous weak-order-two scheme to non-autonomous SDEs. This is the direct reference for the tableau used here; PL1W1 is distinct from the improved Heun scheme above.

In the source's notation, its coefficient arrays and weights are

$$A^{(0)} = B^{(0)} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad A^{(1)} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad B^{(1)} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$
$$\mathbf{c}^{(0)} = (0, 1, 0)^\top, \quad \mathbf{c}^{(1)} = (0, 1, 1)^\top,$$
$$\boldsymbol{\alpha} = (1/2, 1/2, 0), \qquad \boldsymbol{\gamma}^{(1)} = (1/2, 1/4, 1/4), \qquad \boldsymbol{\gamma}^{(2)} = (0, 1/2, -1/2). \tag{32}$$

For a scalar-noise Itô SDE with drift $\mathbf{f}$ and diffusion $\mathbf{b}$, let $\mathbf{f}_1 = \mathbf{f}(\mathbf{x}, t)$, $\mathbf{b}_1 = \mathbf{b}(\mathbf{x}, t)$, $\chi_1 = \Delta W \sim \mathcal{N}(0, h)$, and $\chi_2 = ((\Delta W)^2 - h)/2$. The nontrivial stages are

$$\mathbf{H}_2^{(0)} = \mathbf{x} + h\mathbf{f}_1 + \mathbf{b}_1 \Delta W, \qquad \mathbf{H}_{2,3}^{(1)} = \mathbf{x} + h\mathbf{f}_1 \pm \mathbf{b}_1 \sqrt{h}. \tag{33}$$

Writing $\mathbf{b}_2 = \mathbf{b}(\mathbf{H}_2^{(1)}, t + h)$ and $\mathbf{b}_3 = \mathbf{b}(\mathbf{H}_3^{(1)}, t + h)$, the update is

$$\begin{aligned} \mathbf{x}^+ = \mathbf{x} &+ \tfrac{h}{2}\left[\mathbf{f}_1 + \mathbf{f}(\mathbf{H}_2^{(0)}, t + h)\right] \\ &+ \tfrac{1}{4}(2\mathbf{b}_1 + \mathbf{b}_2 + \mathbf{b}_3)\Delta W + \tfrac{1}{2}(\mathbf{b}_2 - \mathbf{b}_3)\chi_2/\sqrt{h}. \end{aligned} \tag{34}$$

For the frozen SONGS diffusion, $\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3 = \mathbf{g}_k$. Hence the χ_2 term is zero and the diffusion weights sum to one. The resulting implementation is

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}_k(\hat{\mathbf{x}}_{t_k}, t_k), & \mathbf{H}_2 &= \hat{\mathbf{x}}_{t_k} + h_k \mathbf{k}_1 + \mathbf{g}_k \Delta W_k, \\ \mathbf{k}_2 &= \mathbf{f}_k(\mathbf{H}_2, t_{k+1}), & \hat{\mathbf{x}}_{t_{k+1}} &= \hat{\mathbf{x}}_{t_k} + \tfrac{1}{2} h_k (\mathbf{k}_1 + \mathbf{k}_2) + \mathbf{g}_k \Delta W_k. \end{aligned} \tag{35}$$

Thus only two drift evaluations and one scalar Gaussian draw are needed for this specialization, despite the three stochastic stages in the general tableau.

Algorithm 2 SONGS–PL1W1 (frozen-diffusion specialization of Rößler (2006))

Require: $\mathbf{x}_0$, schedule $\{t_k\}_{k=0}^K$, model $\mathbf{v}_\theta$, scales $\{\omega_k\}$, $\varepsilon_{\mathrm{dir}} > 0$
$\hat{\mathbf{x}}_{t_0} \leftarrow \mathbf{x}_0$; $\boldsymbol{\delta} \leftarrow 0$
for $k = 0$ to $K - 1$ **do**
  $h_k \leftarrow t_{k+1} - t_k$; sample $\Delta W_k \sim \mathcal{N}(0, h_k)$
  $\mathbf{v}_1 \leftarrow \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k)$
  $\mathbf{q} \leftarrow \boldsymbol{\delta}$ if $\|\boldsymbol{\delta}\|_2 > 0$, otherwise $\mathbf{q} \leftarrow \mathbf{v}_1$
  $\mathbf{u}_k \leftarrow \mathbf{q}/\max(\|\mathbf{q}\|_2, \varepsilon_{\mathrm{dir}})$; $\mathbf{g}_k \leftarrow \sqrt{2}\,\omega_k \mathbf{u}_k$
  $\mathbf{s}_1 \leftarrow \textsc{ScoreFromVelocity}(\mathbf{v}_1, \hat{\mathbf{x}}_{t_k}, t_k)$
  $\mathbf{k}_1 \leftarrow \mathbf{v}_1 + \omega_k^2 \mathbf{u}_k \mathbf{u}_k^\top \mathbf{s}_1$
  $\mathbf{H}_2 \leftarrow \hat{\mathbf{x}}_{t_k} + h_k \mathbf{k}_1 + \mathbf{g}_k \Delta W_k$
  $\mathbf{v}_2 \leftarrow \mathbf{v}_\theta(\mathbf{H}_2, t_{k+1})$; $\mathbf{s}_2 \leftarrow \textsc{ScoreFromVelocity}(\mathbf{v}_2, \mathbf{H}_2, t_{k+1})$
  $\mathbf{k}_2 \leftarrow \mathbf{v}_2 + \omega_k^2 \mathbf{u}_k \mathbf{u}_k^\top \mathbf{s}_2$
  $\hat{\mathbf{x}}_{t_{k+1}} \leftarrow \hat{\mathbf{x}}_{t_k} + h_k(\mathbf{k}_1 + \mathbf{k}_2)/2 + \mathbf{g}_k \Delta W_k$
  $\boldsymbol{\delta} \leftarrow \mathbf{k}_2 - \mathbf{k}_1$ **available only for the next interval**
end for
Return: $\hat{\mathbf{x}}_{t_K}$

Algorithm 2 retains the velocity-based initialization and zero-discrepancy fallback of this second implementation. Algorithm 1 instead initializes the direction to zero and uses the normalized completed-step discrepancy. These are separate direction conventions; comparisons intended to isolate the integrator must match initialization, normalization, schedules, and noise scales.

Meaning of weak order two. Under the source's regularity and moment assumptions, weak order two bounds $|\mathbb{E}\varphi(\mathbf{X}_T) - \mathbb{E}\varphi(\mathbf{Y}_N)|$ by $C_\varphi h^2$ for sufficiently smooth test functions φ. It does not assert strong order two, nor does it automatically extend to a history-adaptive direction with a zero-discrepancy fallback. The PL1W1 tableau's classical order and the statistical behavior of an adaptive SONGS implementation must therefore be distinguished.

A.3.3 ZERO-NOISE LIMITS AND DIFFERENCES BETWEEN THE IMPLEMENTATIONS

If $\omega_k = 0$ for all steps, Eq. (22) gives $\mathbf{g}_k = 0$ and $\mathbf{f}_k = \mathbf{v}_\theta$. The improved-SRK implementation in Eq. (29) reduces to the explicit two-stage RK family

$$\hat{\mathbf{x}}_{t_{k+1}} = \hat{\mathbf{x}}_{t_k} + h_k \left(1 - \frac{1}{2\xi_k}\right) \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k) + \frac{h_k}{2\xi_k} \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k} + \xi_k h_k \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k), t_k + \xi_k h_k)\,. \quad (36)$$

This family is second-order under the usual smoothness assumptions for admissible ξ_k; it coincides with classical explicit Heun only when $\xi_k = 1$ and with the midpoint RK2 method when $\xi_k = 1/2$. By contrast, the PL1W1 specialization in Eq. (35) reduces to Heun:

$$\hat{\mathbf{x}}_{t_{k+1}} = \hat{\mathbf{x}}_{t_k} + \frac{h_k}{2} \left[\mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k) + \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k} + h_k \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_k}, t_k), t_{k+1})\right]. \quad (37)$$

The direction estimator becomes irrelevant to either zero-noise update because both its drift correction and its noise contribution are multiplied by zero.

At nonzero noise the schemes differ: for improved SRK, $2\xi_k \bar{J}_k = \xi_k \Delta W_k + \xi_k \sqrt{h_k/3}\eta_k$ enters the predictor; PL1W1 uses only ΔW_k. Consequently, for the frozen coefficient, the predictor-noise covariance is $\frac{4}{3}\xi_k^2 h_k \mathbf{g}_k \mathbf{g}_k^\top$ in improved SRK and $h_k \mathbf{g}_k \mathbf{g}_k^\top$ in PL1W1. Both final explicit noise terms are $\mathbf{g}_k \Delta W_k$, but the second drift evaluation generally differs. They are two implementations of the same framework, not two names for the same algorithm.

B PROOFS

B.1 STANDING ASSUMPTIONS

Assumption B.1.

(A1) The ODE in Eq. (1) is well posed on $[0,1]$, induces a marginal probability path p_t, and for each $t \in (0,1)$ the law p_t admits a differentiable density with score $\mathbf{s}(\mathbf{x},t) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.

(A2) The direction field $\mathbf{v}^e : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is Borel measurable in $(\mathbf{x},t)$, C^1 in $\mathbf{x}$ for $t \in (0,1)$, and, when evaluated along an adapted continuous state process $(\mathbf{x}_t)$, the process $\mathbf{v}^e(\mathbf{x}_t,t)$ is progressively measurable and square integrable on every truncated interval $[\delta, 1-\delta] \subset (0,1)$.

(A3) The stochastic scale ω_t is deterministic and bounded on $[0,1]$; in the effective drift, only the regularized product $\omega_t \mathbf{v}^e$ and the derivatives entering $\mathbf{g}_t$ are used.

(A4) The affine Gaussian conversion used in the proposition assumes $\mathbf{x}_0 \sim \mathcal{N}(0, I_d)$, $\alpha, \sigma \in C^1([0,1])$, and $(\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t)\sigma_t \neq 0$ for all $t \in (0,1)$.

(A5) Endpoint singularities are handled only through interior-time evaluations: the standalone score is never required exactly at $t = 0$ or $t = 1$, and any use near an endpoint is accompanied either by a vanishing coefficient in the drift or by truncating the last small time step before that endpoint.

Remark B.2. Assumption B.1(A5) is imposed on the effective ODE/SDE drift used for sampling, not on a bare score formula written in isolation. In FM and stochastic-interpolant formulations, the score typically appears multiplied by a time-dependent coefficient that vanishes near the endpoint, so the product entering the drift can remain bounded even when the standalone score is singular near $t = 0$ or $t = 1$ (Albergo et al., 2025). Practically, one may further damp this coefficient and stop the discrete sampler one small step before the singular endpoint; this is also consistent with the finite-grid implementation viewpoint used in SiT (Ma et al., 2024).

B.2 ERROR FORMULA

To make our proofs easier to understand, we rewrite the celebrated Itô's formula (Evans, 2013; Särkkä & Solin, 2019) by our notation in the main text. Consider the stochastic differential equation (SDE):

$$d\mathbf{x}_t = \tilde{\mathbf{v}} dt + \mathbf{V} d\mathbf{w} \tag{38}$$

where $\mathbf{w}$ is an m-dimensional standard Brownian motion and $\mathbf{V}(\mathbf{x}_t,t) = (\mathbf{V}_{ij}) \in \mathbb{R}^{d\times m}$.

Theorem B.3 (Itô's formula in d-dimensions)**.** *Consider the above SDE* (38)*. Let $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}$ be continuous, with continuous partial derivatives $\frac{\partial f}{\partial t}$, $\frac{\partial f}{\partial \mathbf{x}_i}$, $\frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}$, $(i,j = 1,\dots,d)$. Then*

$$d(f(\mathbf{x}_t,t)) = \frac{\partial f}{\partial t} dt + \sum_{i=1}^{d} \frac{\partial f}{\partial \mathbf{x}_i} d\mathbf{x}_t(i) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j} \sum_{k=1}^{m} \mathbf{V}_{ik}\mathbf{V}_{jk} dt,$$

where $\mathbf{x}_t(i)$ is the ith-element of $\mathbf{x}_t$ and the argument of the partial derivatives of f is $(\mathbf{x}_t,t)$.

Theorem B.3 is the standard Itô formula, for which we refer to (Evans, 2013; Øksendal, 2003).

Remark that Eq. (6) in Lemma 2.2 is the specific case of Eq. (38). Here we directly prove this simple lemma for Eq. (38).

Proof of Lemma 2.2. Note that

$$d\mathbf{x}_t(i) = \tilde{\mathbf{v}}_i dt + (\mathbf{V} d\mathbf{w})_i$$

and the Gram matrix

$$[\mathbf{V}\mathbf{V}^\top]_{ij} := \sum_k \mathbf{V}_{ik}\mathbf{V}_{jk}.$$

By integrating both sides of Itô formulation from t_k to t in Theorem B.3, we obtain that for the differentiable function f defined in $R^d \times [0,1]$,

$$f(\mathbf{x}_t,t) = f(\mathbf{x}_{t_k},t_k) + \int_{t_k}^{t} \mathcal{A}_1(f)(\mathbf{x}_s,s) ds + \sum_j \int_{t_k}^{t} \mathcal{A}_{2,j}(f)(\mathbf{x}_s,s)\, dw_j(s) \tag{39}$$

with linear operators

$$\mathcal{A}_1(\bullet) = \frac{\partial(\bullet)}{\partial t} + \sum_i \frac{\partial(\bullet)}{\partial \mathbf{x}_i}\,\tilde{\mathbf{v}}_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2(\bullet)}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\,[\mathbf{V}\mathbf{V}^\top]_{ij},$$

and

$$\mathcal{A}_{2,j}(\bullet) = \sum_i \frac{\partial(\bullet)}{\partial \mathbf{x}_i}\,\mathbf{V}_{ij}, \qquad \text{for } j = 1, 2, \ldots, m.$$

Let $\Delta\mathbf{w}_k = \mathbf{w}(t_{k+1}) - \mathbf{w}(t_k)$. We integrate both sides of the equation (38) over the interval $[t_k, t_{k+1}]$:

$$\begin{aligned}
\mathbf{x}_{t_{k+1}} &= \mathbf{x}_{t_k} + \int_{t_k}^{t_{k+1}} \tilde{\mathbf{v}}(\mathbf{x}_s, s)\,ds + \int_{t_k}^{t_{k+1}} \mathbf{V}(\mathbf{x}_s, s)\,d\mathbf{w} \\
&\overset{(i)}{=} \mathbf{x}_{t_k} + \int_{t_k}^{t_{k+1}} \Big[\tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k) + \int_{t_k}^{t} \mathcal{A}_1(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\,ds \\
&\qquad + \sum_j \int_{t_k}^{t} \mathcal{A}_{2,j}(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\,dw_j(s)\Big]dt \\
&\quad + \int_{t_k}^{t_{k+1}} \Big[\mathbf{V}(\mathbf{x}_{t_k}, t_k) + \int_{t_k}^{t} \mathcal{A}_1(\mathbf{V})(\mathbf{x}_s, s)\,ds \\
&\qquad + \sum_j \int_{t_k}^{t} \mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_s, s)\,dw_j(s)\Big]d\mathbf{w} \\
&\overset{(ii)}{=} \mathbf{x}_{t_k} + \tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k)\,h_k + \int_{t_k}^{t_{k+1}}\int_{t_k}^{t} \mathcal{A}_1(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\,ds\,dt \\
&\quad + \sum_j \int_{t_k}^{t_{k+1}}\int_{t_k}^{t} \mathcal{A}_{2,j}(\tilde{\mathbf{v}})(\mathbf{x}_s, s)\,d\mathbf{w}_j\,dt + \mathbf{V}(\mathbf{x}_{t_k}, t_k)\,\Delta\mathbf{w}_k \\
&\quad + \int_{t_k}^{t_{k+1}}\int_{t_k}^{t} \mathcal{A}_1(\mathbf{V})(\mathbf{x}_s, s)\,ds\,d\mathbf{w} + \sum_j \int_{t_k}^{t_{k+1}}\int_{t_k}^{t} \mathcal{A}_{2,j}(\mathbf{V})(\mathbf{x}_s, s)\,d\mathbf{w}_j\,d\mathbf{w}.
\end{aligned} \tag{40}$$

where in (i) we directly use the form of (39) with $f = \tilde{\mathbf{v}}$ and $f = \mathbf{V}$, and in (ii) we apply

$$\int_{t_k}^{t_{k+1}} \tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k)\,dt = \tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k)h_k, \quad \int_{t_k}^{t_{k+1}} \mathbf{V}(\mathbf{x}_{t_k}, t_k)\,d\mathbf{w} = \mathbf{V}(\mathbf{x}_{t_k}, t_k)\Delta\mathbf{w}_k. \tag{41}$$

The state $\mathbf{x}_{t_{k+1}}$ is the exact solution of (38) starting from $\mathbf{x}_{t_k}$. Recall the Euler-Maruyama method. We obtain

$$\hat{\mathbf{x}}_{t_{k+1}} = \mathbf{x}_{t_k} + \tilde{\mathbf{v}}(\mathbf{x}_{t_k}, t_k)\,h_k + \mathbf{V}(\mathbf{x}_{t_k}, t_k)\,\Delta\mathbf{w}_k$$

Then, by considering

$$\mathbf{r} = \mathbf{x}_{t_{k+1}} - \hat{\mathbf{x}}_{t_{k+1}},$$

the proof is completed. □

B.3 SDE WITH ONE-DIMENSIONAL BROWNIAN MOTION

Theorem B.4 (Fokker–Planck–Kolmogorov (FPK) equation)**.** *The probability density $p_t(\mathbf{x})$ of the solution of the SDE* (38) *solves the partial differential equation (PDE):*

$$\frac{\partial p_t(\mathbf{x})}{\partial t} = -\sum_i \frac{\partial}{\partial \mathbf{x}_i}\left[\tilde{\mathbf{v}}_i(\mathbf{x}, t) p_t(\mathbf{x})\right] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left\{\left[\mathbf{V}(\mathbf{x}, t)\mathbf{V}^T(\mathbf{x}, t)\right]_{ij} p_t(\mathbf{x})\right\}. \tag{42}$$

Remark B.5. This Partial Differential Equation (PDE) is often called the Fokker–Planck equation in physics, and the forward Kolmogorov equation in stochastics (Särkkä & Solin, 2019). With the probability density $p_{t_0}(\mathbf{x})$ of the initial state $\mathbf{x}_{t_0}$, the PDE is an initial value problem with the given initial condition $p(x, t_0)$ at time $t = t_0$.

Proof. Let $f \in C_c^2(\mathbb{R}^d)$ and write $A(\mathbf{x},t) = \mathbf{V}(\mathbf{x},t)\mathbf{V}(\mathbf{x},t)^\top$. Itô's formula and expectation give

$$\frac{d}{dt}\int f p_t \, d\mathbf{x} = \int \left(\sum_i \tilde{v}_i \partial_i f + \tfrac{1}{2} \sum_{i,j} A_{ij} \partial_i \partial_j f \right) p_t \, d\mathbf{x}.$$

Integrating the drift term once and the diffusion term twice by parts yields

$$\int f \left[-\sum_i \partial_i(\tilde{v}_i p_t) + \tfrac{1}{2} \sum_{i,j} \partial_i \partial_j (A_{ij} p_t) \right] d\mathbf{x}.$$

Compact support removes boundary terms. Since the identity holds for every such f, it gives Eq. (42) in the weak sense and, under sufficient regularity, pointwise. In particular, both derivatives act on the full product $A_{ij}p_t$. □

Proof of Proposition 3.1. The proof uses Assumption B.1(A1)–(A5). More precisely: (A1) guarantees the ODE-induced marginal path and score on interior times; (A2)–(A3) ensure that $\mathbf{a}(\mathbf{x},t) = \omega_t \mathbf{v}^e(\mathbf{x},t)$ is an admissible diffusion coefficient and that the spatial derivatives in $\mathbf{g}_t$ are well defined; the explicit affine Gaussian score-velocity identity additionally uses (A4); and the endpoint caveat is handled through Assumption B.1(A5), so the derivation below is understood on $t \in (0,1)$. By Theorem B.4, the probability path $p_t, t \in [0,1]$ of the ODE (1) satisfies

$$\frac{dp_t(\mathbf{x})}{dt} = -\nabla \cdot (p_t \mathbf{v})(\mathbf{x}).$$

Set

$$\mathbf{a}(\mathbf{x},t) := \omega_t \mathbf{v}^e(\mathbf{x},t).$$

Since ω_t depends only on t, we have

$$\begin{aligned}\frac{1}{2}\nabla \cdot \left[\mathbf{a}(\mathbf{x},t)\mathbf{a}(\mathbf{x},t)^T\right] &= \frac{\omega_t^2}{2}\nabla \cdot \left[\mathbf{v}^e(\mathbf{x},t)\mathbf{v}^e(\mathbf{x},t)^T\right] \\ &= \frac{\omega_t^2}{2}\left[\nabla_{\mathbf{x}}\mathbf{v}^e(\mathbf{x},t) + (\nabla_{\mathbf{x}} \cdot \mathbf{v}^e(\mathbf{x},t))\, I_d\right]\mathbf{v}^e(\mathbf{x},t),\end{aligned}$$

and

$$\frac{1}{2}\mathbf{a}(\mathbf{x},t)\mathbf{a}(\mathbf{x},t)^T\mathbf{s}(\mathbf{x},t) = \frac{\omega_t^2}{2}\left(\mathbf{v}^e(\mathbf{x},t)^\top \mathbf{s}(\mathbf{x},t)\right)\mathbf{v}^e(\mathbf{x},t).$$

Therefore,

$$\frac{1}{2}\nabla \cdot \left[\mathbf{a}(\mathbf{x},t)\mathbf{a}(\mathbf{x},t)^T\right] + \frac{1}{2}\mathbf{a}(\mathbf{x},t)\mathbf{a}(\mathbf{x},t)^T\mathbf{s}(\mathbf{x},t) = \omega_t^2 \mathbf{g}_t \mathbf{v}^e(\mathbf{x},t),$$

where $\mathbf{g}_t$ is the matrix defined in Proposition 3.1. Substituting this identity into the FPK equation gives

$$\begin{aligned}-\nabla \cdot (p_t \mathbf{v})(\mathbf{x}) = &-\nabla \cdot \left(p_t \left[\mathbf{v}(\mathbf{x},t) + \omega_t^2 \mathbf{g}_t \mathbf{v}^e(\mathbf{x},t)\right]\right) \\ &+ \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left[\mathbf{a}_i(\mathbf{x},t)\mathbf{a}_j(\mathbf{x},t)p_t(\mathbf{x})\right].\end{aligned}$$

Applying Theorem B.4 with diffusion coefficient $\mathbf{a} = \omega_t \mathbf{v}^e$ yields (9). The velocity-score identity then follows from the same affine Gaussian calculation as in Section A.1. □

For convenience in implementations such as get_score_from_velocity$(\cdot)$, it is useful to summarize several equivalent parameterizations of the same affine-Gaussian velocity field in the unified form

$$\mathbf{v}_t(\mathbf{x}) = \beta_t \mathbf{x} + \gamma_t f_t(\mathbf{x}).$$

The corresponding coefficients are listed in Table 8.

f_t	β_t	γ_t
velocity	0	1
$\mathbf{x}_1$-prediction	$\frac{\dot{\sigma}_t}{\sigma_t}$	$\frac{\dot{\alpha}_t\sigma_t - \dot{\sigma}_t\alpha_t}{\sigma_t}$
$\mathbf{x}_0$-prediction	$\frac{\dot{\alpha}_t}{\alpha_t}$	$\frac{\dot{\sigma}_t\alpha_t - \dot{\alpha}_t\sigma_t}{\alpha_t}$
score	$\frac{\dot{\alpha}_t}{\alpha_t}$	$-\frac{\dot{\sigma}_t\sigma_t\alpha_t - \dot{\alpha}_t\sigma_t^2}{\alpha_t}$

Table 8: Velocities of common model parameterizations under the affine Gaussian path. Each row gives the coefficients in $\mathbf{v}_t(\mathbf{x}) = \beta_t\mathbf{x} + \gamma_t f_t(\mathbf{x})$.

B.4 Velocity-Score Relation for General Stochastic Interpolants

For completeness, we record a conceptual extension of the velocity-score relation beyond the affine Gaussian path used in the main text. Following Albergo et al. (2025), let

$$I : [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$$

be a general interpolant with $I(0, \mathbf{z}_0, \mathbf{z}_1) = \mathbf{z}_0$ and $I(1, \mathbf{z}_0, \mathbf{z}_1) = \mathbf{z}_1$, where $(\mathbf{z}_0, \mathbf{z}_1) \sim \nu$ has marginals p_0 and p_1. If

$$\mathbf{z}_t := I(t, \mathbf{z}_0, \mathbf{z}_1),$$

then the standard FM target velocity is

$$\mathbf{v}(\mathbf{x}, t) = \mathbb{E}[\partial_t I(t, \mathbf{z}_0, \mathbf{z}_1) \mid \mathbf{z}_t = \mathbf{x}]\,.$$

To regularize the probability path in non-affine or non-Gaussian cases, one may consider the stochastic interpolant

$$\mathbf{x}_t = I(t, \mathbf{z}_0, \mathbf{z}_1) + \gamma_t \mathbf{z}_2, \qquad \mathbf{z}_2 \sim \mathcal{N}(0, I_d),$$

with $\gamma_0 = \gamma_1 = 0$ and $\gamma_t > 0$ for $t \in (0,1)$. Its associated velocity field is

$$\bar{\mathbf{v}}(\mathbf{x}, t) = \mathbb{E}[\partial_t I(t, \mathbf{z}_0, \mathbf{z}_1) \mid \mathbf{x}_t = \mathbf{x}] + \dot{\gamma}_t \mathbb{E}[\mathbf{z}_2 \mid \mathbf{x}_t = \mathbf{x}]\,.$$

Meanwhile, the score of the marginal law of $\mathbf{x}_t$ satisfies

$$\mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -\gamma_t^{-1} \mathbb{E}[\mathbf{z}_2 \mid \mathbf{x}_t = \mathbf{x}]\,.$$

Substituting this identity into the previous display gives

$$\bar{\mathbf{v}}(\mathbf{x}, t) = \mathbb{E}[\partial_t I(t, \mathbf{z}_0, \mathbf{z}_1) \mid \mathbf{x}_t = \mathbf{x}] - \gamma_t \dot{\gamma}_t \mathbf{s}(\mathbf{x}, t).$$

Where $\gamma_t \dot{\gamma}_t \neq 0$, we may equivalently denote

$$\mathbf{v}^\sharp(\mathbf{x}, t) := \mathbb{E}[\partial_t I(t, \mathbf{z}_0, \mathbf{z}_1) \mid \mathbf{x}_t = \mathbf{x}]\,,$$

then

$$\mathbf{s}(\mathbf{x}, t) = \frac{\mathbf{v}^\sharp(\mathbf{x}, t) - \bar{\mathbf{v}}(\mathbf{x}, t)}{\gamma_t \dot{\gamma}_t}.$$

This identity requires the additional conditional field $\mathbf{v}^\sharp$ and is valid only where $\gamma_t \dot{\gamma}_t \neq 0$; it does not recover the score from $\bar{\mathbf{v}}$ alone for a general interpolant. The affine Gaussian setting is emphasized in the main text because the corresponding conditional terms reduce to the closed-form relation used in Proposition 3.1.

This also clarifies the endpoint singularity issue discussed in the main text. Since $\gamma_0 = \gamma_1 = 0$, the standalone score representation

$$\mathbf{s}(\mathbf{x}, t) = -\gamma_t^{-1} \mathbb{E}[\mathbf{z}_2 \mid \mathbf{x}_t = \mathbf{x}]$$

may diverge as $t \to 0$ or $t \to 1$. However, the quantity that enters the induced dynamics is not the naked score itself, but the combination

$$\gamma_t \dot{\gamma}_t\, \mathbf{s}(\mathbf{x}, t) = \mathbf{v}^\sharp(\mathbf{x}, t) - \bar{\mathbf{v}}(\mathbf{x}, t),$$

which can remain finite at the endpoints provided the two velocity fields stay finite there. In plain terms, the score may blow up only because we divided by a vanishing interpolation scale. Once multiplied back by the same endpoint factor, the apparent singularity cancels. This is why the equivalent ODE/SDE sampling dynamics can still be well behaved even though the score formula itself looks singular near the endpoints. In our affine Gaussian setting, the same phenomenon appears in Proposition 3.1: the score formula may be singular at the endpoint, while the FM velocity remains finite and the practical sampling schedule avoids evaluating the singular endpoint directly.

Note that $\nabla \cdot (\mathbf{x}\mathbf{x}^\top) = (d+1)\mathbf{x}$. If $\mathbf{v}^e(\mathbf{x}_t, t) = \mathbf{x}_t$, then SDE (9) can be converted into

$$d\mathbf{x}_t = \mathbf{v}(\mathbf{x}_t, t)\, dt + \frac{\omega_t^2}{2}[(d+1) + \mathbf{x}_t^\top \mathbf{s}(\mathbf{x}_t, t)]\mathbf{x}_t\, dt + \omega_t \mathbf{x}_t\, dW.$$

B.5 Existence and Uniqueness

We denote by $\mathbb{L}^2(0,T)$ the space of all real-valued, progressively measurable stochastic processes $G(\cdot)$ such that

$$\mathbb{E}\left(\int_0^T G^2\, dt\right) < \infty.$$

Theorem B.6 (Existence and Uniqueness Theorem (Evans, 2013; Øksendal, 2003))**.** *Suppose that $\tilde{\mathbf{v}} : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ and $\mathbf{V} : \mathbb{R}^d \times [0,T] \to M^{d\times m}$ are continuous and satisfy the following conditions:*

(a) $|\tilde{\mathbf{v}}(\mathbf{x},t) - \tilde{\mathbf{v}}(\hat{\mathbf{x}},t)| + |\mathbf{V}(\mathbf{x},t) - \mathbf{V}(\hat{\mathbf{x}},t)| \le L|x - \hat{\mathbf{x}}|$ *for all* $0 \le t \le T,\ x, \hat{\mathbf{x}} \in \mathbb{R}^d$,

(b) $|\tilde{\mathbf{v}}(\mathbf{x},t)| + |\mathbf{V}(\mathbf{x},t)| \le L(1+|x|)$ *for all* $0 \le t \le T,\ x \in \mathbb{R}^d$

for some constant L.

(c) Let $\mathbf{x}_0$ be any $\mathbb{R}^d$-valued random variable such that $\mathbb{E}(|\mathbf{x}_0|^2) < \infty$.

(d) $\mathbf{x}_0$ is independent of the future Brownian increments $\mathbf{w}(t) - \mathbf{w}(0),\ t > 0$, where $\mathbf{w}(\cdot)$ is a given m-dimensional Brownian motion.

Then there exists a unique continuous adapted solution with $\mathbb{E}\int_0^T \|\mathbf{x}_t\|^2 dt < \infty$ of the SDE (38) *with given $\mathbf{x}_0$*

Proof of Proposition 3.4. This proof uses Assumption B.1(A1)–(A2) together with the adapted step-process structure in Definition 3.2. In particular, each frozen value $\mathbf{S}_{r,k}$ is $\mathcal{F}(t_k)$-measurable, hence nonanticipative on $[t_k, t_{k+1})$; multiplying by deterministic $\omega_r(t)$ preserves progressive measurability on each sub-interval. The detailed endpoint discussion is postponed to Remark B.2. Let $\tilde{\mathbf{S}}(t) := (\tilde{\mathbf{S}}_r(t))$ be a $\mathbb{R}^{d\times m}$-value stochastic process. Assume that $p_t, t \in [t_0, t_K]$ is the marginal probability path of Eq. (1). Given the $\mathbf{x}_{t_0} \sim p_{t_0}$. First, we consider the case in the interval $[t_0, t_1]$. Then $\tilde{\mathbf{S}}(t) = (\omega_r(t)\mathbf{S}_r(t_0))$ since $\tilde{\mathbf{S}}_r(t) = \omega_r(t)\mathbf{S}_r(t)$ and $\mathbf{S}_r(t)$ is an step process defined on I_K^h. According to Theorem B.6 with $\mathbf{w} := (W_r)$, there exists a unique solution $\mathbf{x}_t$ of Eq. (13) by the conditions on $\tilde{\mathbf{v}}$. By Theorem B.4, the marginal probability density path $p_t, t \in [t_0, t_1]$ satisfies:

$$\begin{aligned}
\frac{\partial p_t(\mathbf{x})}{\partial t} &= -\sum_i \frac{\partial}{\partial \mathbf{x}_i}[\tilde{\mathbf{v}}_i(\mathbf{x},t)p_t(\mathbf{x})] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left\{\left[\tilde{\mathbf{S}}(t)(\tilde{\mathbf{S}}(t))^\top\right]_{ij} p_t(\mathbf{x})\right\} \\
&\overset{(i)}{=} -\sum_i \frac{\partial}{\partial \mathbf{x}_i}\left[\tilde{\mathbf{v}}_i p_t(\mathbf{x}) - \frac{1}{2}[\tilde{\mathbf{S}}(t)\tilde{\mathbf{S}}(t)^\top]_i \nabla p_t(\mathbf{x})\right] \\
&= -\sum_i \frac{\partial}{\partial \mathbf{x}_i}\left[\mathbf{v}_i p_t(\mathbf{x}) + \frac{1}{2}[\tilde{\mathbf{S}}(t)\tilde{\mathbf{S}}(t)^\top]_i \mathbf{s}(\mathbf{x},t)p_t(\mathbf{x}) - \frac{1}{2}[\tilde{\mathbf{S}}(t)\tilde{\mathbf{S}}(t)^\top]_i \nabla p_t(\mathbf{x})\right] \\
&\overset{(ii)}{=} -\nabla \cdot [\mathbf{v}(\mathbf{x},t)p_t(\mathbf{x})]
\end{aligned}$$

where in (i) we use

$$\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left\{\left[\tilde{\mathbf{S}}(t)(\tilde{\mathbf{S}}(t))^\top\right]_{ij} p_t(\mathbf{x})\right\} = \frac{1}{2}\sum_i \frac{\partial}{\partial \mathbf{x}_i}\left[\sum_j \left[\tilde{\mathbf{S}}(t)(\tilde{\mathbf{S}}(t))^\top\right]_{ij} \frac{\partial}{\partial \mathbf{x}_j} p_t(\mathbf{x})\right]$$
$$= \frac{1}{2}\sum_i \frac{\partial}{\partial \mathbf{x}_i}\left\{\left[\tilde{\mathbf{S}}(t)(\tilde{\mathbf{S}}(t))^\top\right]_i \nabla p_t(\mathbf{x})\right\}$$

and (ii) follows directly by

$$\mathbf{s}(\mathbf{x},t)p_t(\mathbf{x}) = (\nabla \log p_t(\mathbf{x}))p_t(\mathbf{x}) = \frac{\nabla p(\mathbf{x},t)}{p_t(\mathbf{x})}p_t(\mathbf{x}).$$

Notice the case of zero diffusion matrix for Theorem B.4 (where Eq. 42 is also called the continuity equation). The solution $\mathbf{x}_t$ of Eq. (13) shares the same marginal probability path of Eq. (1) in $[t_0, t_1]$.

Suppose that there exists a unique solution $\mathbf{x}(\cdot)$ of Eq. (13) in the interval $[t_k, t_{k+1}]$ with $\mathbf{x}(t_k) \sim p_{t_k}$, sharing the marginal probability path of Eq. (1) with the same initial value $\mathbf{x}_{t_k}$ for $k < K-1$, i.e., $\mathbf{x}(t) \sim p_t$. Then in $[t_{k+1}, t_{k+2}]$, by Theorem B.6, we can construct a solution $\tilde{\mathbf{x}}(\cdot)$ of Eq. (13) with $\tilde{\mathbf{x}}(t_{k+1}) = \mathbf{x}(t_{k+1})$. Since $\mathbf{x}(t_{k+1}) \sim p_{t_{k+1}}$, then $\tilde{\mathbf{x}}(t) \sim p_t$ in $[t_{k+1}, t_{k+2}]$ following the above analysis in the interval $[t_0, t_1]$. Let

$$\mathbf{z}(t) := \begin{cases} \mathbf{x}(t), & \text{if} \quad t_k \le t \le t_{k+1}, \\ \tilde{\mathbf{x}}(t), & \text{if} \quad t_{k+1} < t \le t_{k+2}. \end{cases}$$

By Theorem B.4 (the case of the continuity equation), p_t of Eq. 1 is continuous and differentiable with respect to t at t. Hence, given $\mathbf{x}_{t_k}$ as initial value, $\mathbf{z}_t$ is unique solution of Eq. (13) and $\mathbf{z}_t \sim p_t$ in $[t_k, t_{k+2}]$. According to the inductive hypothesis and the case $k = 0$, the proof is completed. □

B.6 Why a Nonzero Stochastic Coefficient Can Help

The additive-noise estimate $O(h^2 + \epsilon^2 h)$ in Appendix A.3.1 is a source SRK discretization result under its regularity and stability assumptions. It is not a proved global bound for history-adaptive SONGS. Its noise-dependent term motivates considering small noise scales; at zero noise the main implementation reduces to the RK2 family in Eq. (36), which is Heun when $\xi_k = 1$, while PL1W1 reduces to Heun in Eq. (37).

Generative sampling, however, is carried out with learned approximations. The following heuristic explains why a small nonzero stochastic coefficient can still be beneficial in the finite-NFE regime. Consider the single-direction case $m = 1$ with a unit step-process direction $\|\mathbf{S}(t)\|_2 = 1$, and write

$$d\mathbf{x}_t = \left[\mathbf{v}(\mathbf{x}_t, t) + \frac{1}{2}\omega_t^2 \mathbf{S}(t)\mathbf{S}(t)^\top \mathbf{s}(\mathbf{x}_t, t)\right] dt + \omega_t \mathbf{S}(t)\, dW.$$

Suppose the learned quantities are $\hat{\mathbf{v}} = \mathbf{v} + e_v$ and $\hat{\mathbf{s}} = \mathbf{s} + e_s$, and let $\hat{p}_t$ denote the marginal law induced by the learned sampler. Then the drift perturbation becomes

$$e_{\text{drift}} = e_v + \frac{1}{2}\omega_t^2 \mathbf{S}(t)\mathbf{S}(t)^\top e_s.$$

Projecting onto the stochastic direction gives

$$\mathbf{S}(t)^\top e_{\text{drift}} = \mathbf{S}(t)^\top e_v + \frac{1}{2}\omega_t^2 \mathbf{S}(t)^\top e_s.$$

For $\omega_t > 0$, define the projected error diagnostic

$$\mathcal{E}_{\|}(\omega) := \frac{1}{2}\int_0^1 \omega_t^{-2} \mathbb{E}_{p_t}\left[\left(\mathbf{S}(t)^\top e_v + \frac{1}{2}\omega_t^2 \mathbf{S}(t)^\top e_s\right)^2\right] dt.$$

This makes the tradeoff explicit: taking ω_t too small suppresses the stochastic correction and enlarges the velocity-error term after rescaling, while taking ω_t too large amplifies the score-error contribution. Using Young's inequality,

$$|\langle \mathbf{S}(t), e_v\rangle \langle \mathbf{S}(t), e_s\rangle| \le \frac{1}{2\omega_t^2}\langle \mathbf{S}(t), e_v\rangle^2 + \frac{\omega_t^2}{2}\langle \mathbf{S}(t), e_s\rangle^2,$$

This inequality motivates examining a nonzero noise scale empirically; it does not prove that a positive scale is optimal. The projected diagnostic is not a full-distribution KL bound: a Girsanov argument for degenerate diffusion additionally requires the entire drift perturbation to lie in the diffusion range, together with the relevant integrability conditions.

This heuristic explains the empirical pattern in the ablations: $\omega = 0$ is a useful deterministic reference point, whereas a modest stochastic coefficient improves finite-budget sample quality; see Figure 3 in the main text and Table 12 in Appendix E.

B.7 Unified Flow Form

Lemma B.7. *Let* $\mathbf{x}_t^{RF} = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ *and* $\mathbf{x}_t = \alpha_t\mathbf{x}_1 + \sigma_t\mathbf{x}_0$ *with* $c_t := \alpha_t + \sigma_t \neq 0$, $\tau(t) := \frac{\alpha_t}{c_t}$ *and* $\tau^{-1}(\tau(t)) = t,\ \dot{\tau}(t) \neq 0$. *Assume that*

$$\mathbf{x}_{\tau(t)}^{RF} := \frac{\mathbf{x}_t}{c_t}$$

Then

$$\mathbf{v}^{RF}(\mathbf{x}_{\tau(t)}^{RF}, \tau(t)) = \frac{\mathbf{v}(\mathbf{x}_t, t)}{c_t\dot{\tau}(t)} - \frac{\dot{c}_t}{c_t^2\dot{\tau}(t)}\mathbf{x}_t$$

where $\mathbf{v}(\mathbf{x}_t, t) := \mathbb{E}[\dot{\alpha}_t\mathbf{x}_1 + \dot{\sigma}_t\mathbf{x}_0|\mathbf{x}_t]$ *and* $\mathbf{v}^{RF}(\mathbf{x}_t, t) := \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0|\mathbf{x}_t]$. *Specifically, if* $\alpha_t \equiv 1$, *then*

$$\mathbf{v}^{RF}(\mathbf{x}_\tau^{RF}, \tau) = -\frac{1+\sigma_t}{\dot{\sigma}_t}\mathbf{v}(\mathbf{x}_t, t) + \mathbf{x}_t$$

Here we directly use the notations in Lemma B.7 for the appendix discussion of unified flow forms.

Proof of the unified flow transformation statement. A direct algebraic manipulation yields

$$\begin{aligned}
\mathbf{x}_t &= \alpha_t\mathbf{x}_1 + \sigma_t\mathbf{x}_0 \\
&= (\alpha_t + \sigma_t)\left(\frac{\alpha_t}{\alpha_t + \sigma_t}\mathbf{x}_1 + \frac{\sigma_t}{\alpha_t + \sigma_t}\mathbf{x}_0\right) \\
&= c_t\left(\tau(t)\mathbf{x}_1 + (1-\tau(t))\mathbf{x}_0\right) \\
&= c_t\mathbf{x}_{\tau(t)}^{RF}.
\end{aligned}$$

This equality holds pointwise and, consequently, also in distribution.

For the marginal velocities, first differentiate the samplewise identity and then condition on $\mathbf{x}_t$. Since $\mathbf{x}_t = c_t\mathbf{x}_{\tau(t)}^{RF}$ is an invertible spatial rescaling,

$$\begin{aligned}
\mathbf{v}(\mathbf{x}_t, t) &= \mathbb{E}[\dot{\alpha}_t\mathbf{x}_1 + \dot{\sigma}_t\mathbf{x}_0 \mid \mathbf{x}_t] \\
&= \dot{c}_t\mathbf{x}_{\tau(t)}^{RF} + c_t\dot{\tau}(t)\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_{\tau(t)}^{RF}] \\
&= \dot{c}_t\mathbf{x}_{\tau(t)}^{RF} + c_t\dot{\tau}(t)\mathbf{v}^{RF}(\mathbf{x}_{\tau(t)}^{RF}, \tau(t)).
\end{aligned}$$

If $\alpha_t \equiv 1$, then $\dot{c}_t = \dot{\sigma}_t$ and

$$\begin{aligned}
\dot{\tau}(t) &= \frac{\dot{\alpha}_t c_t - \alpha_t\dot{c}_t}{c^2(t)} \\
&= -\frac{\dot{\sigma}_t}{c_t^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{v}^{RF}\left(\mathbf{x}_{\tau(t)}^{RF}, \tau(t)\right) &= \frac{\mathbf{v}(\mathbf{x}_t, t)}{-c_t\frac{\dot{\sigma}_t}{c_t^2}} + \frac{\dot{\sigma}_t}{c_t^2\frac{\dot{\sigma}_t}{c_t^2}}\mathbf{x}_t \\
&= -\frac{1+\sigma_t}{\dot{\sigma}_t}\mathbf{v}(\mathbf{x}_t, t) + \mathbf{x}_t.
\end{aligned}$$

This completes the proof. □

We choose the Elucidating Diffusion Model (EDM) (Karras et al., 2022) as an example. For EDM, we have

$$\frac{d\mathbf{x}}{d\sigma} = \frac{x - D(x;\sigma)}{\sigma}.$$

Let $\alpha_t = 1$ and $\sigma_t = \sigma_{max}(1-t)$, then

$$\begin{aligned}
\mathbf{v}(\mathbf{x}_t, t) &= \frac{d\mathbf{x}}{dt} \\
&= \frac{d\mathbf{x}}{d\sigma}\frac{d\sigma}{dt} \\
&= \frac{\mathbf{x}_t - D(\mathbf{x}_t;\sigma_t)}{\sigma_t}(-\sigma_{max}) \\
&= \frac{\sigma_{max}}{\sigma_t}(D(\mathbf{x}_t;\sigma_t) - \mathbf{x}_t)
\end{aligned}$$

where $D(x;\sigma_t)$ is a denoiser function. Then

$$\begin{aligned}
\mathbf{v}^{RF}(\mathbf{x}_\tau^{RF}, \tau) &= -\frac{(1+\sigma_t)\sigma_{max}(D(\mathbf{x}_t;\sigma_t) - \mathbf{x}_t)}{\dot{\sigma}_t\sigma_t} + \mathbf{x}_t \\
&= \frac{(1+\sigma_t)\sigma_{max}(D(\mathbf{x}_t;\sigma_t) - \mathbf{x}_t)}{\sigma_{max}\sigma_t} + \mathbf{x}_t \\
&= \frac{(1+\sigma_t)D(\mathbf{x}_t;\sigma_t) - \mathbf{x}_t}{\sigma_t} \\
&= \frac{D(\mathbf{x}_t;\sigma_t) - \frac{\mathbf{x}_t}{1+\sigma_t}}{\frac{\sigma_t}{1+\sigma_t}} \\
&= \frac{D(\frac{\mathbf{x}_\tau^{RF}}{\tau};\frac{1-\tau}{\tau}) - \mathbf{x}_\tau^{RF}}{1-\tau}
\end{aligned}$$

where $\mathbf{x}_t = \frac{\mathbf{x}_\tau^{RF}}{\tau}$ and

$$t = \tau^{-1}(\tau) = \frac{(1+\sigma_{max})\tau - 1}{\sigma_{max}\tau}$$

for $\tau \in [\frac{1}{1+\sigma_{max}}, 1]$.

B.8 Discussion on Prior Error

As discussed in the appendix on unified flow forms, standard DPMs in practice rely on an approximate prior within the FM formulation. Building on that, this section investigates how to compensate for this approximation, with EDM serving as our running example. Note that for EDM, $\tau(0) = \frac{1}{1+\sigma_{max}}$. Write $Z \sim \mathcal{N}(0, I_d)$ and $Y_0 = \sigma_{max}Z$ for the EDM initial sample. Its distribution is $\mathcal{N}(0, \sigma_{max}^2 I_d)$, and the corresponding RF initialization is

$$\frac{Y_0}{1+\sigma_{max}} \sim \mathcal{N}\left(0, \left(\frac{\sigma_{max}}{1+\sigma_{max}}\right)^2 I_d\right).$$

In the first approximation below, $\mathbf{x}_0 = Z$ denotes the standard Gaussian RF source; in approximations 2–4, $\mathbf{x}_0 = Y_0$ denotes the EDM initial sample. In general, this is different from the ground truth

$$\mathbf{x}_{\tau(0)}^{RF} = \mathbf{x}_0 + \int_0^{\tau(0)} \mathbf{v}^{RF}(\mathbf{x}_\xi^{RF}, \xi)d\xi$$

We discuss four heuristic approximations; these are not exact prior corrections.

1. By a forward Euler approximation, we have

$$\mathbf{x}_{\tau(0)}^{RF} \approx \mathbf{x}_0^{RF} + (\mathbb{E}[\mathbf{x}_1^{RF}] - \mathbf{x}_0^{RF})\tau(0) = \frac{\sigma_{max}}{1+\sigma_{max}}\mathbf{x}_0 + \frac{1}{1+\sigma_{max}}\mathbb{E}[\mathbf{x}_1^{RF}] \approx \frac{\sigma_{max}}{1+\sigma_{max}}\mathbf{x}_0$$

This is used in EDM.

2. A right-endpoint approximation is

$$\begin{aligned}
&\int_0^{\tau(0)} \mathbf{v}^{RF}(\mathbf{x}_\xi^{RF}, \xi) d\xi \\
&\approx \mathbf{v}^{RF}(\mathbf{x}_{\tau(0)}^{RF}, \tau(0))\tau(0) \\
&= (\frac{\mathbf{v}(\mathbf{x}_0, 0)}{c(0)\dot{\tau}(0)} - \frac{\dot{c}(0)}{c^2(0)\dot{\tau}(0)}\mathbf{x}_0)\tau(0) \\
&= (\frac{1+\sigma_{max}}{\sigma_{max}}\mathbf{v}(\mathbf{x}_0, 0) + \mathbf{x}_0)\tau(0) \\
&= ((1+\frac{1}{\sigma_{max}})D(\mathbf{x}_0; \sigma_{max}) - \frac{1}{\sigma_{max}}\mathbf{x}_0)\tau(0),
\end{aligned}$$

3. Since $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_{max}^2 I)$, then

$$\frac{\mathbf{x}_0}{\sigma_{max}} \sim \mathcal{N}(0, I) = p_0.$$

Let $\mathbf{x}_0^{RF} = \frac{\mathbf{x}_0}{\sigma_{max}}$, we have

$$\begin{aligned}
\hat{\mathbf{x}}_0 &= c(0)\mathbf{x}_{\tau(0)}^{RF} \\
&= c(0)(\mathbf{x}_0^{RF} + \int_0^{\tau(0)} \mathbf{v}^{RF}(\mathbf{x}_\xi^{RF}, \xi) d\xi) \\
&\approx (1+\frac{1}{\sigma_{max}})D(\mathbf{x}_0; \sigma_{max}) - \frac{1}{\sigma_{max}}\mathbf{x}_0 + c(0)\mathbf{x}_0^{RF} \\
&= (1+\frac{1}{\sigma_{max}})D(\mathbf{x}_0; \sigma_{max}) + \mathbf{x}_0
\end{aligned}$$

4. Considering $\tau(0)$ is small,

$$\mathbf{v}^{RF}(\mathbf{x}_\tau^{RF}, \tau) = \mathbb{E}[\mathbf{x}_1^{RF} - \mathbf{x}_0^{RF} | \mathbf{x}_\tau^{RF}] \approx \mathbb{E}[\mathbf{x}_1^{RF}] - \mathbf{x}_\tau^{RF} \tag{43}$$

Suppose that $\mathbb{E}(\mathbf{x}_1^{RF}) = 0$. Then $\mathbf{x}_{\tau(0)}^{RF} \approx e^{-\tau(0)}\mathbf{x}_0^{RF}$. So $\hat{\mathbf{x}}_0 \approx e^{-\frac{1}{1+\sigma_{max}}}(1+\frac{1}{\sigma_{max}})\mathbf{x}_0$.

B.9 Resampling by Reverse SDE for the Rectified Flow

If the prior is $\mathcal{N}(0, I)$, consider the reverse-time noising SDE (t decreases and $\bar{\mathbf{w}}$ is a d-dimensional reverse-time Brownian motion):

$$d\mathbf{x}_t = \frac{\mathbf{x}_t}{t}\, dt + \sqrt{2\frac{1-t}{t}} d\bar{\mathbf{w}}$$

Notice that

$$\begin{aligned}
d\frac{\mathbf{x}_t}{t} &= \frac{1}{t}\left(d\mathbf{x}_t - \frac{\mathbf{x}_t}{t}dt\right) \\
&= \frac{1}{t}\sqrt{2\frac{1-t}{t}} d\bar{\mathbf{w}}
\end{aligned}$$

Then for $0 < s < t$,

$$\frac{\mathbf{x}_s}{s} - \frac{\mathbf{x}_t}{t} = \sqrt{\frac{2}{t} - \frac{2}{s} + \frac{1}{s^2} - \frac{1}{t^2}}\, \epsilon \tag{44}$$

where $\epsilon \sim \mathcal{N}(0, I)$. By direct transformation, we have

$$\mathbf{x}_s = \frac{s}{t}\mathbf{x}_t + \sqrt{1 - 2s + \frac{2s^2}{t} - \frac{s^2}{t^2}}\, \epsilon \tag{45}$$

for $0 \le s < t$. The singularity at $s = 0$ in (44) cancels in (45).

Next, we introduce our reSDE update over the interval $[t, t + \Delta t]$, defined as:

$$\mathbf{x}_{t+\Delta t} = \text{reSDE}_\theta(\mathbf{x}_t, \Delta t, \gamma),$$

which inserts a stochastic *forward perturbation* followed by a deterministic *backward process*.

$$\mathbf{x}^\gamma_{t,\Delta_t} = \mathbf{x}_{t-\gamma\Delta t} = (1 - \gamma\frac{\Delta t}{t})\mathbf{x}_t + \mathcal{E}(t \to t - \gamma\Delta t),$$

$$\mathbf{x}_{t+\Delta t} = \text{Predict}_\theta(\mathbf{x}^\gamma_{t,\Delta t}, t - \gamma\Delta t \to t + \Delta t),$$

where

$$\mathcal{E}(t \to t - \gamma\Delta t) \sim \mathcal{N}\left(0, \left(1 - 2s + \frac{2s^2}{t} - \frac{s^2}{t^2}\right) I\right).$$

with $s = t - \gamma\Delta t$ and $0 \le s < t \le 1$.

C More Numerical Methods for the Rectified Flow Form

C.1 Exponential Integrators

Recall the exponential-integrator identities for diffusion ODEs. Here $\lambda_t = \log(\alpha_t/\sigma_t)$ denotes the log signal-to-noise ratio; $\hat{\mathbf{x}}_\theta$ and $\hat{\epsilon}_\theta$ denote data and noise predictions, respectively:

Proposition C.1. *(Exact solution of diffusion ODEs (Lu et al., 2025)). Given an initial value $\mathbf{x}_s$ at time $s > 0$, the solution $\mathbf{x}_t$ at time $t \in [0, s]$ of diffusion ODEs is:*

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s}\mathbf{x}_s + \sigma_t \int_{\lambda_s}^{\lambda_t} e^{\lambda}\hat{\mathbf{x}}_\theta(\hat{\mathbf{x}}_\lambda, \lambda)\, d\lambda. \tag{46}$$

or equivalently,

$$\mathbf{x}_t = \frac{\alpha_t}{\alpha_s}\mathbf{x}_s - \alpha_t \int_{\lambda_s}^{\lambda_t} e^{-\lambda}\hat{\epsilon}_\theta(\hat{\mathbf{x}}_\lambda, \lambda)\, d\lambda, \tag{47}$$

As a direct transformation, we have

Proposition C.2. *(Exact solution of the rectified flow ODE). Given an initial value $\mathbf{x}_s$, the solution $\mathbf{x}_t$ at interior time $s < t < 1$ of the rectified flow ODEs is:*

$$\mathbf{x}_t = \frac{1 - t}{1 - s}\mathbf{x}_s + (1 - t)\int_s^t \frac{\hat{\mathbf{x}}_\theta(\mathbf{x}_\tau, \tau)}{(1 - \tau)^2}\, d\tau. \tag{48}$$

or equivalently,

$$\mathbf{x}_t = \frac{t}{s}\mathbf{x}_s - t\int_s^t \frac{\epsilon_\theta(\mathbf{x}_\tau, \tau)}{\tau^2}\, d\tau, \tag{49}$$

for $s > 0$.

Remark C.3. Note that for noise reparameterization, there exists a singularity at time $s = 0$. This is fine for diffusion models, but harmful for rectified flow.

For an independent source–target coupling, the initial data prediction is $\mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_0] = \mu$. Freezing the initial velocity gives the Euler approximation

$$\hat{\mathbf{x}}_t = (1 - t)\mathbf{x}_0 + t\mu, \qquad \hat{\mathbf{x}}_t \sim \mathcal{N}(t\mu, (1 - t)^2 I_d) \quad \text{if } \mathbf{x}_0 \sim \mathcal{N}(0, I_d).$$

This is a finite-step approximation, not the exact solution for a general data distribution.

Proposition C.4 (Integral identities for a rectified-flow SDE)**.** *On an interval $0 < s < t < 1$, let $\mathbf{u} \in \mathbb{R}^d$ be a fixed direction and consider diffusion coefficient $\omega(t)\mathbf{u}$ with score drift $\frac{1}{2}\omega(t)^2\mathbf{u}\mathbf{u}^\top\mathbf{s}$. Then variation of constants gives*

$$\begin{aligned}\mathbf{x}_t &= \frac{1-t}{1-s}\mathbf{x}_s + (1-t)\int_s^t \frac{\mathbf{x}_\theta(\mathbf{x}_\tau,\tau)}{(1-\tau)^2}\,d\tau \\ &\quad + \frac{1-t}{2}\mathbf{u}\mathbf{u}^\top \int_s^t \frac{\omega(\tau)^2[\tau\mathbf{x}_\theta(\mathbf{x}_\tau,\tau)-\mathbf{x}_\tau]}{(1-\tau)^3}\,d\tau + (1-t)\mathbf{u}\int_s^t \frac{\omega(\tau)}{1-\tau}\,dW_\tau \\ &= \frac{t}{s}\mathbf{x}_s - t\int_s^t \frac{\epsilon_\theta(\mathbf{x}_\tau,\tau)}{\tau^2}\,d\tau \\ &\quad - \frac{t}{2}\mathbf{u}\mathbf{u}^\top \int_s^t \frac{\omega(\tau)^2\epsilon_\theta(\mathbf{x}_\tau,\tau)}{\tau(1-\tau)}\,d\tau + t\mathbf{u}\int_s^t \frac{\omega(\tau)}{\tau}\,dW_\tau.\end{aligned} \tag{50}$$

Here $\mathbf{s} = (t\mathbf{x}_\theta - \mathbf{x})/(1-t)^2 = -\epsilon_\theta/(1-t)$, so the projected score term in the noise-prediction representation has a minus sign. These are integral identities for the specified SDE, not a separate marginal-preservation result for history-dependent directions.

C.2 Single-step Method

Euler Method for Rectified Flow and Denoising Diffusion Implicit Models

Lemma C.5. *For $t > s > 0$, if $\alpha_t := t$ and $\sigma_t := 1 - t$, then the following DDIM (DPM-solver-1) update rule:*

$$\hat{\mathbf{x}}_t = \frac{\alpha_t}{\alpha_s}\hat{\mathbf{x}}_s - \alpha_t\left(\frac{\sigma_s}{\alpha_s} - \frac{\sigma_t}{\alpha_t}\right)\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] \tag{51}$$

is equivalent to

$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_s + (t-s)\underbrace{\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s]}_{\textit{estimated velocity}}. \tag{52}$$

Remark C.6. This conclusion shows that for rectified flow, the plain Euler method sampling is equivalent to DDIM sampling. Note that this equivalence of sampling methods also holds for canonical elucidating diffusion models (EDMs), but does not hold for all flow matching models, such as VP diffusion models.

Proof. Note that

$$\begin{aligned}\hat{\mathbf{x}}_s &= \mathbb{E}[\mathbf{x}_s|\mathbf{x}_s = \hat{\mathbf{x}}_s] \\ &= \mathbb{E}[\alpha_s\mathbf{x}_1 + \sigma_s\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s].\end{aligned}$$

Then

$$\begin{aligned}\hat{\mathbf{x}}_t &= \frac{\alpha_t}{\alpha_s}\mathbb{E}[\mathbf{x}_s|\mathbf{x}_s = \hat{\mathbf{x}}_s] - \alpha_t\left(\frac{\sigma_s}{\alpha_s} - \frac{\sigma_t}{\alpha_t}\right)\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] \\ &= \alpha_t\mathbb{E}[\mathbf{x}_1 + \frac{\sigma_s}{\alpha_s}\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] - \alpha_t\frac{\sigma_s}{\alpha_s}\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] \\ &\quad + \sigma_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] \\ &= \alpha_t\mathbb{E}[\mathbf{x}_1|\mathbf{x}_s = \hat{\mathbf{x}}_s] + \sigma_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s].\end{aligned}$$

Let $\alpha_t := t$ and $\sigma_t := 1 - t$.

$$\begin{aligned}\hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_s + t\mathbb{E}[\mathbf{x}_1|\mathbf{x}_s = \hat{\mathbf{x}}_s] + (1-t)\mathbb{E}[\mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s] - \hat{\mathbf{x}}_s \\ &= \hat{\mathbf{x}}_s + (t-s)\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0|\mathbf{x}_s = \hat{\mathbf{x}}_s]\end{aligned}$$

The proof is completed. □

Two-stage Rectified-Flow Predictor Let $h_i = t_{i+1} - t_i$ and $\hat{t}_{i+1} = t_i + \rho_i h_i$, with $0 < \rho_i \leq 1$ and all times below 1. Given an intermediate state $\mathbf{x}_{\hat{t}_{i+1}}$, form the following weighted data-prediction update. Write $\mathbf{v}_i = \mathbf{v}(\mathbf{x}_{t_i}, t_i)$ and $\mathbf{v}_{mid} = \mathbf{v}(\mathbf{x}_{\hat{t}_{i+1}}, \hat{t}_{i+1})$. Using $\mathbf{x}_1^\theta(\mathbf{x}, t) = \mathbf{x} + (1-t)\mathbf{v}(\mathbf{x}, t)$ yields

$$\begin{aligned}\hat{\mathbf{x}}_{t_{i+1}} &= \frac{1-t_{i+1}}{1-t_i}\mathbf{x}_{t_i} + \frac{h_i}{1-t_i}\left[\left(1 - \frac{1}{2\rho_i}\right)\mathbf{x}_1^\theta(\mathbf{x}_{t_i}, t_i) + \frac{\mathbf{x}_1^\theta(\mathbf{x}_{\hat{t}_{i+1}}, \hat{t}_{i+1})}{2\rho_i}\right] \\ &= \mathbf{x}_{t_i} + h_i\left[\left(1 - \frac{1}{2\rho_i}\right)\mathbf{v}_i + \frac{1-\hat{t}_{i+1}}{2\rho_i(1-t_i)}\mathbf{v}_{mid} + \frac{\mathbf{x}_{\hat{t}_{i+1}} - \mathbf{x}_{t_i}}{2\rho_i(1-t_i)}\right].\end{aligned}$$

The intermediate-state difference cannot be dropped when converting to velocity form. The order of a complete solver additionally depends on how that intermediate state is constructed. Higher-order approximations may also be sensitive to large guidance scales (Lu et al., 2025).

D QUALITATIVE COMPARISON



Figure 4: Euler–Maruyama sampling of the diffusion SDE at 40 NFE.

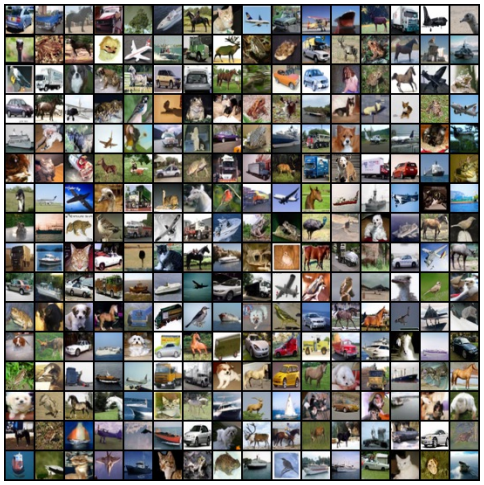
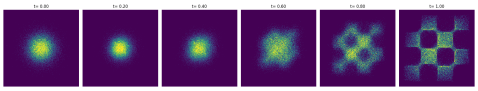
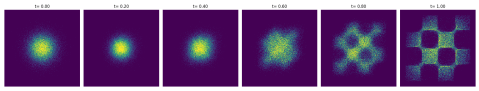


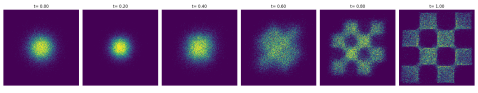
Figure 5: SONGS sampling at 40 NFE.



(a) Euler sampling of the ODE, 5 NFE



(b) Euler–Maruyama sampling of the SDE, 5 NFE



(c) SONGS, 5 NFE

Figure 6: **Comparison at a fixed evaluation budget.** Under the same five-evaluation budget, SONGS produces sharper structure than Euler ODE sampling and Euler–Maruyama SDE sampling. The target p_1 is a two-dimensional checkerboard distribution.



Figure 7: FFHQ 64×64 samples at 5 NFE. Left: DPM-Solver-2. Right: SONGS.



Figure 8: AFHQv2 64×64 samples at 9 NFE. Left: DPM-Solver-2. Right: SONGS.

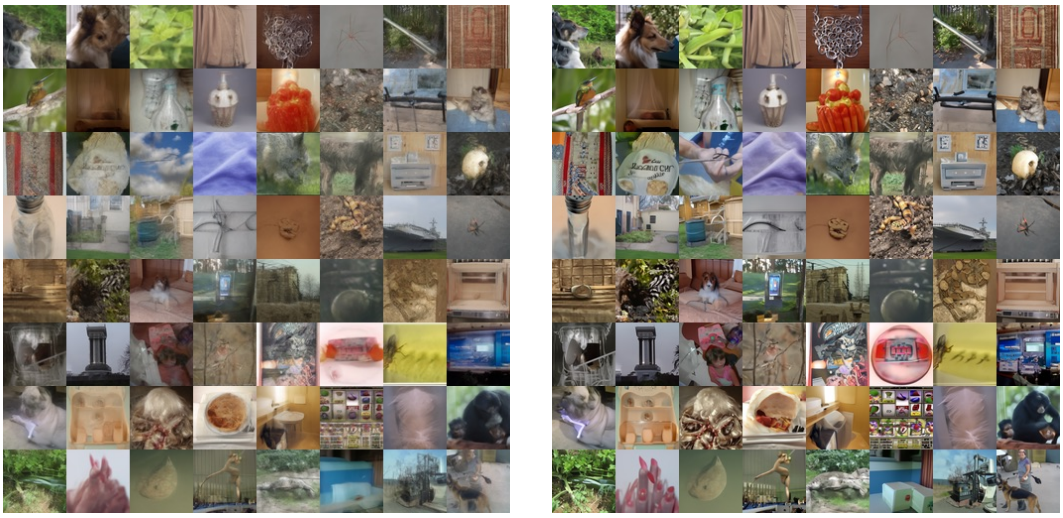


Figure 9: ImageNet 64×64 samples at 9 NFE. Left: DPM-Solver-2. Right: SONGS.

E EXTENSIVE EXPERIMENTS

E.1 HIGHER-NFE RESULTS WITH SONGS–PL1W1

Table 9 evaluates the SONGS–PL1W1 implementation described in Appendix A.3.2, rather than the improved-SRK implementation used in Table 2. At matched NFE, SONGS–PL1W1 improves on its deterministic Heun2 backbone throughout the tested range of 18–40 NFE, with a smaller absolute gap at larger budgets. This comparison evaluates the stochastic correction within the PL1W1 implementation; it does not compare against advanced multistep solvers.

Table 9: Unconditional CIFAR-10 FID (↓) at 18–40 NFE using SONGS–PL1W1 (Algorithm 2) and its deterministic Heun2 backbone.

NFE	18	20	24	32	36	40
Heun2	2.8627	2.4949	2.1690	1.9985	1.9784	1.9672
SONGS–PL1W1	**2.6173**	**2.3317**	**2.0922**	**1.9725**	**1.9596**	**1.9427**

E.2 ADDITIONAL QUANTITATIVE RESULTS

Table 10: FID (↓) from 50K samples on unconditional CIFAR-10 32×32. The table includes additional samplers with publicly reported results.

METHOD	FID ↓ / NFE		
	3	5	7
MULTI-STEP SOLVERS			
SA-SOLVER(DEFAULT) (XUE ET AL., 2023)	305.22	264.04	236.79
SA-SOLVER(LOG-SNR) (XUE ET AL., 2023)	155.31	23.89	7.78
DPM-SOLVER++(3M) (LU ET AL., 2025)	110.00	24.97	6.74
UNIPC (ZHAO ET AL., 2023)	109.60	23.98	5.83
IPNDM (ZHANG & CHEN, 2023)	47.98	13.59	5.08
DPM-SOLVER-V3 (ZHENG ET AL., 2023)	84.17	12.27	5.51
SINGLE-STEP SOLVERS			
IMPROVED SDE (KARRAS ET AL., 2022)	468.28	358.47	135.28
EDM HEUN (KARRAS ET AL., 2022)	468.28	319.80	99.42
DDIM (SONG ET AL., 2021A)	93.36	49.66	27.93
DPM-SOLVER-2 (LU ET AL., 2022)	153.6	43.27	16.69
DPM-SOLVER-V3 (S) (ZHENG ET AL., 2023)	–	21.83	–
LMS (WANG ET AL., 2025)	–	17.28	11.15
SONGS (OURS)	**15.00**	**6.09**	**4.43**

Table 11: Reported peak sampling quality on unconditional CIFAR-10. Models, checkpoints, and evaluation budgets differ across rows, so this table is contextual rather than a controlled comparison.

METHOD	NFE	FID ↓
ODE MODELING		
FLOW MATCHING/OT (LIPMAN ET AL., 2023)	142	6.35
1-RECTIFIED FLOW (LIU ET AL., 2023)	127	2.58
2-RECTIFIED FLOW (LIU ET AL., 2023)	110	3.36
3-RECTIFIED FLOW (LIU ET AL., 2023)	104	3.96
DDIM (SONG ET AL., 2021A)	50	4.67
PFGM (XU ET AL., 2022)	110	2.35
SDE MODELING		
DDPM (HO ET AL., 2020)	1000	3.21
SCORE SDE (SONG ET AL., 2021B)	2000	2.20
VDM (KINGMA ET AL., 2021)	1000	4.0
IDDPM (NICHOL & DHARIWAL, 2021)	4000	2.9
LSGM (VAHDAT ET AL., 2021)	147	2.10
ADVANCED SAMPLERS		
IMPROVED SDE-VE (KARRAS ET AL., 2022)	2048	2.23
IMPROVED SDE-VP (KARRAS ET AL., 2022)	512	2.27
SA-SOLVER (XUE ET AL., 2023)	95	2.63
HEUN (KARRAS ET AL., 2022)	35	1.97
DPM-SOLVER++ (LU ET AL., 2025)	25	2.02
UNIPC (ZHAO ET AL., 2023)	25	2.00
DPM-SOLVER-V3 (ZHENG ET AL., 2023)	25	2.00
SONGS (OURS)	40	**1.94**

Table 12: Ablation over the implementation scale sf at 64 NFE on ImageNet 256×256. We retain the labels recorded with the original measurements because their conversion to the normalized coefficient ω is not documented in the archived experiment files.

SCALE sf	FID ↓	sFID ↓	IS ↑
65	5.20	**17.96**	254.64
75	5.19	17.97	254.99
85	5.19	17.98	255.36
95	5.19	17.98	255.37
105	5.19	17.99	255.28
500	5.18	18.02	255.29
1000	**5.17**	18.03	255.60
1500	5.18	18.04	255.81
2000	5.18	18.04	**255.83**
0	5.24	18.04	252.78

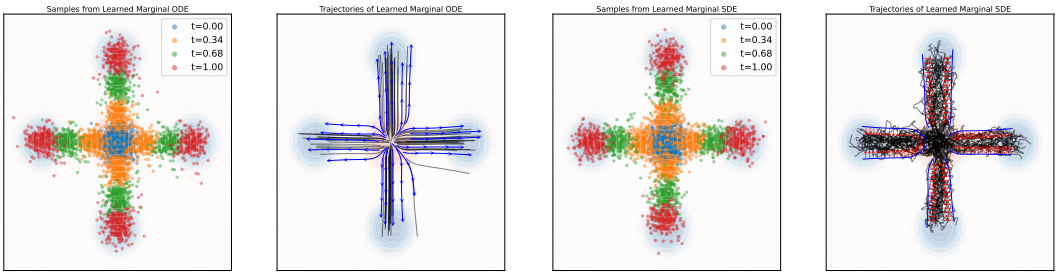
Table 13: SiT FID across sampling budgets and SONGS schedule variants.

Sampler / NFE	16	24	32	48	64	96
ODE(Heun2)	7.46	5.86	5.44	5.26	5.24	5.20
SDE(Euler-Maruyama)	19.32	9.92	7.15	5.70	5.38	5.21
SDE(Heun)	14.71	6.41	5.54	5.15	5.08	5.08
SONGS (Ours, increasing-decreasing)	6.49	5.47	5.22	5.15	5.12	5.12
SONGS (Ours, decreasing)	6.64	5.52	5.26	5.16	5.13	5.12
SONGS (Ours, linear)	6.35	5.41	5.20	5.14	5.13	5.13

Table 14: SiT performance at NFE $= 16$. Percentage changes are computed from the displayed rounded values.

SAMPLER	FID ↓	sFID ↓	IS ↑	PREC. ↑	REC. ↑
ODE (HEUN)	7.46	22.38	211.22	0.78	0.66
SDE (EULER)	19.32	42.89	145.01	0.66	0.66
SONGS (OURS)	**6.33**	**20.24**	**224.21**	**0.80**	**0.67**
IMPROVEMENT OVER BASELINES					
VS ODE (HEUN)	15.15%	9.56%	6.15%	2.56%	1.52%
VS SDE (EULER)	67.24%	52.81%	54.62%	21.21%	1.52%

E.3 SAMPLING-TRAJECTORY ILLUSTRATION



(a) (b)

Figure 10: Samples and trajectories generated by an FM model. (a) ODE sampling. (b) SDE sampling.

F Additional Efficiency Results

F.1 Noise-Generation Microbenchmark

To isolate the scaling of random-number generation, we compare sampling a full latent Gaussian tensor with sampling one scalar Brownian motion (BM) increment per image. We use preallocated CUDA buffers, latent channel count $C = 16$, and $8\times$ spatial downsampling. This is a component-level microbenchmark: end-to-end generation remains dominated by network evaluations, and we do not claim the same speedups for complete sampling. The scalar benchmark measures one Gaussian draw; Algorithm 1 uses two, including an auxiliary draw for its mixed integral.

Table 15: CUDA random-number generation cost per sampling step.

Image	Latent shape	Gaussian vars.	Full BM (ms)	1D BM (ms)	Speedup
2048	$16\times256\times256$	1,048,576	0.0205	0.0091	$2.2\times$
3072	$16\times384\times384$	2,359,296	0.1002	0.0094	$10.7\times$
4096	$16\times512\times512$	4,194,304	0.2675	0.0095	$28.2\times$
6144	$16\times768\times768$	9,437,184	0.6366	0.0145	$43.9\times$