Per-Starting-Point Plausibility and Diversity Bounds for Score-Based Diffusion Models

Nguyen Minh Quang
School of Computing and Information Systems
Singapore Management University
Singapore
mq.nguyen.2023@phdcs.smu.edu.sg

Hady W. Lauw
School of Computing and Information Systems
Singapore Management University
Singapore
hadywlauw@smu.edu.sg

Abstract

Score-based diffusion is well-studied at the mixture level, as results are averaged over the prior. But many practical pipelines (posterior sampling, SDEdit, OOD detection) fix the starting point and care about the per-start output. To our knowledge, this per-starting-point regime has no prior non-asymptotic analysis. We give two coupled bounds: a *plausibility* upper bound on how close the per-start output is to the data distribution, and a complementary *diversity* lower bound on how distinguishable outputs from two different starts remain. Related score-contraction quantities govern plausibility damping and starting-point separation. The diversity bound identifies conditions under which outputs from different starting points remain provably distinguishable over a range of sampling horizons. We illustrate the phenomenon qualitatively on 2D toy datasets, MNIST, and CIFAR-10.

1 Introduction

Score-based diffusion models are the dominant generative framework for continuous data, underlying state-of-the-art results in image, video, molecular, and scientific generation. The foundational formulation treats generation as the time reversal of a forward diffusion that noises the data: the celebrated reverse-time identity of Anderson [1982] expresses the reverse drift in terms of the marginal score $\nabla \ln q_t$, and modern samplers learn a score network $s_\theta(x, t)$ to approximate it. The discrete-time denoising formulation of Ho et al. [2020], the score-matching formulation of Song and Ermon [2019] in the variance-exploding regime, and the continuous SDE unification of Song et al. [2021b] together establish the reverse-SDE viewpoint that also encompasses deterministic samplers such as Song et al. [2021a], the design-space analysis of Karras et al. [2022], flow matching Lipman et al. [2023], rectified flow Liu et al. [2023], and consistency distillation Song et al. [2023]. The forward process in all these cases is an Ornstein–Uhlenbeck-type diffusion in the sense of Uhlenbeck and Ornstein [1930], and training is driven by score matching Hyvärinen and Dayan [2005] or its denoising-autoencoder formulation Vincent [2011].

The mixture gap. Existing non-asymptotic convergence analyses Chen et al. [2023], Benton et al. [2024], Lee et al. [2023], Bortoli [2022], Huang et al. [2021] bound the *mixture* output

$$p_T \;=\; \int \mathcal{H}^{(\mathbf{z})}\, \gamma_d(\mathrm{d}\mathbf{z}),$$

where $\mathcal{H}^{(\mathbf{z})}$ is the law of the discretized reverse SDE initialized at $Y_0 = \mathbf{z}$ and γ_d is the standard Gaussian prior. These bounds are sharp, dimension-efficient, and cover a wide range of regularity assumptions, but they say nothing about any individual component $\mathcal{H}^{(i)}$. In contrast, many practical uses of diffusion models are *per-starting-point*: diffusion posterior sampling fixes a measurement-dependent initialization or conditioning and runs the reverse SDE from there Chung et al. [2023],

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

Kawar et al. [2022], Song et al. [2022], Jalal et al. [2021], Graikos et al. [2022], Chung et al. [2022]; ensemble-style uncertainty quantification Lakshminarayanan et al. [2017], Wenzel et al. [2020] relies on the law of outputs conditioned on a fixed initialization; and diffusion-based OOD detectors Graham et al. [2023] operate on the trajectories seeded from an individual input. For all of these, the object of interest is not the integrated mixture but the law $\mathcal{H}^{(i)}$ seeded at a specific $\mathbf{x}^{(i)}$, and the mixture bound gives no guarantees. The per-starting-point regime has no prior non-asymptotic analysis.

What we call DDS. We study *diffusion from deterministic starting points* (DDS), the setting in which $Y_0 = \mathbf{x}^{(i)} \in \mathbb{R}^d$ is fixed and the reverse SDE is run stochastically until an early-stopping time $T_{\text{stop}} = T - t_{\min}$ (early stopping is unavoidable because $1/\mathfrak{s}_t^2 \to \infty$ as $t \to 0$). Two questions immediately arise and are not answered by mixture-level theory:

1. **Plausibility.** How close is $\mathcal{H}^{(i)}$ to the data distribution q when measured per starting point, not averaged over the prior?
2. **Diversity.** For $\mathbf{x}^{(i)} \neq \mathbf{x}^{(j)}$, how different are $\mathcal{H}^{(i)}$ and $\mathcal{H}^{(j)}$? Is the starting-point information erased by the noise, or does it survive to the data distribution?

1.1 Contributions

- **A W_1 plausibility bound for point-initialized diffusion** (Theorem 1). We decompose $W_1(q, \mathcal{H}^{(i)})$ into three interpretable terms—a Bayesian conditioning cost W_{init} that decays as $\mathcal{O}(\mathfrak{m}_T)$ (forward mean-attenuation), a bridge early-stopping cost W_{stop}, and a synchronous-coupling generation error W_{gen}. The first two terms have no analogue in mixture-level analyses Chen et al. [2023], Benton et al. [2024]: W_{init} requires a Stein–Poincaré transport argument on the Radon–Nikodym derivative of the Bayesian posterior $q^{(i)}$, and W_{gen} is threaded through a *localized* score-error assumption (Assumption 4) so the bound stays informative when $\mathbf{x}^{(i)}$ lies outside the training distribution's typical set—the ingredient that enables Theorem 2 to separate in- from out-of-distribution starts. The bound is stated for the abstract forward-SDE family (1).
- **A TV diversity lower bound with an explicit certificate threshold** (Theorem 2). Synchronous coupling makes the Brownian noise cancel in the difference $Z_t := Y_t^{(i)} - Y_t^{(j)}$. From the resulting deterministic recursion we obtain a Cantelli-based TV lower bound that is in $[-1, 1]$ by construction and flips sign at a threshold $\delta^2 = 4\sigma_V\sigma_U$. This defines a certificate threshold $T^\star$, not a guarantee of starting-point identity erasure.
- **Dual characterization of the plausibility–diversity tradeoff.** Contraction enters the generation-error bound through the uniform true-score rate μ_{eff} (Eq. (3)) and the separation bound through the pairwise learned-score coefficient κ (Eq. (16)). Positive μ_{eff} damps generation error; diversity is certified when the projected mean separation dominates the projected variances. Contraction helps interpret these quantities, but its sign alone does not determine either bound's informativeness. This is the per-starting-point dual at the score level, distinct from the well-studied sample-level quality–diversity tradeoff in generative modelling.

Scope. The bounds are stated for the unguided reverse SDE. Extensions to classifier-free guidance Ho and Salimans [2022] and velocity-field samplers Lipman et al. [2023], Liu et al. [2023] are discussed in Section 6; Section E.4 defines the exploratory class-conditional DDS experiment, outside the formal claims.

2 Related Work

Convergence theory. Non-asymptotic mixture-level bounds between p_T and q have progressed under weakening assumptions: polynomial-time W_2 and TV guarantees under Lipschitz score and finite second moment Chen et al. [2023], near-d-linear rates via stochastic localization Benton et al. [2024], distributions without isoperimetric inequalities Lee et al. [2023], the manifold hypothesis Bortoli [2022], variational and entropic routes Huang et al. [2021], and Schrödinger bridges De Bortoli et al. [2021]. All bound the mixture $p_T = \int \mathcal{H}^{(\mathbf{z})}\gamma_d(\mathrm{d}\mathbf{z})$; none bound any individual $\mathcal{H}^{(i)}$. We reuse the

synchronous-coupling machinery of Chen et al. [2023], Benton et al. [2024] inside W_{gen}, but W_{init} and W_{stop} have no mixture-level analogue and require a new Stein–Poincaré argument (Section B).

Posterior sampling and guidance. Per-start reverse SDEs underlie DPS Chung et al. [2023], DDRM Kawar et al. [2022], medical imaging Song et al. [2022], Jalal et al. [2021], manifold-constrained guidance Chung et al. [2022], plug-and-play priors Graikos et al. [2022], and ILVR Choi et al. [2021]; these inverse-problem pipelines incorporate a physical measurement operator $\mathcal{A}$ (mask, blur, subsampling), which our analysis does *not* directly cover. Gupta et al. [2024] give computational-hardness obstructions orthogonal to our statistical bounds, and classifier-free guidance Ho and Salimans [2022] and the Gaussian-mixture guidance analysis of Wu et al. [2024] remain mixture-level and parametric. What our decomposition does establish is the *specialized* identity $\mathcal{H}^{(i)} \approx q(\cdot \mid \bar{X}_T = \mathbf{x}^{(i)})$ — exactly the posterior induced by treating the forward OU kernel as the likelihood — with W_{init} quantifying that posterior gap. Extending the analysis to arbitrary measurement operators $\mathcal{A}$ (the physical setting of DPS/DDRM/ILVR) requires propagating $\mathcal{A}$ through both the Stein–Poincaré argument for W_{init} and the synchronous coupling for W_{gen}; we mark this as the central open direction.

Alternative samplers and diagnostics. DDIM Song et al. [2021a], flow matching Lipman et al. [2023], rectified flow Liu et al. [2023], consistency models Song et al. [2023], and the EDM design space Karras et al. [2022] reduce function evaluations by different routes. We work throughout with the stochastic reverse SDE Eq. (2): our W_{gen} analysis (Lemma 4) hinges on synchronous coupling of the SDE's Brownian noise, and the probability-flow ODE maps each fixed start to a deterministic output (ODE output std $< 10^{-10}$ on MNIST while the SDE produces diverse digits from the same start; Section E.1); deterministic and velocity-field samplers are out of scope. Foundational score-matching is due to Hyvärinen and Dayan [2005], Vincent [2011]; kernel Kantorovich duality and optimal transport tools come from Villani et al. [2009]. For OOD detection we compare against Graham et al. [2023] and baselines rooted in ensemble intuitions Lakshminarayanan et al. [2017], Wenzel et al. [2020].

3 Setting and Preliminaries

Forward and reverse SDEs. We work with a forward SDE family with linear drift, parametrised by $\rho_{\text{fwd}} : [0,T] \to \mathbb{R}_{\geq 0}$ and $g : [0,T] \to \mathbb{R}_{>0}$:

$$\mathrm{d}\bar{X}_t = -\rho_{\text{fwd}}(t)\,\bar{X}_t\,\mathrm{d}t + g(t)\,\mathrm{d}B_t, \qquad \bar{X}_0 \sim q, \tag{1}$$

forward kernel $\bar{X}_t \mid x_0 \sim \mathcal{N}(\mathfrak{m}_t x_0, \mathfrak{s}_t^2 I_d)$ with $\mathfrak{m}_t := \exp(-\int_0^t \rho_{\text{fwd}})$ and $\mathfrak{s}_t^2 := \int_0^t (\mathfrak{m}_t/\mathfrak{m}_s)^2\, g(s)^2\,\mathrm{d}s$. The abstract family contains OU, VP, Sub-VP, and VE as rows of Table 1; under Assumption 8 our formal bounds apply to OU and VP-fixed-β, while VP-var-β, Sub-VP, and VE are illustrative-only. DDPM Ho et al. [2020] and NCSN Song and Ermon [2019] are discrete-time instantiations of VP and VE. Anderson's reverse Anderson [1982] drives the SGM Song et al. [2021b]

Table 1: Forward-SDE family parameters. Under Assumption 8, g is required constant on $[t_{\min}, T]$, so the formal bounds apply to OU and VP-fixed-β; the time-varying-g rows (VP-var-β, Sub-VP, VE) define the abstract family but lie outside the formal scope of Theorem 1. An extension route is outlined in Appendix A.

SDE	$\rho_{\text{fwd}}(t)$	$g(t)^2$	$\mathfrak{m}_t$	$\mathfrak{s}_t^2$
OU Uhlenbeck and Ornstein [1930]	1	2	e^{-t}	$1-e^{-2t}$
VP Ho et al. [2020]	$\beta(t)/2$	$\beta(t)$	$\sqrt{\bar{\alpha}_t}$	$1-\bar{\alpha}_t$
Sub-VP Song et al. [2021b]	$\beta(t)/2$	$\beta(t)(1-\bar{\alpha}_t^2)$	$\sqrt{\bar{\alpha}_t}$	$(1-\bar{\alpha}_t)^2$
VE Song and Ermon [2019]	0	$2\sigma(t)\dot{\sigma}(t)$	1	$\sigma(t)^2$

with learned $s_\theta \approx \nabla \ln q_\tau$. With $\rho_k := \rho_{\text{fwd}}(T-kh)$, $g_k := g(T-kh)$, the Euler–Maruyama discretisation with drift and diffusion frozen at $T-kh$ on $[kh,(k+1)h]$, step $h = (T-t_{\min})/N$, is

$$Y_{(k+1)h} = (1+\rho_k h)Y_{kh} + g_k^2 h\, s_\theta(Y_{kh}, T-kh) + g_k\sqrt{h}\,\xi_k, \quad \xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d). \tag{2}$$

Early stopping at $T_{\text{stop}} = T - t_{\min}$ is required because $1/\mathfrak{s}_t^2 \to \infty$ as $t \to 0$.

Key distributions. q is the data law (mean μ_q, covariance Σ_q); $\mathcal{H}^{(i)} := \mathrm{law}(Y_{T_{\mathrm{stop}}} \mid Y_0 = \mathbf{x}^{(i)})$ the point-initialized SGM law; $q^{(i)} := q(\cdot \mid \bar{X}_T = \mathbf{x}^{(i)})$ the true Bayesian posterior; and $q^{(i)}_{t_{\min}} := \mathrm{law}(X_{T_{\mathrm{stop}}} \mid X_0 = \mathbf{x}^{(i)})$ the exact-score early-stopped counterpart. The standard prior-initialized SGM output is $p_T = \int \mathcal{H}^{(\mathbf{z})} \gamma_d(\mathrm{d}\mathbf{z})$.

True-score contraction. The following quantity controls the damping in W_{gen} (with $\rho := \rho_{\mathrm{fwd}}(T-t)$, $g := g(T-t)$):

$$\mu_{\mathrm{eff}}(t) := -\rho - g^2 S(t), \quad S(t) := \sup_{x \neq y} \frac{\langle x - y, \nabla \ln q_{T-t}(x) - \nabla \ln q_{T-t}(y) \rangle}{\|x - y\|^2}. \tag{3}$$

Assumptions (full statements, effective score error $\tilde{\varepsilon}_{\mathrm{score}}$, and discussion of necessity in Section A):

#	Name	Formal statement	Used in
1	Lipschitz score	$\exists\, L < \infty :\ \|\nabla \ln q_t(x) - \nabla \ln q_t(y)\| \le L\|x - y\|$, $t \ge t_{\min}$	W_{gen}
2	Time-Lipschitz score	$\exists\, L_t < \infty :\ \|\nabla \ln q_t(x) - \nabla \ln q_s(x)\| \le L_t(1 + \|x\|)\|t - s\|$, $s, t \ge t_{\min}$	W_{gen} (II-b_1)
3	Sub-Gaussian data	$\exists\, K_q < \infty :\ \mathbb{E}_q \exp\big(\|x_0\|^2/(2K_q^2)\big) \le 2$	W_{init}
4	Localized score error	$\sup_{\|x\| \le R} \|s_\theta(x, \tau) - \nabla \ln q_\tau(x)\| \le \varepsilon_{\mathrm{score}}(\tau)$	W_{gen} (II-b_2)
5	s_θ Lipschitz in space	$s_\theta(\cdot, t)$ is L_s-Lipschitz	Thm. 2
6	Step-size cap	$h \lesssim 1/\max(\rho_{\max},\, g_{\max}^2 L,\, g_{\max}^2 L_s)$	C1, B3
7	Poincaré inequality for q	$\exists\, C_P < \infty :\ \mathrm{Var}_q(f) \le C_P\, \mathbb{E}_q \|\nabla f\|^2$	W_{init}
8	SDE regularity	$\rho_{\mathrm{fwd}} \in C^1([t_{\min}, T])$ with $0 \le \rho_{\min} \le \rho_{\mathrm{fwd}} \le \rho_{\max} < \infty$, $\sup \lvert\dot{\rho}_{\mathrm{fwd}}\rvert =: \rho'_{\max} < \infty$; *and g is constant on* $[t_{\min}, T]$, $0 < g_{\min} \le g \le g_{\max} < \infty$ (so $g'_{\max} = 0$)	all

The assumptions above constrain the model. The next definition constrains the *starting point* $\mathbf{x}^{(i)}$ and is what gives the per-start theorems their constants; it is checked per-$\mathbf{x}^{(i)}$ rather than once globally.

Definition 1 (Initialization regularity)**.** *A starting point $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is* initialization-regular *when, with $\eta_T := \mathfrak{m}_T/\mathfrak{s}_T^2$ (OU: $\eta_T = e^{-T}/\mathfrak{s}_T^2$),*

$$\eta_T \, \|\mathbf{x}^{(i)}\| \, K_q \le 1. \tag{4}$$

This bounds the exponential-tilt parameter of the Bayesian posterior $q^{(i)}$ and is what gives Lemma 5 its constants.

4 Plausibility via Coupling

Fig. 1 illustrates the DDS phenomenon on three 2D datasets: the shown per-start laws resemble q; Theorem 1 below separates the three errors governing this comparison.

4.1 Statement

Theorem 1 (Plausibility of point-initialized sampling)**.** *Suppose Assumptions 1 to 8 hold with $L \ge 1$, set $T_{\mathrm{stop}} := T - t_{\min}$, and fix any $\mathbf{x}^{(i)} \in \mathbb{R}^d$ satisfying Eq. (4). Then*

$$\boxed{W_1\big(q, \mathcal{H}^{(i)}\big) \ \le\ W_{\mathrm{init}}(\mathbf{x}^{(i)}) \ +\ W_{\mathrm{stop}} \ +\ W_{\mathrm{gen}}} \tag{5}$$

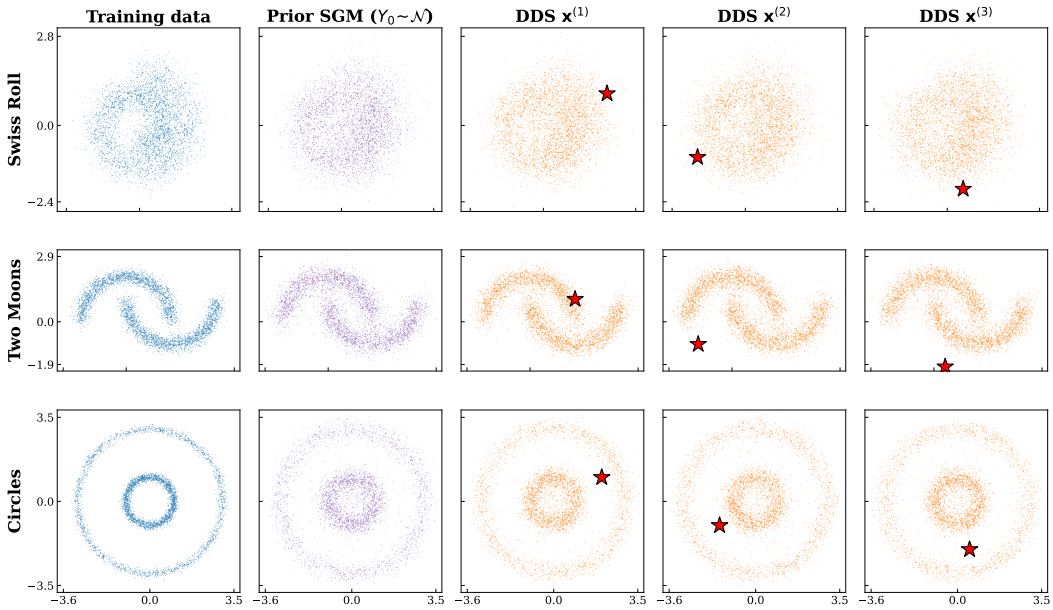

Figure 1: The DDS concept on three 2D datasets (Swiss Roll, Two Moons, Circles). Each column uses a different fixed starting point $\mathbf{x}^{(i)}$ (red star); the columns shown are training data, the prior-initialized SGM output, and three DDS outputs at distinct starts. The shown stochastic outputs resemble the data distribution. Theorem 1 gives a finite-T bound; convergence requires joint control of early stopping, score error, step size, and the generation-error envelope as T grows. The deterministic ODE collapses each per-start law to a single point instead (not shown; see Section E.1, Fig. 4).

with three complementary terms:

$$W_{\text{init}}(\mathbf{x}^{(i)}) \le \eta_T \sqrt{C_P \, (\mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q)^\intercal \Sigma_q \, (\mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q)} + \eta_T^{3/2} \sqrt{C'_R \, C_P} \, \|\mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q\|^{3/2} \, K_q^{3/2} + \eta_T \, \mathfrak{s}_T \sqrt{C''_R \, C_P} \, K_q, \tag{6}$$

$$W_{\text{stop}} \le \sqrt{\left((1 - \mathfrak{a}_{t_{\min}}) \sqrt{M_0^{(i)}} + \mathfrak{b}_{t_{\min}} \|\mathbf{x}^{(i)}\|\right)^2 + \mathfrak{g}^2_{t_{\min}} \, d}, \tag{7}$$

$$W_{\text{gen}}^2 \le 4 \, g_{\max}^4 \, D_\mu \int_0^{T_{\text{stop}}} E(s) \, \mathrm{d}s, \quad D_\mu := \int_0^{T_{\text{stop}}} \exp\Big(-2 \int_s^{T_{\text{stop}}} \mu_{\text{eff}}(u) \, \mathrm{d}u\Big) \mathrm{d}s, \tag{8}$$

The constants are: $\eta_T := \mathfrak{m}_T / \mathfrak{s}_T^2$ (Bayesian envelope); $M_0^{(i)} := \mathbb{E}_{q^{(i)}} \|x_0\|^2 \le C^\star K_q^2$ (posterior second moment); $(\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}})$ are the bridge Gaussian-conditioning parameters (Eq. (35)); and C'_R, C''_R are Taylor-remainder constants (Eq. (27)).

The per-step error budget $E(s)$ (with $\rho_k := \rho_{\text{fwd}}(T - kh)$, $g_k := g(T - kh)$) collects four contributions under Assumption 8 (score error, spatial discretization, temporal frozen-drift, frozen-drift coefficient gap):

$$E(s) := 4\tilde{\varepsilon}^2_{\text{score}} + \big(4(\rho_k/g^2)^2 + 4L^2\big)\big(g_k^2 \, dh + C_{\text{drift}} h^2\big) + 8L_t^2(1 + M_1^Y)h^2 + 4(\rho'_{\max}/g^2_{\min})^2 \, M_1^Y \, h^2, \tag{9}$$

with C_{drift} and M_1^Y defined in Section B, and $\rho'_{\max}$ from Assumption 8 controlling the frozen-drift coefficient gap (active only when ρ_{fwd} is time-varying; $\rho'_{\max} = 0$ for OU and VP-fixed-β). The frozen-diffusion gap and the score-coefficient residual r_g are identically zero under Assumption 8 ($g'_{\max} = 0$) and do not appear in $E(s)$.

Finiteness of the bound. *Under the stated assumptions, Lemma 5 bounds $M_0^{(i)}$, Eqs. (45) and (46) bound M_1^Y, C_{drift}, and $D_\mu \le T_{\text{stop}} e^{2(\rho_{\max} + g^2_{\max} L) T_{\text{stop}}}$. Thus all three costs are finite at each admissible finite horizon. Finiteness alone does not guarantee a small bound.*

Corollary 1 (OU summary)**.** *Under Theorem 1, for OU, $T \geq 1$, $0 < t_{\min} \leq 1/2$, and $\sup_s \tilde{\varepsilon}_{\text{score}}(s) \leq \bar{\varepsilon}$, the bound simplifies to*

$$W_1(q, \mathcal{H}^{(i)}) \lesssim \underbrace{C_{\text{init}} e^{-T}}_{\textit{conditioning}} + \underbrace{t_{\min}(\|\mathbf{x}^{(i)}\| + K_q) + \sqrt{d\, t_{\min}}}_{\textit{early stopping}} + \underbrace{C_{\text{reg}} \sqrt{D_\mu T_{\text{stop}}}(\bar{\varepsilon} + \sqrt{dh} + h)}_{\textit{generation}}.$$

Here C_{init} depends on q and the fixed start; C_{reg} collects the regularity and moment envelopes in Eq. (9). Their dependence is detailed in Section A; in particular, C_{reg} need not stay bounded as $t_{\min} \downarrow 0$ or T grows.

4.2 Proof sketch

Full proof in Section B. The proof inserts two intermediate distributions so the triangle inequality splits the error along physically meaningful cuts:

$$q \xrightarrow[W_{\text{init}}]{\text{conditioning}} q^{(i)} \xrightarrow[W_{\text{stop}}]{\text{early stop}} q^{(i)}_{t_{\min}} \xrightarrow[W_{\text{gen}}]{\text{score + EM}} \mathcal{H}^{(i)}.$$

Each arrow isolates a different mechanism. The first quantifies how much fixing $\bar{X}_T = \mathbf{x}^{(i)}$ tilts the data law q via Bayesian conditioning. The second quantifies the cost of stopping the reverse process at $t_{\min} > 0$, which is unavoidable since the score's Lipschitz constant scales as $1/\mathfrak{s}_t^2$ and diverges at the data end. The third aggregates the SGM-side approximations: localized score error, spatial and temporal regularity gaps, and the frozen-drift gap from Euler–Maruyama on $[kh, (k+1)h]$. The contraction rate μ_{eff} controlling the third arrow is related to the pairwise score diagnostic used to interpret Theorem 2.

Step 1. Conditioning cost (Lemma 2). With $v := \mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q$, Bayes writes $q^{(i)} = w \cdot q$ with Radon–Nikodym derivative $w \propto \exp(A - B)$, where $A := \eta_T \langle v, x_0 - \mu_q \rangle$ and $B := \dfrac{1}{2} \eta_T \mathfrak{m}_T \|x_0 - \mu_q\|^2$ are both $\mathcal{O}(\eta_T)$ under Eq. (4); hence $w - 1 = \mathcal{O}(\eta_T)$. Three successive moves — *(i)* Kantorovich duality, Cauchy–Schwarz, and Poincaré (Assumption 7); *(ii)* a Taylor expansion of $w - 1$ in η_T with cubic/quartic remainders absorbed by sub-Gaussian moments (Assumption 3; [Vershynin, 2018, Prop. 2.5.2]); *(iii)* subadditivity of $\sqrt{\cdot}$ on nonnegatives ($\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$ for $a, b, c \geq 0$) — yield

$$\begin{aligned} W_1(q, q^{(i)}) &\leq \sqrt{C_P\, \mathbb{E}_q[(w-1)^2]} \\ &\leq \sqrt{C_P \big(\eta_T^2 v^\mathsf{T} \Sigma_q v + C'_R \eta_T^3 \|v\|^3 K_q^3 + C''_R \eta_T^2 \mathfrak{s}_T^2 K_q^2\big)} \\ &\leq \eta_T \sqrt{C_P\, v^\mathsf{T} \Sigma_q v} + \eta_T^{3/2} \sqrt{C'_R C_P}\, \|v\|^{3/2} K_q^{3/2} + \eta_T \mathfrak{s}_T \sqrt{C''_R C_P}\, K_q, \end{aligned} \tag{10}$$

which is Eq. (6). The non-decaying-shift term is sub-leading for OU/VP at large T but load-bearing for VE, where $\mathfrak{m}_T = 1$ keeps the tilt's slope from decaying.

Step 2. Early-stopping cost (Lemma 3). The targets $q^{(i)}$ and $q^{(i)}_{t_{\min}}$ are both marginals of the joint Gaussian $(\bar{X}_0, \bar{X}_{t_{\min}}, \bar{X}_T)$ under the abstract forward kernel Eq. (1). Schur-complement projection on the joint covariance gives the closed-form bridge

$$\bar{X}_{t_{\min}} \mid (\bar{X}_0 = x_0, \bar{X}_T = \mathbf{x}^{(i)}) \sim \mathcal{N}\big(\mathfrak{a}_{t_{\min}} x_0 + \mathfrak{b}_{t_{\min}} \mathbf{x}^{(i)},\, \mathfrak{g}^2_{t_{\min}} I_d\big), \tag{11}$$

with parameters $(\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}})$ from Eq. (35). Three successive moves — *(i)* $W_1 \leq W_2$ on the synchronous-$\bar{X}_0$ coupling; *(ii)* substitution of Eq. (11) with $\eta \perp \bar{X}_0$ and $\mathbb{E}\|\eta\|^2 = d$; *(iii)* Cauchy–Schwarz and Jensen on the cross-term, with $M_0^{(i)} := \mathbb{E}_{q^{(i)}} \|x_0\|^2 \leq C^\star K_q^2$ from Lemma 5 — yield

$$\begin{aligned} W_1\big(q^{(i)}, q^{(i)}_{t_{\min}}\big)^2 &\leq \mathbb{E}\|\bar{X}_0 - \bar{X}_{t_{\min}}\|^2 \\ &= (1 - \mathfrak{a}_{t_{\min}})^2 M_0^{(i)} + \mathfrak{b}^2_{t_{\min}} \|\mathbf{x}^{(i)}\|^2 + \mathfrak{g}^2_{t_{\min}} d \\ &\quad - 2(1 - \mathfrak{a}_{t_{\min}}) \mathfrak{b}_{t_{\min}} \big\langle \mathbb{E}_{q^{(i)}} \bar{X}_0,\, \mathbf{x}^{(i)} \big\rangle \\ &\leq \big((1 - \mathfrak{a}_{t_{\min}}) \sqrt{M_0^{(i)}} + \mathfrak{b}_{t_{\min}} \|\mathbf{x}^{(i)}\|\big)^2 + \mathfrak{g}^2_{t_{\min}} d, \end{aligned}$$

which gives Eq. (7) after taking the square root. At small $t_{\min}$, $1-\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}}$ are all $\mathcal{O}(t_{\min})$, so $W_{\text{stop}} = \mathcal{O}(\sqrt{t_{\min}})$.

Step 3. Generation error (Lemmas 4 and 5). Run the true reverse X_t (driven by $\nabla \ln q_{T-t}$) and the SGM Y_t (driven by s_θ) on the same Brownian path, both started at $\mathbf{x}^{(i)}$; the noise increments cancel in $Z_t := X_t - Y_t$ ($Z_0 = 0$). Itô on $\|Z_t\|^2$ gives the Lyapunov inequality $\frac{\mathrm{d}}{\mathrm{d}t}\|Z_t\|^2 \le -2\mu_{\text{eff}}(t)\|Z_t\|^2 + 2\langle Z_t, (\mathrm{II})\rangle$, isolating contraction at rate μ_{eff} (Eq. (3)) from the drift-difference (II). Inserting two pivots $\nabla \ln q_{T-t}(Y_{kh})$ and $\nabla \ln q_{T-kh}(Y_{kh})$ telescopes (II) into four pieces:

$$\begin{aligned}(\mathrm{II}) \;=\; &\underbrace{(\rho/g^2)(Y_t - Y_{kh})}_{\text{frozen-drift gap}} + \underbrace{\nabla \ln q_{T-t}(Y_t) - \nabla \ln q_{T-t}(Y_{kh})}_{\text{spatial Lipschitz}} \\ &+ \underbrace{\nabla \ln q_{T-t}(Y_{kh}) - \nabla \ln q_{T-kh}(Y_{kh})}_{\text{temporal Lipschitz}} + \underbrace{\nabla \ln q_{T-kh}(Y_{kh}) - s_\theta(Y_{kh}, T-kh)}_{\text{localized score error}}.\end{aligned} \tag{12}$$

Three pieces are bounded directly by Assumptions 1, 2 and 4; the frozen-drift gap is the per-step trajectory increment from Lemma 5. Three successive moves — *(i)* synchronous coupling and $W_1 \le W_2$; *(ii)* variation-of-constants on the Lyapunov inequality, smoothed at $Z_0 = 0$, followed by Cauchy–Schwarz to identify the damping integral D_μ; *(iii)* the four-way split Eq. (12) with the assumption-driven and Lemma 5-driven bounds, yielding the per-step error budget $E(s)$ of Eq. (9) — yield

$$\begin{aligned} W_1\big(q^{(i)}_{t_{\min}}, \mathcal{H}^{(i)}\big)^2 &\le \mathbb{E}\|Z_{T_{\text{stop}}}\|^2 \\ &\le D_\mu \int_0^{T_{\text{stop}}} g^4(T-s)\,\mathbb{E}\|(\mathrm{II})\|^2\,\mathrm{d}s \\ &\le 4\, g^4_{\max}\, D_\mu \int_0^{T_{\text{stop}}} E(s)\,\mathrm{d}s, \end{aligned}$$

which is Eq. (8). The localised form of Assumption 4 — requiring score control only on $B_R \ni \mathbf{x}^{(i)}$ rather than globally — is what keeps W_{gen} informative when $\mathbf{x}^{(i)}$ lies outside the typical set of q, allowing Theorem 2 to distinguish in-distribution from out-of-distribution starts at the bound level.

5 Diversity

Previous section tells us that there are a certain theoretical similarity of the DDS's density v.s. the training data distribution. But is there any difference between DDS's densities? Fig. 2 illustrates different starting-point imprints on a 4-mode GMM, motivating the comparison of per-start output laws in Theorem 2.

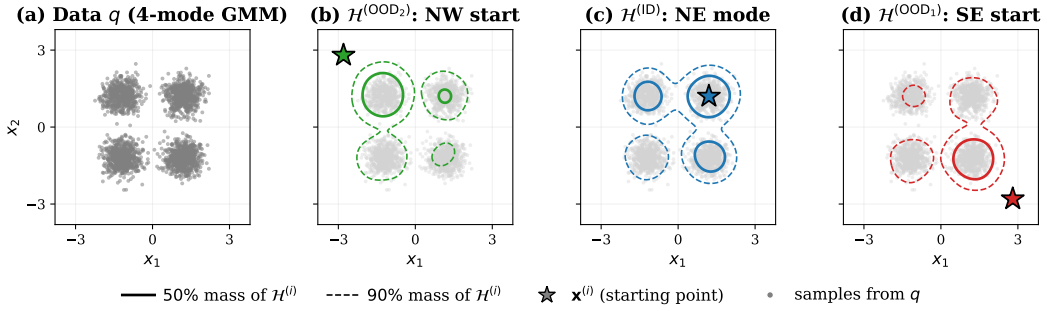

Figure 2: Starting-point imprint on DDS, 4-mode GMM with the analytical true score ($T = 1.5$, reverse OU). **(a)** The data q. **(b)–(d)** Per-starting-point laws $\mathcal{H}^{(i)}$: solid contours enclose 50% of mass, dashed enclose 90%. All three approximately recover q at the 90% level (plausibility), but at the 50% level each concentrates on the mode nearest its starting point $\mathbf{x}^{(i)}$ — two OOD starts at NW (b) and SE (d), one ID start on the NE mode (c). The 50%-region separation across panels is what $\mathrm{TV}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)})$ in Theorem 2 lower-bounds.

5.1 Statement

Run two SGM copies from $\mathbf{x}^{(i)} \neq \mathbf{x}^{(j)}$ with the *same* Brownian motion B_t, and let $Z_{T_{\text{stop}}}(\omega) := Y^{(i)}_{T_{\text{stop}}}(\omega) - Y^{(j)}_{T_{\text{stop}}}(\omega)$. When the mean separation is nonzero, denote by $\bar{e} := \mathbb{E}_\omega[Z_{T_{\text{stop}}}]/\|\mathbb{E}_\omega[Z_{T_{\text{stop}}}]\|$ the direction of maximal mean separation, and let $V := \langle Y^{(i)}_{T_{\text{stop}}}, \bar{e}\rangle$, $U := \langle Y^{(j)}_{T_{\text{stop}}}, \bar{e}\rangle$ be the 1D projections onto $\bar{e}$. Then

$$\delta := \|\mathbb{E}_\omega[Z_{T_{\text{stop}}}]\| = \mathbb{E}[V] - \mathbb{E}[U], \;\; \sigma_V^2 := \mathrm{Var}_\omega(V), \;\; \sigma_U^2 := \mathrm{Var}_\omega(U). \tag{13}$$

Theorem 2 (Diversity of point-initialized sampling)**.** *Under Assumptions 5 and 6, for any $\mathbf{x}^{(i)} \neq \mathbf{x}^{(j)}$ with finite output second moments and $\delta > 0$,*

$$\boxed{\mathrm{TV}\big(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}\big) \;\geq\; 1 - \frac{\sigma_V^2}{\sigma_V^2 + \delta^2/4} - \frac{\sigma_U^2}{\sigma_U^2 + \delta^2/4}} \tag{14}$$

If $\delta = 0$, use the trivial bound $\mathrm{TV} \geq 0$ without normalizing the mean difference. Otherwise, the displayed bound lies in $[-1, 1]$ and is informative (positive) iff $\delta^2 > 4\sigma_V\sigma_U$ (equivalently, $\delta^4 > 16\sigma_V^2\sigma_U^2$).

5.2 Proof sketch

The proof applies Cantelli to the projected output moments to obtain the TV lower bound and its positivity condition; full details are in Section C.

Step 1. Projection onto the mean difference. For $\delta > 0$, choose $\bar{e} := \mathbb{E}_\omega[Z_{T_{\text{stop}}}]/\|\mathbb{E}_\omega[Z_{T_{\text{stop}}}]\|$ and set $V := \langle Y^{(i)}_{T_{\text{stop}}}, \bar{e}\rangle$, $U := \langle Y^{(j)}_{T_{\text{stop}}}, \bar{e}\rangle$. Data processing gives $\mathrm{TV}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) \geq \mathrm{TV}(V, U)$. This direction maximizes the projected mean separation, $\mathbb{E}V - \mathbb{E}U = \delta$, though it need not maximize the variance-dependent Cantelli bound. If $\delta = 0$, use the trivial bound $\mathrm{TV} \geq 0$.

Step 2. Cantelli bounds at the midpoint. Set $m := (\mathbb{E}V + \mathbb{E}U)/2$. Each mean is at distance $\delta/2$ from m, so Cantelli's one-sided Chebyshev inequality gives

$$\mathbb{P}(V \leq m) \;\leq\; \frac{\sigma_V^2}{\sigma_V^2 + \delta^2/4}, \qquad \mathbb{P}(U \geq m) \;\leq\; \frac{\sigma_U^2}{\sigma_U^2 + \delta^2/4}. \tag{15}$$

Step 3. TV lower bound and positivity. Using the event (m, ∞) gives $\mathrm{TV}(V, U) \geq \mathbb{P}(V > m) - \mathbb{P}(U > m) \geq 1 - \mathbb{P}(V \leq m) - \mathbb{P}(U \geq m)$. Substituting Eq. (15) yields Eq. (14). The resulting bound lies in $[-1, 1]$ and is positive exactly when $\delta^2 > 4\sigma_V\sigma_U$. No symmetry assumption is needed; the threshold marks a change in the certificate's sign, without identifying it with contraction or equality of the output laws.

5.3 Three regimes and the plausibility–diversity duality

Theorems 1 and 2 give two seemingly opposing perspectives: the first upper-bounds how far each per-start output law is from the same data distribution q in W_1, while the second lower-bounds how distinguishable two such output laws remain in TV. We examine a possible connection through the score's effect on differences between coupled trajectories.

Lemma 1 (Contraction and separation between starts)**.** *Suppose Assumptions 5, 6 and 8 hold. Run two SGM copies from $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ on the same Brownian path and set $Z_{kh} = Y^{(i)}_{kh} - Y^{(j)}_{kh}$. For $x \neq y$, define*

$$\kappa(x, y, \tau) := -\frac{\langle x - y, g(\tau)^2(s_\theta(x, \tau) - s_\theta(y, \tau))\rangle}{\|x - y\|^2}. \tag{16}$$

κ measures the score's first-order effect on pairwise separation. Write $\rho_k = \rho_{\text{fwd}}(T - kh)$ and $\kappa_{kh} = \kappa(Y^{(i)}_{kh}, Y^{(j)}_{kh}, T - kh)$, with $\kappa_{kh} = 0$ when $Z_{kh} = 0$, and set

$$\Lambda_h^{(ij)}(\omega) := \sum_{k=0}^{N-1} h\big(\kappa_{kh}(\omega) - \rho_k\big), \qquad \varepsilon_{\text{disc}} := \frac{1}{2} T_{\text{stop}} h(\rho_{\max} + g_{\max}^2 L_s)^2. \tag{17}$$

Then every noise realization satisfies

$$\|Z_{T_{\text{stop}}}(\omega)\| \le \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \, \exp\Big[-\Lambda_h^{(ij)}(\omega) + \varepsilon_{\text{disc}}\Big]. \tag{18}$$

For output laws with finite first moments, their mean separation consequently satisfies

$$\begin{aligned} \delta = \|\mathbb{E}_\omega Z_{T_{\text{stop}}}\| &\le \mathbb{E}_\omega \|Z_{T_{\text{stop}}}\| \\ &\le \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \, \mathbb{E}_\omega \exp\Big[-\Lambda_h^{(ij)} + \varepsilon_{\text{disc}}\Big]. \end{aligned} \tag{19}$$

The proof is given in Section C.2.

With other terms fixed, increasing $\kappa_{kh} - \rho_k$ along the coupled paths tightens the upper bound on δ; decreasing it loosens the bound. If this upper bound is at most $2\sqrt{\sigma_V \sigma_U}$, Theorem 2 cannot certify positive TV.

Role of contraction in the plausibility bound. In Theorem 1, contraction controls the error between the ideal trajectory X_t and learned trajectory Y_t from the same start. The two trajectories start together, and score approximation and discretization introduce the errors collected in $E(s)$ (Eq. (9)); the true-score contraction rate μ_{eff} controls their damping. Expanding D_μ in Eq. (8) gives

$$W_{\text{gen}}^2 \le 4 g_{\max}^4 \left(\int_0^{T_{\text{stop}}} E(s)\, \mathrm{d}s \right) \int_0^{T_{\text{stop}}} \exp\left[-2 \int_s^{T_{\text{stop}}} \mu_{\text{eff}}(u)\, \mathrm{d}u \right] \mathrm{d}s. \tag{20}$$

For a step with $x := Y_{kh}^{(i)} \neq y := Y_{kh}^{(j)}$, write $\tau = T - kh$ and define the score error $e_\tau(z) := s_\theta(z,\tau) - \nabla \ln q_\tau(z)$. The definitions in Eqs. (3) and (16) give the exact relation

$$\begin{aligned} \kappa_{kh} - \rho_k = \mu_{\text{eff}}(kh) &+ \underbrace{g_k^2 \left[S(kh) - \frac{\langle x - y, \nabla \ln q_\tau(x) - \nabla \ln q_\tau(y) \rangle}{\|x - y\|^2} \right]}_{A} \\ &- \underbrace{g_k^2 \frac{\langle x - y, e_\tau(x) - e_\tau(y) \rangle}{\|x - y\|^2}}_{B}. \end{aligned} \tag{21}$$

Gaussian is often adopted as initialize sampling Ho et al. [2020], Song et al. [2021a,b], we consider $q_\tau = \mathcal{N}(m_\tau, v_\tau I)$ with $v_\tau > 0$. It results in $\nabla \ln q_\tau(x) - \nabla \ln q_\tau(y) = -(x-y)/v_\tau$, so the pairwise and uniform true-score contraction rates coincide, giving $A = 0$. For distinct $x, y \in B_R$, Assumption 4 gives $\|e_\tau(x) - e_\tau(y)\| \le 2\varepsilon_{\text{score}}(\tau)$, so Cauchy–Schwarz yields

$$|(\kappa_{kh} - \rho_k) - \mu_{\text{eff}}(kh)| = |B| \le \frac{2 g_k^2 \varepsilon_{\text{score}}(\tau)}{\|x - y\|}. \tag{22}$$

Thus, for a well-trained score model whose error $\varepsilon_{\text{score}}(\tau)$ is small enough that $2g_k^2 \varepsilon_{\text{score}}(\tau)/\|x-y\|$ is negligible, $\kappa_{kh} - \rho_k \approx \mu_{\text{eff}}(kh)$.

Three regimes. Let $\bar{\rho}_{\text{fwd}} = T_{\text{stop}}^{-1} \sum_k h \rho_k$ and $\bar{\kappa} = \bar{\rho}_{\text{fwd}} + T_{\text{stop}}^{-1} \mathbb{E}_\omega \Lambda_h^{(ij)}$. These averages summarize the first-order contribution over time and noise realizations (for OU, $\bar{\rho}_{\text{fwd}} = 1$):

- *Contractive* ($\bar{\kappa} > \bar{\rho}_{\text{fwd}}$): favors plausibility under the preceding conditions, but can make Theorem 2 unable to certify that different starting points produce different output distributions.
- *Neutral* ($\bar{\kappa} \approx \bar{\rho}_{\text{fwd}}$): the average net contraction is near zero; this alone establishes neither a small generation error in Theorem 1 nor a positive diversity certificate in Theorem 2.
- *Expansive* ($\bar{\kappa} < \bar{\rho}_{\text{fwd}}$): can weaken the plausibility bound under the preceding conditions; Theorem 2 certifies different output distributions only when $\delta^2 > 4\sigma_V \sigma_U$.

These are diagnostics: the bound on δ depends on $\mathbb{E} e^{-\Lambda_h^{(ij)} + \varepsilon_{\text{disc}}}$, not just the sign of $\mathbb{E}\Lambda_h^{(ij)}$. The output means and variances determine whether the TV certificate is positive.

The threshold horizon $T^\star$. Beyond score contraction, the sampling horizon also affects whether different starting points produce distinguishable output distributions. For a specified family of models and fixed starting points, define

$$T^\star := \inf\{T_{\mathrm{stop}} > 0 : \delta^2 \le 4\sigma_V\sigma_U\}, \tag{23}$$

where $\delta, \sigma_V, \sigma_U$ depend on the sampling horizon, and $\inf \varnothing = +\infty$. For every admissible horizon $0 < T_{\mathrm{stop}} < T^\star$, Theorem 2 certifies $\mathrm{TV}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) > 0$, regardless of the contraction regime. Thus, the output laws cannot become identical before $T^\star$. Reaching or exceeding $T^\star$ does not establish equality or permanent loss of the diversity certificate, since no monotonicity is assumed. This threshold depends on the chosen models and starting points and need not be positive or finite.

On MNIST, stochastic sampling produces diverse recognizable digits from a fixed start, whereas the probability-flow ODE repeats one output (Fig. 4). On CIFAR-10, samples from a fixed noised source image vary in appearance and can change class (Fig. 6). Conditional MNIST further shows that an OOD start can override the requested class at a shorter horizon, with partial recovery of class consistency at a longer horizon (Section E.4). On Swiss Roll, quantitative proxies exhibit the plausibility decomposition's stopping-cost tradeoff and a diversity-certificate crossing while empirical diversity persists (Section E.5).

6 Discussion

Measurement operator $\mathcal{A}$. Extending the Stein–Poincaré argument in Lemma 2 and the synchronous coupling in Lemma 4 to a measurement operator $\mathcal{A}$ (mask, blur, subsampling) would bring the theory to DPS Chung et al. [2023], DDRM Kawar et al. [2022], and ILVR Choi et al. [2021] — the central open direction. Higher-order integrators, classifier-free guidance Ho and Salimans [2022], and velocity-field samplers Lipman et al. [2023], Liu et al. [2023] are concurrent extensions.

Time-varying diffusion coefficient. Under Assumption 8 our formal bounds require g constant on $[t_{\min}, T]$, excluding VP-var-β, Sub-VP, and VE. Appendix A outlines an extension using Gronwall directly on the squared coupling error, retaining the additive Brownian residual in Lemma 4.

The threshold $T^\star$. Eq. (23) marks when the diversity certificate ceases to be positive. Lemma 1 supplies an upper separation envelope, while Theorem 2 follows directly from Cantelli applied to the projected moments. A larger upper envelope does not prove expansion, and a vacuous TV lower bound does not prove identity erasure.

$\bar{\kappa}$ as a test-time diagnostic. For OU and VP-fixed-β, paired runs estimate $\bar{\kappa}$ from two score evaluations per step without q-samples or retraining. Sweep the sampling horizon, inspect the separation envelope, and evaluate $\delta^2/(4\sigma_V\sigma_U)$ to assess the certificate.

Conclusion. Together, Theorems 1 and 2 give the first non-asymptotic theory of point-initialized diffusion — motivated by SDEdit and OOD reconstruction, with measurement-guided posterior sampling an open extension.

Acknowledgments and Disclosure of Funding

Hady W. Lauw gratefully acknowledges the support by the Lee Kong Chian Fellowship awarded by Singapore Management University.

References

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=MhK5aXo3gB. Expert Certification.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14367–14376, 2021.

M Bishop Christopher. Pattern recognition and machine learning, 2006.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2948–2957, 2023.

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022.

Shivam Gupta, Ajil Jalal, Aditya Parulekar, Eric Price, and Zhiyang Xun. Diffusion posterior sampling is computationally intractable. In *International Conference on Machine Learning*, pages 17020–17059. PMLR, 2024.

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

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

Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34: 22863–22876, 2021.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in neural information processing systems*, 34:14938–14954, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

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.

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.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2006.

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 and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vaRCHVj0uGI.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252, 2023.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pages 10248–10259. PMLR, 2020.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *International Conference on Machine Learning*, pages 53291–53327. PMLR, 2024.

A Assumption Details

Where fixed initialization arises. SDEdit Meng et al. [2022] first noises an input image, then denoises it stochastically. Conditional on one noisy realization, its unguided reverse run has the DDS structure with $T = t_{\text{start}}$; applying our bounds additionally requires the stated schedule and regularity assumptions. OOD reconstruction Graham et al. [2023] similarly conditions on an individual test input. Fixing this realization leaves reverse-path randomness, unlike fixing all random seeds. DPS/DDRM/ILVR also incorporate measurement-dependent updates, which the present theorems do not cover.

Gaussian initialization regularity. For $G \sim \mathcal{N}(0, I_d)$, Gaussian concentration gives $\Pr(\|G\| > \sqrt{d} + s) \le e^{-s^2/2}$. In OU, $\eta_T \le 2e^{-T}$ for $T \ge 1$, so

$$T \ge \max\{1, \log(2K_q(\sqrt{d} + s))\} \quad \Longrightarrow \quad \Pr\big(\eta_T \|G\| K_q \le 1\big) \ge 1 - e^{-s^2/2}.$$

With $s = \sqrt{d}$ and $K_q = \Theta(\sqrt{d})$, this sufficient horizon is $\log d + O(1)$ at probability $1 - e^{-d/2}$. It verifies Definition 1, not the other theorem assumptions.

Constants in the OU summary. For Corollary 1, let $R_i = \|\mathbf{x}^{(i)}\| + \|\mu_q\|$. One may take

$$C_{\text{init}} = \sqrt{C_P}\big[2\sqrt{\|\Sigma_q\|_{\text{op}}} R_i + 2^{3/2}\sqrt{C'_R} R_i^{3/2} K_q^{3/2} + 2\sqrt{C''_R} K_q\big].$$

Indeed, substitute $\eta_T \le 2e^{-T}$ into Eq. (6) and use $e^{-3T/2} \le e^{-T}$. The OU bridge coefficients satisfy $1 - \mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}} \lesssim t_{\min}$ on the stated interval. Finally, Eq. (9) gives $E(s) \lesssim C^2_{\text{reg}}(\bar{\varepsilon}^2 + dh + h^2)$, where C_{reg} includes $L, L_t, C_{\text{drift}}, M_1^Y$ and the fixed OU coefficients. Integrating Eq. (8) proves the summary. An asymptotic convergence claim requires all three displayed terms to vanish jointly.

Assumption 1 role of early stopping. Score regularity can deteriorate near the data end: for $q = \delta_0$ under OU, $\nabla \ln q_t(x) = -x/\mathfrak{s}_t^2$ has modulus $1/\mathfrak{s}_t^2$. This is not a universal Lipschitz bound for Gaussian mixtures or all sub-Gaussian laws. For example, $q = \mathcal{N}(0, I_d)$ is stationary under OU and has modulus 1. We assume finite L on $[t_{\min}, T]$ explicitly; smaller $t_{\min}$ can worsen L, L_t and the drift envelope.

Assumption 2 linear-growth factor. The $(1 + \|x\|)$ factor is necessary for non-stationary q: for $q = \delta_0$, $\partial_t \nabla \ln q_t(x) = \dfrac{2\mathfrak{m}_t^2 \rho_{\text{fwd}}(t)}{\mathfrak{s}_t^4} x$ (OU: $\dfrac{2e^{-2t}}{\sigma_t^4} x$), which is linear in $\|x\|$. A constant bound is false.

Assumption 3 sub-Gaussian implication. From $\mathbb{E}_q[\exp(\|x_0\|^2/(2K_q^2))] \le 2$, Markov's inequality and layer-cake integration give $\mathbb{E}_q[\|x_0 - \mu_q\|^p] \lesssim p^{p/2} K_q^p$ for every $p \ge 1$ (the standard sub-Gaussian moment scaling, [Vershynin, 2018, Prop. 2.5.2]; the absolute constant is absorbed into C'_R in Eq. (27)). This is used in the proof of Lemma 2 to bound the cubic and quartic Taylor remainders of $e^{A-B} - 1$, which involve $\mathbb{E}_q[\|x_0 - \mu_q\|^3]$ and $\mathbb{E}_q[\|x_0 - \mu_q\|^4]$.

Assumption 4 Lipschitz-tail inequality. Outside B_R, triangle plus Assumptions 1 and 5 gives $\|s_\theta(x, \tau) - \nabla \ln q_\tau(x)\| \le \varepsilon_{\text{score}}(\tau) + (L + L_s)(\|x\| - R)_+$, so the score error grows at most linearly in the tail. The *effective* score error is

$$\tilde{\varepsilon}_{\text{score}}(\tau)^2 := 2\varepsilon_{\text{score}}(\tau)^2 + 2(L + L_s)^2\, \mathbb{E}[(\|Y_{kh}\| - R)_+^2],\ k = (T - \tau)/h; \tag{24}$$

this localized form is what enables OOD divergence in Theorem 2.

Assumption 5 role and typical values. The spatial Lipschitz constant L_s of the learned score s_θ caps how fast its output can grow in x. It enters: (i) the kappa bound $|\kappa| \le g^2 L_s$ used in Lemma 1 (under Eq. (16)'s g^2-embedded form); (ii) the SGM-moment Gronwall envelope $\alpha = 1 + 6\rho_{\max} + 4g^4_{\max} L_s^2$ in Lemma 4 step (v.b); (iii) the linear-tail extension of Assumption 4 via the $L + L_s$ coefficient in Eq. (24). For a composition of affine layers and globally Lipschitz activations, $L_s \le \prod_\ell \|W_\ell\|_{\text{op}} \operatorname{Lip}(\phi_\ell)$; a residual block $x + f(x)$ contributes at most $1 + \operatorname{Lip}(f)$. Spectral normalization can control the matrix factors. These bounds must hold uniformly in time and do not automatically cover arbitrary attention or normalization layers. Finite L_s permits an admissible step size, but large moduli can make the bound loose. Sampled Jacobian norms are diagnostics, not global certificates. The localized score-error condition Eq. (24) does not remove the global Lipschitz assumptions.

Assumption 7 Poincaré constant for q. This assumption enters only the Kantorovich–Poincaré step Eq. (25) for W_{init}. If $q \propto e^{-V}$ and $\nabla^2 V \succeq mI$ with $m > 0$, then $C_P \le 1/m$; for a nondegenerate Gaussian, $C_P = \|\Sigma_q\|_{\text{op}}$. Compact support alone is insufficient: domain and measure hypotheses are needed for diameter bounds. For multimodal mixtures, component scales, weights, and separation affect C_P, potentially severely. It is a property of q, independent of the sampler. Replacing Poincaré by another transport or functional inequality would require a new conditioning-cost argument; the other two costs and the Cantelli result do not use C_P. We leave this relaxation open.

Dimension dependence. For isotropic sub-Gaussian data with uniformly controlled one-dimensional tails, the norm-moment scale is $K_q = O(\sqrt{d})$. The Gaussian-start condition above then needs a logarithmic horizon. The Taylor-remainder constants C'_R, C''_R in Eq. (27) depend on $\mathfrak{m}_T^2 K_q^2/\mathfrak{s}_T^2$ and can therefore also depend on dimension through K_q. The OU stopping term contains $\sqrt{d\, t_{\min}}$, suggesting $t_{\min} = O(1/d)$ for dimension-uniform stopping control; this can worsen score regularity and the step-size requirement. The dependence of C_P, L, L_t, L_s on dimension and geometry must remain explicit. We do not certify these population or global constants for MNIST/CIFAR-10.

Assumption 8 SDE coverage. The constant-g requirement restricts the stated theorem to constant-diffusion schedules.

- *In scope.* OU ($\rho_{\text{fwd}} \equiv 1$, $g \equiv \sqrt{2}$) and VP-fixed-β ($\rho_{\text{fwd}} \equiv \beta/2$, $g \equiv \sqrt{\beta}$): $\rho'_{\max} = g'_{\max} = 0$, and the frozen-diffusion gap $4d(g'_{\max})^2 h^2$ vanishes from Eq. (9).
- *Out of formal scope (time-varying g).* VP with linear/cosine $\beta(t)$, Sub-VP, and VE ($g^2 = 2\sigma\dot{\sigma}$). The triangle decomposition $W_{\text{init}} + W_{\text{stop}} + W_{\text{gen}}$ still applies, but the generation-error envelope must retain the diffusion and score-coefficient residuals.

$\rho'_{\max}$ enters Eq. (9) only through the deterministic-drift gap $4(\rho'_{\max}/g^2_{\min})^2 M_1^Y h^2$; this is well-handled even when ρ_{fwd} time-varies, so we keep $\rho'_{\max} \ge 0$ in Assumption 8.

Possible extension to time-varying diffusion. A possible route to the remaining rows of Table 1 is to retain the errors caused by varying the diffusion coefficient within each sampling step and apply Gronwall directly to the squared coupling error. This would modify the generation-error bound while preserving the three-term plausibility decomposition. The extension would require suitable schedule regularity; the stated Theorem 1 retains the constant-g assumption.

Assumption 6 cap derivation. The cap $h \le 1/\max(\rho_{\max}, g^2_{\max} L, g^2_{\max} L_s)$ comes from requiring $|\Delta_k| < 1$ in the C1 Euler bracket (Eq. (55)): linearising $(1+\rho_k h)^2 - 2h(1+\rho_k h)\kappa_{kh} + g_k^4 h^2 L_s^2$ gives leading-order $1 + 2h(\rho_k - \kappa_{kh})$, with $|\kappa_{kh}| \le g_k^2 L_s$ from the embedding in Eq. (16) and Assumption 5. OU $\rho_{\max} = 1$, $g^2_{\max} = 2$ gives $h \le 1/\max(1, 2L, 2L_s)$ — a constant factor looser than the original $h \lesssim 1/\max(L, L_s)$ for $L, L_s \ge 1$, invisible at typical step sizes ($h \sim 10^{-3}$).

B Full Proof of Theorem 1

We prove Theorem 1 by combining Lemmas 2 to 4, one per term in the $W_{\text{init}} + W_{\text{stop}} + W_{\text{gen}}$ triangle decomposition; the supporting moment bounds of Lemma 5 feed Lemmas 3 and 4.

Throughout, $\rho := \rho_{\text{fwd}}(T-t)$ and $g := g(T-t)$ denote the slice-evaluated forward-SDE coefficients; on the Euler grid $\rho_k := \rho_{\text{fwd}}(T-kh)$, $g_k := g(T-kh)$. The forward kernel mean-attenuation $\mathfrak{m}_t$, marginal variance $\mathfrak{s}_t^2$, conditional variance $\mathfrak{s}^2_{t_{\min},T} := \int_{t_{\min}}^T (\mathfrak{m}_T/\mathfrak{m}_s)^2 g(s)^2\, \mathrm{d}s$, and bridge parameters

$$\mathfrak{a}_{t_{\min}} = \frac{\mathfrak{m}_{t_{\min}} \mathfrak{s}^2_{t_{\min},T}}{\mathfrak{s}_T^2}, \qquad \mathfrak{b}_{t_{\min}} = \frac{\mathfrak{s}^2_{t_{\min}} (\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})}{\mathfrak{s}_T^2}, \qquad \mathfrak{g}^2_{t_{\min}} = \frac{\mathfrak{s}^2_{t_{\min}} \mathfrak{s}^2_{t_{\min},T}}{\mathfrak{s}_T^2}$$

are derived from the joint Gaussian $(\bar{X}_0, \bar{X}_{t_{\min}}, \bar{X}_T)$ in Lemma 3, step (i). When ρ_{fwd}, g are time-varying, the frozen-drift expansion in Lemma 4 step (i) produces two gap pieces in (II-0): a coefficient gap $((\rho-\rho_k)/g^2) Y_{kh}$ with $|\rho - \rho_k| \le \rho'_{\max} h$, and the Brownian-cancellation residual (II-c) $:= (g(T-t) - g_k)\, \mathrm{d}B_t$ with quadratic variation $\le d(g'_{\max} h)^2$ per step. Both vanish for constant-coefficient schedules ($\rho'_{\max} = g'_{\max} = 0$).

Notation for this proof	
W_{init}	Bayesian-posterior gap, $W_1(q, q^{(i)})$ (Eq. (6))
W_{stop}	early-stopping cost, $W_1(q^{(i)}, q^{(i)}_{t_{\min}})$ (Eq. (7))
W_{gen}	generation error from synchronous coupling (Eq. (8))
D_μ	damping integral (Eq. (8))
$E(s)$	per-step error budget
$\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}}$	bridge Gaussian-conditioning parameters (Eq. (35))
$w(x_0)$	Radon–Nikodym derivative $\mathrm{d}q^{(i)}/\mathrm{d}q = q_T(\mathbf{x}^{(i)} \mid x_0)/q_T(\mathbf{x}^{(i)})$, $\mathbb{E}_q[w] = 1$
A, B	substitution variables in $\ln w$ (proof of Lemma 2), $A := \eta_T\langle \mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q, x_0 - \mu_q\rangle$, $B := \frac{1}{2}\eta_T\mathfrak{m}_T\|x_0 - \mu_q\|^2$
$\ell(x_0)$	linear part of the tilt, $\ell(x_0) := \langle \mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q,\, x_0 - \mu_q\rangle$ (proof of Lemma 2)
$R_w,\, R$	Lagrange and combined Taylor remainders in the expansion of $w - 1$ (proof of Lemma 2)
D	denominator's $\mathcal{O}(\eta_T^2)$ correction, $1 + D := \mathbb{E}_q[e^{A-B}]$ (proof of Lemma 2)
$v_\epsilon(t)$	smoothing surrogate $\sqrt{\|Z_t\|^2 + \epsilon^2}$ used to handle $Z_0{=}0$ (proof of Lemma 4)
C_R, C_w	Taylor remainder constants for $\mathbb{E}_q[R^2]$ and $\mathbb{E}_q[R_w^2]$
$C'_R,\, C''_R$	Taylor remainder constants for $\mathbb{E}_q[(w-1)^2]$ (Eq. (27)); C''_R collects the $\mathfrak{s}_T$-dependent (non-decaying-shift) contributions
C_{drift}	frozen-drift envelope (proof of Lemma 4)
M_1^Y	Gronwall envelope on $\mathbb{E}\|Y_t\|^2$ (proof of Lemma 4)
K_0, α	Gronwall constants in the SGM moment recursion
$\mathfrak{a}_k, \mathfrak{b}_k, \mathfrak{g}_k^2$	per-step bridge coefficients (≤ 1), Eq. (35) with $t_{\min} \to kh$ (proof of Lemma 5)
$\mu_{q^{(i)}},\, \mathrm{Cov}_{q^{(i)}}$	posterior mean / covariance (proof of Lemma 5)
$C^\star$	absolute constant, $\mathbb{E}_{q^{(i)}}\|x_0\|^2 \le C^\star K_q^2$ (Eq. (51))
M_2^X	second-moment envelope $\mathbb{E}\|X_{kh}\|^2 \le 2\|\mathbf{x}\|^2 + 2C^\star K_q^2 + d$
M_{drift}	drift envelope $(2\rho_{\max}^2 + 4g_{\max}^4 L^2)M_2^X + 4g_{\max}^4 \tilde{K}_q^2/\mathfrak{s}^4_{t_{\min}}$

B.1 Lemmas

Lemma 2 (Bayesian-conditioning cost)**.** *Suppose Assumptions 3, 7 and 8 hold and Eq.* (4) *is satisfied. Set* $\eta_T := \mathfrak{m}_T/\mathfrak{s}_T^2$, $v := \mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q$. *Then*

$$W_1(q, q^{(i)}) \le \eta_T\sqrt{C_P\, v^\mathsf{T}\Sigma_q v} \;+\; \eta_T^{3/2}\sqrt{C'_R\, C_P}\,\|v\|^{3/2}\, K_q^{3/2} + \eta_T\,\mathfrak{s}_T\sqrt{C''_R\, C_P}\,K_q,$$

where C'_R, C''_R *are Taylor-remainder constants. We denote the right-hand side* $W_{\text{init}}(\mathbf{x}^{(i)})$ *throughout. The third term is the non-decaying-shift contribution:* $\mathcal{O}(\eta_T)$ *for OU/VP (sub-leading to the* $\eta_T^{3/2}$ *remainder when* $\|v\| \ge 1$*) and* $\mathcal{O}(K_q/\sigma(T))$ *for VE (the genuine VE correction at* $\mathfrak{m}_T = 1$*).*

Proof. **Reduction to a variance bound.** By Bayes, $q^{(i)}(x_0) = w(x_0)q(x_0)$ with $w(x_0) := q_T(\mathbf{x}^{(i)} \mid x_0)/q_T(\mathbf{x}^{(i)})$, so $\mathbb{E}_q[w] = 1$. By Kantorovich duality Villani et al. [2009]:

$$\begin{aligned}
W_1(q, q^{(i)}) &= \sup_{\varphi \in \mathrm{Lip}_1} \left| \mathbb{E}_{q^{(i)}}[\varphi] - \mathbb{E}_q[\varphi] \right| && \text{(Kantorovich)} \\
&= \sup_{\varphi \in \mathrm{Lip}_1} |\mathbb{E}_q[w\,\varphi] - \mathbb{E}_q[\varphi]| && \text{(Bayes)} \\
&= \sup_{\varphi \in \mathrm{Lip}_1} |\mathbb{E}_q[(w-1)\,\varphi]| \\
&= \sup_{\varphi \in \mathrm{Lip}_1} \left| \mathbb{E}_q[(w-1)\,(\varphi - \mathbb{E}_q[\varphi])] + \mathbb{E}_q[\varphi] \cdot \underbrace{\mathbb{E}_q[w-1]}_{=\,0} \right| \\
&= \sup_{\varphi \in \mathrm{Lip}_1} |\mathbb{E}_q[(w-1)\,\bar{\varphi}]| && (\bar{\varphi} := \varphi - \mathbb{E}_q[\varphi]).
\end{aligned}$$

Then:

$$\begin{aligned}
|\mathbb{E}_q[(w-1)\,\bar{\varphi}]| &\le \sqrt{\mathbb{E}_q[(w-1)^2]} \cdot \sqrt{\mathbb{E}_q[\bar{\varphi}^{\,2}]} && \text{(Cauchy–Schwarz)} \\
&= \sqrt{\mathbb{E}_q[(w-1)^2]} \cdot \sqrt{\mathrm{Var}_q[\varphi]} \\
&\le \sqrt{\mathbb{E}_q[(w-1)^2]} \cdot \sqrt{C_P\, \mathbb{E}_q[\|\nabla \varphi\|^2]} && \text{(Assumption 7)} \\
&\le \sqrt{C_P\, \mathbb{E}_q[(w-1)^2]} && (\varphi \in \mathrm{Lip}_1).
\end{aligned}$$

The right-hand side is φ-free, so the sup collapses:

$$W_1(q, q^{(i)}) \;\le\; \sqrt{C_P\, \mathbb{E}_q[(w-1)^2]}. \tag{25}$$

Remark (χ^2 formulation). Here $\mathbb{E}_q[(w-1)^2] = \chi^2(q^{(i)} \| q)$, so Eq. (25) reads $W_1(q, q^{(i)}) \le \sqrt{C_P \chi^2(q^{(i)} \| q)}$. This divergence is well-defined and finite: for $\mathfrak{s}_T^2 > 0$, the Gaussian likelihood lies in $(0, M_T]$, where $M_T = (2\pi\mathfrak{s}_T^2)^{-d/2}$. Thus $Z = q_T(\mathbf{x}^{(i)}) > 0$, $q^{(i)} \ll q$, and $w \le M_T/Z$, giving $\chi^2(q^{(i)} \| q) \le M_T/Z - 1 < \infty$. No Lebesgue density of q is needed; the transport inequality still uses Assumption 7.

The remainder of the proof bounds $\mathbb{E}_q[(w-1)^2]$.

Computing $\mathbb{E}_q[(w-1)^2]$.

(i) Substitution. Define $A := \eta_T\, \ell(x_0)$ and $B := \dfrac{1}{2}\eta_T \mathfrak{m}_T \|x_0 - \mu_q\|^2$ with $\ell(x_0) := \langle \mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q,\, x_0 - \mu_q \rangle$ (recall $\eta_T = \mathfrak{m}_T/\mathfrak{s}_T^2$ from the lemma statement; we use ℓ rather than g to avoid collision with the SDE diffusion coefficient), and let C collect the x_0-independent factors,

$$C \;:=\; \frac{(2\pi\mathfrak{s}_T^2)^{-d/2}}{q_T(\mathbf{x}^{(i)})} \cdot \exp\!\left(-\frac{\|\mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q\|^2}{2\mathfrak{s}_T^2} \right).$$

Plugging in the forward kernel $q_T(\cdot \mid x_0) = \mathcal{N}(\mathfrak{m}_T x_0, \mathfrak{s}_T^2 I_d)$, Bayes gives

$$w(x_0) = \frac{q_T(\mathbf{x}^{(i)} \mid x_0)}{q_T(\mathbf{x}^{(i)})} = \frac{(2\pi\mathfrak{s}_T^2)^{-d/2}}{q_T(\mathbf{x}^{(i)})} \cdot \exp\!\left(-\frac{\|\mathbf{x}^{(i)} - \mathfrak{m}_T x_0\|^2}{2\mathfrak{s}_T^2} \right).$$

Expand the kernel exponent at $\mathfrak{m}_T\mu_q$. With shortcuts $v := \mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q$ and $u := x_0 - \mu_q$ (so $\mathbf{x}^{(i)} - \mathfrak{m}_T x_0 = v - \mathfrak{m}_T u$, $A = \eta_T\langle v, u\rangle$, $B = \frac{1}{2}\eta_T\mathfrak{m}_T\|u\|^2$):

$$\begin{aligned}
-\frac{\|\mathbf{x}^{(i)} - \mathfrak{m}_T x_0\|^2}{2\mathfrak{s}_T^2} &= -\frac{\|v - \mathfrak{m}_T u\|^2}{2\mathfrak{s}_T^2} && \text{(rewrite in shortcuts)}\\
&= -\frac{\|v\|^2 - 2\mathfrak{m}_T\langle v, u\rangle + \mathfrak{m}_T^2\|u\|^2}{2\mathfrak{s}_T^2} && \text{(expand } \|a-b\|^2\text{)}\\
&= -\frac{\|v\|^2}{2\mathfrak{s}_T^2} + \eta_T\langle v, u\rangle - \frac{1}{2}\eta_T\mathfrak{m}_T\|u\|^2 && (\eta_T{=}\mathfrak{m}_T/\mathfrak{s}_T^2;\ \mathfrak{m}_T^2/(2\mathfrak{s}_T^2){=}\frac{1}{2}\eta_T\mathfrak{m}_T)\\
&= -\frac{\|v\|^2}{2\mathfrak{s}_T^2} + A - B. && \text{(def of } A,\ B\text{)}
\end{aligned}$$

Exponentiating and folding the x_0-independent factor into C (the definition above):

$$\begin{aligned}
w(x_0) &= C \cdot \exp(A - B) && \text{(centering identity above)}\\
&= \frac{e^{A-B}}{\mathbb{E}_q[e^{A-B}]} && \big(\mathbb{E}_q[w] = 1 \Rightarrow C = 1/\mathbb{E}_q[e^{A-B}]\big).
\end{aligned}$$

We write the result as

$$w(x_0) \;=\; \frac{e^{A(x_0)-B(x_0)}}{\mathbb{E}_q[e^{A-B}]}. \tag{26}$$

A is linear and B is quadratic in x_0; both are $\mathcal{O}(\eta_T)$, but the extra $\mathfrak{m}_T$ in B makes it strictly subleading whenever $\mathfrak{m}_T \to 0$ (OU/VP at large T). For VE ($\mathfrak{m}_T \equiv 1$) this is no longer subleading; the lemma statement absorbs the B-contribution into the $\eta_T^{3/2}$ remainder under Eq. (4), and the dimension-of-Σ_q correction is recorded in Table 1's VE row. The basic moments are

$$\mathbb{E}_q[A] = 0, \quad \mathbb{E}_q[A^2] = \eta_T^2\,(\mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q)^\mathsf{T}\Sigma_q(\mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q), \quad \mathbb{E}_q[B] = \frac{1}{2}\eta_T\mathfrak{m}_T\,\mathrm{tr}(\Sigma_q).$$

(ii) Linearisation. Taylor-expanding the numerator of (26) and taking $\mathbb{E}_q$ of the same expansion in the denominator,

$$\begin{aligned}
w(x_0) - 1 &= \frac{e^{A-B}}{\mathbb{E}_q[e^{A-B}]} - 1 && \text{(Step (i))}\\
&= \frac{1 + (A-B) + \underbrace{\frac{1}{2}(A-B)^2\,e^{\theta(A-B)}}_{=:\,R_w(x_0)}}{\mathbb{E}_q[e^{A-B}]} - 1 && \text{(Taylor at numerator, } \theta \in [0,1]\text{)}\\
&= \frac{1 + (A-B) + R_w}{\underbrace{1 - \mathbb{E}_q[B] + \mathbb{E}_q[R_w]}_{=:\,1+D,\ D=\mathcal{O}(\eta_T^2)}} - 1 && \text{(}\mathbb{E}_q\text{ of numerator; } \mathbb{E}_q[A] = 0\text{)}\\
&= \frac{(A-B) + R_w - D}{1 + D}\\
&= \big[(A-B) + R_w - D\big]\big(1 - D + \mathcal{O}(D^2)\big) && \text{(geometric, } |D| < 1\text{)}\\
&= (A-B) + R_w - D + \mathcal{O}(\eta_T^3) && (D\cdot[(A-B)+\cdots] = \mathcal{O}(\eta_T^3))\\
&= \eta_T\,\ell(x_0) + \underbrace{\big[-B + R_w - D + \mathcal{O}(\eta_T^3)\big]}_{=:\,R(x_0)} && (A = \eta_T\,\ell).
\end{aligned}$$

The remainder R collects four contributions: (i) the quadratic tilt $-B$; (ii) the Lagrange remainder R_w; (iii) the denominator's $\mathcal{O}(\eta_T^2)$ correction $-D$; (iv) once $w - 1$ is squared (next step), the cross-moment $\mathbb{E}_q[\ell R]$ — non-zero for general sub-Gaussian q, since vanishing would require symmetry of q around μ_q.

(iii) Bound for $\mathbb{E}_q[(w-1)^2]$.

$$\begin{aligned}
&\mathbb{E}_q[\ell^2] \\
&\quad= \mathbb{E}_q[\langle v,u\rangle^2] \qquad \begin{pmatrix} u = x_0 - \mu_q, \\ v = \mathbf{x}^{(i)} - \mathfrak{m}_T\mu_q \end{pmatrix} \\
&\quad= \mathbb{E}_q[v^{\mathsf{T}}uu^{\mathsf{T}}v] \\
&\quad= v^{\mathsf{T}}\mathbb{E}_q[uu^{\mathsf{T}}]v \qquad (v \text{ is fixed}) \\
&\quad= v^{\mathsf{T}}\Sigma_q v. \qquad (\Sigma_q = \mathbb{E}_q[uu^{\mathsf{T}}])
\end{aligned} \tag{a}$$

$$\begin{aligned}
&\mathbb{E}_q[B^2] \\
&\quad= \mathbb{E}_q\Big[\Big(\frac{\eta_T\mathfrak{m}_T}{2}\|u\|^2\Big)^2\Big] \qquad (B = \frac{1}{2}\eta_T\mathfrak{m}_T\|u\|^2) \\
&\quad= \frac{\eta_T^2\mathfrak{m}_T^2}{4}\mathbb{E}_q\|u\|^4 \\
&\quad= \frac{\eta_T^4\mathfrak{s}_T^4}{4}\mathbb{E}_q\|u\|^4 \qquad (\mathfrak{m}_T = \eta_T\mathfrak{s}_T^2) \\
&\quad\le \frac{\eta_T^4\mathfrak{s}_T^4}{4}\mathbb{E}_q\Big[\|u\|^4 e^{2\eta_T\|v\|\|u\|}\Big] \qquad (e^{2\eta_T\|v\|\|u\|} \ge 1) \\
&\quad\le \frac{C_w}{2}\eta_T^4\mathfrak{s}_T^4K_q^4 \le C_w\eta_T^4\mathfrak{s}_T^4K_q^4. \qquad ((33))
\end{aligned} \tag{b}$$

Here C_w is finite and depends only on $\eta_T\|v\|K_q$, as established in (33).

$$\begin{aligned}
&\mathbb{E}_q[R_w^2] \\
&\quad= \mathbb{E}_q\big[(e^{A-B} - 1 - A + B)^2\big] \qquad (R_w = e^{A-B} - 1 - (A-B)) \\
&\quad\le \mathbb{E}_q\big[(e^{A} - 1 - A + B)^2\big] \qquad (B \ge 0,\ e^z \ge 1 + z) \\
&\quad\le \mathbb{E}_q\left[\left(\frac{A^2}{2}e^{|A|} + B\right)^2\right] \qquad (\text{Taylor expansion: } e^A - 1 - A \le \frac{1}{2}A^2e^{|A|}) \\
&\quad= \mathbb{E}_q\left[\left(\frac{\eta_T^2}{2}\langle v,u\rangle^2 e^{\eta_T|\langle v,u\rangle|} + \frac{\eta_T^2\mathfrak{s}_T^2}{2}\|u\|^2\right)^2\right] \qquad (A = \eta_T\langle v,u\rangle) \\
&\quad\le \mathbb{E}_q\left[\left(\frac{\eta_T^2}{2}\|v\|^2\|u\|^2 e^{\eta_T\|v\|\|u\|} + \frac{\eta_T^2\mathfrak{s}_T^2}{2}\|u\|^2\right)^2\right] \qquad (|\langle v,u\rangle| \le \|v\|\|u\|) \\
&\quad= \mathbb{E}_q\left[\left(\frac{\eta_T^2}{2}\|u\|^2\big(\|v\|^2 e^{\eta_T\|v\|\|u\|} + \mathfrak{s}_T^2\big)\right)^2\right] \\
&\quad\le \mathbb{E}_q\left[\left(\frac{\eta_T^2}{2}\big(\|v\|^2 + \mathfrak{s}_T^2\big)\|u\|^2 e^{\eta_T\|v\|\|u\|}\right)^2\right] \qquad (e^z \ge 1\ \forall z \ge 0,\ z = \eta_T\|v\|\|u\|) \\
&\quad= \frac{\eta_T^4}{4}\big(\|v\|^2 + \mathfrak{s}_T^2\big)^2\mathbb{E}_q\Big[\|u\|^4 e^{2\eta_T\|v\|\|u\|}\Big] \\
&\quad\le \frac{C_w\eta_T^4}{2}\big(\|v\|^2 + \mathfrak{s}_T^2\big)^2K_q^4 \qquad ((33)) \\
&\quad\le C_w\eta_T^4\big(\|v\|^4 + \mathfrak{s}_T^4\big)K_q^4. \qquad ((a+b)^2 \le 2(a^2+b^2))
\end{aligned} \tag{c}$$

$$\begin{aligned}
D^2 &= \big(\mathbb{E}_q[R_w - B]\big)^2 \qquad (D = \mathbb{E}_q[R_w - B]) \\
&\le \mathbb{E}_q[(R_w - B)^2] \qquad (\text{Cauchy–Schwarz}) \\
&\le \left(\sqrt{\mathbb{E}_q[R_w^2]} + \sqrt{\mathbb{E}_q[B^2]}\right)^2 \qquad (L^2 \text{ triangle inequality})
\end{aligned}$$

$$\le C_w \eta_T^4 \left(\sqrt{\|v\|^4 + \mathfrak{s}_T^4} + \mathfrak{s}_T^2\right)^2 K_q^4 \qquad \text{((b), (c))}$$
$$\le 4C_w \eta_T^4 \big(\|v\|^4 + \mathfrak{s}_T^4\big) K_q^4. \qquad (\mathfrak{s}_T^2 \le \sqrt{\|v\|^4 + \mathfrak{s}_T^4}) \tag{d}$$

$$\begin{aligned}
&\eta_T \|v\| K_q \\
&\quad \le \eta_T \|\mathbf{x}^{(i)}\| K_q + \eta_T \mathfrak{m}_T \|\mu_q\| K_q \qquad (v = \mathbf{x}^{(i)} - \mathfrak{m}_T \mu_q) \\
&\quad \le 1 + \sqrt{2}\, \eta_T \mathfrak{m}_T K_q^2 \qquad (Eq.\ (4),\ (28)) \\
&\quad = 1 + \sqrt{2}\, \eta_T^2 \mathfrak{s}_T^2 K_q^2. \qquad (\mathfrak{m}_T = \eta_T \mathfrak{s}_T^2)
\end{aligned} \tag{f}$$

Thus C_w can be chosen uniformly over admissible starting points, with dependence only on $\eta_T^2 \mathfrak{s}_T^2 K_q^2 = \mathfrak{m}_T^2 K_q^2 / \mathfrak{s}_T^2$.

$$\begin{aligned}
&\mathbb{E}_q[R^2] \\
&\quad = \mathbb{E}_q[(w - 1 - A)^2] \qquad (R = w - 1 - \eta_T \ell,\ A = \eta_T \ell) \\
&\quad = \mathbb{E}_q\left[\left(\frac{1 + A + R_w - B}{1 + D} - 1 - A\right)^2\right] \qquad (w = e^{A-B}/(1+D)) \\
&\quad = \frac{\mathbb{E}_q[(R_w - B - D - AD)^2]}{(1+D)^2} \\
&\quad \le \frac{\left(\sqrt{\mathbb{E}_q[(R_w - B - D)^2]} + |D| \sqrt{\mathbb{E}_q[A^2]}\right)^2}{(1+D)^2} \qquad (L^2 \text{ triangle inequality}) \\
&\quad \le \frac{\left(\sqrt{\mathbb{E}_q[(R_w - B)^2]} + |D| \sqrt{\mathbb{E}_q[A^2]}\right)^2}{(1+D)^2} \qquad (\text{centering},\ D = \mathbb{E}_q[R_w - B]) \\
&\quad \le \frac{(1 + \eta_T \|v\| \tilde{K}_q)^2}{(1+D)^2} \mathbb{E}_q[(R_w - B)^2] \\
&\qquad\qquad (|D| \le \sqrt{\mathbb{E}_q[(R_w - B)^2]},\ (31),\ \tilde{K}_q := \sqrt{2} K_q) \\
&\quad \le 4C_w \frac{(1 + \eta_T \|v\| \tilde{K}_q)^2}{(1+D)^2} \eta_T^4 \big(\|v\|^4 + \mathfrak{s}_T^4\big) K_q^4 \qquad (\text{(d)}) \\
&\quad \le 4C_w\, 2^{2\eta_T^2 \mathfrak{s}_T^2 K_q^2} \big(1 + \sqrt{2} + 2\eta_T^2 \mathfrak{s}_T^2 K_q^2\big)^2 \eta_T^4 \big(\|v\|^4 + \mathfrak{s}_T^4\big) K_q^4 \\
&\qquad\qquad ((30),\ \text{(f)},\ \tilde{K}_q = \sqrt{2} K_q) \\
&\quad = C_R \eta_T^4 \big(\|v\|^4 + \mathfrak{s}_T^4\big) K_q^4,
\end{aligned} \tag{e}$$

where

$$C_R := 4C_w\, 2^{2\eta_T^2 \mathfrak{s}_T^2 K_q^2} \big(1 + \sqrt{2} + 2\eta_T^2 \mathfrak{s}_T^2 K_q^2\big)^2.$$

Second moment of the density ratio.

$$\begin{aligned}
\mathbb{E}_q[(w-1)^2] &= \eta_T^2 \mathbb{E}_q[\ell^2] + 2\eta_T \mathbb{E}_q[\ell R] + \mathbb{E}_q[R^2] \\
&\le \eta_T^2 v^\mathsf{T} \Sigma_q v + 2\eta_T \sqrt{\mathbb{E}_q[\ell^2]} \sqrt{\mathbb{E}_q[R^2]} + \mathbb{E}_q[R^2] \\
&\le \eta_T^2 v^\mathsf{T} \Sigma_q v + 2\eta_T \|v\| \tilde{K}_q \sqrt{\mathbb{E}_q[R^2]} + \mathbb{E}_q[R^2] \qquad ((31)) \\
&\le \eta_T^2 v^\mathsf{T} \Sigma_q v + 2\sqrt{2C_R}\, \eta_T^3 \|v\| K_q^3 \big(\|v\|^2 + \mathfrak{s}_T^2\big) \\
&\qquad + C_R \eta_T^4 \big(\|v\|^4 + \mathfrak{s}_T^4\big) K_q^4 \\
&= \eta_T^2 v^\mathsf{T} \Sigma_q v + \left[2\sqrt{2C_R} + C_R \eta_T \|v\| K_q\right] \eta_T^3 \|v\|^3 K_q^3 \\
&\qquad + \left[2\sqrt{2C_R}\, \eta_T \|v\| K_q + C_R \eta_T^2 \mathfrak{s}_T^2 K_q^2\right] \eta_T^2 \mathfrak{s}_T^2 K_q^2
\end{aligned}$$

$$\leq \eta_T^2 v^\mathsf{T}\Sigma_q v + \underbrace{\left[2\sqrt{2C_R} + C_R\big(1+\sqrt{2}\,\eta_T^2\mathfrak{s}_T^2K_q^2\big)\right]}_{=:C'_R}\eta_T^3\|v\|^3K_q^3$$
$$+ \underbrace{\left[2\sqrt{2C_R}\big(1+\sqrt{2}\,\eta_T^2\mathfrak{s}_T^2K_q^2\big) + C_R\eta_T^2\mathfrak{s}_T^2K_q^2\right]}_{=:C''_R}\eta_T^2\mathfrak{s}_T^2K_q^2$$
$$= \eta_T^2 v^\mathsf{T}\Sigma_q v + C'_R\eta_T^3\|v\|^3K_q^3 + C''_R\eta_T^2\mathfrak{s}_T^2K_q^2. \tag{27}$$

The constants C_R, C'_R, C''_R depend only on $\mathfrak{m}_T^2K_q^2/\mathfrak{s}_T^2$ and are uniform over starting points satisfying Eq. (4).

Supporting moment estimates.

$$\|\mu_q\|^2 \leq \mathbb{E}_q\|x_0\|^2 \qquad \text{(Jensen)}$$
$$\leq 2K_q^2\log 2 \qquad (Assumption\ 3,\ e^{\mathbb{E}_q\|x_0\|^2/(2K_q^2)} \leq \mathbb{E}_q e^{\|x_0\|^2/(2K_q^2)} \leq 2). \tag{28}$$

$$\mathbb{E}_q\|u\|^2 = \mathbb{E}_q\|x_0\|^2 - \|\mu_q\|^2 \qquad (\mu_q = \mathbb{E}_q x_0)$$
$$\leq 2K_q^2\log 2 \leq 2K_q^2 = \tilde{K}_q^2 \qquad ((28),\ \tilde{K}_q = \sqrt{2}K_q). \tag{29}$$

$$1 + D = \mathbb{E}_q e^{A-B} \qquad (\mathbb{E}_q A = 0,\ D = \mathbb{E}_q[R_w - B])$$
$$\geq e^{\mathbb{E}_q(A-B)} \qquad \text{(Jensen)}$$
$$= e^{-\frac{1}{2}\eta_T^2\mathfrak{s}_T^2\mathbb{E}_q\|u\|^2} \qquad (\mathbb{E}_q A = 0,\ B = \frac{1}{2}\eta_T^2\mathfrak{s}_T^2\|u\|^2)$$
$$\geq 2^{-\eta_T^2\mathfrak{s}_T^2K_q^2} > 0. \qquad ((29)) \tag{30}$$

$$\mathbb{E}_q[\ell^2] \leq \|v\|^2\mathbb{E}_q\|u\|^2 \qquad \text{(Cauchy–Schwarz, } |\langle v,u\rangle| \leq \|v\|\|u\|)$$
$$\leq \|v\|^2\tilde{K}_q^2. \qquad ((29)) \tag{31}$$

$$\mathbb{E}_q e^{\|u\|^2/(8K_q^2)}$$
$$\leq e^{\|\mu_q\|^2/(4K_q^2)}\mathbb{E}_q e^{\|x_0\|^2/(4K_q^2)} \qquad (\|u\|^2 \leq 2\|x_0\|^2 + 2\|\mu_q\|^2)$$
$$\leq e^{\|\mu_q\|^2/(4K_q^2)}\left(\mathbb{E}_q e^{\|x_0\|^2/(2K_q^2)}\right)^{1/2} \qquad \text{(Cauchy–Schwarz)}$$
$$\leq \sqrt{2}\left(\mathbb{E}_q e^{\|x_0\|^2/(2K_q^2)}\right)^{1/2} \qquad ((28),\ \|\mu_q\|^2 \leq 2K_q^2\log 2)$$
$$\leq 2 \qquad (Assumption\ 3,\ \mathbb{E}_q e^{\|x_0\|^2/(2K_q^2)} \leq 2). \tag{32}$$

$$\mathbb{E}_q\left[\|u\|^4 e^{2\eta_T\|v\|\|u\|}\right]$$
$$= \mathbb{E}_q\left[\|u\|^4\exp\left(2(4\eta_T\|v\|K_q)\frac{\|u\|}{4K_q}\right)\right]$$
$$\leq \mathbb{E}_q\left[\|u\|^4\exp\left((4\eta_T\|v\|K_q)^2 + \left(\frac{\|u\|}{4K_q}\right)^2\right)\right] \qquad (2ab \leq a^2+b^2)$$
$$= e^{16\eta_T^2\|v\|^2K_q^2}\mathbb{E}_q\left[\|u\|^4 e^{\|u\|^2/(16K_q^2)}\right]$$
$$\leq C_w K_q^4\,\mathbb{E}_q e^{\|u\|^2/(8K_q^2)} \qquad (t^2e^{-t} \leq 4e^{-2},\ t = \|u\|^2/(16K_q^2))$$
$$\leq 2C_w K_q^4 \qquad ((32)). \tag{33}$$

The moment estimates use Assumption 3; see also [Vershynin, 2018, Prop. 2.5.2].

(iv) Take the square root.

$$W_1(q, q^{(i)}) \leq \sqrt{C_P\,\mathbb{E}_q[(w-1)^2]} \qquad ((25))$$
$$\leq \sqrt{C_P(\eta_T^2 v^\mathsf{T}\Sigma_q v + C'_R\eta_T^3\|v\|^3K_q^3 + C''_R\eta_T^2\mathfrak{s}_T^2K_q^2)} \qquad ((27))$$
$$\leq \eta_T\sqrt{C_P\,v^\mathsf{T}\Sigma_q v} + \eta_T^{3/2}\sqrt{C_P C'_R}\,\|v\|^{3/2}K_q^{3/2} + \eta_T\mathfrak{s}_T\sqrt{C_P C''_R}\,K_q \quad (\sqrt{a+b+c} \leq \textstyle\sum\sqrt{\cdot})$$

which is the lemma statement (6). The third term is the non-decaying-shift contribution: silent for OU/VP (where $\mathfrak{s}_T \le 1$ makes it dominated by the $\eta_T^{3/2}$ remainder under (f)) and load-bearing for VE (where $\mathfrak{s}_T = \sigma(T) \to \infty$).

□

Lemma 3 (Early-stopping cost)**.** *Suppose Assumptions 3 and 8 hold and $t_{\min} \in (0, T)$. Then*

$$W_1(q^{(i)}, q^{(i)}_{t_{\min}}) \le \sqrt{\big((1-\mathfrak{a}_{t_{\min}})\sqrt{M_0^{(i)}} + \mathfrak{b}_{t_{\min}}\|\mathbf{x}^{(i)}\|\big)^2 + \mathfrak{g}^2_{t_{\min}}\, d},$$

with bridge parameters $(\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}}, \mathfrak{g}^2_{t_{\min}})$ from Eq. (35) *and $M_0^{(i)} = \mathbb{E}_{q^{(i)}}\|x_0\|^2 \le C^\star K_q^2$ the posterior second moment (Lemma 5). We denote the right-hand side W_{stop} throughout.*

Proof. **(i) The bridge kernel.** Under the forward SDE (1), the joint $(\bar X_0, \bar X_{t_{\min}}, \bar X_T)$ is Gaussian. Itô on the abstract kernel gives $\bar X_{t_{\min}} = \mathfrak{m}_{t_{\min}} x_0 + V_1$ and $\bar X_T = (\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})\bar X_{t_{\min}} + V_3$ with $V_1 \perp V_3$, $\mathrm{Var}(V_1) = \mathfrak{s}^2_{t_{\min}}$, $\mathrm{Var}(V_3) = \mathfrak{s}^2_{t_{\min},T}$ (so $\mathfrak{s}^2_T = (\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})^2\mathfrak{s}^2_{t_{\min}} + \mathfrak{s}^2_{t_{\min},T}$). Schur-complement on $(\bar X_{t_{\min}}, \bar X_T) \mid \bar X_0$ [Christopher, 2006, §2.3.1]:

$$\bar X_{t_{\min}} \mid (\bar X_0 = x_0,\ \bar X_T = \mathbf{x}^{(i)}) \ \sim\ \mathcal{N}\big(\mathfrak{a}_{t_{\min}}\, x_0 + \mathfrak{b}_{t_{\min}}\, \mathbf{x}^{(i)},\ \mathfrak{g}^2_{t_{\min}}\, I_d\big), \tag{34}$$

$$\mathfrak{a}_{t_{\min}} := \frac{\mathfrak{m}_{t_{\min}}\,\mathfrak{s}^2_{t_{\min},T}}{\mathfrak{s}^2_T}, \quad \mathfrak{b}_{t_{\min}} := \frac{\mathfrak{s}^2_{t_{\min}}(\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})}{\mathfrak{s}^2_T}, \quad \mathfrak{g}^2_{t_{\min}} := \frac{\mathfrak{s}^2_{t_{\min}}\,\mathfrak{s}^2_{t_{\min},T}}{\mathfrak{s}^2_T}. \tag{35}$$

Equivalently, $\bar X_{t_{\min}} = \mathfrak{a}_{t_{\min}}\, x_0 + \mathfrak{b}_{t_{\min}}\, \mathbf{x}^{(i)} + \mathfrak{g}_{t_{\min}}\, \eta$ with $\eta \sim \mathcal{N}(0, I_d) \perp (\bar X_0, \bar X_T)$ (Schur residual); $\mathfrak{a}_{t_{\min}}, \mathfrak{b}_{t_{\min}} \in [0, 1]$. OU specialisation: $\mathfrak{m}_t = e^{-t}$, $\mathfrak{s}^2_t = 1 - e^{-2t}$, $\mathfrak{s}^2_{t_{\min},T} = 1 - e^{-2(T-t_{\min})}$ recover $\mathfrak{a} = \sigma^2_{T-t_{\min}} e^{-t_{\min}}/\mathfrak{s}^2_T$, etc.

(ii) Same-x_0 coupling reduces W_1 to a second moment. Couple via $\bar X_0 \sim q^{(i)}$ then Eq. (34) with the same $\bar X_0$; both targets are marginals of the OU joint $\mid \bar X_T = \mathbf{x}^{(i)}$, so this is valid. Substitute and expand:

$$\begin{aligned}
W_1\big(q^{(i)}, q^{(i)}_{t_{\min}}\big)^2 &\le W_2\big(q^{(i)}, q^{(i)}_{t_{\min}}\big)^2 && (W_1 \le W_2)\\
&\le \mathbb{E}\big\|\bar X_0 - \bar X_{t_{\min}}\big\|^2 && \big(W_2^2 = \inf_\pi \mathbb{E}_\pi\|X-Y\|^2\big)\\
&= \mathbb{E}\big\|(1-\mathfrak{a}_{t_{\min}})\bar X_0 - \mathfrak{b}_{t_{\min}}\mathbf{x}^{(i)} - \mathfrak{g}_{t_{\min}}\eta\big\|^2 && (\text{sub. Eq. (34)})\\
&= (1-\mathfrak{a}_{t_{\min}})^2\,\mathbb{E}_{q^{(i)}}\|\bar X_0\|^2 + \mathfrak{b}^2_{t_{\min}}\|\mathbf{x}^{(i)}\|^2 + \mathfrak{g}^2_{t_{\min}}\, d\\
&\quad - 2(1-\mathfrak{a}_{t_{\min}})\mathfrak{b}_{t_{\min}}\big\langle \mathbb{E}_{q^{(i)}}[\bar X_0], \mathbf{x}^{(i)}\big\rangle. && (\eta \perp \bar X_0,\ \mathbb{E}\eta = 0,\ \mathbb{E}\|\eta\|^2 = d)
\end{aligned} \tag{36}$$

(iii) Cauchy–Schwarz on the cross-term. Recall $M_0^{(i)} := \mathbb{E}_{q^{(i)}}\|\bar X_0\|^2$. Bound the cross-term magnitude:

$$\begin{aligned}
\big|\big\langle \mathbb{E}_{q^{(i)}}[\bar X_0], \mathbf{x}^{(i)}\big\rangle\big| &\le \big\|\mathbb{E}_{q^{(i)}}[\bar X_0]\big\| \cdot \|\mathbf{x}^{(i)}\| && (\text{Cauchy–Schwarz on } \langle\cdot,\cdot\rangle)\\
&\le \sqrt{\mathbb{E}_{q^{(i)}}\|\bar X_0\|^2}\cdot\|\mathbf{x}^{(i)}\| \;=\; \sqrt{M_0^{(i)}}\cdot\|\mathbf{x}^{(i)}\|. && (\text{Jensen})
\end{aligned}$$

Continuing from Eq. (36):

$$\begin{aligned}
W_1\big(q^{(i)}, q^{(i)}_{t_{\min}}\big)^2 &\le (1-\mathfrak{a}_{t_{\min}})^2 M_0^{(i)} + \mathfrak{b}^2_{t_{\min}}\|\mathbf{x}^{(i)}\|^2 + \mathfrak{g}^2_{t_{\min}} d\\
&\quad + 2(1-\mathfrak{a}_{t_{\min}})\mathfrak{b}_{t_{\min}}\big|\big\langle \mathbb{E}_{q^{(i)}}[\bar X_0], \mathbf{x}^{(i)}\big\rangle\big| && (-2x \le 2|x|)\\
&\le (1-\mathfrak{a}_{t_{\min}})^2 M_0^{(i)} + 2(1-\mathfrak{a}_{t_{\min}})\mathfrak{b}_{t_{\min}}\sqrt{M_0^{(i)}}\|\mathbf{x}^{(i)}\|\\
&\quad + \mathfrak{b}^2_{t_{\min}}\|\mathbf{x}^{(i)}\|^2 + \mathfrak{g}^2_{t_{\min}} d && (\text{Cauchy–Schwarz})\\
&= \big((1-\mathfrak{a}_{t_{\min}})\sqrt{M_0^{(i)}} + \mathfrak{b}_{t_{\min}}\|\mathbf{x}^{(i)}\|\big)^2 + \mathfrak{g}^2_{t_{\min}} d.
\end{aligned} \tag{37}$$

Taking the square root yields Eq. (7).

(iv) Small-$t_{\min}$ asymptotics. Each bridge parameter is $\mathcal{O}(t_{\min})$. From $\mathfrak{s}_t^2 = \int_0^t (\mathfrak{m}_t/\mathfrak{m}_s)^2 g(s)^2\,\mathrm{d}s$ and $1-\mathfrak{m}_t = \int_0^t \rho_{\mathrm{fwd}}(u)\,\mathfrak{m}_u\,\mathrm{d}u \le \rho_{\max}\,t$ (Assumption 8):

$$
\begin{aligned}
\mathfrak{s}_{t_{\min}}^2 &\le g_{\max}^2\, t_{\min}, && (\mathfrak{m}_{t_{\min}}/\mathfrak{m}_s \le 1 \text{ for } s \le t_{\min};\ g(s) \le g_{\max})\\
\mathfrak{g}_{t_{\min}}^2 &= \frac{\mathfrak{s}_{t_{\min}}^2\,\mathfrak{s}_{t_{\min},T}^2}{\mathfrak{s}_T^2} && (\text{def, Eq. (35)})\\
&\le \mathfrak{s}_{t_{\min}}^2 && (\mathfrak{s}_{t_{\min},T}^2 \le \mathfrak{s}_T^2 \text{ from } \mathfrak{s}_T^2 = (\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})^2\mathfrak{s}_{t_{\min}}^2 + \mathfrak{s}_{t_{\min},T}^2)\\
\mathfrak{b}_{t_{\min}} &= \frac{\mathfrak{s}_{t_{\min}}^2}{\mathfrak{m}_{t_{\min}}}\cdot\frac{\mathfrak{m}_T}{\mathfrak{s}_T^2} && (\text{def, factor; } \eta_T{=}\mathfrak{m}_T/\mathfrak{s}_T^2)\\
&\le \frac{g_{\max}^2\, t_{\min}}{\mathfrak{m}_{t_{\min}}}\,\eta_T && (\text{line above: } \mathfrak{s}_{t_{\min}}^2 \le g_{\max}^2 t_{\min})\\
&= \mathcal{O}(t_{\min}\,\eta_T). && (\mathfrak{m}_{t_{\min}} \to 1 \text{ as } t_{\min} \to 0)
\end{aligned}
$$

For $1-\mathfrak{a}_{t_{\min}}$:

$$
\begin{aligned}
1-\mathfrak{a}_{t_{\min}} &= \frac{\mathfrak{s}_T^2 - \mathfrak{m}_{t_{\min}}\mathfrak{s}_{t_{\min},T}^2}{\mathfrak{s}_T^2} && (\text{def, Eq. (35); combine})\\
&= \frac{\mathfrak{s}_T^2(1-\mathfrak{m}_{t_{\min}}) + \mathfrak{m}_T^2\,\mathfrak{s}_{t_{\min}}^2/\mathfrak{m}_{t_{\min}}}{\mathfrak{s}_T^2} && (\text{substitute } \mathfrak{s}_T^2{=}(\mathfrak{m}_T/\mathfrak{m}_{t_{\min}})^2\mathfrak{s}_{t_{\min}}^2{+}\mathfrak{s}_{t_{\min},T}^2)\\
&= (1-\mathfrak{m}_{t_{\min}}) + \frac{\mathfrak{m}_T^2\,\mathfrak{s}_{t_{\min}}^2}{\mathfrak{m}_{t_{\min}}\,\mathfrak{s}_T^2} && (\text{distribute } 1/\mathfrak{s}_T^2)\\
&\le \rho_{\max}\,t_{\min} + \frac{\mathfrak{m}_T^2\,\mathfrak{s}_{t_{\min}}^2}{\mathfrak{m}_{t_{\min}}\,\mathfrak{s}_T^2} && (1{-}e^{-x} \le x \text{ on } \mathfrak{m}_{t_{\min}}{=}e^{-\int_0^{t_{\min}}\rho_{\mathrm{fwd}}})\\
&= \rho_{\max}\,t_{\min} + \frac{\eta_T^2\,\mathfrak{s}_T^2\,\mathfrak{s}_{t_{\min}}^2}{\mathfrak{m}_{t_{\min}}} && (\mathfrak{m}_T^2/\mathfrak{s}_T^2 = \eta_T^2\,\mathfrak{s}_T^2)\\
&\le \rho_{\max}\,t_{\min} + \frac{g_{\max}^2\,t_{\min}\,\eta_T^2\,\mathfrak{s}_T^2}{\mathfrak{m}_{t_{\min}}} && (\mathfrak{s}_{t_{\min}}^2 \le g_{\max}^2\,t_{\min})\\
&= \mathcal{O}(t_{\min}). && (T \text{ fixed};\ \mathfrak{m}_{t_{\min}} \to 1)
\end{aligned}
$$

Plugging into Eq. (37), for $t_{\min} \le 1$:

$$
\begin{aligned}
W_{\mathrm{stop}}^2 &\le \big((1-\mathfrak{a}_{t_{\min}})\sqrt{M_0^{(i)}} + \mathfrak{b}_{t_{\min}}\|\mathbf{x}^{(i)}\|\big)^2 + \mathfrak{g}_{t_{\min}}^2 d && (Eq.\ (37))\\
&= \mathcal{O}(t_{\min}^2)\big(\sqrt{M_0^{(i)}} + \|\mathbf{x}^{(i)}\|\big)^2 + \mathcal{O}(t_{\min})\,d && (1{-}\mathfrak{a}_{t_{\min}},\ \mathfrak{b}_{t_{\min}},\ \mathfrak{g}_{t_{\min}}^2 = \mathcal{O}(t_{\min}))\\
&= \mathcal{O}\big(t_{\min}^2(1+\|\mathbf{x}^{(i)}\|^2)\big) + \mathcal{O}(t_{\min})\,d && (M_0^{(i)} = \mathcal{O}(1) \text{ via Lemma 5})\\
&= \mathcal{O}\big(t_{\min}^2\|\mathbf{x}^{(i)}\|^2 + d\,t_{\min}\big). && (t_{\min}^2 \le t_{\min};\ d \ge 1)
\end{aligned}
$$

Hence $W_{\mathrm{stop}} = \mathcal{O}(\sqrt{t_{\min}^2\|\mathbf{x}^{(i)}\|^2 + d\,t_{\min}})$. □

Lemma 4 (Generation envelope)**.** *Suppose Assumptions 1 to 6 and 8 hold with $L \ge 1$ and $t_{\min} \in (0,T)$. Then*

$$
W_1\big(q_{t_{\min}}^{(i)}, \mathcal{H}^{(i)}\big)^2 \;\le\; 4\,g_{\max}^4\,D_\mu \int_0^{T_{\mathrm{stop}}} E(s)\,\mathrm{d}s,
$$

where D_μ is the contraction-damping integral

$$
D_\mu := \int_0^{T_{\mathrm{stop}}} \exp\Big(-2\int_s^{T_{\mathrm{stop}}} \mu_{\mathrm{eff}}(u)\,\mathrm{d}u\Big)\,\mathrm{d}s
$$

and $E(s)$ is the per-step error budget of Eq. (9)*. We denote the right-hand side W_{gen}^2 throughout.*

Proof. Steps (ii), (iv), (v) follow Chen et al. [2023], Benton et al. [2024]; per-start novelties are step (iii) (smoothing $Z_0 = 0$ from shared start $\mathbf{x}^{(i)}$) and the localized $\tilde{\varepsilon}_{\text{score}}$ in step (iv) (replaces the uniform $\bar{\varepsilon}_{\text{score}}$, informative when $\mathbf{x}^{(i)}$ is OOD).

(i) Coupling and difference dynamics. Run the true reverse X_t (Anderson reversal of Eq. (1)) and SGM Y_t (Eq. (2)) with shared Brownian motion on $[0, T_{\text{stop}}]$, both started at $\mathbf{x}^{(i)}$. With $\rho := \rho_{\text{fwd}}(T-t)$, $g := g(T-t)$, $\rho_k := \rho_{\text{fwd}}(T-kh)$, $g_k := g(T-kh)$. Set $Z_t := X_t - Y_t$, $Z_0 = 0$.

$$
\begin{aligned}
\mathrm{d}Z_t &= \mathrm{d}X_t - \mathrm{d}Y_t && \text{(def of } Z_t\text{)} \\
&= \big[\rho X_t + g^2 \nabla \ln q_{T-t}(X_t)\big]\mathrm{d}t + g\,\mathrm{d}B_t \\
&\quad - \big[\rho_k Y_{kh} + g_k^2 s_\theta(Y_{kh}, T-kh)\big]\mathrm{d}t - g_k\,\mathrm{d}B_t && \text{(Anderson Eq. (1); SGM Eq. (2))} \\
&= \big[(\rho X_t - \rho_k Y_{kh}) && (38) \\
&\quad + g^2 \nabla \ln q_{T-t}(X_t) - g_k^2 s_\theta(Y_{kh}, T-kh)\big]\mathrm{d}t + (g-g_k)\,\mathrm{d}B_t. && (39)
\end{aligned}
$$

Linear-drift split (Method 2: $X_t = Z_t + Y_t$, then $\pm\rho_k Y_t$):

$$
\begin{aligned}
\rho X_t - \rho_k Y_{kh} &= \rho(Z_t + Y_t) - \rho_k Y_{kh} && (X_t = Z_t + Y_t) \\
&= \rho Z_t + \rho Y_t - \rho_k Y_{kh} && \text{(distribute)} \\
&= \rho Z_t + (\rho_k Y_t - \rho_k Y_{kh}) + (\rho Y_t - \rho_k Y_t) && (\pm\,\rho_k Y_t) \\
&= \rho Z_t + \rho_k(Y_t - Y_{kh}) + (\rho - \rho_k) Y_t. && \text{(collect)}
\end{aligned}
$$

Score-difference split (factor g^2, $\pm g^2 s_\theta$):

$$
\begin{aligned}
g^2 \nabla \ln q_{T-t}(X_t) - g_k^2 s_\theta(Y_{kh}, T-kh) &= g^2\big[\nabla \ln q_{T-t}(X_t) - s_\theta(Y_{kh}, T-kh)\big] \\
&\quad + (g_k^2 - g^2)\, s_\theta(Y_{kh}, T-kh).
\end{aligned}
$$

Substitute back into Eq. (39) *and factor g^2 from the (II-0)+score group:*

$$
\mathrm{d}Z_t = \big[\rho Z_t + g^2\big(\text{(II-0)} + \nabla \ln q_{T-t}(X_t) - s_\theta(Y_{kh}, T-kh)\big) + r_\rho + r_g\big]\mathrm{d}t + (g-g_k)\,\mathrm{d}B_t, \quad (40)
$$

with the frozen-drift temporal piece (one piece, used identically in step (iv) and Eq. (9))

$$
\text{(II-0)} := (\rho_k / g^2)\,(Y_t - Y_{kh}),
$$

and two SDE-coefficient residuals

$$
r_\rho := (\rho - \rho_k)\, Y_t, \qquad r_g := (g_k^2 - g^2)\, s_\theta(Y_{kh}, T-kh).
$$

On $[kh, (k+1)h]$, Assumption 8 bounds $|\rho - \rho_k| \le \rho'_{\max} h$ and $|g_k^2 - g^2| \le 2 g_{\max} g'_{\max} h$, giving:

$$
\begin{aligned}
&\|r_\rho\| \le \rho'_{\max}\, h\, \|Y_t\| && \text{(drift-coefficient gap)}, \\
&\|r_g\| \le 2 g_{\max}\, g'_{\max}\, h\, \|s_\theta(Y_{kh}, T-kh)\| && \text{(score-coefficient gap)},
\end{aligned}
$$

$$
\text{(II-}c\text{)} := (g-g_k)\,\mathrm{d}B_t, \quad \mathbb{E}[\text{(II-}c\text{)}] = 0, \quad \mathbb{E}\| \textstyle\int_{kh}^{(k+1)h} \text{(II-}c\text{)}\|^2 \le (g'_{\max})^2 h^2 d \quad \text{(Itô isometry)}.
$$

(ii) Contraction decomposition (Lyapunov argument of [Chen et al., 2023, §3]).

Score-difference split. Add and subtract $\nabla \ln q_{T-t}(Y_t)$ inside the $g^2[\cdots]$ bracket of Eq. (40), abbreviating $s_\theta \equiv s_\theta(Y_{kh}, T-kh)$: Define $\text{(I)} := \nabla \ln q_{T-t}(X_t) - \nabla \ln q_{T-t}(Y_t)$ and $\text{(II)} := \text{(II-0)} + \nabla \ln q_{T-t}(Y_t) - s_\theta$. Then

$$
\begin{aligned}
g^2\big[\text{(II-0)} + \nabla \ln q_{T-t}(X_t) - s_\theta\big] &= g^2\big[\text{(II-0)} + \text{(I)} + \nabla \ln q_{T-t}(Y_t) - s_\theta\big] && (\pm \nabla \ln q_{T-t}(Y_t)) \\
&= g^2\big[\text{(I)} + \text{(II)}\big]. && \text{(def of (II))}
\end{aligned}
$$

Substituting into Eq. (40): $\mathrm{d}Z_t = [\rho Z_t + g^2\text{(I)} + g^2\text{(II)} + r_\rho + r_g]\,\mathrm{d}t + (g-g_k)\,\mathrm{d}B_t$.

Lyapunov derivative. Apply Itô's formula to $\|Z_t\|^2 = \langle Z_t, Z_t\rangle$ with $\mathrm{d}Z_t = b_t\,\mathrm{d}t + (g{-}g_k)\,\mathrm{d}B_t$ (noise coefficient $\sigma_t = (g{-}g_k)I_d$, so $\mathrm{tr}(\sigma_t\sigma_t^\top) = (g{-}g_k)^2 d$):

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|Z_t\|^2 &= 2\,\mathbb{E}\langle Z_t, b_t\rangle + (g{-}g_k)^2 d && \text{(Itô; } \mathbb{E}\langle Z_t, \mathrm{d}B_t\rangle{=}0\text{)}\\
&= 2\,\mathbb{E}\langle Z_t,\, \rho Z_t + g^2(\mathrm{I}) + g^2(\mathrm{II}) + r_\rho + r_g\rangle + (g{-}g_k)^2 d && \text{(drift } b_t \text{ from above)}\\
&= 2\rho\,\mathbb{E}\|Z_t\|^2 + 2g^2\,\mathbb{E}\langle Z_t, (\mathrm{I})\rangle + 2g^2\,\mathbb{E}\langle Z_t, (\mathrm{II})\rangle\\
&\quad + 2\,\mathbb{E}\langle Z_t, r_\rho{+}r_g\rangle + (g{-}g_k)^2 d && \text{(distribute } \langle Z_t, \cdot\rangle\text{)}\\
&\le 2\rho\,\mathbb{E}\|Z_t\|^2 + 2g^2 S(t)\,\mathbb{E}\|Z_t\|^2 + 2g^2\,\mathbb{E}\langle Z_t, (\mathrm{II})\rangle\\
&\quad + 2\,\mathbb{E}\langle Z_t, r_\rho{+}r_g\rangle + (g{-}g_k)^2 d && (\langle Z, (\mathrm{I})\rangle \le S(t)\|Z\|^2,\ \textit{Eq.}\ (3))\\
&\le 2\rho\,\mathbb{E}\|Z_t\|^2 + 2g^2 S(t)\,\mathbb{E}\|Z_t\|^2 + 2g^2\,\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]\\
&\quad + 2\,\mathbb{E}[\|Z_t\|(\|r_\rho\|{+}\|r_g\|)] + (g{-}g_k)^2 d && \text{(Cauchy–Schwarz)}\\
&\le 2\rho\,\mathbb{E}\|Z_t\|^2 + 2g^2 S(t)\,\mathbb{E}\|Z_t\|^2 + 2g^2\,\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]\\
&\quad + 2\,\mathbb{E}[\|Z_t\|(\|r_\rho\|{+}\|r_g\|)] + (g'_{\max})^2 h^2 d && (|g{-}g_k| \le g'_{\max}h,\ \textit{Assumption}\ 8)\\
&= -2\mu_{\mathrm{eff}}(t)\,\mathbb{E}\|Z_t\|^2 + 2g^2\,\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]\\
&\quad + 2\,\mathbb{E}[\|Z_t\|(\|r_\rho\|{+}\|r_g\|)] + (g'_{\max})^2 h^2 d. && (\mu_{\mathrm{eff}} = -\rho - g^2 S(t))
\end{aligned} \tag{41}$$

(iii) Smoothing $Z_0 = 0$ (per-start novelty). Dividing Eq. (41) by $\sqrt{\mathbb{E}\|Z_t\|^2}$ fails at $t = 0$ (since $Z_0 = 0$ gives $\mathbb{E}\|Z_0\|^2 = 0$). Regularise via the *deterministic* function

$$v_\epsilon(t) := \sqrt{\mathbb{E}\|Z_t\|^2 + \epsilon^2}, \qquad \epsilon > 0 \text{ fixed (sent to } 0^+ \text{ at step end)},$$

so $v_\epsilon \ge \sqrt{\mathbb{E}\|Z_t\|^2}$, $v_\epsilon \ge \epsilon > 0$, and v_ϵ is differentiable in t; write $\dot v_\epsilon := \mathrm{d}v_\epsilon/\mathrm{d}t$.

$$\begin{aligned}
2v_\epsilon\dot v_\epsilon = \frac{\mathrm{d}}{\mathrm{d}t}v_\epsilon^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|Z_t\|^2 && (\epsilon \text{ constant; chain rule})\\
&\le -2\mu_{\mathrm{eff}}(v_\epsilon^2 - \epsilon^2) + 2g^2\,\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]\\
&\quad + 2\,\mathbb{E}[\|Z_t\|(\|r_\rho\| + \|r_g\|)] + (g'_{\max})^2 h^2 d. && (\textit{Eq.}\ (41);\ \|Z_t\|^2 = v_\epsilon^2 - \epsilon^2)
\end{aligned}$$

Divide the previous bound by $2v_\epsilon$ (v_ϵ deterministic, so $1/v_\epsilon$ pulls out of expectations); use $(v_\epsilon^2{-}\epsilon^2)/v_\epsilon = v_\epsilon - \epsilon^2/v_\epsilon$ to rearrange:

$$\begin{aligned}
\dot v_\epsilon + \mu_{\mathrm{eff}}\, v_\epsilon &= \mu_{\mathrm{eff}}\,\frac{\epsilon^2}{v_\epsilon} + g^2\,\frac{\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]}{v_\epsilon}\\
&\quad + \frac{\mathbb{E}[\|Z_t\|(\|r_\rho\|{+}\|r_g\|)]}{v_\epsilon} + \frac{(g'_{\max})^2 h^2 d}{2v_\epsilon}\\
&\le |\mu_{\mathrm{eff}}|\,\epsilon + g^2\,\frac{\mathbb{E}[\|Z_t\|\|(\mathrm{II})\|]}{v_\epsilon}\\
&\quad + \frac{\mathbb{E}[\|Z_t\|(\|r_\rho\|{+}\|r_g\|)]}{v_\epsilon} + \frac{(g'_{\max})^2 h^2 d}{2v_\epsilon}\\
&\le |\mu_{\mathrm{eff}}|\,\epsilon + g^2\sqrt{\mathbb{E}\|(\mathrm{II})\|^2}\\
&\quad + \sqrt{\mathbb{E}(\|r_\rho\|{+}\|r_g\|)^2} + \frac{(g'_{\max})^2 h^2 d}{2v_\epsilon},
\end{aligned} \tag{42}$$

where the first inequality uses $v_\epsilon \ge \epsilon$, and the second uses Cauchy–Schwarz: for any non-negative random W, $\mathbb{E}[\|Z_t\|\,W] \le \sqrt{\mathbb{E}\|Z_t\|^2}\sqrt{\mathbb{E}W^2} \le v_\epsilon\sqrt{\mathbb{E}W^2}$, so $\mathbb{E}[\|Z_t\|\,W]/v_\epsilon \le \sqrt{\mathbb{E}W^2}$.

Define $\Phi(s) := e^{-\int_s^{T_{\mathrm{stop}}}\mu_{\mathrm{eff}}}$ and $|\mu_{\mathrm{eff}}|_\infty := \sup_{t\in[0,T_{\mathrm{stop}}]}|\mu_{\mathrm{eff}}(t)|$ (finite via $|\mu_{\mathrm{eff}}|_\infty \le \rho_{\max} + g_{\max}^2 L$, Assumptions 1 and 8), so $\Phi(s) \le e^{|\mu_{\mathrm{eff}}|_\infty T_{\mathrm{stop}}}$. Let $f(t)$ denote the RHS of Eq. (42).

Multiplying by the integrating factor $e^{\int_0^t \mu_{\text{eff}}}$, integrating $0 \to T_{\text{stop}}$, then multiplying by $\Phi(0)$, with $v_\epsilon(0) = \sqrt{0+\epsilon^2} = \epsilon$ (since $Z_0 = 0$):

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[v_\epsilon\, e^{\int_0^t \mu_{\text{eff}}}\big] \le f(t)\, e^{\int_0^t \mu_{\text{eff}}} \qquad (Eq.\ (42) \times e^{\int_0^t \mu_{\text{eff}}};\ \text{product rule on LHS})$$

$$v_\epsilon(T_{\text{stop}})\, e^{\int_0^{T_{\text{stop}}} \mu_{\text{eff}}} - \epsilon \le \int_0^{T_{\text{stop}}} f(s)\, e^{\int_0^s \mu_{\text{eff}}}\, \mathrm{d}s \qquad (\int_0^{T_{\text{stop}}};\ v_\epsilon(0){=}\epsilon)$$

$$v_\epsilon(T_{\text{stop}}) \le \epsilon\, \Phi(0) + \int_0^{T_{\text{stop}}} \Phi(s)\, f(s)\, \mathrm{d}s \quad (\times \Phi(0);\ \Phi(s) = e^{-\int_s^{T_{\text{stop}}} \mu_{\text{eff}}})$$

Substituting $f(s) =$ RHS of Eq. (42):

$$v_\epsilon(T_{\text{stop}}) \le \epsilon\, \Phi(0) + \int_0^{T_{\text{stop}}} \Phi(s) \Big[g(T{-}s)^2 \sqrt{\mathbb{E}\|(\mathrm{II})(s)\|^2} + \sqrt{\mathbb{E}(\|r_\rho(s)\| + \|r_g(s)\|)^2} \Big]\, \mathrm{d}s + \epsilon \int_0^{T_{\text{stop}}} \Phi(s)\, |\mu_{\text{eff}}(s)|\, \mathrm{d}s + \mathcal{O}((g'_{\max} h)^2 T_{\text{stop}}).$$

The ϵ-coefficient integral is bounded: $\int_0^{T_{\text{stop}}} \Phi(s) |\mu_{\text{eff}}(s)|\, \mathrm{d}s \le C_\mu := |\mu_{\text{eff}}|_\infty\, T_{\text{stop}}\, e^{|\mu_{\text{eff}}|_\infty T_{\text{stop}}}$.
Send $\epsilon \to 0^+$; $v_\epsilon(T_{\text{stop}}) \to \sqrt{\mathbb{E}\|Z_{T_{\text{stop}}}\|^2}$ and the bound is ϵ-independent:

$$\sqrt{\mathbb{E}\|Z_{T_{\text{stop}}}\|^2} \;\le\; \int_0^{T_{\text{stop}}} \Phi(s) \Big[g(T{-}s)^2 \sqrt{\mathbb{E}\|(\mathrm{II})(s)\|^2} + \sqrt{\mathbb{E}(\|r_\rho(s)\| + \|r_g(s)\|)^2} \Big]\, \mathrm{d}s + \mathcal{O}((g'_{\max} h)^2 T_{\text{stop}}). \tag{43}$$

(iv) 4-way decomposition of (II) (mirrors Chen et al. [2023], Benton et al. [2024]; localized $\tilde{\varepsilon}_{\text{score}}$ is the per-start novelty). Insert pivots $\nabla \ln q_{T-t}(Y_{kh}), \nabla \ln q_{T-kh}(Y_{kh})$ in the score difference:

$$\begin{aligned}(\mathrm{II}) &= (\rho_k/g^2)\,(Y_t - Y_{kh}) + \nabla \ln q_{T-t}(Y_t) - s_\theta(Y_{kh}, T{-}kh) \\ &= \underbrace{(\rho_k/g^2)(Y_t - Y_{kh})}_{\text{(II-0): frozen-drift gap, from step (i)}} + \underbrace{\nabla \ln q_{T-t}(Y_t) - \nabla \ln q_{T-t}(Y_{kh})}_{\text{(II-}a\text{): spatial Lipschitz}} \\ &\quad + \underbrace{\nabla \ln q_{T-t}(Y_{kh}) - \nabla \ln q_{T-kh}(Y_{kh})}_{\text{(II-}b_1\text{): temporal Lipschitz}} + \underbrace{\nabla \ln q_{T-kh}(Y_{kh}) - s_\theta(Y_{kh}, T{-}kh)}_{\text{(II-}b_2\text{): localized score error}}.\end{aligned}$$

Term-wise bounds (sub-bounds (a)–(d)):

$$\begin{aligned}&\|(\text{II-0})\| \le (\rho_{\max}/g^2_{\min})\|Y_t - Y_{kh}\|, \\ &\|(\text{II-}a)\| \le L\, \|Y_t - Y_{kh}\| && (Assumption\ 1), \\ &\|(\text{II-}b_1)\| \le L_t(1 + \|Y_{kh}\|)\, h && (Assumption\ 2;\ |t - kh| \le h), \\ &\|(\text{II-}b_2)\| \le \tilde{\varepsilon}_{\text{score}}(T{-}kh) && (Eq.\ (24);\ \text{localized}).\end{aligned}$$

Then, with $M_1^Y := \sup_t \mathbb{E}\|Y_t\|^2$ the SGM second-moment envelope (closed in step (v)):

$$\begin{aligned}\mathbb{E}\|(\mathrm{II})\|^2 &= \mathbb{E}\big\|(\text{II-0}) + (\text{II-}a) + (\text{II-}b_1) + (\text{II-}b_2)\big\|^2 \\ &\le 4\,\mathbb{E}\big[\|(\text{II-0})\|^2 + \|(\text{II-}a)\|^2 + \|(\text{II-}b_1)\|^2 + \|(\text{II-}b_2)\|^2\big] \quad \big(\big\|\textstyle\sum a_i\big\|^2 \le 4 \sum \|a_i\|^2;\ \text{exact}\big) \\ &\le 4\,\mathbb{E}\big[(\rho_k/g^2)^2 \|Y_t - Y_{kh}\|^2 + L^2 \|Y_t - Y_{kh}\|^2 \\ &\qquad + L_t^2 (1 + \|Y_{kh}\|)^2\, h^2 + \tilde{\varepsilon}_{\text{score}}(T{-}kh)^2\big] \qquad (\text{(a)–(d) squared}) \\ &= 4\big((\rho_k/g^2)^2 + L^2\big)\, \mathbb{E}\|Y_t - Y_{kh}\|^2 \\ &\quad + 4L_t^2\, \mathbb{E}(1 + \|Y_{kh}\|)^2\, h^2 + 4\,\tilde{\varepsilon}_{\text{score}}(T{-}kh)^2 \qquad (\text{merge (II-0) + (II-}a\text{)}) \\ &\le 4\big((\rho_k/g^2)^2 + L^2\big)\, \mathbb{E}\|Y_t - Y_{kh}\|^2 \\ &\quad + 8\, L_t^2\, (1 + M_1^Y)\, h^2 + 4\,\tilde{\varepsilon}_{\text{score}}(T{-}kh)^2. \qquad \big(\mathbb{E}(1 + \|Y_{kh}\|)^2 \le 2(1 + M_1^Y)\big)\end{aligned} \tag{44}$$

(v) SGM step-size + Gronwall envelope. Two sub-bounds via Young + Tweedie + discrete Gronwall.

(a) Frozen-drift increment. Eq. (2) integrates exactly on $[kh,(k{+}1)h]$: $Y_t - Y_{kh} = (t{-}kh)[\rho_k Y_{kh} + g_k^2 s_\theta(Y_{kh}, T{-}kh)] + g_k(B_t - B_{kh})$. Tweedie Efron [2011] gives $\|s_\theta(0,\tau)\|^2 \le 2\tilde{K}_q^2/\mathfrak{s}_{t_{\min}}^4 + 2\bar{\varepsilon}_{\text{score}}^2$ ($\tilde{K}_q := \sqrt{2}K_q$ from Lemma 5; origin score + score-error budget):

$$\begin{aligned}
\mathbb{E}\|Y_t - Y_{kh}\|^2 &= (t{-}kh)^2\,\mathbb{E}\|\rho_k Y_{kh} + g_k^2 s_\theta(Y_{kh}, T{-}kh)\|^2 + g_k^2 d(t{-}kh) && \text{(drift } \perp \text{ noise)}\\
&\le (t{-}kh)^2\big[2\rho_k^2\,\mathbb{E}\|Y_{kh}\|^2 + 2g_k^4\,\mathbb{E}\|s_\theta\|^2\big] + g_k^2 d(t{-}kh) && \text{(Young } \eta{=}1)\\
&\le (t{-}kh)^2\big[(2\rho_k^2 + 4g_k^4 L_s^2)\,\mathbb{E}\|Y_{kh}\|^2 + 4g_k^4\|s_\theta(0,T{-}kh)\|^2\big] + g_k^2 d(t{-}kh)\\
&\le (t{-}kh)^2\, C_{\text{drift}}(kh) + g_k^2 d(t{-}kh) && \text{(Tweedie)}\\
&\le C_{\text{drift}}(kh)\, h^2 + g_k^2 dh. && (t{-}kh \le h)
\end{aligned} \tag{45}$$

with

$$C_{\text{drift}}(kh) := (2\rho_k^2 + 4g_k^4 L_s^2)\, M_1^Y(kh) + 8g_k^4\tilde{K}_q^2/\mathfrak{s}_{t_{\min}}^4 + 8g_k^4\bar{\varepsilon}_{\text{score}}^2.$$

(b) SGM 2nd-moment envelope $M_1^Y(kh)$. LHS = $\mathbb{E}\|Y_{(k+1)h}\|^2$ throughout. Set $K_0^2 := 2\tilde{K}_q^2/\mathfrak{s}_{t_{\min}}^4 + 2\bar{\varepsilon}_{\text{score}}^2$ and $\alpha := 1 + 6\rho_{\max} + 4g_{\max}^4 L_s^2$:

$$\begin{aligned}
&\mathbb{E}\|Y_{(k+1)h}\|^2 = \mathbb{E}\|(1+\rho_k h)Y_{kh} + g_k^2 h\, s_\theta\|^2 + g_k^2 dh && (\xi_k \perp \text{drift};\ \mathbb{E}\|g_k\sqrt{h}\,\xi_k\|^2{=}g_k^2 dh)\\
&\le (1{+}h)(1{+}\rho_k h)^2\,\mathbb{E}\|Y_{kh}\|^2 + (1{+}h)g_k^4 h\,\mathbb{E}\|s_\theta\|^2 + g_k^2 dh && (\text{Young},\ \eta = h)\\
&\le \big(1{+}h(1{+}6\rho_{\max})\big)\mathbb{E}\|Y_{kh}\|^2 + 2g_{\max}^4 h\,\mathbb{E}\|s_\theta\|^2 + g_{\max}^2 dh && ((1{+}\rho_k h)^2 \le 1{+}3\rho_k h;\ 1{+}h \le 2)\\
&\le (1 + h\,\alpha)\,\mathbb{E}\|Y_{kh}\|^2 + h\,(4g_{\max}^4 K_0^2 + g_{\max}^2 d). && (\|s_\theta\|^2{\le}2L_s^2\|Y\|^2{+}2K_0^2;\ \text{collect})
\end{aligned}$$

Set $a_k := \mathbb{E}\|Y_{kh}\|^2$, $\beta := 4g_{\max}^4 K_0^2 + g_{\max}^2 d$. Discrete Gronwall [Quarteroni et al., 2006, Ch. 11] (iterate, geometric sum, $(1+\alpha h)^k \le e^{\alpha kh}$):

$$\begin{aligned}
&a_{k+1} \le (1+\alpha h)\, a_k + \beta h && \text{(recursion above)}\\
\Rightarrow\ &a_k \le (1+\alpha h)^k\, a_0 + \beta h \sum_{j=0}^{k-1}(1+\alpha h)^j && \text{(iterate from } j=0 \text{ to } k-1)\\
&\le (1+\alpha h)^k\,\|\mathbf{x}^{(i)}\|^2 + \beta h\,\frac{(1+\alpha h)^k - 1}{\alpha h} && (a_0 = \|\mathbf{x}^{(i)}\|^2;\ \text{geometric sum})\\
&\le e^{\alpha kh}\,\|\mathbf{x}^{(i)}\|^2 + \frac{e^{\alpha kh}-1}{\alpha}\,\beta. && ((1+\alpha h)^k \le e^{\alpha kh})
\end{aligned}$$

Define $M_1^Y(kh)$ as this closed-form upper bound:

$$\mathbb{E}\|Y_{kh}\|^2 \ \le\ M_1^Y(kh) \ :=\ e^{\alpha kh}\,\|\mathbf{x}^{(i)}\|^2 + \frac{e^{\alpha kh}-1}{\alpha}\,(4g_{\max}^4 K_0^2 + g_{\max}^2 d). \tag{46}$$

Plugging $M_1^Y := M_1^Y(T)$ (uniform majorant; $kh \mapsto M_1^Y(kh)$ increasing for $\alpha > 0$) into Eq. (45) closes the recursion.

(vi) Assemble. Set $C_{\text{drift}} := C_{\text{drift}}(T)$, $\tilde{\varepsilon}_{\text{score}}^2 \equiv \tilde{\varepsilon}_{\text{score}}(T{-}kh)^2$, $k = \lfloor s/h \rfloor$. Bound $E\|(\text{II})(s)\|^2$ by substituting step (v.a) into Eq. (44), then add residuals from step (i):

$$\begin{aligned}
E(s) \ :=\ &\underbrace{4\tilde{\varepsilon}_{\text{score}}^2}_{\text{(II-}b_2\text{): score error}} + \underbrace{4\big((\rho_k/g^2)^2{+}L^2\big)\big(g_k^2 dh{+}C_{\text{drift}}h^2\big)}_{\text{(II-0)+(II-}a\text{): Eq. (44)+step (v.a)}}\\
&+ \underbrace{8L_t^2(1{+}M_1^Y)h^2}_{\text{(II-}b_1\text{): temporal Lip}} + \underbrace{4d(g'_{\max})^2 h^2}_{\text{(II-}c\text{): frozen-diffusion}} + \underbrace{4(\rho'_{\max}/g^2_{\min})^2 M_1^Y h^2}_{r_\rho\text{: drift-coeff gap}},
\end{aligned} \tag{47}$$

matching Eq. (9). Set $\Psi(s) := g(T-s)^2\sqrt{\mathbb{E}\|(\mathrm{II})\|^2} + \sqrt{\mathbb{E}(\|r_\rho\| + \|r_g\|)^2}$ — the integrand of Eq. (43). Square and apply Cauchy–Schwarz with $(a+b)^2 \le 2(a^2+b^2)$ to handle residuals:

$$
\begin{aligned}
W_1\big(q^{(i)}_{t_{\min}}, \mathcal{H}^{(i)}\big)^2 \le W_2^2 &\le \mathbb{E}\|Z_{T_{\mathrm{stop}}}\|^2 && (W_1 \le W_2;\ \text{synchronous coupling})\\
&\le 2\Big(\int_0^{T_{\mathrm{stop}}} \Phi(s)\,\Psi(s)\,\mathrm{d}s\Big)^2 + 2\,\mathcal{O}((g'_{\max}h)^4 T_{\mathrm{stop}}^2) && (\text{square } Eq.\ (43))\\
&\le 2D_\mu \int_0^{T_{\mathrm{stop}}} \Psi(s)^2\,\mathrm{d}s + 2\mathcal{O}((g'_{\max}h)^4 T_{\mathrm{stop}}^2) && (\text{C–S})\\
&\le 4D_\mu \int_0^{T_{\mathrm{stop}}} \big[g^4(T-s)\mathbb{E}\|(\mathrm{II})\|^2 + \mathbb{E}(\|r_\rho\|+\|r_g\|)^2\big]\mathrm{d}s\\
&\quad + 2\mathcal{O}((g'_{\max}h)^4 T_{\mathrm{stop}}^2) && ((c+d)^2 \le 2(c^2+d^2)\ \text{on}\ \Psi^2)\\
&\le 4g_{\max}^4 D_\mu \int_0^{T_{\mathrm{stop}}} E(s)\,\mathrm{d}s\\
&\quad + 2\mathcal{O}((g'_{\max}h)^4 T_{\mathrm{stop}}^2). && (g^4 \le g_{\max}^4;\ Eqs.\ (44) and\ (47))
\end{aligned}
$$

□

Remark 1 (Tight specialization for constant-coefficient schedules)**.** *When $\rho'_{\max} = g'_{\max} = 0$ (OU, VP-fixed-β), residuals and $\mathcal{O}$ vanish in the proof of Lemma 4; $\Psi(s) = g(T-s)^2\sqrt{\mathbb{E}\|(\mathrm{II})\|^2}$ alone and the two $(a+b)^2$ splits become equalities, so the bound tightens to:*

$$
\begin{aligned}
W_1^2 \le \mathbb{E}\|Z_{T_{\mathrm{stop}}}\|^2 &\le \Big(\int_0^{T_{\mathrm{stop}}} \Phi(s)\, g(T-s)^2\sqrt{\mathbb{E}\|(\mathrm{II})\|^2}\,\mathrm{d}s\Big)^2 && (\textit{square Eq. (43); } \mathcal{O}{=}0)\\
&\le D_\mu \int_0^{T_{\mathrm{stop}}} g(T-s)^4\,\mathbb{E}\|(\mathrm{II})\|^2\,\mathrm{d}s && (\textit{C–S})\\
&\le g_{\max}^4 D_\mu \int_0^{T_{\mathrm{stop}}} E(s)\,\mathrm{d}s, && (g^4 \le g_{\max}^4;\ Eqs.\ (44) and\ (47))
\end{aligned}
$$

which is a factor of 4 sharper than the lemma's general bound $4\,g_{\max}^4\,D_\mu \int E(s)\,\mathrm{d}s$, the gain coming from the two $(a+b)^2 \le 2(a^2+b^2)$ splits collapsing to equalities when residuals vanish.

Lemma 5 (Conditional moment bounds)**.** *Let $X_{kh} \overset{d}{=} \bar{X}_{\tau_k} \mid \bar{X}_T = \mathbf{x}^{(i)}$ (write $\mathbf{x} := \mathbf{x}^{(i)}$, $\tau_k := T - kh$). Set $\tilde{K}_q := \sqrt{2}\,K_q$ (any $K \ge K_q$ also satisfies Assumption 3; this rescaled proxy absorbs the $\sqrt{2}$ from the Tweedie origin bound, see proof of (ii)). Under Assumptions 1, 3 and 8 and Eq. (4):*

(i) $\mathbb{E}\|X_{kh}\|^2 \le 2\|\mathbf{x}\|^2 + 2C^\star K_q^2 + d \;=:\; M_2^X$, *with $C^\star$ the absolute constant of Eq. (51).*

(ii) $\mathbb{E}\|\nabla \ln q_{\tau_k}(X_{kh})\|^2 \le 2\tilde{K}_q^2/\mathfrak{s}_{\tau_k}^4 + 2L^2 M_2^X$.

(iii) $\mathbb{E}\|X_{(k+1)h} - X_{kh}\|^2 \le 2M_{\mathrm{drift}}h^2 + 2g_{\max}^2 dh$, *with* $M_{\mathrm{drift}} := (2\rho_{\max}^2 + 4g_{\max}^4 L^2)\,M_2^X + 4g_{\max}^4 \tilde{K}_q^2/\mathfrak{s}_{t_{\min}}^4$.

Proof of Lemma 5. **(i) Second moment of X_{kh}.** Set $v := \mathbf{x} - \mathfrak{m}_T\mu_q$ throughout this proof. By Bayes, $q^{(i)}(x_0) = w(x_0)q(x_0)$ with $w \propto e^{A-B}$, $A := \eta_T\langle v,\, x_0 - \mu_q\rangle$ (linear), $B := \dfrac{1}{2}\eta_T\mathfrak{m}_T\|x_0 - \mu_q\|^2$ (quadratic). The posterior is an exponential tilt of the sub-Gaussian q (Assumption 3) by $A - B$.

Sub-bound (a): covariance. Set $Z := \mathbb{E}_q[e^{A-B}]$. First, lower-bound Z:

$$
\begin{aligned}
\log Z &\ge \mathbb{E}_q[A - B] && (\text{Jensen: } e^x \text{ convex})\\
&= -\mathbb{E}_q[B] && (\mathbb{E}_q[x_0{-}\mu_q] = 0 \Rightarrow \mathbb{E}_q A = 0)\\
&= -\frac{1}{2}\eta_T\mathfrak{m}_T\,\mathbb{E}_q\|x_0 - \mu_q\|^2 && (\text{def of } B)\\
&\ge -\frac{1}{2}\,c_0\,\eta_T\mathfrak{m}_T K_q^2 && (Assumption\ 3:\ \mathbb{E}_q\|x_0{-}\mu_q\|^2 \le c_0 K_q^2)\\
&\ge -\frac{1}{2}\,c_0\,c_*^2. && (\eta_T\mathfrak{m}_T K_q^2 = (\mathfrak{m}_T K_q/\mathfrak{s}_T)^2 \le c_*^2,\ \eta_T\mathfrak{s}_T K_q \le c_*:\ Assumption\ 8)
\end{aligned}
$$

Hence

$$Z^{-1} \;\le\; \exp\!\big(\frac{1}{2}\, c_0\, c_*^2\big). \tag{48}$$

By Bayes change of measure and Cauchy–Schwarz:

$$\begin{aligned}
\mathrm{tr}(\mathrm{Cov}_{q^{(i)}}) &\le \mathbb{E}_{q^{(i)}}\|x_0-\mu_q\|^2 && (\text{Cov} \le \text{2nd moment about } \mu_q)\\
&= Z^{-1}\,\mathbb{E}_q\big[\|x_0-\mu_q\|^2\, e^{A-B}\big] && (\text{Bayes change of measure})\\
&\le Z^{-1}\,\sqrt{\mathbb{E}_q\|x_0-\mu_q\|^4}\,\sqrt{\mathbb{E}_q e^{2(A-B)}} && (\text{Cauchy–Schwarz under } q)\\
&\le Z^{-1}\cdot c_1 K_q^2\cdot\sqrt{\mathbb{E}_q e^{2(A-B)}} && (Assumption\ 3:\ \mathbb{E}_q\|x_0-\mu_q\|^4 \le c_1^2 K_q^4)\\
&\le Z^{-1}\cdot c_1 K_q^2\cdot\sqrt{\mathbb{E}_q e^{2A}}. && (-B\le 0 \Rightarrow e^{2(A-B)}\le e^{2A})
\end{aligned}$$

The remaining factor $\sqrt{\mathbb{E}_q e^{2A}}$ via the sub-Gaussian linear MGF bound (Assumption 3 + [Vershynin, 2018, Prop. 2.5.2]: $\mathbb{E}_q e^{\langle\lambda, x_0-\mu_q\rangle} \le 2\, e^{\|\lambda\|^2 K_q^2/2}$ for any $\lambda\in\mathbb{R}^d$). With $A=\eta_T\langle v, x_0-\mu_q\rangle$, set $\lambda := 2\eta_T v$ so $\|\lambda\|^2 = 4\eta_T^2\|v\|^2$, giving

$$\mathbb{E}_q\, e^{2A} \;\le\; 2\, e^{\|\lambda\|^2 K_q^2/2} \;=\; 2\, e^{2\eta_T^2\|v\|^2 K_q^2}. \tag{49}$$

Substituting into the trace bound:

$$\begin{aligned}
\mathrm{tr}(\mathrm{Cov}_{q^{(i)}}) &\le Z^{-1}\cdot c_1 K_q^2\cdot\sqrt{2\, e^{2\eta_T^2\|v\|^2 K_q^2}} && (Eq.\ (49))\\
&\le Z^{-1}\cdot c_1 K_q^2\cdot\sqrt{2}\, e^4 && (Eq.\ (\text{f}):\ \eta_T\|v\|K_q \le 2 \Rightarrow \eta_T^2\|v\|^2 K_q^2 \le 4)\\
&\le c\, K_q^2. && (Eq.\ (48)) \qquad (50)
\end{aligned}$$

Sub-bound (b): mean shift. By Bayes change of measure and $\mathbb{E}_q[w]=1$,

$$\mu_{q^{(i)}} - \mu_q = \mathbb{E}_q[x_0\, w] - \mu_q\,\mathbb{E}_q[w] = \mathbb{E}_q[(x_0-\mu_q)\, w].$$

Direct Cauchy–Schwarz on the magnitude (the directional information $\eta_T\Sigma_q v$ is unused downstream — only $\|\mu_{q^{(i)}}\|$ enters via $M_0^{(i)}$ in Lemma 3 step (iii)):

$$\begin{aligned}
\|\mu_{q^{(i)}}-\mu_q\| &= \|\mathbb{E}_q[(x_0-\mu_q)\, w]\| && (\text{Bayes}{+}\mathbb{E}_q[w]{=}1)\\
&\le \sqrt{\mathbb{E}_q\|x_0-\mu_q\|^2}\,\sqrt{\mathbb{E}_q[w^2]} && (\text{Cauchy–Schwarz})\\
&\le \sqrt{c_0}\, K_q\,\sqrt{\mathbb{E}_q[w^2]} && (Assumption\ 3;\ \mathbb{E}_q\|x_0-\mu_q\|^2 \le c_0 K_q^2)\\
&= \sqrt{c_0}\, K_q\, Z^{-1}\sqrt{\mathbb{E}_q[e^{2(A-B)}]} && (w = e^{A-B}/Z)\\
&\le \sqrt{c_0}\, K_q\, Z^{-1}\sqrt{\mathbb{E}_q[e^{2A}]} && (-2B\le 0)\\
&\le \sqrt{c_0}\, K_q\, Z^{-1}\sqrt{2}\, e^{\eta_T^2\|v\|^2 K_q^2} && (Eq.\ (49)\text{ at }\lambda = 2\eta_T v)\\
&\le \sqrt{c_0}\, K_q\cdot\sqrt{2}\, e^4\cdot Z^{-1} && (Eq.\ (\text{f});\ \eta_T^2\|v\|^2 K_q^2 \le 4)\\
&\le \sqrt{c_0}\, K_q\cdot\sqrt{2}\, e^4\cdot\exp\!\big(\frac{1}{2}c_0 c_*^2\big) && (Eq.\ (48))\\
&= C_\mu\, K_q, && \big(C_\mu := \sqrt{2c_0}\, e^4\, e^{c_0 c_*^2/2}\big)
\end{aligned}$$

all constants explicit in c_0, c_* from Assumptions 3 and 8. Triangle: $\|\mu_{q^{(i)}}\| \le \|\mu_q\| + \|\mu_{q^{(i)}}-\mu_q\| \le (\sqrt{2}+C_\mu)K_q$, hence $\|\mu_{q^{(i)}}\|^2 = \mathcal{O}(K_q^2)$.

Combine.

$$\mathbb{E}_{q^{(i)}}\|x_0\|^2 = \|\mu_{q^{(i)}}\|^2 + \mathrm{tr}(\mathrm{Cov}_{q^{(i)}}) \;\le\; C^\star K_q^2. \tag{51}$$

Sub-bound (c): bridge to X_{kh}. Bridge decomposition (Eq. (35) at $t_{\min} \to kh$) gives $\bar{X}_{\tau_k} = \mathfrak{a}_k x_0 + \mathfrak{b}_k \mathbf{x} + \mathfrak{g}_k \eta$ with $\eta \sim \mathcal{N}(0, I_d) \perp x_0$ and $\mathfrak{a}_k, \mathfrak{b}_k, \mathfrak{g}_k^2 \le 1$. Marginalising:

$$\begin{aligned}
\mathbb{E}\|X_{kh}\|^2 &= \mathbb{E}\big\|\mathfrak{a}_k x_0 + \mathfrak{b}_k \mathbf{x} + \mathfrak{g}_k \eta\big\|^2 && (Eq.\ (35)) \\
&= \mathbb{E}\big\|\mathfrak{a}_k x_0 + \mathfrak{b}_k \mathbf{x}\big\|^2 + \mathfrak{g}_k^2\, \mathbb{E}\|\eta\|^2 && (\eta \perp x_0,\ \mathbb{E}\eta = 0 \Rightarrow \text{cross term vanishes}) \\
&= \mathbb{E}\big\|\mathfrak{a}_k x_0 + \mathfrak{b}_k \mathbf{x}\big\|^2 + \mathfrak{g}_k^2\, d && (\mathbb{E}\|\eta\|^2 = d \text{ for } \eta \sim \mathcal{N}(0, I_d)) \\
&\le 2\mathfrak{a}_k^2\, \mathbb{E}_{q^{(i)}}\|x_0\|^2 + 2\mathfrak{b}_k^2\, \|\mathbf{x}\|^2 + \mathfrak{g}_k^2\, d && (\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2) \\
&\le 2\mathfrak{a}_k^2\, C^\star K_q^2 + 2\mathfrak{b}_k^2\, \|\mathbf{x}\|^2 + \mathfrak{g}_k^2\, d && (\mathbb{E}_{q^{(i)}}\|x_0\|^2 \le C^\star K_q^2 :\ Eq.\ (51)) \\
&\le 2\, C^\star K_q^2 + 2\|\mathbf{x}\|^2 + d && (\mathfrak{a}_k, \mathfrak{b}_k, \mathfrak{g}_k^2 \le 1) \\
&=: M_2^X. && (\text{name the bound})
\end{aligned}$$

(ii) Score moment.

Lipschitz reduction. Assumption 1 (L-Lipschitz score) with $y = 0$:

$$\begin{aligned}
\|\nabla \ln q_t(x) - \nabla \ln q_t(0)\| &\le L\,\|x\| && (Assumption\ 1,\ y = 0) \\
\|\nabla \ln q_t(x)\| &\le \|\nabla \ln q_t(0)\| + L\,\|x\|. && (\text{triangle})
\end{aligned}$$

Tweedie origin bound. Tweedie's formula Efron [2011]: $\nabla \ln q_t(x) = \big(\mathfrak{m}_t\, \mathbb{E}_q[x_0 \mid \bar{X}_t = x] - x\big)/\mathfrak{s}_t^2$. By Bayes, the posterior at $x = 0$ is

$$q^{(0)}(x_0) := q(x_0 \mid \bar{X}_t = 0) = \frac{1}{Z}\, q(x_0)\, e^{-c\,\|x_0\|^2}, \qquad c := \frac{\mathfrak{m}_t^2}{2\mathfrak{s}_t^2} \ge 0,\ \ Z := \mathbb{E}_q[e^{-c\|x_0\|^2}].$$

Chebyshev correlation inequality: for any real Y, f non-decreasing and g non-increasing, $\mathbb{E}[f(Y)g(Y)] \le \mathbb{E}[f(Y)]\,\mathbb{E}[g(Y)]$. Apply with $Y = \|x_0\|^2$, $f(Y) = Y$, $g(Y) = e^{-cY}$ ($c \ge 0$):

$$\mathbb{E}_q\big[\|x_0\|^2 e^{-c\|x_0\|^2}\big] \ \le\ \mathbb{E}_q\|x_0\|^2 \cdot \mathbb{E}_q\big[e^{-c\|x_0\|^2}\big] = \mathbb{E}_q\|x_0\|^2 \cdot Z. \tag{52}$$

Now bound $\|\nabla \ln q_t(0)\|$:

$$\begin{aligned}
\|\nabla \ln q_t(0)\| &= \frac{\mathfrak{m}_t\, \big\|\mathbb{E}_{q^{(0)}}[x_0]\big\|}{\mathfrak{s}_t^2} && (\text{Tweedie at } x = 0) \\
&\le \frac{\mathfrak{m}_t\, \sqrt{\mathbb{E}_{q^{(0)}}\|x_0\|^2}}{\mathfrak{s}_t^2} && (\text{Jensen on convex } \|\cdot\|^2) \\
&= \frac{\mathfrak{m}_t}{\mathfrak{s}_t^2} \sqrt{\frac{\mathbb{E}_q[\|x_0\|^2 e^{-c\|x_0\|^2}]}{Z}} && (\text{Bayes def of } q^{(0)}) \\
&\le \frac{\mathfrak{m}_t\, \sqrt{\mathbb{E}_q\|x_0\|^2}}{\mathfrak{s}_t^2} && (Eq.\ (52);\ \text{cancel } Z) \\
&\le \frac{\mathfrak{m}_t\, \sqrt{2K_q^2}}{\mathfrak{s}_t^2} = \frac{\mathfrak{m}_t\, \tilde{K}_q}{\mathfrak{s}_t^2} && (\mathbb{E}_q\|x_0\|^2 \le 2K_q^2 :\ Assumption\ 3;\ \tilde{K}_q := \sqrt{2}K_q) \\
&\le \frac{\tilde{K}_q}{\mathfrak{s}_t^2}. && (\mathfrak{m}_t \le 1)
\end{aligned}$$

Main bound. Square the Lipschitz reduction, take expectation, plug in:

$$\begin{aligned}
\mathbb{E}\|\nabla \ln q_{\tau_k}(X_{kh})\|^2 &\le \mathbb{E}\big(\|\nabla \ln q_{\tau_k}(0)\| + L\|X_{kh}\|\big)^2 && (\text{Lipschitz reduction; square}) \\
&\le 2\,\|\nabla \ln q_{\tau_k}(0)\|^2 + 2L^2\,\mathbb{E}\|X_{kh}\|^2 && ((a+b)^2 \le 2a^2 + 2b^2) \\
&\le \frac{2\tilde{K}_q^2}{\mathfrak{s}_{\tau_k}^4} + 2L^2\, M_2^X. && (\text{Tweedie origin: } \|\nabla \ln q_t(0)\|^2 \le \tilde{K}_q^2/\mathfrak{s}_t^2;\ \text{Sub-bound (c)})
\end{aligned}$$

(iii) Increment bound. Set $\Delta_{kh} := X_{(k+1)h} - X_{kh}$. The Anderson reversal of Eq. (1) with $\rho_s := \rho_{\mathrm{fwd}}(T-s)$, $g_s := g(T-s)$ splits the increment into drift and diffusion integrals:

$$\Delta_{kh} \;=\; \underbrace{\int_{kh}^{(k+1)h} \big[\rho_s X_s + g_s^2\, \nabla \ln q_{T-s}(X_s)\big]\,\mathrm{d}s}_{=:\ \text{drift}} \;+\; \underbrace{\int_{kh}^{(k+1)h} g_s\,\mathrm{d}B_s}_{=:\ \text{diffusion}}.$$

Sub-bound (a): drift integral.

$$\begin{aligned}
\mathbb{E}\big\|\textstyle\int \text{drift}\big\|^2 &\le h\,\mathbb{E}\textstyle\int_{kh}^{(k+1)h} \|\rho_s X_s + g_s^2 \nabla \ln q_{T-s}(X_s)\|^2\,\mathrm{d}s && \text{(Cauchy–Schwarz on } L^2[kh,(k+1)h]) \\
&\le h^2 \sup_s \mathbb{E}\big\|\rho_s X_s + g_s^2\,\nabla \ln q_{T-s}(X_s)\big\|^2 && \text{(integrand} \le \sup) \\
&\le h^2\big[2\rho_{\max}^2 \sup_s \mathbb{E}\|X_s\|^2 + 2g_{\max}^4 \sup_s \mathbb{E}\|\nabla \ln q_{T-s}\|^2\big] && ((a{+}b)^2 \le 2a^2{+}2b^2;\ \rho_s \le \rho_{\max},\ g_s \le g_{\max}) \\
&\le h^2\big[2\rho_{\max}^2\, M_2^X + 2g_{\max}^4\big(\frac{2\tilde K_q^2}{\mathfrak{s}_{t_{\min}}^4} + 2L^2\, M_2^X\big)\big] && \text{(lemma items (i),(ii); } \mathfrak{s}_s \ge \mathfrak{s}_{t_{\min}}) \\
&= \big[(2\rho_{\max}^2 + 4g_{\max}^4 L^2)\, M_2^X + \frac{4g_{\max}^4 \tilde K_q^2}{\mathfrak{s}_{t_{\min}}^4}\big]\, h^2 && \text{(distribute)} \\
&=: M_{\mathrm{drift}}\, h^2. && \text{(name } M_{\mathrm{drift}})
\end{aligned}$$

Sub-bound (b): diffusion integral.

$$\begin{aligned}
\mathbb{E}\big\|\textstyle\int \text{diffusion}\big\|^2 &= \mathbb{E}\Big\|\textstyle\int_{kh}^{(k+1)h} g_s\,\mathrm{d}B_s\Big\|^2 \\
&= \int_{kh}^{(k+1)h} g_s^2 \cdot d\,\mathrm{d}s && \text{(Itô isometry; } \mathrm{d}B_s \sim \mathcal{N}(0, I_d\,\mathrm{d}s)) \\
&\le g_{\max}^2\, d\, h. && (g_s \le g_{\max};\ \text{interval length } h)
\end{aligned}$$

Sub-bound (c): combine.

$$\begin{aligned}
\mathbb{E}\|\Delta_{kh}\|^2 &\le 2\,\mathbb{E}\big\|\textstyle\int \text{drift}\big\|^2 + 2\,\mathbb{E}\big\|\textstyle\int \text{diffusion}\big\|^2 && ((a{+}b)^2 \le 2a^2{+}2b^2) \\
&\le 2\, M_{\mathrm{drift}}\, h^2 + 2g_{\max}^2\, d\, h. && \text{(Sub-bounds (a),(b))}
\end{aligned}$$

For OU ($\rho_{\max} = 1, g_{\max}^2 = 2, g_{\max}^4 = 4$), $M_{\mathrm{drift}} = (2{+}16L^2)M_2^X + 16\,\tilde K_q^2/\mathfrak{s}_{t_{\min}}^4$ and the diffusion term reduces to $4dh$. □

B.2 Proof of Theorem 1

Proof of Theorem 1. By the triangle inequality,

$$W_1(q, \mathcal{H}^{(i)}) \le \underbrace{W_1(q, q^{(i)})}_{\le W_{\mathrm{init}} \text{ by Lemma 2}} + \underbrace{W_1(q^{(i)}, q^{(i)}_{t_{\min}})}_{\le W_{\mathrm{stop}} \text{ by Lemma 3}} + \underbrace{W_1(q^{(i)}_{t_{\min}}, \mathcal{H}^{(i)})}_{\le W_{\mathrm{gen}} \text{ by Lemma 4}}. \tag{53}$$

□

C Full Proof of Theorem 2

We prove Theorem 2 in Section C.1 and Lemma 1 in Section C.2.

Notation for this proof	
$Y_t^{(i)}, Y_t^{(j)}$	two SGM copies from $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ with the same Brownian motion
Z_t	difference $Y_t^{(i)} - Y_t^{(j)}$; deterministic for each noise realization
$\Delta s_{\theta,kh}$	per-step score difference $s_\theta(Y_{kh}^{(i)}, T-kh) - s_\theta(Y_{kh}^{(j)}, T-kh)$
κ_{kh}	local contraction coefficient (Eq. (16))
$\Lambda_h^{(ij)}$	discrete contraction sum $\sum_k h(\kappa_{kh}-\rho_k)$ (Eq. (17)); OU $\rho_k \equiv 1$ recovers $\sum h(\kappa-1)$
δ	projected mean separation, $\delta = \|\mathbb{E}[Z_{T_{\text{stop}}}]\|$
$\bar{e}$	projection direction, $\bar{e} = \mathbb{E}[Z_{T_{\text{stop}}}]/\|\mathbb{E}[Z_{T_{\text{stop}}}]\|$
V, U	1-D projections of $Y_{T_{\text{stop}}}^{(i)}, Y_{T_{\text{stop}}}^{(j)}$ onto $\bar{e}$
σ_V^2, σ_U^2	variances of V and U

C.1 Proof of Theorem 2

Proof of Theorem 2. If $\delta = 0$, the trivial bound $\mathrm{TV} \geq 0$ applies. For $\delta > 0$, let V, U, σ_V, σ_U be as in Eq. (13); the assumed finite output second moments ensure that both projected variances are finite. Set $c := \mathbb{E}[V] - \delta/2 = \mathbb{E}[U] + \delta/2$.

$$
\begin{aligned}
\mathrm{TV}(\mathcal{H}^{(i)}, \mathcal{H}^{(j)}) &\geq \mathrm{TV}(V, U) \\
&\geq \Pr(V \geq c) - \Pr(U \geq c) && \text{(def TV; test set } \{x \geq c\}) \\
&= 1 - \Pr(V < c) - \Pr(U \geq c) && (\Pr(V \geq c) = 1 - \Pr(V < c)) \\
&= 1 - \Pr\big(\mathbb{E}[V] - V > \frac{\delta}{2}\big) - \Pr\big(U - \mathbb{E}[U] \geq \frac{\delta}{2}\big) && (c \text{ from } \delta) \\
&\geq 1 - \frac{\sigma_V^2}{\sigma_V^2 + \delta^2/4} - \frac{\sigma_U^2}{\sigma_U^2 + \delta^2/4}. && (\text{Cantelli, } a = \frac{\delta}{2})
\end{aligned}
$$

Each fraction $\in [0,1] \Rightarrow$ bound $\in [-1,1]$; positive iff sum of fractions $< 1 \Leftrightarrow \delta^4 > 16\sigma_V^2\sigma_U^2 \Leftrightarrow \delta^2 > 4\sigma_V\sigma_U$. □

C.2 Proof of Lemma 1

Proof of Lemma 1.

$$
\Delta s_{\theta,kh} := s_\theta(Y_{kh}^{(i)}, T-kh) - s_\theta(Y_{kh}^{(j)}, T-kh), \qquad g_k := g(T-kh).
$$

$$
C := \rho_{\max} + g_{\max}^2 L_s, \qquad \Delta_k := 2h(\rho_k - \kappa_{kh}) + h^2(\rho_k^2 - 2\rho_k\kappa_{kh} + g_k^4 L_s^2),
$$

$$
\|\Delta s_{\theta,kh}\| \leq L_s\|Z_{kh}\|, \qquad |\kappa_{kh}| \leq g_k^2 L_s, \qquad g_k^2\langle Z_{kh}, \Delta s_{\theta,kh}\rangle = -\kappa_{kh}\|Z_{kh}\|^2.
$$

$$
Z_{(k+1)h} = (1+\rho_k h)Z_{kh} + g_k^2 h\, \Delta s_{\theta,kh}. \tag{54}
$$

$$
\begin{aligned}
\|Z_{(k+1)h}\|^2 &= (1+\rho_k h)^2\|Z_{kh}\|^2 + 2g_k^2 h(1+\rho_k h)\langle Z_{kh}, \Delta s_{\theta,kh}\rangle + g_k^4 h^2\|\Delta s_{\theta,kh}\|^2 \\
&\leq \big[(1+\rho_k h)^2 - 2h(1+\rho_k h)\kappa_{kh} + g_k^4 h^2 L_s^2\big]\|Z_{kh}\|^2 \\
&= (1+\Delta_k)\|Z_{kh}\|^2 \leq e^{\Delta_k}\|Z_{kh}\|^2 \\
&\leq \exp\big[-2h(\kappa_{kh} - \rho_k) + h^2 C^2\big]\,\|Z_{kh}\|^2.
\end{aligned} \tag{55}
$$

$$
\begin{aligned}
\|Z_{T_{\text{stop}}}\| &\leq \|Z_0\| \prod_{k=0}^{N-1} \exp\Big[-h(\kappa_{kh} - \rho_k) + \frac{1}{2}h^2 C^2\Big] \\
&= \|Z_0\| \exp\left[-\sum_{k=0}^{N-1} h(\kappa_{kh} - \rho_k) + \frac{1}{2}Nh^2C^2\right] \\
&= \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \exp\Big[-\Lambda_h^{(ij)} + \varepsilon_{\text{disc}}\Big], \qquad Nh = T_{\text{stop}}.
\end{aligned} \tag{56}
$$

$$
\begin{aligned}
\delta = \|\mathbb{E}_\omega Z_{T_{\text{stop}}}\| &\leq \mathbb{E}_\omega\|Z_{T_{\text{stop}}}\| && \text{(Jensen)} \\
&\leq \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|\, \mathbb{E}_\omega \exp\Big[-\Lambda_h^{(ij)} + \varepsilon_{\text{disc}}\Big].
\end{aligned}
$$

□

D SDE Instances (Theorems 1 and 2 specialised)

Both Theorems 1 and 2 are stated for the abstract forward family (1): Theorem 1 via $(\rho_{\mathrm{fwd}}, g, \mathfrak{m}_t, \mathfrak{s}_t^2, \eta_T, \mu_{\mathrm{eff}}, \mathfrak{a}, \mathfrak{b}, \mathfrak{g})$, and Theorem 2 via the g^2-embedded κ (Eq. (16)), the abstract Z-recursion of Lemma 1, and the threshold $\bar{\kappa} \gtrless \bar{\rho}_{\mathrm{fwd}}$ in Eq. (17). Each in-scope SDE is one row of Table 1 (in §3); specialising the row's $(\rho_{\mathrm{fwd}}, g, \mathfrak{m}_t, \mathfrak{s}_t^2)$ yields the per-SDE bounds for both theorems directly.

For Gaussian q with covariance Σ_q, the closed-form μ_{eff} per row is

$$\mu_{\mathrm{eff}}^{\mathrm{OU}}(t) = -1 - 2S(t) = \frac{2}{e^{-2(T-t)}\|\Sigma_q\|_{\mathrm{op}} + \sigma_{T-t}^2} - 1,$$
$$\mu_{\mathrm{eff}}^{\mathrm{VP}}(t) = \frac{\beta(T-t)}{\bar{\alpha}_{T-t}\|\Sigma_q\|_{\mathrm{op}} + 1 - \bar{\alpha}_{T-t}} - \frac{\beta(T-t)}{2},$$
$$\mu_{\mathrm{eff}}^{\mathrm{VE}}(t) = \frac{2\sigma(T-t)\dot{\sigma}(T-t)}{\sigma(T-t)^2 + \|\Sigma_q\|_{\mathrm{op}}} \geq 0.$$

VE non-decaying-shift contribution. For VE the third term of Eq. (6), $\eta_T \mathfrak{s}_T \sqrt{C_R'' C_P}\, K_q$, specialises at $\mathfrak{m}_T = 1$, $\mathfrak{s}_T = \sigma(T)$, $\eta_T = 1/\sigma(T)^2$ to $\sqrt{C_R'' C_P}\, K_q/\sigma(T)$ — the genuine non-decaying-shift correction at scale $K_q/\sigma(T)$. For OU/VP, $\mathfrak{s}_T \leq 1$ keeps the third term sub-leading to the $\eta_T^{3/2}$ remainder; Lemma 2 produces all three regimes from the same bound.

E Empirical Observations and Compute

E.1 ODE degeneracy / DDS witness

We observe empirically that the probability-flow ODE collapses each per-start law $\mathcal{H}_{\mathrm{ODE}}^{(i)}$ to a (near-)point mass, while the shown SDE outputs retain stochastic variability. Witnesses:

Setting	Witness	Result
2D toy (3 datasets)	Fig. 1	SDE recovers q from a point start
4-mode GMM, true score	Fig. 2	per-start "imprint" visible
MNIST (d=784)	Fig. 4	ODE std $< 10^{-10}$; SDE diverse digits
CIFAR-10 (d=3072)	Fig. 6	pre-trained DDPM, SDE half only

E.2 Score-network architectures and training hyperparameters

The figures of Section E.1 come from four experimental settings; Table 2 lists each. The trained 2D and MNIST SDE experiments use OU ($\rho_{\mathrm{fwd}} \equiv 1$, $g \equiv \sqrt{2}$), with $T \in \{0.5, 1, 2, 3, 5\}$ in the 2D sweeps and $T = 2.0$ on MNIST. MNIST and CIFAR-10 are loaded via standard `torchvision` pipelines; 2D synthetic datasets use $n = 10000$ samples per dataset with the recorded experiment seeds. CIFAR-10 instead uses the pretrained discrete DDPM schedule, outside the constant-g formal scope.

Image inference setup. The MNIST witness displays five fixed starts and eight stochastic outputs per start, plus an ODE comparison. For CIFAR-10, five test images are noised to step 200 of the pre-trained DDPM's 1000 steps, then five stochastic outputs are shown per source, without retraining. These are qualitative illustrations; the CIFAR-10 schedule is outside the constant-g theorem scope.

E.3 Compute and broader impacts

Compute. Final numbers: 3×RTX 3090 (CUDA 12.8). 2D training ∼15 min across 3 GPUs; MNIST UNet ∼15 min on one GPU; conditional MNIST ∼20 min; CIFAR-10 demonstration uses a pre-trained model.

Table 2: Architecture and training hyperparameters per experimental setting. The conditional MNIST score net is trained ahead of time with the listed architecture; for that row we report inference-time configuration only.

Setting	Score net	Optimizer	Epochs
2D toy GMM (d=2)	MLP score net: hidden 128, 3 layers, time-embed dim 64, sinusoidal+embedding time encoding; SiLU activations. VE override: hidden 256, 4 layers, $\sigma_{\max}$=10	Adam, lr $2{\cdot}10^{-3}$, batch 512	5000 (VE: 10000)
MNIST flat (d=784)	UNet score net: channels $[32, 64, 128, 256]$, 2 res-blocks per scale, embed dim 128, time-embed dim 64, attention at the lowest scale	Adam, lr $2{\cdot}10^{-4}$, batch 128, cosine schedule	30
Conditional MNIST	Conditional UNet: channels $[32, 64, 128]$, 2 res-blocks, embed dim 96, 100 noise steps, 10 classes, p_{uncond}=0.1, dropout 0.05; CFG-trained	inference-time only (pre-computed checkpoint)	—
CIFAR-10 (d=3072)	Pre-trained `google/ddpm-cifar10-32` (`UNet2DModel`, `DDPMScheduler`, 1000 timesteps); inference-only	—	—

Broader impacts. This paper is a theoretical analysis of existing diffusion samplers and does not introduce new models, datasets, or applications. No human-subjects data is used; public datasets (MNIST CC BY-SA 3.0, CIFAR-10 MIT) and a publicly released pre-trained model (Apache 2.0) are used as-is.

E.4 Conditional DDS Supplement

Here *conditional DDS* fixes both a starting realization and a class label y, then samples with $s_\theta(x, t, y)$. The MNIST UNet uses unconditional-label dropout 0.1 and guidance $s_w = (1 + w)s_\theta(x, t, y) - ws_\theta(x, t, \varnothing)$; thus $w = 0$ denotes pure conditional sampling. With $y = 3$, ID (digit 3) and OOD (digit 7) starts produce different outcomes (Figs. 3 and 5). As discussed in Section 6, this exploratory conditional extension is outside the unguided Theorems 1 and 2.

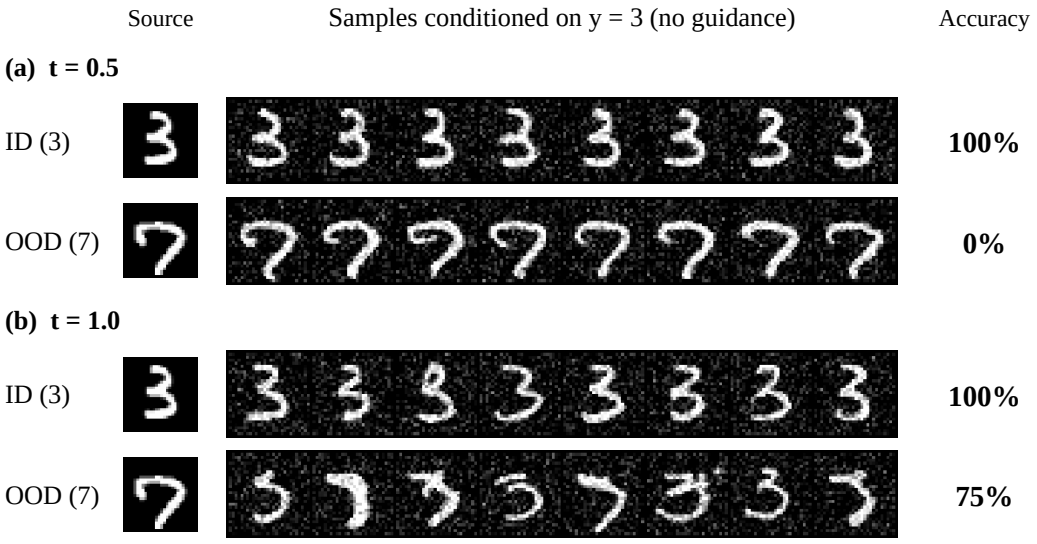

Figure 3: Conditional DDS reveals starting-point inconsistency. Conditioning $y = 3$; ID start (a 3) vs. OOD start (a 7) at $t_{\text{start}} = 0.5$ (top) and $t_{\text{start}} = 1.0$ (bottom). At $t_{\text{start}} = 0.5$ the OOD start overrides conditioning; at $t_{\text{start}} = 1.0$ the stochastic reverse SDE partly restores class consistency.

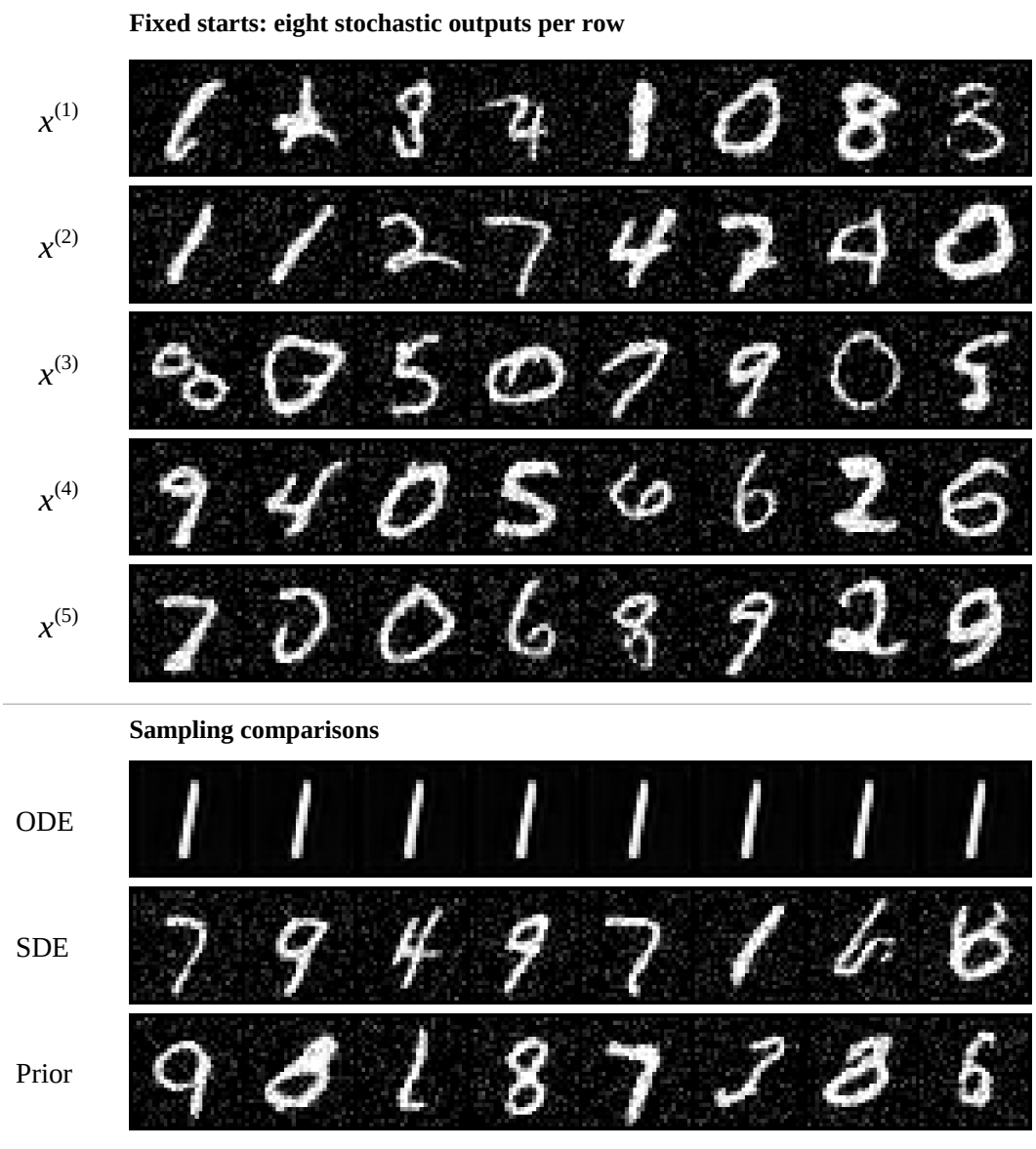

Figure 4: DDS on MNIST ($d = 784$). Rows $\mathbf{x}^{(1)}$–$\mathbf{x}^{(5)}$: each fixed starting point produces 8 diverse recognizable digits via the stochastic reverse SDE. ODE row: the probability-flow ODE produces a single constant digit per start. SDE/Prior rows: standard stochastic and prior-initialized sampling for comparison.

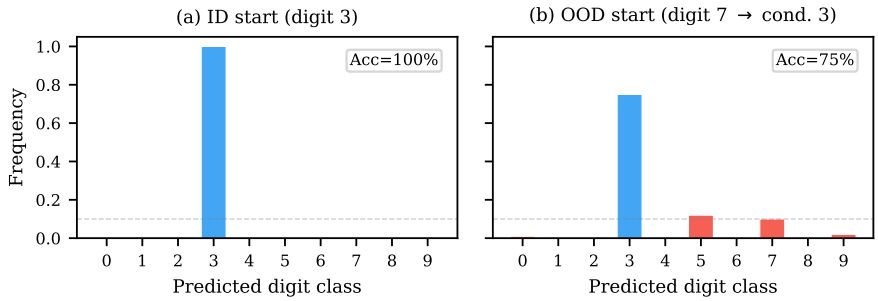

Figure 5: Digit class distribution of conditional DDS samples at $t = 1.0$, CFG$= 0$. (a) ID start (digit 3): near-delta at class 3. (b) OOD start (digit 7 conditioned on 3): confused mixture dominated by digits $3, 5, 7$.

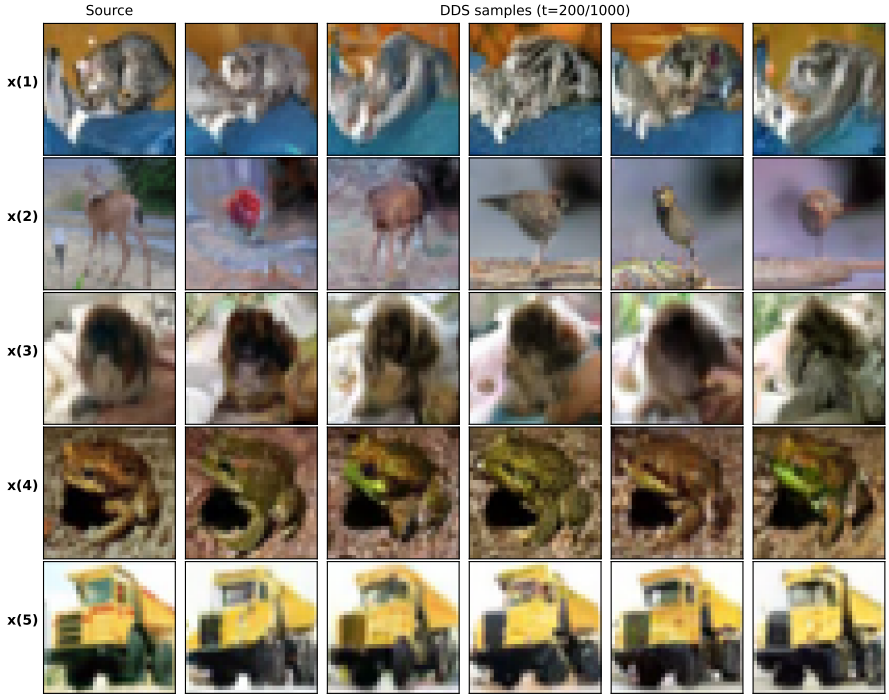

Figure 6: DDS on CIFAR-10 ($d = 3072$) with a pretrained DDPM. Each row shows a source test image forward-diffused to step $200/1000$, followed by five stochastic samples. Outputs vary around the source and may change class. This is a qualitative witness outside the constant-g scope; visual appearance does not certify a small W_1 or membership in the data distribution's support.

E.5 Quantitative bound diagnostics

We report the quantitative comparisons presented in the author response, using the trained OU score on Swiss Roll with 10000 data samples. The horizon sweep fixes $t_{\min} = 0.05$, uses $N = 200$ Euler–Maruyama steps with $h = (T - t_{\min})/N$, and draws 2000 samples per start; the two starts have separation 4.

Estimation. $\widehat{W}_1$ denotes empirical sliced-W_1 with 128 random projections; $\widehat{\mathrm{TV}}$ is a kNN-classifier proxy with $k = 5$. For the conditioning-cost proxy, we evaluate the χ^2 form $\sqrt{C_P \chi^2(q^{(i)} \| q)}$ of Eq. (25), using $C_P = 4\|\widehat{\Sigma}_q\|_{\mathrm{op}}$ as a covariance-scale proxy and self-normalised Monte Carlo over the 10000 data samples for χ^2; reported bootstrap deviations are at most 3.3%. The stopping-cost proxy uses Eq. (7) with $M_0^{(i)}$ estimated by the same posterior Monte Carlo. The generation-cost proxy is the synchronous-coupling cost from 2000 paired trajectories with a 500-point empirical marginal-score reference. These numerical proxies do not certify the Poincaré constant or evaluate all constants in the analytic generation bound.

For diversity, we estimate $\delta, \sigma_V, \sigma_U$ from 2000 paired trials, projecting onto the empirical mean-difference direction, and substitute them into Eq. (14). We also estimate $\mathrm{env} := \mathbb{E}\|Z_{T_{\mathrm{stop}}}\|$ using the shared-noise recursion, so that $\delta = \|\mathbb{E} Z_{T_{\mathrm{stop}}}\| \le \mathrm{env}$. The resulting TV right-hand side is a plug-in diagnostic, without a finite-sample confidence guarantee.

Table 3: Horizon sweep on Swiss Roll. The three cost columns and their sum are empirical proxies; TV RHS is Eq. (14) evaluated at empirical moments.

T	$\widehat{W}_1$	W_{init}	W_{stop}	W_{gen}	Sum	δ	env	$\widehat{\mathrm{TV}}$	TV RHS
0.5	0.63	2.24	0.55	0.04	2.83	2.19	2.20	0.89	+0.52
1.0	0.37	1.21	0.47	0.05	1.73	1.29	1.30	0.52	−0.22
2.0	0.13	0.44	0.45	0.06	0.95	0.46	0.46	0.14	−0.88
3.0	0.06	0.16	0.44	0.07	0.67	0.17	0.17	0.06	−0.98
5.0	0.07	0.02	0.44	0.07	0.53	0.02	0.02	0.00	−1.00

The decomposition proxy exceeds observed sliced-W_1 throughout Table 3 and decreases with T; at longer horizons, the fixed stopping cost dominates. The diversity diagnostic becomes negative at $T = 1$, although $\widehat{\mathrm{TV}} = 0.52$: loss of a positive certificate does not mean loss of diversity.

Certificate crossing. On a grid local to the crossing, the intersection of log-linear fits to δ^2 and $4\sigma_V\sigma_U$ gives $T^\star = 0.786 \pm 0.002$. Replacing δ by the empirical envelope env in the same calculation gives 0.790 ± 0.002, a 0.5% gap. Each $\pm$ is the standard deviation across three seed-level fitted roots. This compares the measured crossing with an empirical envelope-based prediction.

Early stopping. At $T = 5$, a master grid with $h = T/1000 = 0.005$ is truncated at each requested $t_{\min}$. In Table 4, the stopping-cost proxy decreases approximately as $\sqrt{t_{\min}}$, while the generation-cost proxy increases modestly. Their tradeoff reduces the total proxy from 0.541 to 0.275; observed sliced-W_1 lies in $[0.048, 0.056]$.

Table 4: Early-stopping sweep on Swiss Roll. Sum includes the conditioning-cost proxy; values are rounded independently.

$t_{\min}$	W_{stop} proxy	W_{gen} proxy	Sum
0.05	0.441	0.078	0.541
0.02	0.281	0.094	0.398
0.01	0.199	0.104	0.325
0.005	0.141	0.112	0.275

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state scope and contributions; supporting analysis is in Sections 4 and 5.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 discusses scope and open directions.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Assumptions are listed in Section 3; full proofs are in Sections B and C.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Architectures and hyperparameters for the qualitative demonstrations are listed in Section E; sampling grids, estimators, and sample counts for the quantitative diagnostics are given in Section E.5.

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

   Answer: [N/A]

   Justification: This is a theoretical work releasing no new datasets, code, or model checkpoints. Datasets used (MNIST, CIFAR-10) are standard public benchmarks; the CIFAR-10 demonstration uses the publicly released pretrained model `google/ddpm-cifar10-32`.

6. **Experimental setting/details**

   Answer: [Yes]

   Justification: See Section E.

7. **Experiment statistical significance**

   Answer: [Yes]

   Justification: Section E.5 reports bootstrap deviations for the conditioning-cost proxy and standard deviations across three seeds for the fitted certificate crossings. The remaining diagnostic table entries are point estimates without confidence intervals.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: Compute budget reported in Section E; results reproducible on a single consumer GPU.

9. **Code of ethics**

   Answer: [Yes]

   Justification: Theoretical work; no ethical concerns.

10. **Broader impacts**

    Answer: [Yes]

    Justification: Broader impacts discussed in Section E.3.

11. **Safeguards**

    Answer: [N/A]

    Justification: No new high-risk assets released.

12. **Licenses for existing assets**

    Answer: [Yes]

    Justification: MNIST and CIFAR-10 used under their standard licenses; pre-trained CIFAR-10 DDPM from HuggingFace (Apache 2.0).

13. **New assets**

    Answer: [N/A]

    Justification: No new datasets or pre-trained models released.

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

    Answer: [N/A]

    Justification: Not applicable.

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

    Answer: [N/A]

    Justification: Not applicable.

16. **Declaration of LLM usage**

    Answer: [N/A]

    Justification: LLMs are not part of the core methods.