OPTIMIZING NOISE SCHEDULES IN DENOISING DIFFUSION PROBABILISTIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how to choose noise schedules in denoising diffusion probabilistic models by analyzing the error that can remain even with the best Gaussian reverse kernels. We derive an exact identity for the averaged one-step KL error and use it to obtain an explicit upper bound for every step after the first, assuming only a finite second moment of the data distribution. With the first-step noise and terminal signal level fixed, we minimize the resulting bound in closed form. The unique optimal schedule has geometrically growing noise-to-signal odds, or equivalently, linearly decreasing log-SNR. Across seven image datasets and class-conditional CIFAR-10, the resulting schedules improve the variational bound on negative log-likelihood compared with linear and cosine baselines, while also improving sample-quality metrics such as FID in many cases.

1 INTRODUCTION

Diffusion and flow-based generative models are the state of the art in image (Dhariwal & Nichol, 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Peebles & Xie, 2023; Esser et al., 2024) and video (Ho et al., 2022; Blattmann et al., 2023; Polyak et al., 2024) generation. For diffusion models, the choice of noise schedule can substantially affect performance (Nichol & Dhariwal, 2021; Chen, 2023; Lin et al., 2024; Hoogeboom et al., 2023). It is therefore very important to choose the right noise schedule.

This paper studies the question of choosing noise schedules for denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020). DDPMs generate samples by reversing a noise-adding Markov chain. The forward chain progressively corrupts data with Gaussian noise, and the reverse chain undoes this by applying a learned sequence of Gaussian denoising steps. Even when the denoiser is perfect – that is, when each reverse Gaussian transition is the best possible Gaussian fit to the true reverse conditional – some error remains because the true reverse conditional is not exactly Gaussian. We refer to this irreducible mismatch as *model error*, and it is the focus of this paper.

Problem Setup. We work in the standard DDPM framework of Ho et al. (2020). The forward process is a Markov chain $(X_t)_{t=0}^T$ in $\mathbb{R}^d$ defined by

$$X_t = \sqrt{1-\beta_t}\, X_{t-1} + \sqrt{\beta_t}\, Z_t, \qquad Z_t \sim \mathcal{N}(0, I_d), \tag{1}$$

where $\beta_1, \dots, \beta_T \in (0,1)$ are the noise parameters. Writing $\bar{\alpha}_t := \prod_{i=1}^t (1-\beta_i)$, with $\bar{\alpha}_0 := 1$, we have the explicit representation

$$X_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1-\bar{\alpha}_t}\, Z, \qquad Z \sim \mathcal{N}(0, I_d).$$

Thus X_t is a progressively Gaussianized version of X_0.

For each $t \in \{1, \dots, T\}$, write

$$q_t(\cdot \mid y) := P(X_{t-1} \in \cdot \mid X_t = y)$$

for the true reverse kernel, and let $g_t^\star(\cdot \mid y)$ be the Gaussian density minimizing $D_{\mathrm{KL}}(q_t(\cdot \mid y) \,\|\, g)$ over all Gaussian densities g on $\mathbb{R}^d$. The *Gaussian reverse chain* starts from $\widetilde{X}_T \sim \mathcal{N}(0, I_d)$ and applies the kernels $g_t^\star$ in reverse; we write $\widetilde{P}_0$ for the resulting law at time 0.

The question we study in this paper is:

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

How should the schedule $\beta_1, \ldots, \beta_T$ be chosen in order to minimize $D_{\mathrm{KL}}(P_0 \parallel \widetilde{P}_0)$?

In practice, the Gaussian reverse kernels are learned using neural networks and therefore need not equal $g_t^\star$, introducing additional learning error. We set this learning error aside and focus on choosing the noise schedule to minimize the model error. This yields a clean mathematical formulation that isolates the contribution of the noise schedule.

Following Kingma et al. (2021), we call $\bar{\alpha}_t/(1-\bar{\alpha}_t)$ the signal-to-noise ratio of step t, write $\lambda_t := \log\big(\bar{\alpha}_t/(1-\bar{\alpha}_t)\big)$ for its logarithm, the log-SNR, and call $\lambda_{t-1} - \lambda_t \geq 0$ the log-SNR increment of step t. A schedule is equivalently a decreasing sequence of log-SNR values.

Our Contributions.

- **An exact and interpretable characterization of Gaussianization error.** We characterize the averaged KL error that remains when each true reverse conditional is replaced by a Gaussian with the same mean and covariance (Proposition 2 and Theorem 1). Our characterization is exact and interpretable, and the proofs are elementary.
- **Closed-form schedule optimization.** We obtain an explicit upper bound on the Gaussianization error for every reverse step after the first (Theorems 2 and 3). With the first-step noise and terminal signal level fixed, we minimize the accumulated bulk bound in closed form, obtaining a schedule that is linear in log-SNR (Theorem 4). This schedule has been used in practice (Lu et al., 2022), and here we show that it arises from minimizing an upper bound on the error due to the Gaussian approximation. At fixed endpoints, the optimized bulk bound is of order $1/T^2$, with the first-step and terminal errors treated separately (Remark 3). These bounds require only the weak assumption that X_0 has a finite second moment.
- **Empirical results.** We compare our schedule against the linear schedule of Ho et al. (2020) and the cosine schedule of Nichol & Dhariwal (2021), matching the endpoints of our schedule to those of each baseline. Holding these endpoints and all training settings fixed within each comparison, the optimized schedule improves the variational bound on negative log-likelihood in all 30 matched comparisons across seven unconditional datasets and class-conditional CIFAR-10, under both variational and hybrid training objectives (Sections 6.2 and G.2). It also improves perceptual metrics such as FID and CLIP-MMD in many cases.

Paper Structure. Section 2 sets up the pathwise KL decomposition, Section 3 derives the exact derivative formula and the one-step bound, Section 4 applies the one-step bound to the DDPM chain, Section 5 solves the schedule optimization problem, and Section 6 reports the experiments. We discuss some related work below.

1.1 Related Work

We briefly review prior work on choosing noise schedules for diffusion models to place our contributions in the context of prior literature. A fuller discussion appears in Section A in the appendix.

Noise schedules for diffusion models were initially chosen heuristically on the basis of empirical performance. Early examples include the linear schedule of DDPM (Ho et al., 2020) and the cosine schedule of Nichol & Dhariwal (2021). Later work explored other schedule choices and their effects on sample quality (Karras et al., 2022; Chen, 2023; Hoogeboom et al., 2023).

Other papers choose noise schedules by optimizing an explicit criterion; our paper belongs to this category. Aghapour et al. (2026) show that the the linear log-SNR schedule can be derived by minimizing an entropy-based bound on reverse-SDE discretization error. Another related work is Kong et al. (2026), who use Fisher information to formulate schedule design as an optimal-control problem. Both study the error of a specified reverse-SDE discretization. Our approach instead characterizes exactly the per-step error that remains even with the best Gaussian reverse transition kernels, isolating the effect of the Gaussian approximation. Our resulting bound requires only a finite second moment, without any additional entropy or distributional assumptions, and we compute its unique minimizer in closed form.

Other relevant work studies variational training objectives and schedule learning (Kingma et al., 2021; Kingma & Gao, 2023), numerical solvers and sampling grids (Watson et al., 2021; Lu et al.,

2022; Sabour et al., 2024; Williams et al., 2024), and how schedules affect sampling error (Strasman et al., 2025; Benita et al., 2025; Sun et al., 2026). Work on reverse transitions includes optimal Gaussian means and covariances (Bao et al., 2022b;a), convergence guarantees for moment-matched Gaussian kernels (Liang et al., 2025), and more expressive alternatives to Gaussian transitions (Xiao et al., 2022; Guo et al., 2023).

2 PATHWISE KL DECOMPOSITION

Let P_t denote the law of X_t under the forward chain (1). Our goal is to bound $D_{\mathrm{KL}}(P_0 \parallel \widetilde{P}_0)$, the KL divergence between the true data distribution and the output of the Gaussian reverse chain. However, this quantity is difficult to analyze directly, because the marginal divergence does not decompose into per-step contributions. To obtain such a decomposition, we lift the comparison from marginals to full path measures on $(X_T, \dots, X_0)$:

- the *true reverse law* P_{rev}, with $X_T \sim P_T$ and $X_{t-1} \mid X_t \sim q_t(\cdot \mid X_t)$;
- the *Gaussian reverse law* $\widetilde{P}_{\mathrm{rev}}$, with $\widetilde{X}_T \sim \mathcal{N}(0, I_d)$ and $\widetilde{X}_{t-1} \mid \widetilde{X}_t \sim g_t^\star(\cdot \mid \widetilde{X}_t)$.

Since P_0 and $\widetilde{P}_0$ are the time-0 marginals of these two path measures, the data processing inequality (Cover & Thomas, 2006, Theorem 2.8.1) gives

$$D_{\mathrm{KL}}(P_0 \parallel \widetilde{P}_0) \le D_{\mathrm{KL}}(P_{\mathrm{rev}} \parallel \widetilde{P}_{\mathrm{rev}}). \tag{2}$$

The chain rule for KL divergence (Cover & Thomas, 2006, Theorem 2.5.3) then decomposes the right-hand side into per-step terms:

$$D_{\mathrm{KL}}(P_{\mathrm{rev}} \parallel \widetilde{P}_{\mathrm{rev}}) = \underbrace{D_{\mathrm{KL}}(P_T \parallel \mathcal{N}(0, I_d))}_{\text{terminal}} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[D_{\mathrm{KL}}(q_t(\cdot \mid X_t) \parallel g_t^\star(\cdot \mid X_t))]}_{\text{per-step Gaussianization cost}}. \tag{3}$$

The terminal term measures how far the forward chain is from pure noise at time T. Each per-step term is the expected KL cost of replacing the true reverse conditional at step t by its best Gaussian approximation. The remainder of the paper is devoted to bounding these individual terms and optimizing the schedule.

3 ONE-STEP GAUSSIAN APPROXIMATION: EXACT IDENTITY AND BOUND

The pathwise decomposition (3) reduces the problem to the error of approximating one reverse conditional by a Gaussian. We first study this error in a general additive-noise model, and then specialize the result to DDPM in the next section.

3.1 THE ONE-STEP APPROXIMATION PROBLEM

Let U be a random vector in $\mathbb{R}^d$ with $\mathbb{E}\|U\|^2 < \infty$, let $W, Z \sim \mathcal{N}(0, I_d)$ be independent of each other and of U, and fix $\eta > 0$. Define

$$S := U + \sqrt{\eta}\, W, \qquad Y_u := S + \sqrt{u}\, Z, \qquad u \ge 0. \tag{4}$$

Here η is the Gaussian smoothing already present in the signal, and u is the variance of the new observation noise. For $u > 0$, let

$$q_u(\cdot \mid y) := \mathcal{L}(S \mid Y_u = y), \qquad g_u^\star(\cdot \mid y) := \mathcal{N}(\mathbb{E}[S \mid Y_u = y], \mathrm{Cov}(S \mid Y_u = y)).$$

Here, $\mathcal{L}(S \mid Y_u = y)$ denotes the conditional distribution of S given $Y_u = y$. We study the averaged error of this moment-matched Gaussian approximation:

$$\mathcal{K}(u) := \mathbb{E}_{Y_u}\Big[D_{\mathrm{KL}}(q_u(\cdot \mid Y_u) \parallel g_u^\star(\cdot \mid Y_u))\Big]. \tag{5}$$

We set $\mathcal{K}(0) := 0$; continuity at zero will be established below. Moment matching is the appropriate choice because it minimizes forward KL over Gaussian distributions.

Lemma 1 (The KL-best Gaussian is the moment-matched Gaussian; proof in Appendix C.1.1)**.** *Let Q be any probability measure on $\mathbb{R}^d$ with finite second moment. Among all Gaussian distributions g on $\mathbb{R}^d$, the quantity $D_{\mathrm{KL}}(Q \parallel g)$ is minimized by the Gaussian with the same mean and covariance as Q.*

To analyze $\mathcal{K}$, let p_u be the density of Y_u and write

$$\ell_u := \log p_u, \qquad H_u := \nabla^2 \ell_u, \qquad A_u := I_d + uH_u.$$

Note that Gaussian smoothing makes p_u positive and smooth. The proposition below expresses the mean and covariance of the reverse conditional distribution in terms of these quantities.

Proposition 1 (Posterior mean and covariance in the generic setup; proof in Appendix C.1.2)**.** *For the setup* (4)*, the KL-best Gaussian posterior is*

$$g_u^\star(\cdot \mid y) = N\big(m_u(y), \Sigma_u(y)\big),$$

where $m_u(y) = y + u\nabla \ell_u(y)$ and $\Sigma_u(y) = uA_u(y) = u\big(I_d + u\nabla^2 \ell_u(y)\big)$. Moreover,

$$A_u(y) \succeq \frac{\eta}{\eta + u} I_d \succ 0 \qquad \textit{for all } u > 0,\ y \in \mathbb{R}^d.$$

3.2 An entropy representation

For each observation y, the KL error equals the entropy of the matched Gaussian minus the entropy of the true posterior. Averaging over Y_u and applying the entropy chain rule gives the following identity.

Proposition 2 (Entropy representation of $\mathcal{K}(u)$; proof in Appendix C.1.3)**.** *For every $u > 0$,*

$$\mathcal{K}(u) = h(Y_u) - h(S) + \frac{1}{2}\,\mathbb{E}\big[\log\det A_u(Y_u)\big]. \tag{6}$$

Lemma 2 (Right-continuity of $\mathcal{K}$ at the origin; proof in Appendix C.1.4)**.** *If $\mathbb{E}\|U\|^2 < \infty$, then $\mathcal{K}(u) \to 0$ as $u \to 0^+$.*

3.3 The exact derivative formula

Differentiating the entropy representation isolates the part of the reverse conditional that is not captured by its mean and covariance. We write $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm; for a third-order tensor T, $\|T\|_{\mathrm{F}}^2 = \sum_{i,j,k} T_{ijk}^2$.

Theorem 1 (Exact derivative formula for averaged Gaussianization; full proof in Appendix C.1.5)**.** *Assume the setup* (4)*, with $\eta > 0$ and $\mathbb{E}\|U\|^2 < \infty$. For every $u > 0$, the function $\mathcal{K}(u)$ is differentiable and*

$$\mathcal{K}'(u) = \frac{1}{4}\,\mathbb{E}\left[\sum_{r=1}^{d} \big\| A_u(Y_u)^{-1/2}\,(\partial_r A_u)(Y_u)\,A_u(Y_u)^{-1/2}\big\|_F^2\right]. \tag{7}$$

Here, ∂_r denotes differentiation with respect to the r-th coordinate of y. Equivalently, writing $T_{u,r} := \partial_r H_u$ for the r-th spatial derivative of the Hessian,

$$\mathcal{K}'(u) = \frac{u^2}{4}\,\mathbb{E}\left[\sum_{r=1}^{d} \big\| A_u(Y_u)^{-1/2}\,T_{u,r}(Y_u)\,A_u(Y_u)^{-1/2}\big\|_F^2\right]. \tag{8}$$

In particular, $\mathcal{K}'(u) \geq 0$.

Proof sketch. Write the quadratic expression in (8) as

$$R(u) := \mathbb{E}\left[\sum_{r} \|A_u(Y_u)^{-1/2} T_{u,r}(Y_u) A_u(Y_u)^{-1/2}\|_F^2\right].$$

The two terms in the entropy representation satisfy

$$\frac{d}{du}h(Y_u) = -\frac{1}{2}\mathbb{E}[\mathrm{tr}H_u(Y_u)],$$
$$\frac{d}{du}\mathbb{E}[\log\det A_u(Y_u)] = \mathbb{E}[\mathrm{tr}H_u(Y_u)] + \frac{u^2}{2}R(u).$$

The trace terms cancel, leaving $\mathcal{K}'(u) = u^2R(u)/4$. The surviving term involves third spatial derivatives, which vanish for a Gaussian density because its log-density is quadratic. The full proof in Appendix C.1.5 derives the log-determinant evolution and justifies the calculation under only a finite second moment. □

3.4 A BOUND FROM PRE-EXISTING SMOOTHING

The covariance lower bound controls the inverse matrices in the derivative identity. The averaged third-derivative bound in Proposition 4 then controls its remaining factor. Together they give an explicit bound in the ratio u/η of new noise to pre-existing smoothing.

Theorem 2 (One-step Gaussianization bound with pre-existing smoothing)**.** *Assume the setup* (4)*. For every $u \geq 0$,*

$$\mathcal{K}(u) \leq 16\, d(d+2)(d+4)\,\Psi\left(\frac{u}{\eta}\right), \qquad \Psi(r) := \frac{r^2}{2} - r + \log(1+r). \tag{9}$$

Proof. Write $C_d := 16d(d+2)(d+4)$. By the covariance bound in Proposition 1,

$$\|A_u(y)^{-1}\|_{\mathrm{op}} \leq (\eta+u)/\eta.$$

Also, Y_u has the same law as $U + \sqrt{\eta+u}\,G$, with G standard Gaussian and independent of U. Thus Proposition 4 gives

$$\mathbb{E}\big[\|\nabla^3\ell_u(Y_u)\|_{\mathrm{F}}^2\big] \leq \frac{4C_d}{(\eta+u)^3}.$$

Using $\|BMB\|_F \leq \|B\|_{\mathrm{op}}^2\|M\|_F$ in the exact derivative identity, and summing the slices of the third-derivative tensor, yields

$$\mathcal{K}'(u) \leq \frac{u^2}{4}\mathbb{E}\big[\|A_u(Y_u)^{-1}\|_{\mathrm{op}}^2\|\nabla^3\ell_u(Y_u)\|_{\mathrm{F}}^2\big] \tag{10}$$
$$\leq \frac{u^2}{4}\left(\frac{\eta+u}{\eta}\right)^2\mathbb{E}\big[\|\nabla^3\ell_u(Y_u)\|_{\mathrm{F}}^2\big] \leq C_d\frac{u^2}{\eta^2(\eta+u)}. \tag{11}$$

Since $\mathcal{K}(0) = 0$, integration and the substitution $x = s/\eta$ give

$$\mathcal{K}(u) \leq C_d\int_0^u \frac{s^2}{\eta^2(\eta+s)}\,ds = C_d\int_0^{u/\eta}\frac{x^2}{1+x}\,dx = C_d\Psi(u/\eta).$$

□

Remark 1. *The bound quantifies the intuition that a small reverse step is approximately Gaussian. For fixed $\eta > 0$, the derivative bound is of order u^2 near zero, so its integral is of order u^3. The relevant small parameter is u/η; the bulk diffusion steps have $\eta > 0$, whereas the final denoising transition at $t = 1$ requires separate treatment.*

4 APPLICATION TO THE DIFFUSION CHAIN

We now specialize the generic result to the DDPM forward chain (1). The idea is the following: at each step $t \geq 2$, the signal entering the step already contains Gaussian noise accumulated from all previous steps. This accumulated noise plays the role of the "pre-existing smoothing" η in the generic setup, and the new noise β_t plays the role of u.

Define for each $t \geq 2$ the rescaled previous state $S_t := \sqrt{1-\beta_t}\,X_{t-1}$. Then $X_t = S_t + \sqrt{\beta_t}\,Z_t$. Because

$$X_{t-1} = \sqrt{\bar\alpha_{t-1}}X_0 + \sqrt{1-\bar\alpha_{t-1}}\,W, \qquad W \sim \mathcal{N}(0, I_d),$$

we also have

$$S_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{\eta_t}\, W, \qquad \eta_t := (1-\beta_t)(1-\bar{\alpha}_{t-1}). \tag{12}$$

Thus the signal entering step t already contains Gaussian smoothing of variance η_t.

Theorem 3 (Bulk one-step Gaussianization bound; proof in Appendix C.2.1)**.** *For each $t \geq 2$, define*

$$\mathcal{K}_t := \mathbb{E}\Big[D_{\mathrm{KL}}(q_t(\cdot \mid X_t) \parallel g_t^\star(\cdot \mid X_t))\Big].$$

Then

$$\mathcal{K}_t \leq 16\, d(d+2)(d+4)\, \Psi\left(\frac{\beta_t}{\eta_t}\right) = 16\, d(d+2)(d+4)\, \Psi\left(\frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_{t-1})}\right). \tag{13}$$

Remark 2 (Why the first step is special)**.** *At $t = 1$, there is no pre-existing Gaussian smoothing, so $\eta_1 = 0$ and Theorem 3 does not apply. Therefore, we will leave the first-step error $\mathcal{K}_1$ explicit in the global bound presented later. However, if the data distribution itself has Gaussian smoothing, meaning $X_0 = U_0 + \sqrt{\eta_0}\, W_0$ for some $\eta_0 > 0$ and a standard Gaussian W_0 independent of U_0, then Theorem 2 gives*

$$\mathcal{K}_1 \leq 16\, d(d+2)(d+4)\, \Psi\left(\frac{\beta_1}{(1-\beta_1)\eta_0}\right).$$

A related distinction appears in practice: DDPM treats $t = 1$, the final reverse step, as a separate reconstruction term, using a discretized Gaussian decoder for pixel data (Ho et al., 2020; Nichol & Dhariwal, 2021).

The terminal term is easier to bound and follows from the contraction properties of the Ornstein–Uhlenbeck semigroup.

Proposition 3 (Terminal KL contraction; proof in Appendix C.2.2)**.** *For any initial law P_0,*

$$D_{\mathrm{KL}}(P_T \parallel \mathcal{N}(0, I_d)) \leq \bar{\alpha}_T\, D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d)).$$

Let $\mathcal{K}_1 := \mathbb{E}[D_{\mathrm{KL}}(q_1(\cdot \mid X_1) \parallel g_1^\star(\cdot \mid X_1))]$. We leave this first-step term explicit, for the reason explained in Remark 2. Putting everything together, we get the following upper bound on the model error.

Corollary 1 (Global upper bound; proof in Appendix C.2.3)**.** *Assume $D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d)) < \infty$. Then*

$$D_{\mathrm{KL}}(P_0 \parallel \widetilde{P}_0) \leq \bar{\alpha}_T\, D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d)) + \mathcal{K}_1 + 16\, d(d+2)(d+4) \sum_{t=2}^{T} \Psi\left(\frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_{t-1})}\right).$$

5 Schedule optimization

The upper bound in Corollary 1 consists of three terms. The first is $\bar{\alpha}_T D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d))$. Since $D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d))$ is independent of the noise schedule, this term is fixed once $\bar{\alpha}_T$ is fixed. The second is $\mathcal{K}_1$, the first-step error. Because $\mathcal{K}_1$ depends only on β_1, it is fixed once β_1 is fixed. Therefore, among schedules with prescribed values of β_1 and $\bar{\alpha}_T$, the only remaining schedule-dependent term is the bulk sum over $t \geq 2$, which we minimize below.

The optimization problem. Fix $T \geq 2$, $\beta_1^\star \in (0, 1)$, and $\bar{\alpha}_T^\star \in (0, 1 - \beta_1^\star)$. We solve

$$\begin{aligned} \underset{\beta_1, \ldots, \beta_T \in (0,1)}{\text{minimize}} \quad & \sum_{t=2}^{T} \Psi\left(\frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_{t-1})}\right) \\ \text{subject to} \quad & \beta_1 = \beta_1^\star, \qquad \bar{\alpha}_T = \bar{\alpha}_T^\star. \end{aligned} \tag{14}$$

Here $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. We omit the constant factor $16\, d(d+2)(d+4)$, which does not affect the minimizer.

Theorem 4 (Optimal schedule)**.** *Define*

$$q := \left(\frac{(1-\bar{\alpha}_T^\star)(1-\beta_1^\star)}{\bar{\alpha}_T^\star \beta_1^\star} \right)^{1/(T-1)}.$$

The unique minimizer of (14) *is*

$$\beta_1 = \beta_1^\star, \qquad \beta_t = \frac{(q-1)\beta_1^\star q^{t-2}}{1-\beta_1^\star + \beta_1^\star q^{t-1}}, \qquad t = 2, \dots, T. \tag{15}$$

The solution has a simple interpretation. Its noise-to-signal odds $z_t := (1-\bar{\alpha}_t)/\bar{\alpha}_t = e^{-\lambda_t}$ satisfy

$$z_t = z_1 q^{t-1}, \qquad \lambda_t = \lambda_1 - (t-1)\log q, \qquad t = 1, \dots, T, \tag{16}$$

where $z_1 = \beta_1^\star/(1-\beta_1^\star)$. We call this the geometric-odds schedule because its noise-to-signal odds are multiplied by the same factor q at every bulk step. Equivalently, the log-SNR decreases by the same amount at each step.

Proof. Let $s_t := \lambda_{t-1} - \lambda_t > 0$ and $\Lambda := \lambda_1 - \lambda_T = (T-1)\log q$. Writing $r_t := \beta_t/\eta_t$ and using $\bar{\alpha}_t = (1-\beta_t)\bar{\alpha}_{t-1}$ gives

$$1 + r_t = \frac{\bar{\alpha}_{t-1}(1-\bar{\alpha}_t)}{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})} = \frac{z_t}{z_{t-1}} = e^{s_t}. \tag{17}$$

The endpoint constraint becomes $\sum_{t=2}^T s_t = \Lambda$. Conversely, any positive increments with this sum give an admissible schedule by setting $\bar{\alpha}_t = (1+e^{-\lambda_t})^{-1}$ and $\beta_t = 1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}$ for $t \geq 2$.

The objective is therefore $\sum_{t=2}^T F(s_t)$, where $F(s) := \Psi(e^s - 1)$. Since $F''(s) = 2e^s(e^s-1) > 0$ for $s > 0$, Jensen's inequality gives

$$\sum_{t=2}^T F(s_t) \geq (T-1) F\left(\frac{\Lambda}{T-1} \right),$$

with equality only when every $s_t = \Lambda/(T-1) = \log q$. This gives (16). Finally,

$$\beta_t = 1 - \frac{1+z_{t-1}}{1+z_t} = \frac{(q-1)z_{t-1}}{1+qz_{t-1}}.$$

Substituting $z_{t-1} = \beta_1^\star q^{t-2}/(1-\beta_1^\star)$ gives (15). □

Remark 3 (Dependence on the number of steps)**.** *Let $\Lambda := \lambda_1 - \lambda_T$. For the optimal schedule, $r_t = e^{\Lambda/(T-1)} - 1$, so Theorem 3 gives*

$$\sum_{t=2}^T \mathcal{K}_t \leq 16\, d(d+2)(d+4)(T-1)\Psi\Big(e^{\Lambda/(T-1)} - 1\Big) = O\big((T-1)^{-2}\big)$$

at fixed endpoints, since $\Psi(e^s - 1) = s^3/3 + O(s^4)$ as $s \to 0$. This rate concerns the bulk error; the first-step and terminal terms remain separate.

In the experiments, we choose $(\beta_1^\star, \bar{\alpha}_T^\star)$ to match each baseline schedule, so that the comparisons isolate the effect of schedule shape.

6 Empirical evaluation

We evaluate the geometric-odds schedule of Theorem 4 in trained diffusion models, comparing variational likelihood bounds and FID under variational and hybrid training objectives.

We compare two standard baselines, the linear schedule of Ho et al. (2020) and the cosine schedule of Nichol & Dhariwal (2021), against their corresponding geometric-odds schedules. We call these `geometric-linear` (`g-lin`) and `geometric-cosine` (`g-cos`). Each uses the same endpoints $(\beta_1, \bar{\alpha}_T)$ as its corresponding baseline. Within each matched pair, all training settings are identical; only the schedule shape varies. Figure 1 plots the four schedules used in our experiments ($T = 4000$).

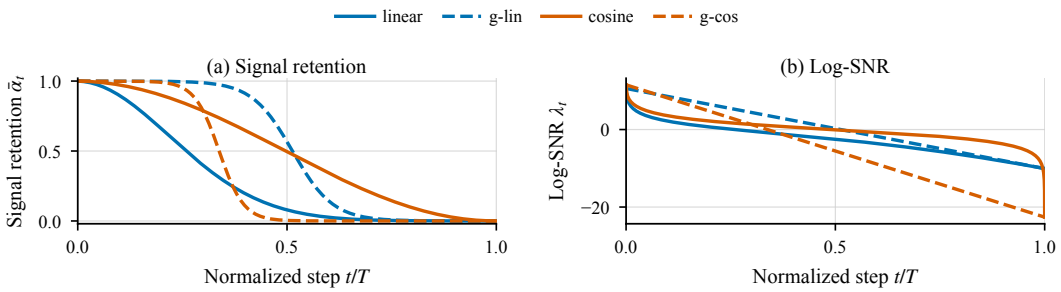

Figure 1: Schedules used in our experiments ($T = 4000$). Solid curves are the linear and cosine baselines; dashed curves are their geometric-odds versions with matching endpoints. **(a)** Cumulative signal retention $\bar{\alpha}_t$. **(b)** Log-SNR, which decreases linearly for the geometric-odds schedules.

6.1 Setup

Datasets and training. We train unconditional models on seven datasets: SVHN, CIFAR-10, and CIFAR-100 at 32×32 resolution, and CelebA-64, LSUN-Bedroom-64, LSUN-Church-64, and ImageNet-64 at 64×64. We also train a class-conditional model on CIFAR-10. Our implementation builds on the released code and U-Net architecture of Nichol & Dhariwal (2021), with added geometric-odds schedules and $T = 4000$ steps in both the forward and reverse chains. Dataset sizes and detailed architecture and training settings are given in Appendix F.

Schedules. We compare linear and cosine baselines against geometric-odds schedules with matching endpoints. The DDPM linear schedule (Ho et al., 2020) spaces β_t linearly between $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$ for $T = 1000$; for other T, both endpoints are scaled by $1000/T$. Thus our $T = 4000$ experiments use $\beta_1 = 2.5 \times 10^{-5}$ and $\beta_T = 5 \times 10^{-3}$. The cosine schedule (Nichol & Dhariwal, 2021) sets $\bar{\alpha}_t = f(t)/f(0)$, where $f(t) = \cos^2\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right)$ and $s = 0.008$. It then recovers $\beta_t = 1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}$, clipped at 0.999.

Training objectives. We test two objectives. The first is L_{vlb}, the negative ELBO (a variational upper bound on NLL), which is the objective most closely aligned with our KL-based analysis. The second is $L_{\mathrm{hybrid}} = L_{\mathrm{simple}} + \lambda L_{\mathrm{vlb}}$ with $\lambda = 0.001$ (Nichol & Dhariwal, 2021), where L_{simple} is the DDPM MSE on the noise prediction (Ho et al., 2020). We include the hybrid objective to test the schedule under a training loss commonly used to balance likelihood and sample quality (Nichol & Dhariwal, 2021).

Metrics. The NLL bound is the full L_{vlb} variational upper bound in bits per dimension. FID compares Inception features of 50K generated samples against the training set. We also report CLIP-MMD and KID as distributional-distance measures, and density and coverage as manifold-fidelity measures; Appendix F.3 gives estimator and sample-size details.

6.2 Matched-endpoint comparison

Table 1 reports the NLL bound and FID for the two $\{\text{baseline}, \text{geometric}\}$ schedule pairs across all (dataset, objective) combinations.

Replacing each baseline by its matched geometric-odds schedule lowers the variational NLL bound in all 14 comparisons under each training objective, for 28 improvements in total (Table 1). Geometric-cosine attains the lowest bound of all four schedules on every dataset under both objectives. Our models learn diagonal Gaussian reverse kernels, so these improvements can reflect changes in both Gaussianization error and the additional error from learning kernels within this covariance family.

The FID results depend on the training objective. Under L_{vlb}, the geometric schedules improve FID in 12 of 14 comparisons: all seven for geometric-cosine and five of seven for geometric-linear. Under

Table 1: Matched-endpoint comparison on seven datasets: holding $(\beta_1, \bar{\alpha}_T)$ and all training hyperparameters fixed within each pair, and varying only the schedule shape. Bold marks the better of each {baseline, geometric} pair (linear vs. `g-lin`, cosine vs. `g-cos`); an underline marks the best of the four schedules in the row, across both endpoint families. Variational NLL bound in bits/dim; FID against the training set from 50K generated samples. Lower is better for both metrics.

	NLL bound (bits/dim)				FID			
Dataset	linear	`g-lin`	cosine	`g-cos`	linear	`g-lin`	cosine	`g-cos`
Variational objective L_{vlb}								
SVHN	1.715	**1.710**	1.664	**1.619**	**3.729**	3.766	3.921	**2.811**
CIFAR-10	3.090	**3.045**	2.981	**2.925**	10.524	**9.858**	11.756	**10.694**
CIFAR-100	3.128	**3.058**	3.000	**2.962**	**16.266**	17.017	19.568	**16.906**
CelebA-64	2.374	**2.324**	2.393	**2.294**	4.715	**4.652**	7.990	**5.109**
LSUN-Bedroom-64	2.514	**2.457**	2.544	**2.409**	14.808	**12.524**	12.023	**9.739**
LSUN-Church-64	2.768	**2.703**	2.789	**2.663**	12.691	**11.612**	10.818	**8.555**
ImageNet-64	3.341	**3.315**	3.357	**3.310**	38.378	**37.085**	60.109	**41.036**
Hybrid objective L_{hybrid}								
SVHN	1.888	**1.733**	1.820	**1.644**	2.936	**2.789**	**1.630**	1.922
CIFAR-10	3.296	**3.024**	3.207	**2.921**	8.421	**6.510**	**5.475**	9.038
CIFAR-100	3.317	**3.047**	3.238	**2.945**	16.622	**10.851**	**11.613**	15.668
CelebA-64	2.572	**2.327**	2.488	**2.316**	**2.160**	3.066	**3.436**	4.545
LSUN-Bedroom-64	2.801	**2.478**	2.703	**2.425**	**2.699**	4.188	**2.997**	5.049
LSUN-Church-64	3.066	**2.719**	2.981	**2.686**	**2.772**	4.229	**2.562**	5.436
ImageNet-64	3.398	**3.318**	3.367	**3.316**	**28.173**	32.214	**25.796**	37.308

L_{hybrid}, they improve FID in only 3 of 14 comparisons, all for geometric-linear; the cosine baseline outperforms geometric-cosine on all seven datasets. Thus the likelihood-bound improvements hold under both objectives, while the sample-quality improvements occur mainly under variational training.

6.3 Additional evaluations

The geometric schedules generally improve CLIP-MMD, KID, density, and coverage over their matched baselines under L_{vlb}, whereas results under L_{hybrid} often favor the baselines. Full results for all seven datasets are in Appendix G.1.

For class-conditional CIFAR-10, geometric-cosine lowers the NLL bound relative to cosine under both objectives, without re-tuning. It also improves FID, CLIP-MMD, and KID under L_{vlb}, whereas cosine performs better on these metrics under L_{hybrid}. This matches the unconditional CIFAR-10 pattern; see Appendix G.2 for the full results.

7 Conclusion

We analyzed the model error that remains when the true reverse conditionals of a discrete-time diffusion process are replaced by their KL-optimal Gaussian approximations. An exact identity for the averaged one-step Gaussianization error led to a universal upper bound for every step after the first, requiring only finite moments. Once the first-step noise and terminal signal level are fixed, the bulk bound is minimized by a closed-form schedule whose noise-to-signal odds grow geometrically — equivalently, whose log-SNR decreases linearly.

In the experiments, both `geometric-linear` and `geometric-cosine` lower the variational NLL bound relative to their matched baselines on every tested dataset under both L_{vlb} and L_{hybrid}. These results support the practical usefulness of the schedule, while reflecting both Gaussianization and learned Gaussian fitting errors. Results for FID and CLIP-MMD are mixed and often favor the baselines under L_{hybrid}, so improvements in the likelihood bound do not transfer uniformly to sample-quality metrics. Accounting for the fitting error under different training objectives, and developing schedule analyses for perceptual criteria, are natural directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with developing proofs, identifying related work, polishing the writing, and adapting the experimental code. Our implementation builds on the code released by Nichol & Dhariwal (2021), with modifications to implement the noise schedules studied here. The authors manually checked the proofs and references. AI tools were also used to run unit tests and examine experimental outputs and training logs. The authors take responsibility for the paper's contents.

ETHICS STATEMENT

This work studies noise schedules through mathematical analysis and experiments on existing image datasets. Its applications share the broader ethical considerations of image generation.

REPRODUCIBILITY STATEMENT

The proofs and supporting lemmas for our theoretical results are provided in the main text and Appendices B–D. Section 5 gives the optimal schedule in closed form, with additional schedule details in Appendix E. Section 6.1 describes the baseline schedules, training objectives, and matched-endpoint comparisons. Appendix F provides the datasets, model architecture, training settings, and metric estimation procedures used in our experiments.

REFERENCES

Ahmad Aghapour, Erhan Bayraktar, and Ziqing Zhang. Entropy-based dimension-free convergence and loss-adaptive schedules for diffusion models. *arXiv preprint arXiv:2601.21943*, 2026.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1555–1584. PMLR, 2022a.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022b.

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. In *Advances in Neural Information Processing Systems*, volume 38, pp. 2073–2127, 2025.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

Yifan Chen, Eric Vanden-Eijnden, and Jiawei Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Valentin De Bortoli, Romuald Elie, Anna Kazeykina, Zhenjie Ren, and Jiacheng Zhang. Dimension-free error estimate for diffusion model and optimal scheduling. *arXiv preprint arXiv:2512.01820*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Alex Dytso, H Vincent Poor, and Shlomo Shamai (Shitz). A general derivative identity for the conditional mean estimator in Gaussian noise and some applications. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1183–1188, 2020.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Hanzhong Guo, Cheng Lu, Fan Bao, Tianyu Pang, Shuicheng Yan, Chao Du, and Chongxuan Li. Gaussian mixture solvers for diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 2023.

Samuel Hurault, Thomas Moreau, and Gabriel Peyré. Geometry-aware discretization error of diffusion models. *arXiv preprint arXiv:2605.08392*, 2026.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14569–14589. PMLR, 2023.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Yuchen Jiao, Yuchen Zhou, and Gen Li. Optimal convergence analysis of DDPM for general distributions. *arXiv preprint arXiv:2510.27562*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707, 2021.

Seo Taek Kong, Weina Wang, and R. Srikant. Noise schedule design for diffusion models: An optimal control perspective. *arXiv preprint arXiv:2605.21911*, 2026.

Gen Li and Yuling Yan. O(d/T) convergence theory for diffusion probabilistic models under minimal assumptions. *Journal of Machine Learning Research*, 26(292):1–55, 2025.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Broadening target distributions for accelerated diffusion models via a novel analysis approach. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=reZKq6hjOZ.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411, 2024.

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

Peter McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, London, 1987.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42947–42975. PMLR, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B*, 57(4):749–760, 1995.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Aart J Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Qiang Sun, H. Vincent Poor, and Wenyi Zhang. A Gaussian perspective for distributional discrepancy in generative diffusion models. *arXiv preprint arXiv:2601.13602*, 2026.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107960–107983, 2024.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022.

A Related Literature

A.1 Diffusion models, likelihoods, and schedule parametrizations

The forward-noising/reverse-denoising construction originates in nonequilibrium diffusion models Sohl-Dickstein et al. (2015) and was made effective for image generation by DDPMs (Ho et al., 2020). The latter use a linear β_t schedule, while Nichol & Dhariwal (2021) introduced the widely used cosine schedule, which spreads corruption more evenly so that useful signal survives further into the chain, and showed that schedule shape, variance learning, and timestep allocation materially affect likelihood and sample quality. In the continuous-time variational formulation of Kingma et al. (2021), the idealized variational objective depends on the log-SNR endpoints rather than the time parametrization between them. Finite-step implementations break that invariance, which is precisely why allocating a fixed number of transitions is substantive here.

Several practical works directly parameterize or redistribute log-SNR. Sigmoid schedules and adaptive computation were studied by Jabri et al. (2023); Chen (2023); shifted log-SNR schedules were used for high-resolution generation by Hoogeboom et al. (2023); and Lu et al. (2022) places solver evaluations uniformly in $\lambda = \frac{1}{2}\log \mathrm{SNR}$. Endpoint selection is a separate design issue: Lin et al. (2024) show that nonzero terminal SNR can create a training–inference mismatch and advocate rescaling schedules to zero terminal SNR. Our optimizer belongs to this broad sigmoid/log-SNR family. Its distinction is not the functional form alone, but its derivation from the Gaussianization error of the finite DDPM reverse kernels with fixed endpoints.

A.2 Theoretical analyses of schedule-dependent error

Theoretical work on diffusion sampling typically separates score-estimation, initialization, and numerical discretization errors. Continuous-time analyses establish polynomial convergence guarantees under broad data assumptions (Chen et al., 2023; Benton et al., 2024), while direct DDPM analyses obtain rates under weak moment conditions (Li & Yan, 2025) and sharper KL rates under a relaxed score-smoothness condition (Jiao et al., 2025). These results treat the schedule as prescribed by the proof rather than as the decision variable. Their conditions nevertheless give an informative connection to our optimizer. Writing $z_t = (1-\bar{\alpha}_t)/\bar{\alpha}_t$, one has exactly

$$\frac{z_t}{z_{t-1}} - 1 = \frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_{t-1})} = \frac{\beta_t}{\eta_t}.$$

The schedule restriction in (Jiao et al., 2025, Eq. (10)) controls a closely related, but not identical, ratio:

$$s_t := \frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_t)}.$$

In fact, that condition is $s_t \le c\log(T)/T$. If $s_t \le \varepsilon < 1$, then the relation $\bar{\alpha}_t = (1-\beta_t)\bar{\alpha}_{t-1}$ implies the exact bound $\beta_t/\eta_t \le \varepsilon/(1-\varepsilon)$. Related coefficient restrictions appear in Li & Yan (2025). Thus those analyses control the same odds-growth scale to first order as a stability device, whereas here

a constant odds-growth factor is derived as the exact minimizer of a stated fixed-endpoint design problem.

The closest work in objective is Kong et al. (2026), which derives a KL upper bound for continuous-time score-based SDE sampling and formulates schedule design as an optimal-control problem. Its scalar state is Fisher information, governed by an ODE obtained from Bochner's formula. The optimized trajectory gives an $\widetilde{O}(d/n)$ sampling-error guarantee under regularity conditions. To obtain tractable closed forms, that work further analyzes strongly log-concave targets and isotropic Gaussian mixtures, producing a tunable family of affine-coupled schedules. The paper then evaluates those schedules with pretrained score networks on CIFAR-10 and ImageNet over neural-function-evaluation budgets from 10 to 50, reporting FID improvements after budget-specific tuning.

The present paper differs along three axes. First, our object is the non-Gaussianity of an exact discrete reverse conditional after projection onto the Gaussian family, rather than the error from numerically discretizing a reverse SDE. Second, we derive an exact identity for the averaged one-step projection error before bounding it, whereas the optimal-control formulation starts from a variational sampling-error bound. Third, our closed-form schedule solves the fixed-endpoint finite-sequence problem and needs no log-concavity or mixture structure for the bulk bound; Kong et al. (2026) instead optimizes a continuous Fisher-information trajectory and uses additional distributional structure for its explicit schedule parametrization. Our exact identity and bulk bound require only a finite second moment: the bounded-input calculation is extended by truncation, using Gaussian-noise moments to control the limit. We do not require log-concavity, score smoothness, or mixture structure. The conclusions are therefore complementary: both support nonuniform schedules, but for different approximations and under different design constraints.

There is also a rapidly developing literature that derives continuous-time or inference-time schedules from other explicit criteria. Examples include schedule-dependent KL bounds optimized within a parametric family (Strasman et al., 2025), dimension-free weak-error bounds with a variational time grid (De Bortoli et al., 2025), entropy-based loss-adaptive discretization (Aghapour et al., 2026), exact-score first-order discretization expansions (Hurault et al., 2026), and interpolation schedules chosen to improve drift regularity after establishing a path-space KL equivalence (Chen et al., 2025). For Gaussian sources, Sun et al. (2026) obtains a covariance-dependent tangent-law schedule by minimizing the leading KL discretization term via calculus of variations. This is a particularly close schedule-optimization neighbor, but its Gaussian reverse conditionals have zero Gaussianization error, so its objective is different from ours. These works optimize numerical discretization or continuous interpolation, not the finite-chain Gaussian projection error studied here. At a structural level, however, the closest comparison is an equalization principle. Specifically, the optimizer of Kong et al. (2026) holds $2f - \dot{\mathcal{J}}/\mathcal{J}$ constant along its optimized Fisher-information trajectory, whereas our optimizer holds $r_t = \beta_t/\eta_t$, and hence the bulk cost $\Psi(r_t)$, constant across the discrete steps.

A separate inference-time literature selects a sampling grid for a fixed trained model. Watson et al. (2021) uses dynamic programming over DDPM KL terms, Sabour et al. (2024) optimizes solver- and model-specific schedules, and Williams et al. (2024) adapts discretization points using score evaluations. Those methods leave the trained model's forward schedule fixed; our question is upstream, namely how to choose that forward schedule.

A.3 Expressiveness of Gaussian reverse transitions

The limits of a single Gaussian reverse kernel at coarse step counts have also been studied directly. Xiao et al. (2022) attributes the difficulty of few-step diffusion sampling to multimodal denoising conditionals and replaces the Gaussian kernel with a conditional GAN. Guo et al. (2023) likewise documents violations of the Gaussian assumption for mixture data and proposes Gaussian-mixture transition kernels fitted through higher-order moments.

A complementary line remains within the Gaussian family. Analytic-DPM (Bao et al., 2022b) derives the likelihood-optimal isotropic reverse variance from the score, while Bao et al. (2022a) treats diagonal and full covariance choices and corrects them when the reverse mean is imperfect. Our idealized reverse kernel already uses the unrestricted moment-matched Gaussian, so our model error is the residual after optimizing within that family. The distinct contribution our work is to express and

bound this residual and then optimize the forward schedule against the resulting bulk bound. We do not claim any priority over nearby schedule-design and Gaussian-kernel results.

A.4 Gaussian approximation and information identities

Classical Laplace and cumulant expansions quantify Gaussian approximations to posterior distributions, often through local asymptotics or concentration regimes (Tierney & Kadane, 1986; Shun & McCullagh, 1995; McCullagh, 1987). Our reverse kernel is instead the global moment-matched Gaussian, the information projection that minimizes forward KL. The calculation is enabled by Gaussian channel identities related to Tweedie's formula and higher-order conditional derivative identities (Efron, 2011; Dytso et al., 2020), together with de Bruijn/Stam information relations under Gaussian smoothing (Stam, 1959). The new step is to combine these tools into an exact derivative formula for the averaged Gaussian-projection error and then exploit the smoothing already accumulated along the diffusion chain.

B Gaussian-smoothing identities and bounds

This appendix proves the Gaussian-smoothing identities used in Section 3. The conditional moment-generating function yields the posterior mean and covariance formulas and the averaged third-derivative bound.

B.1 A conditional mgf identity

Let V be an arbitrary random vector in $\mathbb{R}^d$, let $Z \sim \mathcal{N}(0, I_d)$ be independent of V, fix $\sigma > 0$, and define

$$Y := V + \sigma Z.$$

Since Y is a Gaussian convolution, its density $p(y) = \mathbb{E}[\phi_\sigma(y - V)]$ (where ϕ_σ is the density of $\mathcal{N}(0, \sigma^2 I_d)$) is everywhere positive and C^∞. In particular, $\ell := \log p$ is a well-defined smooth function on all of $\mathbb{R}^d$. We use the posterior given by Bayes' formula, which is defined for every y because $p(y) > 0$.

Lemma 3 (Conditional mgf of the Gaussian noise)**.** *For every $u \in \mathbb{R}^d$ and every $y \in \mathbb{R}^d$,*

$$\mathbb{E}\big[e^{u^\top Z} \mid Y = y\big] = \exp\left(\frac{1}{2}\|u\|^2\right) \frac{p(y - \sigma u)}{p(y)}. \tag{18}$$

Consequently,

$$\mathbb{E}[Z \mid Y = y] = -\sigma\, \nabla \ell(y), \tag{19}$$

$$\mathrm{Cov}(Z \mid Y = y) = I_d + \sigma^2 \nabla^2 \ell(y), \tag{20}$$

$$\kappa_3(Z \mid Y = y) = -\sigma^3 \nabla^3 \ell(y). \tag{21}$$

Here $\kappa_3(Z \mid Y = y)$ is the conditional third cumulant tensor, defined as the third-order tensor with entries

$$\kappa_3(Z \mid Y = y)_{ijk} := \mathbb{E}\big[(Z_i - \mu_i)(Z_j - \mu_j)(Z_k - \mu_k) \mid Y = y\big], \qquad \mu := \mathbb{E}[Z \mid Y = y].$$

Proof. Write P_V for the law of V. Bayes' formula gives

$$P(V \in dv \mid Y = y) = \frac{\phi_\sigma(y - v)}{p(y)}\, P_V(dv).$$

Since $Z = (y - V)/\sigma$ conditional on $Y = y$, completing the square gives

$$\begin{aligned}\mathbb{E}\big[e^{u^\top Z} \mid Y = y\big] &= \frac{1}{p(y)} \int_{\mathbb{R}^d} e^{u^\top (y - v)/\sigma} \phi_\sigma(y - v)\, P_V(dv) \\ &= \frac{e^{\|u\|^2/2}}{p(y)} \int_{\mathbb{R}^d} \phi_\sigma(y - \sigma u - v)\, P_V(dv) = e^{\|u\|^2/2} \frac{p(y - \sigma u)}{p(y)}.\end{aligned}$$

This proves (18) for every u and y.

Taking logarithms gives the conditional cumulant generating function

$$\log \mathbb{E}\big[e^{u^\top Z} \mid Y = y\big] = \tfrac{1}{2}\|u\|^2 + \ell(y - \sigma u) - \ell(y).$$

Differentiating at $u = 0$ yields (19)–(21). □

Since $V = Y - \sigma Z$ given Y, the preceding lemma immediately yields the posterior mean and covariance of V.

Corollary 2 (Posterior mean and covariance)**.** *With the notation above,*

$$\mathbb{E}[V \mid Y = y] = y + \sigma^2 \nabla \ell(y),$$
$$\mathrm{Cov}(V \mid Y = y) = \sigma^2\big(I_d + \sigma^2 \nabla^2 \ell(y)\big).$$

Moreover,

$$\kappa_3(V \mid Y = y) = \sigma^6 \nabla^3 \ell(y). \tag{22}$$

Proof. Since $V = y - \sigma Z$ given $Y = y$, we have $\mathbb{E}[V \mid Y = y] = y - \sigma \mathbb{E}[Z \mid Y = y]$, $\mathrm{Cov}(V \mid Y = y) = \sigma^2 \mathrm{Cov}(Z \mid Y = y)$, and $\kappa_3(V \mid Y = y) = (-\sigma)^3 \kappa_3(Z \mid Y = y)$. Substituting (19)–(21) gives the result. □

B.2 An averaged bound for the third derivative of $\log p$

We write $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm of a matrix or tensor. For a third-order tensor $T = (T_{ijk})$, this means

$$\|T\|_{\mathrm{F}}^2 := \sum_{i,j,k=1}^{d} T_{ijk}^2.$$

Proposition 4 (An averaged third-derivative bound for a Gaussian convolution)**.** *Let $Y = V + \sigma Z$ with $\sigma > 0$ and $Z \sim \mathcal{N}(0, I_d)$ independent of V. Let p be the density of Y, and write $\ell = \log p$. Then*

$$\mathbb{E}\big[\|\nabla^3 \ell(Y)\|_{\mathrm{F}}^2\big] \le \frac{64\, d(d+2)(d+4)}{\sigma^6}. \tag{23}$$

Proof. By (21),

$$\sigma^3 \nabla^3 \ell(Y) = -\kappa_3(Z \mid Y).$$

Since third cumulants are translation-invariant, $\kappa_3(Z \mid Y)$ equals the third moment tensor of $Z - \mathbb{E}[Z \mid Y]$ given Y:

$$\kappa_3(Z \mid Y) = \mathbb{E}\big[(Z - \mathbb{E}[Z \mid Y])^{\otimes 3} \mid Y\big].$$

Applying Jensen's inequality (with the convex function $\|\cdot\|_{\mathrm{F}}^2$) to the conditional expectation, and using the identity $\|x^{\otimes 3}\|_{\mathrm{F}}^2 = \|x\|^6$, gives

$$\|\kappa_3(Z \mid Y)\|_{\mathrm{F}}^2 \le \mathbb{E}\big[\|(Z - \mathbb{E}[Z \mid Y])^{\otimes 3}\|_{\mathrm{F}}^2 \mid Y\big] = \mathbb{E}\big[\|Z - \mathbb{E}[Z \mid Y]\|^6 \mid Y\big].$$

Taking expectations and using the triangle inequality followed by $(a+b)^6 \le 32(a^6 + b^6)$,

$$\begin{aligned}\mathbb{E}\big[\|\kappa_3(Z \mid Y)\|_{\mathrm{F}}^2\big] &\le \mathbb{E}\|Z - \mathbb{E}[Z \mid Y]\|^6 \\ &\le 32\,\mathbb{E}\|Z\|^6 + 32\,\mathbb{E}\|\mathbb{E}[Z \mid Y]\|^6.\end{aligned}$$

For the second term, Jensen's inequality gives $\|\mathbb{E}[Z \mid Y]\|^6 \le \mathbb{E}[\|Z\|^6 \mid Y]$, so $\mathbb{E}\|\mathbb{E}[Z \mid Y]\|^6 \le \mathbb{E}\|Z\|^6$. Therefore

$$\mathbb{E}\big[\|\kappa_3(Z \mid Y)\|_{\mathrm{F}}^2\big] \le 64\,\mathbb{E}\|Z\|^6 = 64\, d(d+2)(d+4).$$

Dividing by σ^6 proves (23). □

C Proofs of results from the main text

Some of the proofs below use supporting results from Sections B and D.

C.1 Proofs for Section 3: One-step Gaussian approximation

C.1.1 Proof of Lemma 1

Proof. Write $G_\star := N(m_Q, \Sigma_Q)$ for the moment-matched Gaussian (which may be degenerate) and let $A := m_Q + \mathrm{Ran}(\Sigma_Q)$. First, $Q(A) = 1$: if $v \in \ker(\Sigma_Q)$, then $\mathrm{Var}(v^\top X) = v^\top \Sigma_Q v = 0$, so $v^\top X = v^\top m_Q$ a.s., which forces $X - m_Q \in \mathrm{Ran}(\Sigma_Q)$ a.s. Moreover, A is the smallest affine subspace carrying full Q-mass.

Let G be any Gaussian measure on $\mathbb{R}^d$. If $D_{\mathrm{KL}}(Q \parallel G) = +\infty$ there is nothing to prove. Otherwise $Q \ll G$. Let B be the affine support of G, so $G(B) = 1$ and hence $Q(B) = 1$. By minimality of A, we have $A \subseteq B$. If $A \subsetneq B$, then A is a proper affine subspace of B, so $G(A) = 0$, contradicting $Q \ll G$ and $Q(A) = 1$. Therefore $B = A$.

Thus both G and $G_\star$ have strictly positive densities g and $g_\star$ with respect to Lebesgue measure λ_A on A. Since $Q \ll G \ll \lambda_A$, write $q := dQ/d\lambda_A$. Now $\log(g_\star/g)$ is a polynomial of degree at most two on A, and Q and $G_\star$ share the same mean and covariance, so

$$\int_A \log \frac{g_\star}{g}\, dQ = \int_A \log \frac{g_\star}{g}\, dG_\star = D_{\mathrm{KL}}(G_\star \parallel G).$$

Hence

$$D_{\mathrm{KL}}(Q \parallel G) = \int_A \log \frac{q}{g_\star}\, dQ + \int_A \log \frac{g_\star}{g}\, dQ = D_{\mathrm{KL}}(Q \parallel G_\star) + D_{\mathrm{KL}}(G_\star \parallel G) \geq D_{\mathrm{KL}}(Q \parallel G_\star).$$

If $D_{\mathrm{KL}}(Q \parallel G_\star) < \infty$, equality holds if and only if $G = G_\star$. □

C.1.2 Proof of Proposition 1

Proof. Apply Corollary 2 with $V = S$, $\sigma = \sqrt{u}$, and $Y = Y_u$. This gives the stated posterior mean and covariance formulas. For the lower bound, write $Y_u = U + \sqrt{\eta + u}\, G$ in distribution, where G is an independent standard Gaussian. By (20),

$$A_u(y) = \frac{\eta}{\eta + u} I_d + \frac{u}{\eta + u} \mathrm{Cov}(G \mid Y_u = y) \succeq \frac{\eta}{\eta + u} I_d.$$

□

C.1.3 Proof of Proposition 2

Proof. Gaussian smoothing and the finite second moment make $h(S)$ and $h(Y_u)$ finite. The covariance lower bound in Proposition 1 bounds the negative part of $\log\det A_u(Y_u)$; its positive part is bounded by $\mathrm{tr} A_u(Y_u)$, whose expectation is finite by the conditional covariance formula. Thus the entropy terms used below are integrable.

For a Gaussian density g, $\log g(x)$ is a quadratic polynomial in x, so $\int \log g\, dQ$ depends on Q only through its mean and covariance. When G is the moment-matched Gaussian (same mean and covariance as Q), this gives $\int \log g\, dQ = \int \log g\, dG = -h(G)$, and therefore $D_{\mathrm{KL}}(Q \parallel G) = -h(Q) + h(G)$. Applying this pointwise with $Q = q_u(\cdot \mid y)$ and $G = g_u^\star(\cdot \mid y)$:

$$D_{\mathrm{KL}}(q_u(\cdot \mid y) \parallel g_u^\star(\cdot \mid y)) = h\big(g_u^\star(\cdot \mid y)\big) - h\big(q_u(\cdot \mid y)\big).$$

Average over Y_u. The Gaussian entropy formula gives

$$h\big(g_u^\star(\cdot \mid y)\big) = \frac{1}{2} \log\Big((2\pi e)^d \det \Sigma_u(y)\Big).$$

Using the covariance formula in Proposition 1,

$$\det \Sigma_u(y) = u^d \det A_u(y).$$

Therefore

$$\mathbb{E}\big[h(g_u^\star(\cdot \mid Y_u))\big] = \frac{d}{2} \log(2\pi e u) + \frac{1}{2} \mathbb{E}\big[\log\det A_u(Y_u)\big].$$

On the other hand, by the definition of conditional entropy,

$$\mathbb{E}\big[h(q_u(\cdot \mid Y_u))\big] = h(S \mid Y_u).$$

Finally, applying the chain rule for entropy gives

$$h(S \mid Y_u) = h(S) + h(Y_u \mid S) - h(Y_u),$$

and because $Y_u = S + \sqrt{u}Z$,

$$h(Y_u \mid S) = h(\sqrt{u}Z) = \frac{d}{2}\log(2\pi e u).$$

Substituting these identities into the definition of $\mathcal{K}(u)$ proves (6). □

C.1.4 Proof of Lemma 2

Proof. For $u > 0$, compare the optimal Gaussian posterior with the candidate $N(Y_u, uI_d)$. By Lemma 1 and the conditional entropy identity,

$$0 \le \mathcal{K}(u) \le \mathbb{E}[D_{\mathrm{KL}}(q_u(\cdot \mid Y_u) \parallel N(Y_u, uI_d))] = h(Y_u) - h(S).$$

Indeed, the averaged cross-entropy against this candidate equals $\frac{d}{2}\log(2\pi e u)$, since $\mathbb{E}\|S - Y_u\|^2 = ud$. By (19) and conditional Jensen, $J(Y_s) := \mathbb{E}\|\nabla \ell_s(Y_s)\|^2 \le d/(\eta + s)$. De Bruijn's identity therefore gives

$$0 \le \mathcal{K}(u) \le \frac{1}{2}\int_0^u J(Y_s)\, ds \le \frac{d}{2}\log(1 + u/\eta) \longrightarrow 0.$$

□

C.1.5 Proof of Theorem 1

Proof. We first assume that U is bounded. We differentiate the entropy representation (6) using an operator that combines the heat equation with integration by parts. The two terms in the entropy representation will then cancel except for a nonnegative quadratic form. After this calculation, we remove boundedness by L^2 truncation.

Justification of the calculus. Put $v = \eta + u$. The Gaussian convolution has a positive density, smooth jointly in (u, y), and

$$\ell_u(y) = -\frac{d}{2}\log(2\pi v) - \frac{\|y\|^2}{2v} + \log \mathbb{E}\exp\left(\frac{y^\top U}{v} - \frac{\|U\|^2}{2v}\right).$$

Spatial derivatives of the last term are conditional cumulants of the bounded input. Consequently the score grows at most linearly in y, while the Hessian and higher spatial derivatives are bounded, locally uniformly in u. The heat equation gives polynomial growth for their u-derivatives. Moreover, Proposition 1 gives $A_u(y) \succeq \eta/(\eta + u)I_d$, so A_u^{-1} is uniformly bounded in y on compact intervals of u. These bounds and the Gaussian tails of p_u and its derivatives justify the differentiations and integrations by parts below, including the application of Lemma 4 to $\log\det A_u$.

An operator for expectations along the heat flow. Write $b_u := \nabla \ell_u$, and define

$$\mathcal{D} := \partial_u - \frac{1}{2}\Delta - b_u \cdot \nabla,$$

acting componentwise on matrix-valued functions. The reason for this choice is that, for the smooth functions used below, integration by parts gives

$$\mathbb{E}\big[(\Delta f_u + b_u \cdot \nabla f_u)(Y_u)\big] = \int_{\mathbb{R}^d} \nabla \cdot (p_u \nabla f_u)\, dy = 0.$$

Thus Lemma 4 can be written as

$$\frac{d}{du}\mathbb{E}[f_u(Y_u)] = \mathbb{E}\big[(\partial_u + \tfrac{1}{2}\Delta) f_u(Y_u)\big] = \mathbb{E}\big[\mathcal{D} f_u(Y_u)\big]. \tag{24}$$

This absorbs the score terms that otherwise appear separately in the derivative of an expectation.

Hessian and log-determinant evolution. The heat equation $\partial_u p_u = \frac{1}{2}\Delta p_u$ implies

$$\partial_u \ell_u = \frac{1}{2}\Delta \ell_u + \frac{1}{2}\|b_u\|^2.$$

To differentiate this twice in space, observe that

$$\frac{1}{2}\partial_{ij}\|b_u\|^2 = \sum_{r=1}^{d}(H_u)_{ir}(H_u)_{jr} + \sum_{r=1}^{d}(b_u)_r\,\partial_r(H_u)_{ij}.$$

Since H_u is symmetric, the first sum is $(H_u^2)_{ij}$. Therefore

$$\partial_u H_u = \frac{1}{2}\Delta H_u + b_u \cdot \nabla H_u + H_u^2, \qquad \text{or equivalently} \qquad \mathcal{D}H_u = H_u^2. \tag{25}$$

Using $A_u = I_d + uH_u$, we obtain the particularly simple identity

$$\mathcal{D}A_u = H_u + uH_u^2 = A_u H_u.$$

For a positive-definite matrix B, the first two directional derivatives of $F(B) := \log\det B$, in a symmetric direction M, are

$$DF(B)[M] = \mathrm{tr}(B^{-1}M), \qquad D^2F(B)[M,M] = -\mathrm{tr}(B^{-1}MB^{-1}M).$$

Applying the matrix chain rule to $\mathcal{D}F(A_u)$ now gives

$$\begin{aligned}\mathcal{D}\log\det A_u &= DF(A_u)[\mathcal{D}A_u] - \frac{1}{2}\sum_{r=1}^{d} D^2F(A_u)[\partial_r A_u, \partial_r A_u] \\ &= \mathrm{tr}H_u + \frac{1}{2}\sum_{r=1}^{d}\mathrm{tr}\big(A_u^{-1}(\partial_r A_u)A_u^{-1}(\partial_r A_u)\big).\end{aligned} \tag{26}$$

The positive sign of the quadratic term comes from the negative Laplacian in $\mathcal{D}$ and the negative second derivative of $\log\det$. The first term simplifies because $\mathrm{tr}(A_u^{-1}\mathcal{D}A_u) = \mathrm{tr}(A_u^{-1}A_uH_u) = \mathrm{tr}H_u$.

Set $\Phi(u) := \mathbb{E}[\log\det A_u(Y_u)]$ and $T_{u,r} := \partial_r H_u$. Since $\partial_r A_u = uT_{u,r}$, (24) and (26) yield

$$\Phi'(u) = \mathbb{E}\big[\mathrm{tr}H_u(Y_u)\big] + \frac{u^2}{2}\mathbb{E}\left[\sum_{r=1}^{d}\mathrm{tr}\big(A_u^{-1}T_{u,r}A_u^{-1}T_{u,r}\big)(Y_u)\right]. \tag{27}$$

Cancellation with the entropy derivative. By de Bruijn's identity (Stam, 1959) and integration by parts,

$$\frac{d}{du}h(Y_u) = \frac{1}{2}\mathbb{E}\|b_u(Y_u)\|^2 = -\frac{1}{2}\mathbb{E}\big[\mathrm{tr}H_u(Y_u)\big]. \tag{28}$$

Indeed, $p_u b_u = \nabla p_u$, so $\int p_u\|b_u\|^2 = \int b_u\cdot\nabla p_u = -\int p_u\,\nabla\cdot b_u$. Differentiating (6) and using (27) cancels the two trace-Hessian terms:

$$\mathcal{K}'(u) = \frac{d}{du}h(Y_u) + \frac{1}{2}\Phi'(u) = \frac{u^2}{4}\mathbb{E}\left[\sum_{r=1}^{d}\mathrm{tr}\big(A_u^{-1}T_{u,r}A_u^{-1}T_{u,r}\big)(Y_u)\right].$$

Finally, for symmetric T and positive-definite A, cyclicity of the trace gives

$$\mathrm{tr}(A^{-1}TA^{-1}T) = \mathrm{tr}\big((A^{-1/2}TA^{-1/2})^2\big) = \big\|A^{-1/2}TA^{-1/2}\big\|_F^2.$$

This proves (8); using $\partial_r A_u = uT_{u,r}$ gives (7) as well. The identity is now established for bounded U.

Extension to second moments. Now assume only $\mathbb{E}\|U\|^2 < \infty$. Let $U_n = U \min\{1, n/\|U\|\}$, with the multiplier defined as 1 at $U = 0$. Then $U_n \to U$ almost surely and in L^2, and $\|U_n\| \le \|U\|$. Use a subscript n for the quantities in (4) constructed with input U_n. For marginal calculations we use a common independent standard Gaussian G and couple

$$Y_{n,s} = U_n + \sqrt{\eta + s}\, G, \qquad Y_s = U + \sqrt{\eta + s}\, G.$$

Set

$$d_n(s, y) := \frac{s^2}{4} \sum_{r=1}^{d} \left\| A_{n,s}(y)^{-1/2} T_{n,s,r}(y) A_{n,s}(y)^{-1/2} \right\|_F^2,$$

$$D_n(s) := \mathbb{E}[d_n(s, Y_{n,s})], \qquad T_{n,s,r} := \partial_r H_{n,s},$$

and define d, D analogously for U, with $d_n(0, y) = d(0, y) = 0$.

Convergence of the derivative expressions. With $v = \eta + s$, the conditional Gaussian identities give

$$A_{n,s}(Y_{n,s}) = \frac{\eta}{v} I_d + \frac{s}{v} \mathrm{Cov}(G \mid Y_{n,s}) \succeq \frac{\eta}{v} I_d,$$

$$\nabla^3 \ell_{n,s}(Y_{n,s}) = -v^{-3/2} \kappa_3(G \mid Y_{n,s}).$$

The conditional third cumulant satisfies $\|\kappa_3(G \mid Y_{n,s})\|_{\mathrm{F}} \le 8\mathbb{E}[\|G\|^3 \mid Y_{n,s}]$, by the triangle inequality and conditional Jensen. Therefore, for every $p \ge 1$,

$$\mathbb{E}[d_n(s, Y_{n,s})^p] \le \left(\frac{16 s^2}{\eta^2(\eta + s)} \right)^p \mathbb{E}\|G\|^{6p}. \tag{29}$$

The same estimate holds for d. For $p > 1$, it gives uniform integrability over n and s in any bounded interval. For $p = 1$, it gives

$$0 \le D_n(s), D(s) \le 16 d(d + 2)(d + 4) \frac{s^2}{\eta^2(\eta + s)}, \tag{30}$$

an integrable bound near $s = 0$ that uses no higher moments of U.

Gaussian kernels and their derivatives are bounded and equicontinuous when their variance lies in a compact subset of $(0, \infty)$. Consequently, $p_{n,s}$ and their spatial derivatives through order three converge locally uniformly in (s, y) to those of p_s. The limiting density is positive, and the covariance lower bound controls the matrix inverses. Hence $d_n \to d$ locally uniformly. Under the coupling above this gives $d_n(s, Y_{n,s}) \to d(s, Y_s)$ almost surely, and uniform integrability yields $D_n(s) \to D(s)$. Applying the same argument to $s_k \to s$ with the input law fixed shows that D is continuous for $s > 0$; the bound (30) gives continuity at zero as well.

Convergence of the Gaussianization errors. For each fixed $s \ge 0$, the densities satisfy

$$|\log p_{n,s}(y)| \le C_s(1 + \|y\|^2) \qquad \text{uniformly in } n.$$

To see this, choose a ball centered at the origin carrying positive probability under U. Radial truncation preserves that mass in the ball, so the convolution formula gives a Gaussian lower bound on $p_{n,s}$, uniformly in n; its upper bound is $(2\pi(\eta + s))^{-d/2}$. Local convergence of the densities and $\|Y_{n,s}\| \le \|U\| + \sqrt{\eta + s}\|G\|$ then allow dominated convergence, giving $h(Y_{n,s}) \to h(Y_s)$, including $h(S_n) \to h(S)$ at $s = 0$.

For fixed $u > 0$, the negative part of $\log\det A_{n,u}(Y_{n,u})$ is bounded by $d \log(1 + u/\eta)$, and its positive part is bounded by

$$\mathrm{tr} A_{n,u}(Y_{n,u}) \le d + \frac{u}{\eta + u} \mathbb{E}[\|G\|^2 \mid Y_{n,u}].$$

Conditional Jensen gives uniform L^p bounds for any finite p. Local convergence and uniform integrability therefore give convergence of the expected log determinants. The entropy representation (6) now implies $\mathcal{K}_n(u) \to \mathcal{K}(u)$.

For bounded U_n, the calculation above, Lemma 2, and (30) give

$$\mathcal{K}_n(u) = \int_0^u D_n(s)\, ds.$$

Passing to the limit by dominated convergence yields $\mathcal{K}(u) = \int_0^u D(s)\, ds$. Since D is continuous, $\mathcal{K}'(u) = D(u)$ for every $u > 0$, proving both exact derivative formulas under only the stated second-moment assumption.

□

C.2 Proofs for Section 4: Application to the diffusion chain

C.2.1 Proof of Theorem 3

Proof. By (12), the rescaled variable S_t and the observation $X_t = S_t + \sqrt{\beta_t}\, Z_t$ fit the generic setup (4) with

$$U = \sqrt{\bar{\alpha}_t}\, X_0, \qquad \eta = \eta_t, \qquad u = \beta_t.$$

Applying Theorem 2 gives

$$\mathbb{E}\Big[D_{\mathrm{KL}}(\mathcal{L}(S_t \mid X_t) \parallel \widehat{g}_t^\star(\cdot \mid X_t))\Big] \le 16\, d(d+2)(d+4)\, \Psi\left(\frac{\beta_t}{\eta_t}\right), \tag{31}$$

where $\mathcal{L}(S_t \mid X_t)$ denotes the conditional distribution of S_t given X_t, and $\widehat{g}_t^\star(\cdot \mid X_t)$ is its moment-matched Gaussian approximation.

We would like to replace the left-hand side of (31) with $\mathcal{K}_t$. The map $\varphi(x) = \sqrt{1-\beta_t}\, x$ is an invertible affine bijection whose push-forward sends $\mathcal{L}(X_{t-1} \mid X_t)$ to $\mathcal{L}(S_t \mid X_t)$. Because affine maps preserve the first two moments, it also sends $g_t^\star(\cdot \mid X_t)$ to the moment-matched Gaussian of $\mathcal{L}(S_t \mid X_t)$, namely $\widehat{g}_t^\star(\cdot \mid X_t)$. By the invariance of KL under invertible affine maps,

$$D_{\mathrm{KL}}(q_t(\cdot \mid X_t) \parallel g_t^\star(\cdot \mid X_t)) = D_{\mathrm{KL}}(\mathcal{L}(S_t \mid X_t) \parallel \widehat{g}_t^\star(\cdot \mid X_t)).$$

Taking expectations identifies the left-hand side of (31) with $\mathcal{K}_t$, which proves (13).

□

C.2.2 Proof of Proposition 3

Proof. Write $\gamma := \mathcal{N}(0, I_d)$ throughout this proof. We first establish a one-step contraction. Let μ be any probability measure on $\mathbb{R}^d$, let $X \sim \mu$, and set

$$Y = \sqrt{1-\beta}\, X + \sqrt{\beta}\, Z, \qquad Z \sim \gamma \text{ independent of } X.$$

Writing $\tau = -\frac{1}{2}\log(1-\beta)$,

$$Y = e^{-\tau} X + \sqrt{1 - e^{-2\tau}}\, Z,$$

which is exactly one step of the Ornstein–Uhlenbeck semigroup at time τ started at μ. Let $(\mu_s)_{s \ge 0}$ denote the resulting OU trajectory, so that $\mu_0 = \mu$ and $\mu_\tau = \mathcal{L}(Y)$. Then

$$\frac{d}{ds} D_{\mathrm{KL}}(\mu_s \parallel \gamma) = -I(\mu_s \| \gamma),$$

where $I(\cdot\|\gamma)$ is the relative Fisher information. The Gaussian log-Sobolev inequality (Bakry et al., 2014, Prop. 5.5.1) says

$$I(\mu_s \| \gamma) \ge 2 D_{\mathrm{KL}}(\mu_s \parallel \gamma),$$

so

$$\frac{d}{ds} D_{\mathrm{KL}}(\mu_s \parallel \gamma) \le -2 D_{\mathrm{KL}}(\mu_s \parallel \gamma),$$

and Grönwall's inequality yields

$$D_{\mathrm{KL}}(\mu_\tau \parallel \gamma) \le e^{-2\tau} D_{\mathrm{KL}}(\mu \parallel \gamma) = (1-\beta) D_{\mathrm{KL}}(\mu \parallel \gamma).$$

Applying this contraction to each forward step of the diffusion chain, with $\mu = P_{t-1}$, $\mu_\tau = P_t$, and $\beta = \beta_t$, gives $D_{\mathrm{KL}}(P_t \parallel \gamma) \le (1-\beta_t) D_{\mathrm{KL}}(P_{t-1} \parallel \gamma)$. Iterating from $t = 1$ to $t = T$,

$$D_{\mathrm{KL}}(P_T \parallel \gamma) \le \prod_{t=1}^{T} (1-\beta_t)\, D_{\mathrm{KL}}(P_0 \parallel \gamma) = \bar{\alpha}_T\, D_{\mathrm{KL}}(P_0 \parallel \mathcal{N}(0, I_d)).$$

□

C.2.3 Proof of Corollary 1

Proof. Insert Theorem 3 and Proposition 3 into the pathwise decomposition (3), and then use (2). □

D TECHNICAL LEMMAS

Lemma 4 (Heat expectation identity). *Let p_u be the density of $Y_u = U + \sqrt{\eta + u}\, W'$, so that p_u satisfies the heat equation $\partial_u p_u = \frac{1}{2}\Delta p_u$. Let $f_u : \mathbb{R}_{>0} \times \mathbb{R}^d \to \mathbb{R}$ be C^1 in u and C^2 in y, and assume that $f_u, \partial_u f_u, \nabla f_u, \Delta f_u$ all have at most polynomial growth in y, locally uniformly in u. Assume further that U has enough moments that every such polynomial bound is integrable against p_u. Then for every $u > 0$,*

$$\frac{d}{du}\mathbb{E}[f_u(Y_u)] = \mathbb{E}[\partial_u f_u(Y_u)] + \frac{1}{2}\mathbb{E}[\Delta f_u(Y_u)]. \tag{32}$$

Proof. Use a common standard Gaussian G, independent of U, to write $Y_u = U + \sqrt{\eta + u}\, G$. On any compact interval of $u > 0$, the chain rule gives

$$\frac{d}{du} f_u(Y_u) = \partial_u f_u(Y_u) + \frac{1}{2\sqrt{\eta + u}}\, G^\top \nabla f_u(Y_u).$$

The polynomial growth assumptions bound its absolute value by $C(1+\|U\|^m+\|G\|^m)(1+\|G\|)$ for some C, m. This is integrable by the moment assumption and independence, so we may differentiate inside the expectation:

$$\frac{d}{du}\mathbb{E}[f_u(Y_u)] = \mathbb{E}[\partial_u f_u(Y_u)] + \frac{1}{2\sqrt{\eta + u}}\mathbb{E}[G^\top \nabla f_u(Y_u)].$$

Conditional on U, Gaussian integration by parts gives

$$\mathbb{E}[G^\top \nabla f_u(Y_u) \mid U] = \sqrt{\eta + u}\, \mathbb{E}[\Delta f_u(Y_u) \mid U].$$

Indeed, this follows by integrating the divergence of $\phi(z)\nabla f_u(U + \sqrt{\eta + u}\, z)$, where ϕ is the standard Gaussian density; the polynomial growth bounds and Gaussian tails make the boundary term vanish. Averaging over U and substituting above proves (32). □

Remark 4 (Use in the derivative proof). *We apply Lemma 4 only during the bounded-input calculation in the proof of Theorem 1; truncation then extends the exact identity to finite-second-moment inputs.*

E ADDITIONAL SCHEDULE DETAILS

E.1 RATES FOR GENERAL SCHEDULES

For any admissible schedule, let $s_t := \lambda_{t-1} - \lambda_t > 0$, $\Lambda := \lambda_1 - \lambda_T$, and $\delta := \max_{2 \le t \le T} s_t$. Then $\sum_{t=2}^{T} s_t = \Lambda$, and (17) gives $\beta_t/\eta_t = e^{s_t} - 1$. Set $C_d := 16d(d+2)(d+4)$ and $F(s) := \Psi(e^s - 1)$, so that Theorem 3 gives $\mathcal{K}_t \le C_d F(s_t)$. As shown in the proof of Theorem 4, F is convex and $F(0) = 0$. Hence $F(s_t) \le (s_t/\delta)F(\delta)$, and therefore

$$\sum_{t=2}^{T} \mathcal{K}_t \le C_d \sum_{t=2}^{T} F(s_t) \le C_d\, \Lambda\, \frac{F(\delta)}{\delta}.$$

This bound holds for every $\delta > 0$. Since $F(\delta)/\delta = \delta^2/3 + O(\delta^3)$ as $\delta \to 0$, at fixed endpoints any sequence of schedules whose largest log-SNR increment is $O(1/T)$ has bulk error $O(1/T^2)$. The first-step and terminal terms are not included in this rate.

E.2 ENDPOINT SELECTION

Remark 5 (Endpoint selection). *For fixed T, the optimal schedule is determined by $(\beta_1, \bar{\alpha}_T)$, so endpoint tuning requires only a two-dimensional search. One can reduce this to a single parameter by prescribing an intermediate noise level. For example, let $T \ge 4$, $m = \lfloor T/2 \rfloor$, and $0 < \beta_1 < 1/2$. The midpoint condition $\bar{\alpha}_m = 1/2$ sets the SNR to one. Since $z_m = z_1 q^{m-1} = 1$, it fixes*

$$\bar{\alpha}_T = \left[1 + \left(\frac{1 - \beta_1}{\beta_1}\right)^{(T-m)/(m-1)}\right]^{-1}.$$

For fixed β_1, this tends to β_1 as $T \to \infty$. Our experiments instead inherit both endpoints from each baseline.

F EXPERIMENTAL DETAILS

This appendix provides the dataset, training, and evaluation details for the experiments in Section 6.

F.1 DATASETS

We use seven datasets, three at 32×32 resolution and four at 64×64. The 32×32 group consists of SVHN ($\approx$ 73K training images), CIFAR-10, and CIFAR-100 (50K each); the 64×64 group consists of CelebA-64 ($\approx$ 163K), LSUN-Bedroom-64 ($\approx$ 3.03M), LSUN-Church-64 ($\approx$ 126K), and ImageNet-64 ($\approx$ 1.28M). Together they range from street-number crops to diverse natural scenes, with training sets spanning 50K to $\approx$ 3M images. All seven are modeled unconditionally. As a generalization test, we also train a class-conditional model on CIFAR-10.

We use the official training and test splits of SVHN, CIFAR-10, CIFAR-100, and CelebA; for SVHN, we use only the training split, without the extra split. ImageNet-64 uses the official downsampled 64×64 training and validation data, and the two LSUN datasets use the official training and validation LMDBs. CelebA aligned images are center-cropped to a square and resized to 64×64 using Lanczos resampling. LSUN images are resized using BOX resampling so that their shorter side is 64, then center-cropped to 64×64. All images are converted to RGB and scaled to $[-1, 1]$.

F.2 ARCHITECTURE AND TRAINING

All experiments use the U-Net of Nichol & Dhariwal (2021) with 128 base channels, three residual blocks per resolution, and attention at resolutions 16 and 8; the channel multipliers are $(1, 2, 2, 2)$ for the 32×32 datasets and $(1, 2, 3, 4)$ for the 64×64 datasets (ImageNet-64, CelebA-64, LSUN-Bedroom-64, and LSUN-Church-64). Models are trained with $T = 4000$ forward steps, AdamW with zero weight decay, initial learning rate 10^{-4} linearly decayed over training, batch size 128, and EMA rate 0.9999. The 32×32 models are trained for 500K iterations and the 64×64 models for 200K iterations. Following Nichol & Dhariwal (2021), the linear and geometric-linear models use dropout 0.1, while the cosine and geometric-cosine models use dropout 0.3. Sampling uses the full $T = 4000$-step reverse chain. The class-conditional CIFAR-10 model reuses the 32×32 configuration and adds a learned class-label embedding to the timestep embedding (10 classes).

Under both objectives, the model predicts the added noise and learns a diagonal reverse covariance using the learned-range parameterization of Nichol & Dhariwal (2021). Hybrid training samples timesteps uniformly and stops gradients through the noise prediction in the variational term. The sampled per-step variational term is multiplied by $T\lambda$, with $\lambda = 0.001$. Variational training uses loss-second-moment timestep sampling with importance weights that preserve the summed variational objective.

F.3 METRIC ESTIMATION

The columns labeled NLL bound report the full L_{vlb} variational upper bound on NLL in bits per dimension, evaluated on 10K held-out samples; for the two LSUN datasets, whose official validation splits contain only 300 images, we use those. We evaluate on the test splits of SVHN, CIFAR-10, CIFAR-100, and CelebA, and on the validation splits of ImageNet-64 and LSUN. The bound includes the discretized Gaussian decoder term at the first step, the intermediate Gaussian KL terms, and the terminal prior KL. For each evaluated image, we draw one forward-noise sample at each timestep to estimate the decoder and intermediate terms; the terminal term is computed analytically. Both likelihood evaluation and sampling clip the predicted clean image to $[-1, 1]$.

For every dataset, FID is computed against the training set from 50K generated samples. We also report four richer metrics. CLIP-MMD is a variant of CMMD (Jayasumana et al., 2024), using L2-normalized OpenCLIP ViT-B/32 embeddings with the `laion2b_s34b_b79k` pretrained weights and an RBF kernel. The squared bandwidth is the median nonzero squared distance among a subsample of 2048 pooled real and generated features, selected separately for each comparison. KID is an MMD on Inception-V3 features with a degree-three polynomial kernel, averaged over 100 subsets of size 1000. Density and coverage are $k = 5$ nearest-neighbor manifold-fidelity scores, also on Inception-V3 features. CLIP-MMD, density, and coverage use 10K real and 10K generated images.

KID uses the full available feature sets; for ImageNet-64, all four schedules use 50K real and 50K generated features.

G ADDITIONAL EXPERIMENTAL RESULTS

G.1 BEYOND THE NLL BOUND AND FID: DISTRIBUTIONAL FIT AND MANIFOLD FIDELITY

Tables 2 and 3 report the four richer metrics on the seven datasets, one table per endpoint family. CLIP-MMD and KID measure distributional distance; density and coverage measure manifold fidelity.

Table 2: Richer-metric comparison on the matched `cosine`/`g-cos` pair. CLIP-MMD and KID (polynomial-kernel inception MMD) are reported in units of 10^{-3}; density and coverage are manifold-fidelity scores in $[0, 1+]$. Bold marks the better value in each pair. CLIP-MMD and KID lower is better; density and coverage higher is better.

	CLIP-MMD ($\times 10^3$)		KID ($\times 10^3$)		density		coverage	
Dataset	cosine	g-cos	cosine	g-cos	cosine	g-cos	cosine	g-cos
Variational objective L_{vlb}								
SVHN	7.881	**4.866**	3.119	**2.180**	**1.097**	1.079	0.953	**0.954**
CIFAR-10	20.394	**18.345**	9.679	**8.516**	0.745	**0.747**	0.755	**0.759**
CIFAR-100	29.365	**26.074**	17.455	**14.055**	0.648	**0.668**	0.616	**0.639**
CelebA-64	12.728	**8.514**	5.238	**2.596**	1.135	**1.189**	0.869	**0.929**
LSUN-Bedroom-64	47.291	**33.588**	11.865	**6.756**	0.754	**0.860**	0.768	**0.855**
LSUN-Church-64	37.467	**25.872**	8.590	**5.039**	**0.980**	0.812	0.788	**0.838**
ImageNet-64	64.34	**52.13**	73.94	**39.45**	0.275	**0.312**	0.239	**0.321**
Hybrid objective L_{hybrid}								
SVHN	3.015	**2.848**	**0.906**	1.128	**1.164**	1.096	**0.973**	0.963
CIFAR-10	**6.570**	16.528	**2.723**	6.811	**1.103**	0.814	**0.930**	0.799
CIFAR-100	**12.849**	24.591	**6.947**	12.767	**0.975**	0.717	**0.855**	0.665
CelebA-64	**3.404**	4.030	**2.242**	2.601	1.322	**1.417**	**0.963**	0.957
LSUN-Bedroom-64	**15.674**	24.231	**1.917**	3.370	**1.221**	1.211	**0.962**	0.919
LSUN-Church-64	**7.541**	17.622	**1.477**	3.520	**1.112**	1.108	**0.956**	0.899
ImageNet-64	**21.24**	35.64	**30.79**	43.86	**0.591**	0.407	**0.557**	0.396

Table 3: Same richer-metric comparison as Table 2, now on the matched `linear`/`g-lin` pair. Units and bolding conventions are identical.

	CLIP-MMD ($\times 10^3$)		KID ($\times 10^3$)		density		coverage	
Dataset	linear	g-lin	linear	g-lin	linear	g-lin	linear	g-lin
Variational objective L_{vlb}								
SVHN	**8.166**	8.945	3.371	**3.327**	1.059	**1.092**	**0.955**	0.955
CIFAR-10	18.929	**18.325**	7.551	**7.017**	0.742	**0.777**	0.768	**0.793**
CIFAR-100	**26.432**	28.205	**12.644**	13.279	0.678	**0.700**	**0.643**	0.637
CelebA-64	8.653	**7.523**	2.264	**2.106**	1.044	**1.064**	0.914	**0.921**
LSUN-Bedroom-64	39.427	**33.941**	9.734	**8.221**	0.673	**0.697**	0.818	**0.836**
LSUN-Church-64	34.581	**31.517**	7.624	**6.973**	0.680	**0.718**	0.800	**0.820**
ImageNet-64	47.00	**45.13**	35.59	**35.54**	0.330	**0.353**	0.348	**0.371**
Hybrid objective L_{hybrid}								
SVHN	8.003	**6.676**	**2.357**	2.431	**1.171**	1.092	**0.970**	0.965
CIFAR-10	11.839	**10.633**	4.990	**4.007**	**0.992**	0.889	**0.905**	0.876
CIFAR-100	19.076	**17.524**	11.208	**7.732**	**1.012**	0.772	**0.826**	0.760
CelebA-64	**2.433**	3.360	**1.076**	1.578	1.230	**1.251**	**0.970**	0.956
LSUN-Bedroom-64	**12.768**	18.496	**2.002**	3.125	1.020	**1.033**	**0.950**	0.929
LSUN-Church-64	**9.113**	13.127	**1.764**	2.892	**1.087**	1.054	**0.946**	0.920
ImageNet-64	**24.69**	30.17	**33.60**	37.28	**0.525**	0.480	**0.510**	0.468

Under L_{vlb}, the geometric schedule improves upon its baseline counterpart for almost all of the datasets and all of the richer metrics. However, under L_{hybrid}, the baseline schedule is usually better than its geometric version.

G.2 CLASS-CONDITIONAL CIFAR-10

To test whether the empirical schedule-shape pattern persists with class-label conditioning, we re-run the cosine/geometric-cosine comparison on a class-conditional CIFAR-10 model.

Table 4: Class-conditional CIFAR-10, matched-endpoint comparison. Variational NLL bound in bits/dim; CLIP-MMD and KID in units of 10^{-3}. Bold marks the better of the pair.

	NLL bound		FID		CLIP-MMD ($\times 10^3$)		KID ($\times 10^3$)	
Objective	cosine	g-cos	cosine	g-cos	cosine	g-cos	cosine	g-cos
L_{vlb}	2.977	**2.930**	9.418	**8.140**	14.78	**13.63**	7.578	**6.228**
L_{hybrid}	3.214	**2.920**	**5.166**	6.992	**5.704**	11.23	**2.848**	5.059

The schedule-shape pattern carries over without re-tuning (Table 4). Geometric-cosine lowers the variational NLL bound under both objectives, and improves every other reported metric under L_{vlb}. Under L_{hybrid}, the cosine baseline keeps the edge on FID, CLIP-MMD, and KID. Their agreement with the unconditional CIFAR-10 rows of Tables 1 and 2 shows that the same empirical pattern persists under conditioning.