ENTROPY-BASED BOUNDS FOR FLOW AND DIFFUSION MODELS VIA STOCHASTIC THERMODYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We derive a lower bound on the negative log-likelihood of diffusion and flow models from the entropy rates of their noising processes. Our main theoretical result is an exact decomposition of a plug-in surrogate of this quantity into data and noise entropies, a model-induced entropy rate integral, and a nonnegative score-error term. This yields a thermodynamically interpretable lower bound on the surrogate that complements variational upper bounds. The same decomposition gives a model-based upper bound on the differential entropy of the data, and both bounds are tight in the exact-score limit. We validate these results on Swiss roll data and finite Ising systems, where independent entropy estimates or analytic targets are available, and evaluate pretrained image models, including convolutional and transformer score-based diffusion models and a rectified flow, on CIFAR-10, ImageNet-32, FFHQ, CelebA-HQ, and LSUN. Improved likelihood is consistently accompanied by increased integrated entropy production, revealing an accuracy–dissipation relation within variance-preserving and variance-exploding model families. We also introduce the entropy reduction ratio, a likelihood-free measure of fit computed from the score alone. Together, these results establish entropy rates as both theoretical limits and empirical diagnostics for diffusion and flow models, linking likelihood, data entropy, and irreversibility within a unified stochastic thermodynamic framework.

1 INTRODUCTION

Generative modeling is increasingly understood as a problem in nonequilibrium statistical inference: a model must transform an easy reference distribution into structured data by learning the score, control, or transport that reverses an entropy-increasing destruction of information. This viewpoint was implicit in the nonequilibrium thermodynamic origins of diffusion probabilistic models (Sohl-Dickstein et al., 2015; Ho et al., 2020) and became operational in score-based stochastic differential equations (SDEs) and probability flow ordinary differential equations (ODEs) (Song et al., 2021b), likelihood-weighted score matching (Song et al., 2021a), variational diffusion objectives (Huang et al., 2021), Schrödinger bridges (Bortoli et al., 2021), and flow matching (Lipman et al., 2022). The area is now broad enough to be treated in its own right: Welling, Lu, and Holdijk's *Generative AI and Stochastic Thermodynamics* (Welling et al., 2026) frames variational free energy as a bridge between generative models, optimal transport, entropy production, and fluctuation theorems. This paper contributes to that program by turning the free energy/entropy rate viewpoint into a quantitative likelihood limit for score-based generative models: an exact decomposition of the negative log-likelihood (NLL) into data and noise entropies, a learned entropy rate integral, and a nonnegative score-error term.

Score-based diffusion models (Song et al., 2021b) have driven much of the empirical progress in this program, with physical connections to optimal transport (Kwon et al., 2022; Lipman et al., 2022), critical damping (Dockhorn et al., 2022), heat dissipation (Rissanen et al., 2023), Maxwell's demon (Premkumar, 2025), and thermodynamic hardware (Coles et al., 2023). Existing variational and Kullback–Leibler (KL) analyses (Huang et al., 2021; Premkumar, 2025) bound NLL from above, and stochastic thermodynamic speed–accuracy relations (Ikeda et al., 2025) and energy–speed–accuracy limits (Klinger & Rotskoff, 2025) constrain sampling speed and dissipated work rather than likelihood. Our contribution is the complementary lower-bound inequality on NLL itself, with a gap equal to the integrated denoising score matching error.

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.

Under a consistent plug-in convention in which system entropy rates $\dot{S}_{\boldsymbol{\theta}}(t)$ are computed from the learned score, we derive a thermodynamic lower bound on NLL,

$$\mathrm{NLL} \;\geq\; \frac{S_{\text{data}} + S_{\text{noise}}}{2} - \frac{1}{2}\int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\, dt,$$

where S_{data} is the entropy of the data and S_{noise} that of the equilibrium distribution. The trivial bound $\mathrm{NLL} \geq S_{\text{data}}$ follows from $\mathrm{KL} \geq 0$. Our result complements it via a different mechanism: a D-weighted, integrated score norm correction, with $D(t) := \frac{1}{2}g(t)^2$ the diffusion coefficient of the forward SDE (§2.2). The slack between NLL and this bound is exactly an integrated score-error term.

Our main contributions are as follows:

- **Thermodynamic likelihood limit.** We derive an exact entropy rate decomposition of a plug-in surrogate of NLL into data entropy, noise entropy, the learned system entropy rate integral, and a nonnegative score-error term, yielding a thermodynamically interpretable lower bound that complements existing upper bounds from the evidence lower bound (ELBO). At the exact score the surrogate equals the data entropy, and the actual NLL exceeds it only by the final marginal's KL divergence from the prior.
- **Data entropy estimate from the trained score.** Rearranging the decomposition gives a model-based upper bound on the differential entropy of the data distribution, useful when S_{data} is not known a priori. The bound is tight in the exact-score limit and is empirically within ~ 0.04 bits per dimension (bpd) of the nonparametric references on the Swiss roll. On the 5×5 Ising sweep where every term is computable, the bound holds on all 60 measured checkpoints on both likelihoods, 0.03–0.06 bpd above the analytic entropy.
- **Accuracy–dissipation diagnostics at scale.** Across controllable synthetic systems and pre-trained image score models, improved likelihood is consistently accompanied by increased integrated entropy production, exposing a measurable accuracy–dissipation relation.

2 Background: Stochastic Thermodynamics of Generative Models

2.1 Generative Models as Nonequilibrium Dynamics

Many modern generative models can be interpreted as controlled nonequilibrium dynamics between probability distributions. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b) reverse an entropy-increasing forward process. Probability flow ODEs and continuous normalizing flows (Chen et al., 2018) accumulate likelihood along a learned transport. Flow matching methods (Lipman et al., 2022) learn velocity fields between probability paths. Schrödinger bridges (Bortoli et al., 2021) are entropy-regularized transports on path space. Variational latent variable models such as variational autoencoders (Kingma & Welling, 2014) and normalizing flows (Rezende & Mohamed, 2015) optimize variational objectives that perform thermodynamic computations. The shared structure is that generation is a controlled process in which the system entropy of a tractable reference is reshaped to match a structured target. Stochastic thermodynamics supplies physically interpretable state variables: system entropy, entropy flow between system and surroundings, intrinsic entropy production, and probability currents (Seifert, 2012). Appendix A provides a concept dictionary mapping each modeling object below to its stochastic thermodynamic counterpart.

Time conventions. The forward (noising) process uses time $t \in [0, T]$, while generation runs in reverse with $\tau = T - t$. We use the same symbol $\dot{S}_{\boldsymbol{\theta}}$ for both time variables (clocks) and let the argument disambiguate: $\dot{S}_{\boldsymbol{\theta}}(t)$ denotes the system entropy rate in the forward noising clock (the object that enters the NLL decomposition), and $\dot{S}_{\boldsymbol{\theta}}(\tau)$ denotes the rate in the controlled generation clock (the object in §3.2). On the driftless SDE used in §3.2, the two are related by a sign flip, $\dot{S}_{\boldsymbol{\theta}}(\tau) = -\dot{S}_{\boldsymbol{\theta}}(T - \tau)$. Intrinsic entropy production $\dot{S}^i_{\boldsymbol{\theta}}$ is nonnegative under either clock. The integrated intrinsic production $\int_0^1 \dot{S}^i_{\boldsymbol{\theta}}(\tau)\, d\tau$ plotted in Figures 3a and 4 coincides with the forward-clock $\int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\, dt$ of Theorem 1 only for driftless processes (Appendix B, Table 4).

2.2 SCORE-BASED GENERATIVE MODELS

We adopt the score-based diffusion setup of Song et al. (2021b). The forward Itô SDE is $d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\, dt + \mathbf{G}(\mathbf{x}_t, t)\, d\mathbf{w}_t$ with $t \in [0, 1]$ (we fix $T{=}1$ throughout), $\mathbf{x}_t \in \mathbb{R}^d$, and standard Wiener process $\mathbf{w}_t$. With $\mathbf{x}_0 \sim p_{\text{data}}$ and $(\mathbf{f}, \mathbf{G})$ chosen so that $p_1 \approx \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, the time-reversed SDE (Anderson, 1982; Haussmann & Pardoux, 1986; Song et al., 2021b) depends only on the score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, which we approximate by a neural network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ trained by denoising score matching (DSM) (Vincent, 2011):

$$\min_{\boldsymbol{\theta}} \; \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t | \mathbf{x}_0} \left[\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{x}_0) \|_2^2 \right]. \tag{1}$$

Throughout we use the standard form $\mathbf{f}(\mathbf{x}, t) = f(t)\, \mathbf{x}$ and $\mathbf{G}(\mathbf{x}, t) = g(t)\, \mathbf{I}_d$, and adopt the scalar diffusion coefficient $D(t) := \frac{1}{2} g(t)^2$ for this isotropic case (the general matrix form $\frac{1}{2}\mathbf{G}\mathbf{G}^\top$ reduces to this in our setting). Appendix D recalls the explicit reverse-time SDE and its controlled-forward reparameterization in $\tau = 1 - t$, which we use in Section 3.2.

2.3 ENTROPY RATES IN STOCHASTIC THERMODYNAMICS

In stochastic thermodynamics, irreversibility is quantified by entropy production along a stochastic process. Throughout, the entropies and rates below are mathematical quantities defined from the model's densities and the learned score, evaluated under the time conventions of §2.1. Let p_t denote the marginal density of $\mathbf{x}_t$ and define the differential entropy

$$S(t) \;:=\; -\mathbb{E}_{p_t}\big[\log p_t(\mathbf{x}_t)\big], \qquad \dot{S}(t) \;=\; \frac{d}{dt} S(t). \tag{2}$$

For a stochastic process, the system entropy rate $\dot{S}(t)$ decomposes as (Seifert, 2012)

$$\dot{S}(t) \;=\; \dot{S}^{\,i}(t) \;+\; \dot{S}^{\,e}(t), \tag{3}$$

where $\dot{S}^{\,i}(t) \geq 0$ is the intrinsic entropy production rate (dissipation or irreversibility), and $\dot{S}^{\,e}(t)$ is an exchange entropy rate (entropy flow between the system and its surroundings).

For an overdamped SDE with homogeneous scalar noise, $d\mathbf{x}_t = \mathbf{f}\,(\mathbf{x}_t, t)\, dt + g(t) d\mathbf{w}_t$, the probability current is defined as $\mathbf{J}(\mathbf{x}, t) = p_t(\mathbf{x})\big(\mathbf{f}(\mathbf{x}, t) - D(t)\mathbf{s}(\mathbf{x}, t)\big)$ and the intrinsic entropy production rate has the form

$$\dot{S}^{\,i}(t) = \int \frac{\|\mathbf{J}(\mathbf{x}, t)\|^2}{D(t)\, p_t(\mathbf{x})}\, d\mathbf{x} = \frac{1}{D(t)}\, \mathbb{E}_{p_t}\Big[\big\| \mathbf{f}(\mathbf{x}_t, t) - D(t)\, \mathbf{s}(\mathbf{x}_t, t) \big\|^2 \Big], \tag{4}$$

which grows when the dynamics drive large probability currents.

A key identity we use later is a simplified expression for the system entropy rate:

$$\dot{S}(t) = \mathbb{E}_{p_t}\big[\nabla_{\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}_t, t)\big] + D(t)\, \mathbb{E}_{p_t}\big[\|\mathbf{s}(\mathbf{x}_t, t)\|^2\big]. \tag{5}$$

In generative models, the true score $\mathbf{s}(\mathbf{x}, t)$ is approximated by a neural network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$. We therefore define the model-based entropy rates by substituting $\mathbf{s}_{\boldsymbol{\theta}}$ into the above expressions, e.g.

$$\dot{S}_{\boldsymbol{\theta}}(t) := \mathbb{E}_{p_t}\big[\nabla_{\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}_t, t)\big] + D(t)\, \mathbb{E}_{p_t}\big[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|^2\big], \tag{6}$$

and similarly for $\dot{S}^{\,i}_{\boldsymbol{\theta}}(t)$ and $\dot{S}^{\,e}_{\boldsymbol{\theta}}(t)$. These quantities are directly computable from the trained score network and appear in our NLL bound and experiments. The rate $\dot{S}_{\boldsymbol{\theta}}$ is a plug-in functional of the learned score under the true marginals p_t, not the entropy rate of any learned density.

3 MAIN RESULTS

3.1 LOWER BOUND ON NEGATIVE LOG-LIKELIHOOD

Convention. We work under the *plug-in convention*: the model-induced entropy rates of §2.3 are evaluated against the true forward noising marginals p_t rather than the model-flow marginals $q_t^{\boldsymbol{\theta}}$ produced by integrating the learned probability flow ODE from p_{data}. Under this convention Eq. 8 is

exact for the surrogate $\mathrm{NLL}_{\mathrm{plug}}(\boldsymbol{\theta}) := S_{\mathrm{noise}} + \int_0^1 \mathbb{E}_{p_t}[-\nabla\cdot\mathbf{f} + D(t)\nabla\cdot\mathbf{s}_{\boldsymbol{\theta}}]\,dt$. The probability flow likelihood differs by a residual, $\mathrm{NLL}_{\mathrm{actual}} = \mathrm{NLL}_{\mathrm{plug}} + R_{\boldsymbol{\theta}}$, which reduces to the constant K_1 of Corollary 1 at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$ and which we measure on every system (Appendix C.4). It has no universal sign or scale and is carried explicitly below. The surrogate is a functional of the learned score, not a code length: it is unbounded below (Remark 2), so no claim about achievable likelihood rests on it alone.

Theorem 1 (Plug-in entropy rate decomposition)**.** *Under the assumptions of Appendix C, the plug-in NLL admits the exact decomposition*

$$\mathrm{NLL}_{\mathrm{plug}} \;=\; \frac{S_{data} + S_{noise}}{2} \;-\; \frac{1}{2}\int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\,dt \;+\; \frac{1}{2}\int_0^1 D(t)\,\mathbb{E}_{p_t}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) - \mathbf{s}_{\mathrm{true}}(\mathbf{x}_t,t)\|^2\right] dt. \tag{7}$$

Corollary 1 (NLL plug-in lower bound)**.** *Since the score-error term in Theorem 1 is nonnegative,*

$$\boxed{\mathrm{NLL}_{\mathrm{plug}} - S_{data} \geq \frac{1}{2}\left[S_{noise} - S_{data} - \int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\,dt\right],} \tag{8}$$

with equality at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$, in which case $\mathrm{NLL}_{\mathrm{plug}} = S_{\mathrm{data}}$ and $\mathrm{NLL}_{\mathrm{actual}} = S_{\mathrm{data}} + K_1$, where $K_1 = D_{\mathrm{KL}}(p_1\|p_{\mathrm{prior}})$ is the cost of starting the reverse integration from the chosen prior rather than from p_1 (Appendix C.3).

Corollary 1 is a statement about $\mathrm{NLL}_{\mathrm{plug}}$, where it cannot be violated because its slack is a squared norm. Transported to the probability flow likelihood it reads $\mathrm{NLL}_{\mathrm{actual}} \geq \mathrm{LB} + R_{\boldsymbol{\theta}}$, where LB is the lower bound, with the consequences for the entropy bound given in Section 4.

Proof sketch (Theorem 1). Let $\tilde{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{x},t) := \mathbf{f}(\mathbf{x},t) - D(t)\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x},t)$ denote the probability flow ODE vector field associated with the forward SDE Song et al. (2021b). By the instantaneous change-of-variables formula Chen et al. (2018),

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_0) \;=\; \log p_1(\mathbf{x}_1) + \int_0^1 \left(\nabla\cdot\mathbf{f}(\mathbf{x}_t,t) - D(t)\,\nabla\cdot\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)\right) dt. \tag{9}$$

Evaluating the time integrand under the true forward marginals p_t defines $\mathrm{NLL}_{\mathrm{plug}}$. The actual probability flow NLL is $\mathrm{NLL}_{\mathrm{actual}} = \mathrm{NLL}_{\mathrm{plug}} + R_{\boldsymbol{\theta}}$, where $R_{\boldsymbol{\theta}}$ is a prior constant at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$ (Appendix C). Applying Stein's identity $\mathbb{E}_{p_t}[\nabla\cdot\mathbf{s}_{\boldsymbol{\theta}}] = -\mathbb{E}_{p_t}[\mathbf{s}_{\boldsymbol{\theta}}\cdot\mathbf{s}_{\mathrm{true}}]$ to the score divergence and completing the square in the resulting cross term with $\|\mathbf{s}_{\mathrm{true}}\|^2$ yields Eq. 7; full proof in Appendix C.

3.2 Connection to Maxwell's demon and entropy rates

The Maxwell's demon thought experiment posits an external controller that observes a system and selectively acts on it to drive its entropy below what unaided dynamics would allow. Score-based generative models behave the same way: during training the network "measures" marginals of the noised process by minimizing the DSM loss in Eq. 1. At generation time it uses what it has learned to drive the system entropy of the controlled-forward process below its uncontrolled value. We make this precise on the driftless SDE $d\mathbf{x}_t = g(t)\,d\mathbf{w}_t$, using the controlled-forward parameterization $\tau = T - t$ with drift $\tilde{\mathbf{f}}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau,\tau) = g(T-\tau)^2\,\mathbf{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau, T-\tau)$ (Appendix D.2). Substituting this drift into the intrinsic and exchange entropy rate definitions of Appendix B (which apply to any drift, not only to the true score) and using $\dot{S}_{\boldsymbol{\theta}}(\tau) = \dot{S}^i_{\boldsymbol{\theta}}(\tau) + \dot{S}^e_{\boldsymbol{\theta}}(\tau)$ yields

$$\dot{S}^i_{\boldsymbol{\theta}}(\tau) = \tfrac{g(T-\tau)^2}{2}\,\mathbb{E}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau, T-\tau)\|^2\right], \quad \dot{S}^e_{\boldsymbol{\theta}}(\tau) = -2\,\dot{S}^i_{\boldsymbol{\theta}}(\tau), \quad \dot{S}_{\boldsymbol{\theta}}(\tau) = -\dot{S}^i_{\boldsymbol{\theta}}(\tau). \tag{10}$$

The system entropy rate of the controlled-forward process is exactly the negative of its intrinsic entropy production: a network capable of generation must drive a strictly negative system entropy rate, exchanging twice that magnitude with the surroundings. The full derivation is in Appendix D.

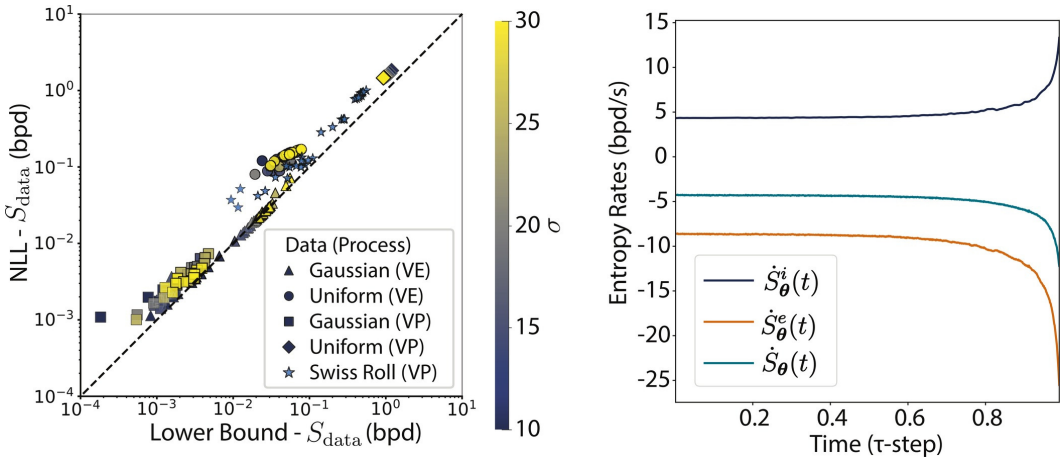


Figure 1: NLL versus the theoretical lower bound. (Left) NLL against the bound of Eq. 8 (dashed) for Gaussian, Uniform, and Swiss roll data. Marker shape encodes distribution and noising process, color encodes $\sigma \in [10, 30]$. (Right) Entropy rates ($\dot{S}^i_{\boldsymbol{\theta}}$, $\dot{S}^e_{\boldsymbol{\theta}}$, $\dot{S}_{\boldsymbol{\theta}}$) for the best-NLL Uniform (variance-exploding, VE) score network, confirming the driftless prediction $\dot{S}^e_{\boldsymbol{\theta}} = -2\,\dot{S}^i_{\boldsymbol{\theta}}$.

4 Numerical Results

We validate the theory on the noisy 2D Swiss roll, Gaussian, and synthetic 8-bit uniform-pixel data. For these systems $\mathrm{NLL}_{\mathrm{plug}}$ is evaluated by quadrature under the closed-form forward kernel, so Figures 1, 2a, and 3 and Table 1a are statements about the surrogate and carry no residual. The Ising and image experiments report the probability flow $\mathrm{NLL}_{\mathrm{actual}}$ and carry $R_{\boldsymbol{\theta}}$ (Appendix C.4). Appendix H gives the toy distribution protocols. Figure 1 (left) sweeps $\sigma \in \{10, 15, 20, 25, 30\}$ with 10 runs per setting (Gaussian, Uniform) and a range of model sizes (Swiss roll). The bound holds across all configurations, with the slack between NLL and the lower bound tracking the squared score-error term $\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2$ in Eq. 7: tighter bounds correlate with better-fitting models (Pearson r=0.803, Spearman r_s=0.909, both p<0.001). Figure 1 (right) shows that the empirical entropy rates of the best-performing Uniform-VE score network reproduce the driftless predictions of Eq. 10: $\dot{S}^i_{\boldsymbol{\theta}}(\tau)$ stays strictly positive, $\dot{S}^e_{\boldsymbol{\theta}}(\tau) = -2\,\dot{S}^i_{\boldsymbol{\theta}}(\tau)$ holds throughout, and the system rate $\dot{S}_{\boldsymbol{\theta}}(\tau) = -\dot{S}^i_{\boldsymbol{\theta}}(\tau)$ is uniformly negative, in agreement with the demon-like role of Section 3.2.

S_{data} is in closed form for Uniform/Gaussian data but unknown for structured distributions such as the Swiss roll. Rearranging Eq. 7 gives a model-based upper bound (UB) that tightens monotonically as the score-error slack shrinks and equals S_{data} in the exact-score limit:

$$S_{\mathrm{data}} \;\le\; \widehat{S}^{\mathrm{UB}}_{\mathrm{plug}} \;:=\; 2\,\mathrm{NLL}_{\mathrm{plug}}(\boldsymbol{\theta}) + \int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\,dt - S_{\mathrm{noise}}. \tag{11}$$

The slack $\widehat{S}^{\mathrm{UB}}_{\mathrm{plug}} - S_{\mathrm{data}}$ is exactly the score-error integral $\int_0^1 D(t)\,\mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2 dt$, twice the last term of Eq. 7. Evaluated with the probability flow likelihood instead, the bound reads $\widehat{S}^{\mathrm{UB}}_{\mathrm{actual}} = \widehat{S}^{\mathrm{UB}}_{\mathrm{plug}} + 2R_{\boldsymbol{\theta}}$ and remains an upper bound only while the score-error integral exceeds $2|R_{\boldsymbol{\theta}}|$, a property of the model that we test rather than assume (Appendix I). We compare the bound to nonparametric estimators (Kozachenko–Leonenko k-nearest-neighbor (k-NN) (Kozachenko & Leonenko, 1987) and a 2D histogram) and to the exact finite-lattice Ising entropy (Onsager, 1944; Ferdinand & Fisher, 1969). Estimator details are in Appendices N.2.2 and I. Figure 2 shows that the model-based UB tightens with score-network width and tracks the independent references. We report the smallest UB across checkpoints as a point estimate of S_{data}, not as a bound: a minimum over noisy estimates is biased low, and $\widehat{S}^{\mathrm{UB}}$ rewards low dissipation as well as low NLL, so the Ising sweep selects on a validation split and evaluates on fresh exact samples.

Table 1 compares the model-based UB against ground truth (nonparametric for the Swiss roll; exact enumeration over the 2^{25} configurations of the 5×5 noisy Ising model) and against Uniform

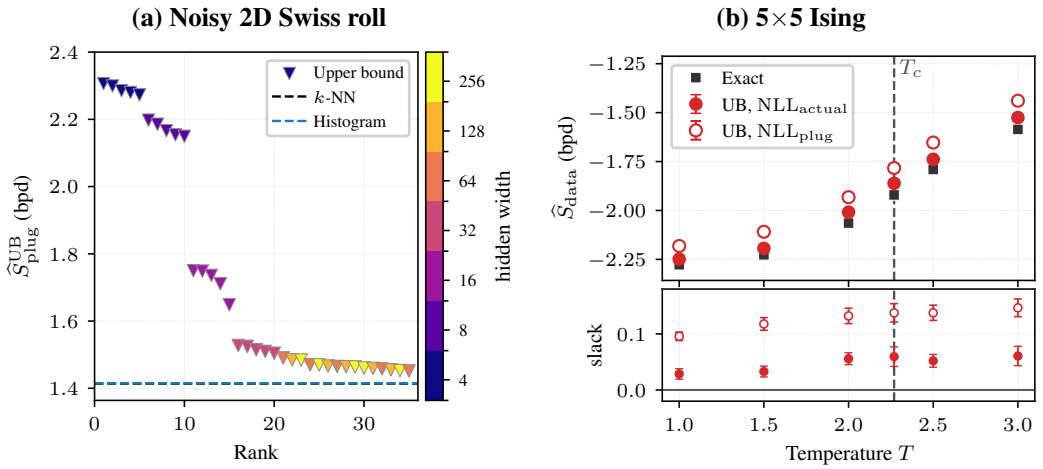


Figure 2: **Model-based upper bound on the data entropy. (a)** Noisy 2D Swiss roll: each marker is a trained score network (color = hidden width). Dashed lines mark k-NN and 2D-histogram nonparametric estimates of S_{data}. **(b)** 5×5 Ising with uniform noise of half-width 0.1, swept over $T \in [1, 3]$. Filled: $\widehat{S}^{\text{UB}}_{\text{actual}}$; hollow: $\widehat{S}^{\text{UB}}_{\text{plug}}$; both from a periodic-convolutional score network, mean±95% confidence interval (CI) over 10 seeds per T (Appendix I); lower panel: slack, UB − exact (bpd). Black: analytic differential entropy (Onsager, 1944; Ferdinand & Fisher, 1969). Dashed line: $T_c \approx 2.269$. Plug-in slack = score error + truncation smoothing. The actual slack adds $2R_{\boldsymbol{\theta}} < 0$, measured per temperature in Table 5.

(bounding-box) and moment-matched Gaussian maximum entropy references. On the Swiss roll both nonparametric estimates sit just below the model UB and both maximum entropy references far exceed it. The minimum is attained at H=64, and the three widest networks give indistinguishable bounds, as expected once the score stops improving with width. The UB is an estimate of S_{data}, not a substitute for it in any inequality: writing $\widehat{S} = S_{\text{data}} + \Delta$, the reported KL becomes $D_{\text{KL}}(p_{\text{data}} \| p_{\boldsymbol{\theta}}) - \Delta$ and the right-hand side of Eq. 8 rises with Δ, so $\Delta > 0$ is optimistic on both counts. Only a guaranteed lower bound on S_{data} may be substituted, and neither $\widehat{S}^{\text{UB}}$ nor an external cross-entropy is one.

Figure 3a shows the accuracy–dissipation relation on the Swiss roll. We sweep the hidden width of a 3-layer multilayer perceptron (MLP) score network (5 seeds each) and plot $\text{NLL}_{\text{plug}} - S_{\text{data}}$, the plug-in surrogate for $D_{\text{KL}}(p_0 \| p_{\boldsymbol{\theta}})$, and our entropic lower bound, computed from the forward-clock

Table 1: **Data entropy estimates (bpd).** (a) Noisy 2D Swiss roll. (b) 5×5 noisy Ising at T=3, noise = 0.1. Each panel reports the model-based UB, an independent ground truth (nonparametric for (a); exact enumeration for (b)), and Uniform/Gaussian maximum entropy references. Swiss roll UB: $\widehat{S}^{\text{UB}}_{\text{plug}}$, smallest across checkpoints (H=64, Appendix N.2.2); k-NN at k=10. Ising UB: the network of Fig. 2b at T=3 (10-seed mean), plug-in NLL; slack +0.147 bpd = score error 0.102 + truncation smoothing 0.045; the probability flow value −1.525 sits lower by $2R_{\boldsymbol{\theta}} = -0.086$ bpd (Appendix I). The nonparametric references are estimates, not guaranteed bounds: $\widehat{S}^{\text{UB}}$ may be compared against them but not substituted into Eq. 8, whose right-hand side increases with the entropy supplied.

(a) Noisy 2D Swiss roll

Estimator / reference	$\widehat{S}_{\text{data}}$ (bpd)
Model UB	1.452
k-NN (Kozachenko–Leonenko)	1.414
2D discretized histogram	1.415
Uniform (reference)	2.472
Gaussian (reference)	2.487

(b) 5×5 Ising model (noise = 0.1, T = 3)

Estimator / reference	$\widehat{S}_{\text{data}}$ (bpd)
Model UB (plug-in NLL)	−1.439
Exact (differential, analytic)	−1.586
Uniform (reference)	1.138
Gaussian (reference)	2.049

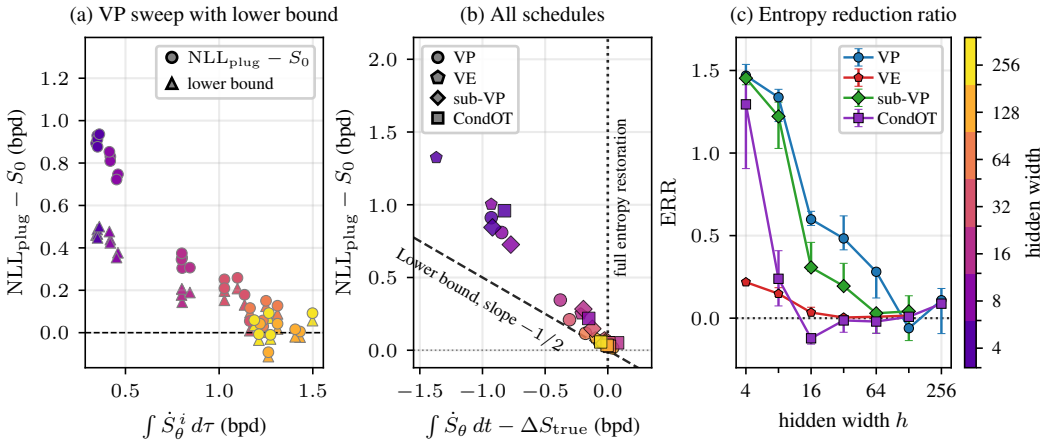


Figure 3: **Accuracy–dissipation relation on the Swiss roll.** The plug-in surrogate $\mathrm{NLL}_{\mathrm{plug}} - S_{\mathrm{data}}$ for $D_{\mathrm{KL}}(p_0 \| p_{\boldsymbol{\theta}})$ (bpd). Color is hidden width. **(a)** Variance-preserving (VP) sweep, one point per seed, against the intrinsic rate: plug-in KL (circles) and the entropic lower bound (up-triangles), which holds at all 35 checkpoints. **(b)** Per-width medians for VP, VE, sub-variance-preserving (sub-VP), and conditional optimal transport (CondOT) against the system rate shifted by each family's $\Delta S_{\mathrm{true}} = S_{\mathrm{noise}} - S_{\mathrm{data}}$: the dotted line is full entropy restoration ($\mathrm{ERR} = 0$), attained at the exact score, and the dashed line is Corollary 1, the same line of slope $-\frac{1}{2}$ for every family. **(c)** The entropy reduction ratio (Eq. 12) against width; medians with interquartile bars over seeds.

system rate (Appendix B), against the integrated intrinsic dissipation $\int_0^1 \dot{S}^i_{\boldsymbol{\theta}}(\tau)\, d\tau$. The lower bound is a theorem for the plug-in surrogate and holds at all 35 checkpoints. We do not plot the probability flow likelihood, for which it is not a theorem: transported to the probability flow KL, the bound holds on 26 of the 32 converged checkpoints against the nonparametric reference, failing on three each at $H \in \{32, 64\}$. The sufficient condition of Appendix C.4, which needs no reference entropy, establishes it on 11 of the 32, never falsely. Greater width yields lower KL at higher dissipation, with the entropic-bound gap tightening as the score improves.

The relation transfers to pretrained image models (Figure 4; details in Appendix J). On CIFAR-10, moving from standard to deep or from Fréchet inception distance (FID) selected to likelihood-weighted (LW) checkpoints lowers NLL while raising $\int_0^1 \dot{S}^i_{\boldsymbol{\theta}}\, d\tau$. On ImageNet-32, restricting to the eight LW DDPM++ score networks of Song et al. (2021a) matches the 3.76 bpd LW-trained reference and avoids the mismatched-objective DSM-vs-LW gap. The entropic lower bound tracks the achieved NLL from below at every point.

The same probe applies unchanged to four FID-optimized NCSN++ VE checkpoints from Song et al. (2021b) at 256×256 on held-out splits (Table 2). For VE the system and intrinsic rates coincide. External likelihoods are cross-entropies and hence upper bounds on S_{data}, so the Efficient-VDVAE values (Hazami et al., 2022) serve only as entropic proxies. The entropy bounds carry $2R_{\boldsymbol{\theta}} \in [-1.84, -0.86]$ bpd, so the FFHQ bound of 2.53 bpd does not undercut the 2.17 bpd proxy. Against $\mathrm{NLL}_{\mathrm{plug}}$, the quantity the bound is about, the lower bound holds by 0.18 bpd on FFHQ and 0.93 bpd on CelebA-HQ (against $\mathrm{NLL}_{\mathrm{actual}}$ the FFHQ bound would read as violated). On LSUN it equals $\mathrm{NLL}_{\mathrm{plug}}$ by construction. The N_t=50 rate quadrature is unconverged by at least 1.43 bpd (Appendix J). Correcting it changes no conclusion but leaves these entropy bounds above the models' own NLL, hence uninformative. Dissipation lies in $[12.3, 13.4]$ bpd across datasets, set by $\sigma_{\max}$=380, while NLL ranges from 3.15 to 4.04 bpd.

Finally, Figure 3b places the VP, VE, sub-VP, and CondOT Swiss roll sweeps on the system axis on which Corollary 1 is stated, shifted by each family's exact-score value $\Delta S_{\mathrm{true}} = S_{\mathrm{noise}} - S_{\mathrm{data}}$ so that all families share one reference. There the corollary is a single line of slope $-\frac{1}{2}$ through the origin (Appendix L): every model lies above it, small models underproduce entropy relative to the exact process ($\int \dot{S}_{\boldsymbol{\theta}} dt < \Delta S_{\mathrm{true}}$), and with width every family moves down the line toward the corner

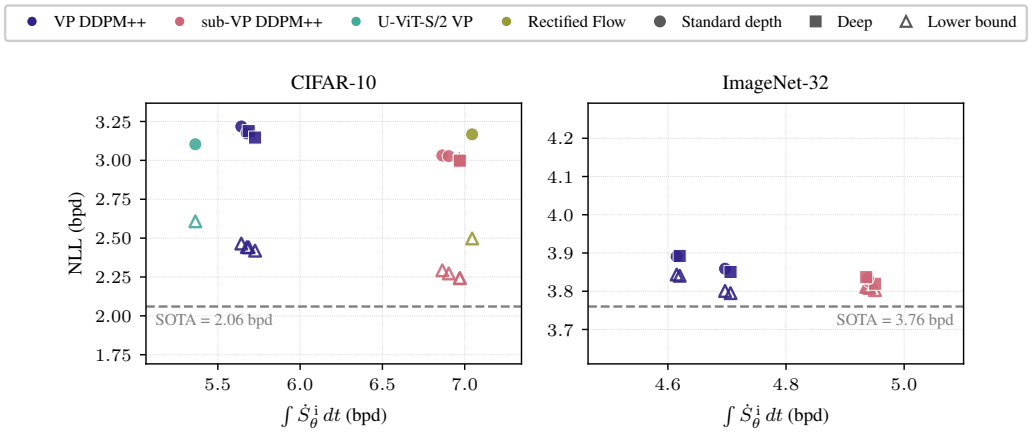


Figure 4: **Accuracy–dissipation relation on real-image score-based models.** NLL (filled) and entropic lower bound (hollow triangles) versus $\int_0^1 \dot{S}^{i}_{\boldsymbol{\theta}}\, d\tau$ (bpd). Left: CIFAR-10 VP/sub-VP DDPM++ (Song et al., 2021b) (circles = standard, squares = deep), U-ViT-S/2 (Bao et al., 2023), and Rectified Flow (Liu et al., 2023), the VP network trained on a flow matching path (Appendix J). Arrows connect checkpoints within each (SDE, depth) family. Right: ImageNet-32 VP/sub-VP DDPM++, likelihood-weighted (LW) variants only (plain LW and LW with importance-weighted t-sampling). Both quantities are evaluated from the probability flow NLL and carry $R_{\boldsymbol{\theta}}$ (Appendix C.4). Dashed lines: external NLLs used as entropy proxies (2.06 bpd CIFAR-10 (Xu et al., 2025); 3.76 bpd ImageNet-32 (Song et al., 2021a)), which upper-bound S_{data}, so a bound computed from them is a proxy quantity. VP and sub-VP positions are not comparable across families: the sub-VP dissipation integral diverges logarithmically at the lower endpoint, so its value is set by the cutoff (Appendix J). Within-family arrows are unaffected. The plotted axis is the intrinsic rate. For VP and sub-VP it differs from the system rate that enters Theorem 1 by 0.25–0.53 bpd (Appendix B).

where the bound is attained. This suggests a measure of fit, the *entropy reduction ratio*

$$\mathrm{ERR}(\boldsymbol{\theta}) \;=\; 1 - \frac{\int_0^1 \dot{S}_{\boldsymbol{\theta}}\, dt}{\Delta S_{\mathrm{true}}} \;=\; \frac{\Delta S_{\mathrm{true}} - \int_0^1 \dot{S}_{\boldsymbol{\theta}}\, dt}{S_{\mathrm{noise}} - S_{\mathrm{data}}}, \tag{12}$$

the fraction of the noising schedule's entropy change that the model fails to restore, computed from score forward passes alone (Algorithm 1; Appendix M). At the exact score $\mathrm{ERR} = 0$, and Corollary 1 reads $\mathrm{NLL}_{\mathrm{plug}} - S_{\mathrm{data}} \geq \frac{1}{2}\,\Delta S_{\mathrm{true}}\,\mathrm{ERR}$. ERR is not sign-definite, since a model can overproduce entropy, but in our sweeps it is positive at small width and plateaus at $|\mathrm{ERR}| \leq 0.1$ from $h = 16$ (VE), 32 (CondOT), 64 (sub-VP), and 128 (VP) (Figure 3c). Beyond that width added capacity restores no further entropy, which gives a stopping rule for model size that needs no held-out likelihood. Within each family, dissipation and NLL are both monotone in capacity, so this sweep cannot separate a dissipation effect from a capacity effect, and we treat the relation as observational (Appendix K.5).

Algorithm 1 Entropy reduction ratio (ERR) of a trained model

Input: score $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ (a flow model's velocity converted as in Appendix K.6); schedule $f(t)$, $D(t) = g(t)^2/2$, $m(t)$, $\sigma(t)$; data $\{\mathbf{x}_0^{(i)}\}_{i=1}^n$; data-entropy reference S_{data}; quadrature nodes and weights $\{(t_j, w_j)\}$

$\Delta S_{\mathrm{true}} \leftarrow \frac{d}{2}\log\big(2\pi e\,\sigma^2(1)\big) - S_{\mathrm{data}}$ ▷ $S_{\mathrm{noise}} - S_{\mathrm{data}}$

for each node t_j **do**

  $\mathbf{x}^{(i)} \leftarrow m(t_j)\,\mathbf{x}_0^{(i)} + \sigma(t_j)\,\mathbf{z}^{(i)}, \quad \mathbf{z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

  $a_j \leftarrow \frac{1}{n}\sum_i \|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, t_j)\|^2$ ▷ score forward passes only

end for

$\Delta S_{\boldsymbol{\theta}} \leftarrow d\int f(t)\,dt + \sum_j w_j\, D(t_j)\, a_j$ ▷ $= \int \dot{S}_{\boldsymbol{\theta}}\, dt$, Eq. 5

return $\mathrm{ERR} = 1 - \Delta S_{\boldsymbol{\theta}}/\Delta S_{\mathrm{true}}$

Table 2: **Cross-dataset evaluation at 256×256.** NCSN++ VE checkpoints from Song et al. (2021b) on held-out splits (n=500 images per dataset; LSUN Bedroom 300). Columns: integrated system entropy rate; probability flow NLL±standard error; measured residual $R_{\boldsymbol{\theta}} = \mathrm{NLL}_{\mathrm{actual}} - \mathrm{NLL}_{\mathrm{plug}}$; $\mathrm{NLL}_{\mathrm{plug}}$; the Corollary 1 floor with the external proxy substituted for S_{data} (blank when none exists), a proxy quantity that overstates the floor and bounds only the continuous-density NLL (Appendix J); the external entropic proxy (Hazami et al., 2022); and $\widehat{S}^{\mathrm{UB}}_{\mathrm{plug}} = 2\,\mathrm{NLL}_{\mathrm{plug}} - S_{\mathrm{noise}} + \int \dot{S}_{\boldsymbol{\theta}}\, dt$. S_{noise}=17.617 bpd; 8-bit corrected bpd, offset 7.0.

Dataset	$\int \dot{S}_{\boldsymbol{\theta}}\, dt$ (bpd)	$\mathbf{NLL}_{\mathbf{actual}}$ (bpd)	$R_{\boldsymbol{\theta}}$ (bpd)	$\mathbf{NLL}_{\mathbf{plug}}$ (bpd)	**Floor (proxy, cont.)** (bpd)	**Proxy** S_{data} (bpd)	$\widehat{S}^{\mathrm{UB}}_{\mathrm{plug}}$ (bpd)
CelebA-HQ	12.344	3.602 ± 0.016	-0.635	4.237	3.312	1.35	3.201
LSUN Bedroom	12.857	3.390 ± 0.029	-0.920	4.310	–	–	3.860
FFHQ	12.978	3.154 ± 0.015	-0.432	3.586	3.404	2.17	2.533
LSUN Church	13.401	4.044 ± 0.021	-0.560	4.604	–	–	4.993

5 Discussion

Contributions. Variational analyses connect score matching to ELBO-style upper bounds (Huang et al., 2021; Song et al., 2021a). Eq. 7 complements them with a plug-in lower-side identity whose slack is the integrated score error. The two bound different objects, the ELBO the reverse-SDE likelihood and Corollary 1 the plug-in surrogate, so they do not form a two-sided bound. Rearranging the same identity makes the data entropy an estimable quantity: it yields a model-based upper bound on S_{data} whose slack is the score-error integral, tight in the exact-score limit. Across synthetic, finite-Ising, and pretrained image score models, improved NLL is accompanied by larger integrated intrinsic entropy production, complementing speed/accuracy (Ikeda et al., 2025) and energy/speed/accuracy (Klinger & Rotskoff, 2025) analyses by measuring the NLL side directly.

Limitations. The decomposition is exact for $\mathrm{NLL}_{\mathrm{plug}}$. The probability flow NLL differs by a residual $R_{\boldsymbol{\theta}}$ of no fixed sign, measured on every system (Appendix C.4). Transported to $\mathrm{NLL}_{\mathrm{actual}}$, Corollary 1 and Eq. 11 hold only while the score-error integral exceeds $2|R_{\boldsymbol{\theta}}|$ (38 of 60 Ising checkpoints), so a guaranteed statement about actual NLL needs an exact S_{data} or the plug-in surrogate. On natural images neither S_{data} nor the absolute slack is observable. Evaluation depends on dequantization (noise that makes discrete data continuous), solver tolerances, rate quadrature, and proxy entropy references. Likelihood, fidelity, and coverage measure different things (Theis et al., 2016; Kynkäänniemi et al., 2019). The accuracy–dissipation relation is an association across checkpoints: capacity drives both axes in our sweeps. Within a single width the dissipation spread is a small fraction of the between-width range (Appendix K.5), so a causal version would require training matched-likelihood pairs that differ only in a dissipation penalty. Theorem 1 assumes affine drift and scalar diffusion (Appendix C) and covers any parameterization that maps to the score of Gaussian forward marginals. Schrödinger bridges (Bortoli et al., 2021), nonaffine drift, and discrete-state diffusion are outside its scope, and our empirical conclusions cover only the models evaluated here.

Impact and future directions. Exact likelihoods of large diffusion and flow models need an ODE solve with trace estimation per sample. The entropy reduction ratio needs only score evaluations on noised data (Algorithm 1). It can track fit during training and across model scales where likelihood evaluation is impractical, and its plateau marks the scale beyond which added capacity restores no further entropy. Treating $\int \dot{S}^i_{\boldsymbol{\theta}}\, d\tau$ as a design variable makes well-posed the search for target-NLL solutions at minimal dissipation (International Energy Agency, 2025). On stochastic/thermodynamic hardware (Coles et al., 2023) these rates acquire physical meaning, extending the Maxwell's demon framing of Premkumar (2025). In digital experiments they remain mathematical diagnostics. In physics, biology, climate, and inverse problems (Watson et al., 2023; Price et al., 2025; Daras et al., 2024), generative surrogates target distributions whose entropy and irreversibility are scientifically meaningful. The Ising experiments show how to calibrate entropy diagnostics against analytic targets before deploying them where exact references are unavailable.

AI USE STATEMENT

During the preparation of this work, the authors used Claude Opus 4.7, 4.8, and 5 and Claude Code (Anthropic) to improve language and readability, and as coding assistants in developing the research software. All AI-assisted code was written under the authors' direction and reviewed and tested by the authors before use in any reported experiment. Large language models were not used to design the study, derive the theory, or interpret the results. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the published article.

REFERENCES

Brian D O Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12:313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679, 2023.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

Ricky T Q Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Patrick J Coles, Collin Szczepanski, Denis Melanson, Kaelan Donatella, Antonio J Martinez, and Faris Sbahi. Thermodynamic ai and the fluctuation frontier, 2023. URL https://arxiv.org/abs/2302.06584.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2005. ISBN 9780471241959. doi: 10.1002/047174882X.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024. URL https://arxiv.org/abs/2410.00083.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 4 2022.

Arthur E. Ferdinand and Michael E. Fisher. Bounded and inhomogeneous ising models. i. specific-heat anomaly of a finite lattice. *Physical Review*, 185:832–846, 1969. doi: 10.1103/PhysRev.185.832.

U G Haussmann and E Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14:1188 – 1205, 1986. doi: 10.1214/aop/1176992362. URL https://doi.org/10.1214/aop/1176992362.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-VDVAE: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Peter Holderrieth and Ezra Erives. An introduction to flow matching and diffusion models, 2025. URL https://arxiv.org/abs/2506.02070.

Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. In M Ranzato, A Beygelzimer, Y Dauphin, P S Liang, and J Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22863–22876. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c11abfd29e4d9b4d4b566b01114d8486-Paper.pdf.

Kotaro Ikeda, Tomoya Uda, Daisuke Okanohara, and Sosuke Ito. Speed-accuracy relations for diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport. *Physical Review X*, 15:031031, 7 2025. doi: 10.1103/x5vj-8jq9. URL https://link.aps.org/doi/10.1103/x5vj-8jq9.

International Energy Agency. Energy and AI. Technical report, International Energy Agency, Paris, 2025. URL https://www.iea.org/reports/energy-and-ai.

Boris Kastening. Simplifying Kaufman's solution of the two-dimensional Ising model. *Physical Review E*, 64:066106, 2001. doi: 10.1103/PhysRevE.64.066106.

Bruria Kaufman. Crystal statistics. ii. partition function evaluated by spinor analysis. *Physical Review*, 76:1232–1243, 1949. doi: 10.1103/PhysRev.76.1232.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Jérémie Klinger and Grant M Rotskoff. Universal energy-speed-accuracy trade-offs in driven nonequilibrium systems. *Physical Review E*, 111:014114, 1 2025. doi: 10.1103/PhysRevE.111.014114. URL https://link.aps.org/doi/10.1103/PhysRevE.111.014114.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23:95–101, 6 1987.

Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2022. ISBN 9781713871088.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Qiang Liu and Dilin Wang. Stein variational gradient descent: a general purpose bayesian inference algorithm. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2378–2386. Curran Associates Inc., 2016. ISBN 9781510838819.

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

Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65:117–149, 1944. doi: 10.1103/PhysRev.65.117.

Akhil Premkumar. Neural entropy, 2025. URL https://arxiv.org/abs/2409.03817.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 2025. doi: 10.1038/s41586-024-08252-9.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations*, 5 2023.

Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75:126001, 11 2012. doi: 10.1088/0034-4885/75/12/126001. URL https://dx.doi.org/10.1088/0034-4885/75/12/126001.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265. PMLR, 2 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 2021a.

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.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.01844.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23:1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Max Welling, Cheng Lu, and Lars Holdijk. *Generative AI and Stochastic Thermodynamics: A Tale of Free Energies*. Cambridge University Press, 2026.

Lu Xu, Tsai Hor Chan, Kwok Fai Lam, Lequan Yu, and Guosheng Yin. Variational Polya tree. *arXiv preprint arXiv:2510.22651*, 2025.

APPENDIX CONTENTS

Part I Theory

Generative modeling and stochastic thermodynamics . . . . . **A, p. 14**

Entropy rates for score-based diffusion models . . . . . **B, p. 14**
Identity relating current, score, and drift . . . . . B.1, p. 14
Intrinsic entropy production rate . . . . . B.2, p. 15
Exchange (medium) entropy flow rate . . . . . B.3, p. 15
System entropy rate . . . . . B.4, p. 15
Three entropy rate objects and their clocks . . . . . B.5, p. 15

Lower bound for negative log-likelihood . . . . . **C, p. 16**
Log-likelihood from probability flow ODE . . . . . C.1, p. 16
NLL lower bound . . . . . C.2, p. 17
Three terminal quantities . . . . . C.3, p. 18
Plug-in surrogate vs. actual NLL . . . . . C.4, p. 18
Stein's identity . . . . . C.5, p. 19

Maxwell's demon in the controlled-forward process . . . . . **D, p. 20**
Reverse process . . . . . D.1, p. 20
Controlled-forward process . . . . . D.2, p. 20
Entropy rates of the controlled-forward process . . . . . D.3, p. 21

General continuous-time diffusion processes . . . . . **E, p. 21**
Variance-exploding (VE) SDE . . . . . E.1, p. 21
Variance-preserving (VP) SDE . . . . . E.2, p. 21
Constructing the VP schedule from σ . . . . . E.3, p. 22

Standard Gaussian data . . . . . **F, p. 23**

Exact score of the uniform and normal distributions . . . . . **G, p. 23**

Part II Numerical Methods

Toy distribution experimental details . . . . . **H, p. 25**

Ising experimental details . . . . . **I, p. 26**

Image dataset experimental details . . . . . **J, p. 28**

Swiss roll experimental details . . . . . **K, p. 30**
Swiss roll dataset specifications . . . . . K.1, p. 30
VP schedule specifications . . . . . K.2, p. 31
Model specifications . . . . . K.3, p. 31
Training specifications . . . . . K.4, p. 31
Capacity confound . . . . . K.5, p. 32
CondOT flow matching models . . . . . K.6, p. 32

Reference lines of the accuracy–dissipation figure . . . . . **L, p. 32**

Computing the entropy reduction ratio . . . . . **M, p. 34**

Numerical Estimates of Exact NLL Terms . . . . . **N, p. 35**
Equilibrium entropy S_{noise} . . . . . N.1, p. 35
Dataset entropy S_{data} . . . . . N.2, p. 35
Squared norm of the model score . . . . . N.3, p. 36
Squared difference term, I_{diff} . . . . . N.4, p. 36

Part I Theory

A Generative modeling and stochastic thermodynamics

Table 3 summarizes the correspondence between objects in score-based generative modeling and their stochastic thermodynamic counterparts as used in this paper. The mapping is intended as a reading aid for Section 2.1 and Section 3, not as an assertion that the two columns are physically identical. In digital experiments the right column is a mathematical model quantity, while on dedicated thermodynamic hardware it can become a physical observable.

Table 3: Correspondence between objects of score-based generative modeling and their stochastic thermodynamic counterparts as used in this paper.

Generative modeling object	**Stochastic thermodynamic counterpart**	**Role in this paper**
Data distribution p_{data}	Far-from-equilibrium state	Initial structured distribution
Forward marginal p_1; prior p_{prior}	Relaxed state; equilibrium reference state	Terminal law of noising vs. the law the sampler starts from; they differ by $K_1 = D_{\mathrm{KL}}(p_1 \Vert p_{\mathrm{prior}})$ (§C.3)
Forward noising SDE	Entropy-increasing relaxation	Defines marginals p_t
Score $\nabla \log p_t$	Thermodynamic force (control field)	Determines reverse dynamics
Learned score $\mathbf{s}_{\boldsymbol{\theta}}$	Approximate controller (demon)	Induces model entropy rates
Probability current $\mathbf{J}$	Irreversible flux	Defines intrinsic entropy production
System entropy rate $\dot{S}_{\boldsymbol{\theta}}$	Time derivative of differential entropy	Enters the NLL lower bound
Score error	Excess control cost	Nonnegative slack in the decomposition
$\mathrm{NLL}_{\mathrm{actual}}$; $\mathrm{NLL}_{\mathrm{plug}}$	Cross-entropy (coding cost); plug-in functional of the score, not a code length	Reported likelihood; the object Corollary 1 bounds

Terminology. *Plug-in*: the learned score $\mathbf{s}_{\boldsymbol{\theta}}$ is substituted into the entropy rate formulas of §2.3 and expectations are taken under the true forward marginals p_t, obtained by noising data with the closed-form kernel, rather than under the model flow marginals $q_t^{\boldsymbol{\theta}}$. The resulting surrogate is $\mathrm{NLL}_{\mathrm{plug}}$. The likelihood reported for the Ising and image models is the probability flow $\mathrm{NLL}_{\mathrm{actual}} = \mathrm{NLL}_{\mathrm{plug}} + R_{\boldsymbol{\theta}}$, with the residual $R_{\boldsymbol{\theta}}$ defined and measured in §C.4. The toy and Swiss roll experiments report $\mathrm{NLL}_{\mathrm{plug}}$ itself. *Nats* are natural log units; *bpd* is bits per dimension, $\mathrm{nats}/(d \log 2)$ plus the 8-bit offset of §J. Every inequality we report compares terms that share one support, dequantization, prior, and unit convention. *Observational* denotes an association measured across trained checkpoints, objectives, schedules, or solvers; *causal* denotes a claim that would survive intervening on one factor with the others held fixed.

B Entropy rates for score-based diffusion models

Seifert's original formulation (Seifert, 2012) and subsequent applications to diffusion models (Ikeda et al., 2025) motivate the mathematical framework we adopt here.

B.1 Identity relating current, score, and drift

For the general overdamped SDE $d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\, dt + g(t) d\mathbf{w}_t$, the probability current is

$$\mathbf{J}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t) p_t(\mathbf{x}) - D(t) \nabla_{\mathbf{x}} p_t(\mathbf{x}) = p_t(\mathbf{x}) [\mathbf{f}(\mathbf{x}, t) - D(t) \mathbf{s}(\mathbf{x}, t)],$$

where we have used $g(t)^2 = 2D(t)$, $\mathbf{s}(\mathbf{x}) = \nabla \log p_t(\mathbf{x})$ and $\nabla p_t(\mathbf{x}) = p_t(\mathbf{x}) \nabla \log p_t(\mathbf{x}) = p_t(\mathbf{x}) \mathbf{s}(\mathbf{x})$. The local velocity field is defined as

$$\mathbf{v}(\mathbf{x}, t) := \mathbf{f}(\mathbf{x}, t) - D(t) \mathbf{s}(\mathbf{x}, t) = \mathbf{J}(\mathbf{x}, t) / p_t(\mathbf{x})$$

B.2 Intrinsic entropy production rate

Seifert (2012)'s original expression for the intrinsic entropy production rate is given by

$$\dot{S}^i(t) = \int \frac{\|\mathbf{J}(\mathbf{x},t)\|^2}{D(t)p_t(\mathbf{x})}\mathrm{d}\mathbf{x} = \frac{1}{D(t)}\int p_t(\mathbf{x})\,\|\mathbf{v}(\mathbf{x},t)\|^2\,d\mathbf{x}.$$

In expectation notation,

$$\begin{aligned}\dot{S}^i(t) &= \frac{1}{D(t)}\mathbb{E}\left[\|\mathbf{v}(\mathbf{x},t)\|^2\right] = \frac{2}{g(t)^2}\mathbb{E}\left[\|\mathbf{f}(\mathbf{x},t) - \frac{g(t)^2}{2}\mathbf{s}(\mathbf{x},t)\|^2\right] \\ &= \frac{1}{2g(t)^2}\mathbb{E}\left[\|2\mathbf{f}(\mathbf{x},t) - g(t)^2\mathbf{s}(\mathbf{x},t)\|^2\right]\end{aligned}$$

B.3 Exchange (medium) entropy flow rate

Seifert defines the entropy component of the medium surrounding a system (related to the heat dissipated into that medium) through the work done by the force $F(\mathbf{x},t)$ on the system at some time-dependent temperature $T(t)$,

$$\dot{S}^m(t) = \frac{1}{T(t)}\int \mathbf{F}(\mathbf{x},t)\cdot\mathbf{J}(\mathbf{x},t)d\mathbf{x} = \frac{1}{D(t)}\int \mathbf{f}(\mathbf{x},t)\cdot\mathbf{J}(\mathbf{x},t)d\mathbf{x}$$

In order to make the analogy between the diffusion algorithm and a physical system, we imagine a mobility (inverse friction) constant μ and corresponding Einstein relation $D(t) = \mu T(t)$, allowing us to write the drift term $\mathbf{f} = \mu\mathbf{F}$ in the SDE $d\mathbf{x}_t = \mu\mathbf{F}(\mathbf{x},t) + g(t)d\mathbf{w}_t$.

The exchange/flow rate of entropy into the system is just the negative of the one into the medium,

$$\begin{aligned}\dot{S}^e(t) = -\dot{S}^m(t) &= -\frac{1}{D(t)}\mathbb{E}\left[\mathbf{f}(\mathbf{x},t)\cdot(\mathbf{f}(\mathbf{x},t) - D(t)\mathbf{s}(\mathbf{x},t))\right] \\ &= -\frac{1}{D(t)}\mathbb{E}\left[\|\mathbf{f}(\mathbf{x},t)\|^2\right] + \mathbb{E}[\mathbf{f}(\mathbf{x},t)\cdot\mathbf{s}(\mathbf{x},t)] \\ &= -\frac{2}{g(t)^2}\mathbb{E}\left[\|\mathbf{f}(\mathbf{x},t)\|^2\right] + \mathbb{E}[\mathbf{f}(\mathbf{x},t)\cdot\mathbf{s}(\mathbf{x},t)],\end{aligned}$$

where in the first line we have used the fact that $\mathbf{J}(\mathbf{x},t) = \mathbf{v}(\mathbf{x},t)p_t(\mathbf{x})$.

B.4 System entropy rate

Combining the expressions for $\dot{S}^i$ and $\dot{S}^e$, expanding the square and canceling terms gives the simplified equation for the (total) system entropy rate:

$$\begin{aligned}\dot{S}(t) &= \dot{S}^i(t) + \dot{S}^e(t) \\ &= \frac{1}{D(t)}\mathbb{E}\left[\|\mathbf{f}(\mathbf{x},t) - D(t)\mathbf{s}(\mathbf{x},t)\|^2\right] - \frac{1}{D(t)}\mathbb{E}\left[\|\mathbf{f}(\mathbf{x},t)\|^2\right] + \mathbb{E}[\mathbf{f}(\mathbf{x},t)\cdot\mathbf{s}(\mathbf{x},t)] \\ &= -\mathbb{E}[\mathbf{f}(\mathbf{x},t)\cdot\mathbf{s}(\mathbf{x},t)] + D(t)\mathbb{E}\left[\|\mathbf{s}(\mathbf{x},t)\|^2\right] \\ &= \mathbb{E}[\nabla_{\mathbf{x}}\cdot\mathbf{f}(\mathbf{x},t)] + \frac{g(t)^2}{2}\mathbb{E}\left[\|\mathbf{s}(\mathbf{x},t)\|^2\right].\end{aligned}$$

where we have used Stein's identity for $\mathbb{E}[\nabla_{\mathbf{x}}\cdot\mathbf{f}(\mathbf{x},t)] = -\mathbb{E}[\mathbf{f}(\mathbf{x},t)\cdot\mathbf{s}(\mathbf{x},t)]$ (see Sec. C.5).

B.5 Three entropy rate objects and their clocks

The first and third coincide only for driftless processes. Under VP the measured difference $\int \dot{S}_{\boldsymbol{\theta}}\,dt - \int \dot{S}^i_{\boldsymbol{\theta}}\,d\tau$ is $+0.53$ bpd on the CIFAR-10 VP checkpoints, $+0.33$ on sub-VP, and $+0.25$ to $+0.50$ on ImageNet-32, and it is exactly zero on every VE row. On the Swiss roll VP sweep the two can even differ in sign at low capacity. The probability flow ODE along which likelihood is evaluated is deterministic and produces no entropy of its own. The dissipation reported in this paper is that of the

Table 4: **Three entropy rate objects and their clocks.**

Object	Clock	Where it appears
$\dot{S}_{\boldsymbol{\theta}}(t)$, system rate under p_t with the learned score	forward (noising) t	Theorem 1, Corollary 1, Eq. 11
$\dot{S}_{\boldsymbol{\theta}}(\tau)$, system rate of the controlled-forward process	reverse (generation) τ	§3.2, driftless case only
$\int_0^1 \dot{S}^i_{\boldsymbol{\theta}}(\tau)\, d\tau$, integrated intrinsic production	reverse (generation) τ	Figures 3 and 4

forward SDE whose marginals the plug-in convention borrows. Two further cautions. $\dot{S}_{\boldsymbol{\theta}}$ is a plug-in functional of the learned score under the true marginals, not the entropy rate of any learned density. And the relation $\dot{S}_{\boldsymbol{\theta}} = \dot{S}^i_{\boldsymbol{\theta}} + \dot{S}^e_{\boldsymbol{\theta}}$ holds algebraically for any $\mathbf{s}_{\boldsymbol{\theta}}$, whereas the divergence form of Eq. 5 uses Stein's identity and coincides with the sum form only at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\text{true}}$.

C Lower bound for negative log-likelihood

Assumptions. The derivation below uses the score-based diffusion setup of §2.2: (i) the forward SDE has the affine form $\mathbf{f}(\mathbf{x}, t) = f(t)\,\mathbf{x}$ and $\mathbf{G}(\mathbf{x}, t) = g(t)\,\mathbf{I}_d$ on $t \in [0, 1]$, with terminal $p_1 \approx \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$; (ii) the data and prior have finite differential entropy and $p_t, \mathbf{s}_\theta, \mathbf{s}_{\text{true}}$ are sufficiently regular for Stein's identity (Sec. C.5) and the integration-by-parts in $\dot{S}(t)$ to apply, i.e. boundary contributions at $\|\mathbf{x}\| \to \infty$ vanish; (iii) the probability flow ODE admits unique solutions on $[0, 1]$ (Lipschitz $\mathbf{s}_\theta$ in $\mathbf{x}$ for each fixed t, integrable in t). **Convention.** The expectations $\mathbb{E}_{p_t}[\,\cdot\,]$ in the derivation below are taken under the true forward noising marginals p_t, not under the model flow marginals $q^{\boldsymbol{\theta}}_t$ produced by integrating the learned probability flow ODE from p_{data}. Eq. 8 is therefore an identity for the plug-in surrogate NLL_{plug}. The actual NLL equals $\text{NLL}_{\text{plug}} + R_{\boldsymbol{\theta}}$, where $R_{\boldsymbol{\theta}}$ reduces at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\text{true}}$ to the terminal prior term $H_\times - S[p_{\text{prior}}]$, zero when $p_{\text{prior}} = p_1$ (§C.4 below). **Scope.** Assumption (i) constrains the noise process, not the model: any parameterization that maps to the score of these Gaussian forward marginals inherits the decomposition. Nonaffine or state-dependent drift, matrix-valued diffusion, Schrödinger bridges, and discrete-state diffusion are outside its scope.

C.1 Log-likelihood from probability flow ODE

For all diffusion processes, there exists a corresponding deterministic process called the probability flow ODE whose trajectories share the same marginal probability densities $\{p_t(\mathbf{x})\}_{t=0}^T$ as the SDE Song et al. (2021b). For the case of $d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + g(t)\mathrm{d}\mathbf{w}_t$, where $g(t) = \sigma^t$, the probability flow ODE is

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] dt.$$

The probability flow ODE has the following form when we approximate the score with the score neural network model $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$:

$$d\mathbf{x}_t = \underbrace{\left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right]}_{=:\tilde{\mathbf{f}}_\theta(\mathbf{x}, t)} dt.$$

With the instantaneous change of variables formula (Chen et al., 2018), we can compute the log-likelihood of $p_0(\mathbf{x})$ using

$$\log p_0(\mathbf{x}_0) = \log p_T(\mathbf{x}_T) + \int_0^T \nabla \cdot \tilde{\mathbf{f}}_\theta(\mathbf{x}_t, t)\mathrm{d}t$$

where $\mathbf{x}(t)$ as a function of t can be obtained by solving the probability flow ODE. Using $T = 1$ and the definition of $\tilde{\mathbf{f}}_{\boldsymbol{\theta}}$ above, the log-likelihood is

$$\log p_0(\mathbf{x}_0) = \log p_1(\mathbf{x}_1) + \int_0^1 \left[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t) - \frac{g(t)^2}{2} \nabla \cdot \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right] dt.$$

C.2 NLL LOWER BOUND

Evaluating the change-of-variables integrand under the true forward marginals p_t defines the plug-in surrogate

$$-\mathrm{NLL}_{\mathrm{plug}}(\boldsymbol{\theta}) \;:=\; \mathbb{E}_{p_1}[\log p_1(\mathbf{x}_1)] + \int_0^1 \mathbb{E}_{p_t}\left[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t) - \frac{g(t)^2}{2} \nabla \cdot \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right] dt. \tag{13}$$

The actual data-average log-likelihood $\mathbb{E}_{p_{\mathrm{data}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_0)] = -\mathrm{NLL}_{\mathrm{actual}}(\boldsymbol{\theta})$ differs by the residual $-R_{\boldsymbol{\theta}}$ defined in §C.4. The two coincide at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$. The first term above is the entropy at $t = 1$, $S_{\mathrm{noise}} = -\mathbb{E}_{p_1}[\log p_1]$. We reexpress the divergence of the score term using $\mathbb{E}_{p_t}[\nabla \cdot \mathbf{s}_{\boldsymbol{\theta}}] = -\mathbb{E}_{p_t}[\mathbf{s}_{\boldsymbol{\theta}} \cdot \mathbf{s}_{\mathrm{true}}]$ (Stein's identity (Liu & Wang, 2016)), which gives:

$$\mathrm{NLL}_{\mathrm{plug}} = S_{\mathrm{noise}} + \int_0^1 \left[-\mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t)] - \frac{g(t)^2}{2} \mathbb{E}_{p_t}[\mathbf{s}_{\boldsymbol{\theta}} \cdot \mathbf{s}_{\mathrm{true}}]\right] dt$$

Using one of the polarization identities, $\mathbf{s}_{\boldsymbol{\theta}} \cdot \mathbf{s}^{\mathrm{true}} = \frac{1}{2}\left(\|\mathbf{s}_{\boldsymbol{\theta}}\|^2 + \|\mathbf{s}_{\mathrm{true}}\|^2 - \|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2\right)$, gives

$$\begin{aligned}\mathrm{NLL}_{\mathrm{plug}} = S_{\mathrm{noise}} &- \int_0^1 \mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t)]\, dt - \frac{1}{2}\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[||\mathbf{s}_{\boldsymbol{\theta}}||^2\right] dt \\ &- \frac{1}{2}\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[||\mathbf{s}_{\mathrm{true}}||^2\right] dt + \frac{1}{2}\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[||\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}||^2\right] dt.\end{aligned}$$

The remaining $\|\mathbf{s}_{\mathrm{true}}\|^2$ integral is fixed by the entropy rate of the true forward process. Applying Eq. 5 to the true marginals p_t gives $\dot{S}_{\mathrm{true}}(t) = \mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t)] + \frac{g(t)^2}{2}\mathbb{E}_{p_t}[\|\mathbf{s}_{\mathrm{true}}\|^2]$, and integrating both sides from 0 to 1 together with the fundamental theorem of calculus $\int_0^1 \dot{S}_{\mathrm{true}}(t)\, dt = S(1) - S(0) = S_{\mathrm{noise}} - S_{\mathrm{data}}$ yields

$$\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[\|\mathbf{s}_{\mathrm{true}}\|^2\right] dt \;=\; \left(S_{\mathrm{noise}} - S_{\mathrm{data}}\right) \;-\; \int_0^1 \mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t)]\, dt. \tag{14}$$

Substituting Eq. 14 into the previous display gives

$$\begin{aligned}\mathrm{NLL}_{\mathrm{plug}} = \frac{S_{\mathrm{data}} + S_{\mathrm{noise}}}{2} &- \frac{1}{2}\int_0^1 \mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}(\mathbf{x}_t, t)]\, dt - \frac{1}{2}\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[||\mathbf{s}_{\boldsymbol{\theta}}||^2\right] dt \\ &+ \frac{1}{2}\int_0^1 \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[||\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}||^2\right] dt.\end{aligned}$$

Using $\dot{S}_{\boldsymbol{\theta}}(t) = \mathbb{E}[\nabla_{\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}, t)] + \frac{g(t)^2}{2}\mathbb{E}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)\|^2\right]$ and the nonnegativity of the squared-difference term, we find that the plug-in negative log-likelihood obeys the lower bound

$$\boxed{\mathrm{NLL}_{\mathrm{plug}} \geq \frac{S_{\mathrm{data}} + S_{\mathrm{noise}}}{2} - \frac{1}{2}\int_0^1 \dot{S}_{\boldsymbol{\theta}}(t) dt} \tag{15}$$

and the bound is tight when $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$, in which case $\mathrm{NLL}_{\mathrm{plug}} = S_{\mathrm{data}}$. The corresponding statement for the probability flow likelihood carries the terminal prior term, $\mathrm{NLL}_{\mathrm{actual}} = S_{\mathrm{data}} + K_1$ (§C.3).

For the driftless diffusion process, $\mathbf{f}(\mathbf{x}_t, t) = 0$ and the system entropy rate is

$$\dot{S}_{\boldsymbol{\theta}}(t) = \dot{S}^i_{\boldsymbol{\theta}}(t) = \frac{g(t)^2}{2} \mathbb{E}_{p_t}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)\|^2\right]$$

so the lower bound, in terms of entropies, is

$$\boxed{\mathrm{NLL}_{\mathrm{plug}} \geq \frac{S_{\mathrm{data}} + S_{\mathrm{noise}}}{2} - \frac{1}{2}\int_0^1 \dot{S}^i_{\boldsymbol{\theta}}(t) dt.} \tag{16}$$

C.3 THREE TERMINAL QUANTITIES

Write p_{prior} for the tractable Gaussian the sampler is initialized from and p_1 for the exact forward marginal at $t = 1$. Three constants must be kept apart:

$$S[p_1], \qquad S[p_{\mathrm{prior}}], \qquad H_\times := -\mathbb{E}_{p_1}[\log p_{\mathrm{prior}}] = S[p_1] + K_1, \qquad K_1 := D_{\mathrm{KL}}(p_1 \| p_{\mathrm{prior}}) \geq 0.$$

The derivation above produces $S[p_1]$ in its boundary term (Stein's identity is applied to $\nabla \log p_1$, not to $\nabla \log p_{\mathrm{prior}}$). The probability flow likelihood produces the cross-entropy of the model's distribution at $t = 1$ against p_{prior}, because the sampler's terminal density is the prior it was handed. The constant our experiments evaluate as S_{noise} (Appendix N) is $S[p_{\mathrm{prior}}]$, the closed-form Gaussian entropy. Three consequences follow.

(i) The forward marginal at $t = 1$ is a convolution, not the prior. For the driftless VE process, $\mathrm{Cov}(p_1) = \mathrm{Cov}(p_0) + v(1)\mathbf{I}$: p_1 is the data smoothed by the terminal kernel, and since that kernel is p_{prior} the gap is a mutual information, $\Delta := S[p_1] - S[p_{\mathrm{prior}}] = I(\mathbf{x}_0; \mathbf{x}_1) \geq 0$. Both $\mathrm{NLL}_{\mathrm{plug}}$ and $\widehat{S}^{\mathrm{UB}}_{\mathrm{plug}}$ are built from the same constant, so substituting $S[p_{\mathrm{prior}}]$ for $S[p_1]$ does not cancel: $\widehat{S}^{\mathrm{UB}}$ is understated by Δ and the slack of Theorem 1 by $\Delta/2$, the unsafe direction for both. Per dimension, Δ is 0.033 bpd for the σ=10 Gaussian toy and 0.005 at σ=30, 0.003 for the σ=10 Uniform data, below 10^{-4} bpd for the image VE schedules ($\sigma_{\max} \geq 50$), and about 3×10^{-5} for the VP and Ising settings. Only the VE toy sweep is materially affected, and the figures we state for it include this allowance.

(ii) At the exact score the likelihood is $S_{\mathrm{data}} + K_1$, not S_{data}. At $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$ the probability flow map transports p_{data} to p_1 exactly and D_{KL} is invariant under that bijection, so $\mathrm{NLL}_{\mathrm{actual}} = S_{\mathrm{data}} + K_1$: the model still pays for starting the reverse integration from p_{prior} rather than from p_1. The exact-score equalities stated for Corollary 1 are therefore matched-prior statements, exact as printed only when $p_{\mathrm{prior}} = p_1$.

(iii) $S[p_{\mathrm{prior}}]$ cannot silently stand in for $S[p_1]$. For a Gaussian prior of variance v, $S[p_{\mathrm{prior}}] - H_\times = \big(d\,v - \mathbb{E}_{p_1}\|\mathbf{x}\|^2\big)/(2v)$, a second moment mismatch of no fixed sign. The safe practice is to carry $H_\times$, which is exactly computable from $\mathbb{E}\|\mathbf{x}_1\|^2 = m(1)^2\,\mathbb{E}\|\mathbf{x}_0\|^2 + d\,v(1)$. On the score-SDE image schedules $H_\times - S[p_{\mathrm{prior}}] \leq 5.4 \times 10^{-5}$ bpd on every row, and it is -1.4×10^{-3} bpd on the Rectified Flow row, whose integrals stop at $1 - 10^{-3}$ (Appendix J).

C.4 PLUG-IN SURROGATE VS. ACTUAL NLL

The chain of identities above replaces the model flow expectations $\mathbb{E}_{q_t^{\boldsymbol{\theta}}}[\,\cdot\,]$ that strictly arise from integrating the change-of-variables formula along $\tilde{\mathbf{f}}_{\boldsymbol{\theta}} = \mathbf{f} - D\mathbf{s}_{\boldsymbol{\theta}}$ from p_{data}, with the true forward expectations $\mathbb{E}_{p_t}[\,\cdot\,]$ required by Stein's identity (since the score appearing in Stein's identity is $\mathbf{s}_{\mathrm{true}} = \nabla \log p_t$). It also replaces the model flow terminal cross-entropy $-\mathbb{E}_{q_1^{\boldsymbol{\theta}}}[\log p_{\mathrm{prior}}]$, which the likelihood pays because the sampler starts from p_{prior}, with $S_{\mathrm{noise}} = S[p_{\mathrm{prior}}]$ (§C.3). The path expectations coincide at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$, where $q_t^{\boldsymbol{\theta}} \equiv p_t$ for all $t \in [0, 1]$, a condition that is sufficient but not necessary (Remark 1). The terminal replacement leaves $H_\times - S[p_{\mathrm{prior}}]$ even then. For trained scores, applying the change-of-variables formula to $\log p_{\boldsymbol{\theta}}(\mathbf{x}_0)$ along the model flow gives

$$\mathrm{NLL}_{\mathrm{actual}}(\boldsymbol{\theta}) = -\mathbb{E}_{q_1^{\boldsymbol{\theta}}}[\log p_{\mathrm{prior}}(\mathbf{x}_1)] + \int_0^1 \mathbb{E}_{q_t^{\boldsymbol{\theta}}}\big[-\nabla\cdot\mathbf{f}(\mathbf{x}_t, t) + D(t)\,\nabla\cdot\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big]\,dt, \tag{17}$$

while the plug-in surrogate evaluates the same integrand under p_t and replaces the boundary term by $S[p_{\mathrm{prior}}]$. Subtracting, and adding and subtracting $H_\times = -\mathbb{E}_{p_1}[\log p_{\mathrm{prior}}]$, gives the formal residual

$$\begin{aligned} R_{\boldsymbol{\theta}} &:= \mathrm{NLL}_{\mathrm{actual}}(\boldsymbol{\theta}) - \mathrm{NLL}_{\mathrm{plug}}(\boldsymbol{\theta}) \\ &= \big(\mathbb{E}_{p_1} - \mathbb{E}_{q_1^{\boldsymbol{\theta}}}\big)[\log p_{\mathrm{prior}}] + \big(H_\times - S[p_{\mathrm{prior}}]\big) \\ &\quad + \int_0^1 \big(\mathbb{E}_{p_t} - \mathbb{E}_{q_t^{\boldsymbol{\theta}}}\big)\big[\nabla\cdot\mathbf{f}(\mathbf{x}_t, t) - D(t)\,\nabla\cdot\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big]\,dt, \end{aligned} \tag{18}$$

where $\mathrm{NLL}_{\mathrm{actual}} = -\mathbb{E}_{p_{\mathrm{data}}}[\log p_{\boldsymbol{\theta}}]$ is the probability flow NLL reported for the Ising and image experiments (the toy and Swiss roll experiments report $\mathrm{NLL}_{\mathrm{plug}}$ itself). Every $\mathbb{E}_{p_t} - \mathbb{E}_{q_t^{\boldsymbol{\theta}}}$ difference vanishes in particular when $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$, since then $q_t^{\boldsymbol{\theta}} \equiv p_t$. Hence, in the exact-score limit, $R_{\boldsymbol{\theta}}$ equals

the terminal prior term $H_\times - S[p_{\text{prior}}]$, a second moment mismatch that vanishes when $p_{\text{prior}} = p_1$ (§C.3(iii)). A vanishing path term does not imply an exact score (Remark 1). Every reported $R_{\boldsymbol{\theta}}$ is the measured difference of the two likelihoods and therefore includes the terminal prior term.

Measurement. We do not bound $R_{\boldsymbol{\theta}}$ a priori. The hypotheses under which one might (a bounded $\log p_1$ and a bounded divergence integrand, transferred to a total variation distance controlled by the score error) fail for the Gaussian priors used here, whose log density is unbounded below. We know of no argument that transfers a score matching error to probability flow total variation without further assumptions on the learned field. We instead measure $R_{\boldsymbol{\theta}}$ as the difference of two independently evaluated likelihoods: $\text{NLL}_{\text{actual}}$ by integrating the probability flow ODE along the model flow from data samples, and NLL_{plug} by quadrature of the same divergence integrand under the closed-form forward marginals p_t. Across 411 checkpoints (215 Gaussian and Uniform toys, 110 Swiss roll, 60 Ising, 26 image models) the residual has no universal sign or scale: it lies in $[+0.09, +0.15]$ bpd on the CIFAR-10 checkpoints, in $[-0.03, -0.007]$ bpd on the likelihood-weighted ImageNet-32 checkpoints, in $[-0.92, -0.43]$ bpd on the four 256×256 VE models, and in $[-0.058, -0.020]$ bpd on the Ising sweep. On the Swiss roll its sign varies with the noise schedule and the network width. Where a figure or table reports $\text{NLL}_{\text{actual}}$ (Figures 2b and 4, Tables 1b and 2), the entropy upper bound of Eq. 11 evaluated from it equals the plug-in bound plus $2R_{\boldsymbol{\theta}}$, and the lower bound of Corollary 1 transported to $\text{NLL}_{\text{actual}}$ carries $+R_{\boldsymbol{\theta}}$.

Certification. Transported to the probability flow likelihood, Corollary 1 and Eq. 11 hold if and only if the margin $m_{\boldsymbol{\theta}} := \widehat{S}^{\text{UB}}_{\text{actual}} - S_{\text{data}} = I_{\text{diff}} + 2R_{\boldsymbol{\theta}} = 2D_{\text{KL}}(p_{\text{data}}\|p_{\boldsymbol{\theta}}) + A$ is nonnegative, where $A := \Delta S_{\boldsymbol{\theta}} - \Delta S_{\text{true}}$ is the score norm gap. Deciding this needs S_{data}. Two conditions certify it, that is, guarantee it without a reference entropy, from the checkpoint alone. If $R_{\boldsymbol{\theta}} \geq 0$, then $\widehat{S}^{\text{UB}}_{\text{actual}} \geq \widehat{S}^{\text{UB}}_{\text{plug}} \geq S_{\text{data}}$ by Theorem 1. If $\widehat{S}^{\text{UB}}_{\text{actual}} \geq \text{NLL}_{\text{actual}}$, a comparison of two measured quantities that equals $D_{\text{KL}} + A \geq 0$, then $m_{\boldsymbol{\theta}} \geq D_{\text{KL}} + A \geq 0$. Either condition is sufficient and neither can be satisfied when the margin is negative. When both fail the procedure abstains rather than refutes, and the gap between the test and the claim is exactly the unknown D_{KL}. The condition is a statement about the theorem's slack rather than about model quality: with $\langle \mathbf{u}, \mathbf{v} \rangle := \int_0^1 D\, \mathbb{E}_{p_t}[\mathbf{u}\cdot\mathbf{v}]\, dt$ one has $R_{\boldsymbol{\theta}} = \langle \mathbf{s}_{\text{true}}, \mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\text{true}} \rangle + D_{\text{KL}}$, first order in the score error, while I_{diff} is second order. Hence $m_{\boldsymbol{\theta}}$ must turn negative for a sufficiently accurate score whose error has a negative component along $\mathbf{s}_{\text{true}}$ (Remark 2 gives the aligned case).

Remark 1 ($R_{\boldsymbol{\theta}} = 0$ does not certify an exact score)**.** *Take the stationary VP process $\mathbf{f}(\mathbf{x}, t) = -\mathbf{x}$, $D \equiv 1$, initialized at its stationary law, so that $p_t = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for every t and $\mathbf{s}_{\text{true}} = -\mathbf{x}$. Let $\mathbf{s}_{\boldsymbol{\theta}} = -\mathbf{x} + \boldsymbol{\Omega}\mathbf{x}$ with $\boldsymbol{\Omega}$ a constant skew-symmetric matrix. The induced probability flow field $\mathbf{f} - D\mathbf{s}_{\boldsymbol{\theta}} = -\boldsymbol{\Omega}\mathbf{x}$ is a rigid rotation, which preserves the standard Gaussian. Hence $q^{\boldsymbol{\theta}}_t = p_t$ for all t, the generative ODE is exact, and $R_{\boldsymbol{\theta}} = 0$. Yet the score-error integral $\int_0^1 D\, \mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\text{true}}\|^2 dt = \|\boldsymbol{\Omega}\|_F^2$ and the excess entropy rate integral $\int_0^1 (\dot{S}_{\boldsymbol{\theta}} - \dot{S}_{\text{true}})\, dt = \|\boldsymbol{\Omega}\|_F^2$ are arbitrarily large. A vanishing residual is evidence about the marginals of the model flow, not about the score.*

Remark 2 (The plug-in surrogate is not a coding cost)**.** NLL_{plug} *is not the negative log-likelihood of any density, and it is unbounded below. For the driftless VE process with $d = 1$, $p_{\text{data}} = \mathcal{N}(0, 1)$, $\text{std}(t)^2 = t$ and $D \equiv \frac{1}{2}$, take the aligned rescaling $\mathbf{s}_{\boldsymbol{\theta}} = \alpha\, \mathbf{s}_{\text{true}}$. Stein's identity gives $\text{NLL}_{\text{plug}} = S_{\text{noise}} - \alpha \int_0^1 D\, \mathbb{E}_{p_t}\|\mathbf{s}_{\text{true}}\|^2 dt$, and for a driftless process $S_{\text{noise}} - S_{\text{data}} = \int_0^1 D\, \mathbb{E}_{p_t}\|\mathbf{s}_{\text{true}}\|^2 dt$, so*

$$\text{NLL}_{\text{plug}} \;=\; S_{\text{data}} + (1-\alpha)\int_0^1 D\, \mathbb{E}_{p_t}\|\mathbf{s}_{\text{true}}\|^2 dt \;\xrightarrow[\alpha\to\infty]{}\; -\infty,$$

while $\text{NLL}_{\text{actual}} \geq S_{\text{data}}$ for every $\boldsymbol{\theta}$, being a cross-entropy. Here $S_{\text{data}} = \frac{1}{2}\log(2\pi e) = 1.419$ nats and $\alpha = 1000$ gives $\text{NLL}_{\text{plug}} = -344.8$ nats. Corollary 1 is therefore a limit on the surrogate, and the only bridge from it to a reported likelihood is the measured $R_{\boldsymbol{\theta}}$. A model whose $\text{NLL}_{\text{actual}}$ appears to violate the lower bound is exhibiting a large residual, not a broken inequality.

C.5 Stein's identity

In (Liu & Wang, 2016), Stein's identity states that for sufficiently regular ϕ, we have

$$\mathbb{E}_{x\sim p}\left[\mathcal{A}_p\boldsymbol{\phi}(x)\right] = 0, \quad \text{where} \quad \mathcal{A}_p\boldsymbol{\phi}(x) = \boldsymbol{\phi}(x)\nabla_x \log p(x)^\top + \nabla_x \boldsymbol{\phi}(x), \tag{19}$$

where $\mathcal{A}_p$ is called the Stein operator, which acts on function ϕ and yields a zero mean function $\mathcal{A}_p\phi(x)$ under $x \sim p$. Expanding this identity coordinatewise, it is exactly the statement:

$$\mathbb{E}_p[\nabla \cdot \phi] = -\mathbb{E}_p[\phi \cdot s].$$

With the true score $\mathbf{s}_{\text{true}} = \nabla \log p(\mathbf{x})$, we have:

$$\mathbb{E}_p[\nabla \cdot \mathbf{s}_{\text{true}}] = -\mathbb{E}_p\left[\|\mathbf{s}_{\text{true}}\|^2\right].$$

For an approximate score $\mathbf{s}_{\boldsymbol{\theta}}$:

$$\mathbb{E}_p\left[\nabla \cdot \mathbf{s}_{\boldsymbol{\theta}}\right] = -\mathbb{E}_p\left[\mathbf{s}_{\boldsymbol{\theta}} \cdot \mathbf{s}_{\text{true}}\right]$$

which equals $-\mathbb{E}_p\left[\|\mathbf{s}_{\boldsymbol{\theta}}\|^2\right]$ only if $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\text{true}}$.

D MAXWELL'S DEMON IN THE CONTROLLED-FORWARD PROCESS

Song et al. (2021b) use the notation of Haussmann–Pardoux / Anderson (Haussmann & Pardoux, 1986; Anderson, 1982)

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \quad \text{(Forward)} \tag{20}$$

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}, t) - g(t)^2\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)]\mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}}_t \quad \text{(Reverse)} \tag{21}$$

where $\overline{\mathbf{w}}_t$ is the standard Wiener process when time is run backwards. **Note**, the reverse SDE above is usually integrated from T down to 0, making dt negative.

D.1 REVERSE PROCESS

We have chosen the forward SDE to be

$$d\mathbf{x}_t = \sigma^t d\mathbf{w}_t, \quad t \in [0, 1].$$

To sample from our time-dependent score-based model $s_{\boldsymbol{\theta}}(\mathbf{x}, t)$, we first draw a sample from the prior distribution $p_1 \approx \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \frac{\sigma^2-1}{2\ln\sigma}\mathbf{I}\right)$ (consistent with $v(1) = (\sigma^2 - 1)/(2\ln\sigma)$ used in App. N.2.2), and then solve the reverse-time SDE with numerical methods. In particular, using our time-dependent score-based model, the reverse-time SDE can be approximated by

$$d\mathbf{x}_t = -\sigma^{2t} s_{\boldsymbol{\theta}}(\mathbf{x}, t)dt + \sigma^t d\overline{\mathbf{w}}_t$$

Next, one can use numerical methods to solve for the reverse-time SDE, such as the Euler–Maruyama approach. It is based on a simple discretization to the SDE, replacing dt with $\Delta t > 0$ and $d\mathbf{w}$ with $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, g^2(t)\Delta t\mathbf{I}\right)$. When applied to our reverse-time SDE, we can obtain the following iteration rule

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t + \sigma^{2t} s_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right)\Delta t + \sigma^t\sqrt{\Delta t}\mathbf{z}_t$$

where $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

D.2 CONTROLLED-FORWARD PROCESS

Time always runs forward in the real world: one can achieve a physical realization of the generative process by defining a clock

$$\tau := T - t, \quad 0 \le \tau \le T,$$

such that integrating forward in τ is the same as integrating backward in t. We plug in $t = T - \tau$, $dt = -d\tau$, and $d\bar{\mathbf{w}}_t = d\mathbf{w}_\tau$ to reparameterize the reverse SDE (the second equation above in App. D) as a controlled-forward process

$$\mathrm{d}\bar{\mathbf{x}}_\tau = [-\mathbf{f}(\bar{\mathbf{x}}_\tau, T - \tau) + g(T - \tau)^2\mathbf{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau, T - \tau)]\mathrm{d}\tau + g(T - \tau)\mathrm{d}\mathbf{w}_\tau \tag{22}$$

where $\bar{\mathbf{x}}_\tau := \mathbf{x}_t$.

D.3 Entropy rates of the controlled-forward process

For the controlled-forward formulation, the drift becomes $\tilde{\mathbf{f}}(\bar{\mathbf{x}}_\tau, \tau) = g(T-\tau)^2 \mathbf{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau, T-\tau)$ (Eq. 22). The diffusion coefficient and score are evaluated at forward time $T-\tau$ because the noise schedule $g(t)$ is not symmetric under $t \mapsto T-t$ for VP/VE. Substituting this drift into the general intrinsic and exchange entropy rate expressions of Appendix B (which require no Stein's identity step and so apply for any $\mathbf{s}_{\boldsymbol{\theta}}$, exact or learned) yields

$$\dot{S}^i_{\boldsymbol{\theta}}(\tau) = \frac{g(T-\tau)^2}{2}\,\mathbb{E}\big[\|\mathbf{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_\tau, T-\tau)\|^2\big], \quad \dot{S}^e_{\boldsymbol{\theta}}(\tau) = -2\dot{S}^i_{\boldsymbol{\theta}}(\tau), \quad \dot{S}_{\boldsymbol{\theta}}(\tau) = -\dot{S}^i_{\boldsymbol{\theta}}(\tau),$$

where the system rate is taken as the algebraic sum $\dot{S}_{\boldsymbol{\theta}} = \dot{S}^i_{\boldsymbol{\theta}} + \dot{S}^e_{\boldsymbol{\theta}}$. The equivalent divergence form expression $\dot{S}_{\boldsymbol{\theta}} = \mathbb{E}[\nabla \cdot \tilde{\mathbf{f}}_{\boldsymbol{\theta}}] + D\,\mathbb{E}[\|\mathbf{s}_{\boldsymbol{\theta}}\|^2]$ (Eq. 5) requires Stein's identity to bring the divergence of $\mathbf{s}_{\boldsymbol{\theta}}$ to a squared-norm form, and so coincides with the sum form rate above only when $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\text{true}}$. We use the sum form throughout here to avoid that conditional step. Thus, in the controlled-forward process the system entropy rate is exactly the negative of the intrinsic entropy production rate. Note that this $\dot{S}_{\boldsymbol{\theta}}(\tau)$ is the rate in the controlled generation clock: it is signed differently from the rate $\dot{S}_{\boldsymbol{\theta}}(t)$ in the forward noising clock that enters Eq. 8. On the driftless SDE the two are related by $\dot{S}_{\boldsymbol{\theta}}(\tau) = -\dot{S}_{\boldsymbol{\theta}}(T-\tau)$, consistent with the Time conventions paragraph in §2.1.

E General continuous-time diffusion processes

Song et al. (2021b) showed that score-based generative models can be formulated in terms of a general Itô SDE of the form

$$d\mathbf{x}_t = \mathbf{f}\left(\mathbf{x}_t, t\right) dt + g(t) d\mathbf{w}_t$$

where $\mathbf{f}\left(\mathbf{x}_t, t\right)$ is the drift, $g(t)$ the diffusion coefficient, and $\mathbf{w}_t$ a standard Wiener process. Two canonical instantiations of this framework correspond to the variance exploding (VE) and variance preserving (VP) processes.

E.1 Variance-exploding (VE) SDE

The VE process is defined by

$$d\mathbf{x}_t = \sqrt{\frac{d}{dt}\sigma^2(t)}\, d\mathbf{w}_t$$

with $\sigma^2(t)$ a nondecreasing variance schedule. Here the drift vanishes, $\mathbf{f}\left(\mathbf{x}_t, t\right) = 0$, while the diffusion coefficient is chosen so that the marginal variance of $\mathbf{x}_t$ increases monotonically in t. As $t \to T$, the variance diverges (hence "exploding"), and the distribution approaches a Gaussian prior. This setting is natural when starting from bounded data distributions, since the forward process progressively washes out structure by injecting unbounded noise.

E.2 Variance-preserving (VP) SDE

In contrast, the VP process includes both drift and diffusion terms:

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

where $\beta(t)$ is a positive noise rate schedule. The drift pulls $\mathbf{x}_t$ toward the origin at a rate proportional to $\beta(t)$, while the diffusion injects noise of matching strength. This balance ensures that the overall variance of the process remains bounded (and can be normalized to unity) for all t. Thus, the forward diffusion maps data smoothly into an isotropic Gaussian prior without the variance diverging.

Both SDEs fit seamlessly into the score-based generative modeling framework. In each case, the reverse-time dynamics introduce an additional score-dependent drift term,

$$d\mathbf{x}_\tau = \left[-\mathbf{f}\left(\mathbf{x}_\tau, T-\tau\right) + g^2(T-\tau)\nabla_{\mathbf{x}_\tau} \log p_{T-\tau}\left(\mathbf{x}_\tau\right)\right] d\tau + g(T-\tau) d\mathbf{w}_\tau$$

where the score $\nabla_{\mathbf{x}_\tau} \log p_{T-\tau}\left(\mathbf{x}_\tau\right)$ (forward-time marginal evaluated at the controlled-forward state) is approximated by a neural network. The VE and VP choices thus represent two distinct, yet

complementary, continuous-time noise injection schemes, both of which reduce to the driftless case when $f = 0$ and variance is allowed to grow freely. They provide the practical foundation for most modern diffusion models, differing primarily in how variance is managed over time and, correspondingly, in their tradeoffs between sample quality and likelihood.

In addition to the VE process considered in the main text, we evaluate the VP process used in denoising diffusion probabilistic models. The VP process is governed by the forward SDE

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

where $\beta(t)$ is the variance schedule and w_t is standard Brownian motion. This process interpolates between the data distribution at $t = 0$ and an isotropic Gaussian prior at $t = 1$ while preserving variance at each time step. The reverse process is parameterized by the learned score network, and the associated entropy production integrals are estimated analogously to the VE case.

E.3 Constructing the VP schedule from σ

To specify $\beta(t)$, we set a desired terminal noise scale σ, which encodes how much Gaussian noise is injected by the end of the forward process.

E.3.1 Integrated noise budget

The mean-scaling factor of the VP process is

$$\alpha(t) = \exp\left(-\frac{1}{2}\int_0^t \beta(u)du\right)$$

so at terminal time $t = 1$,

$$\alpha(1)^2 = \exp(-B), \quad B := \int_0^1 \beta(u)du$$

The variance contributed by the noise term is

$$\sigma(1)^2 = 1 - \alpha(1)^2.$$

Requiring $\sigma(1)^2 = \sigma^2/(1+\sigma^2)$ yields the condition

$$B = \log\left(1+\sigma^2\right)$$

Thus the entire VP schedule is determined by the integrated noise budget B.

E.3.2 Linear schedule construction

A common choice is to make $\beta(t)$ linear in t:

$$\beta(t) = \beta_{\min} + t\,(\beta_{\max} - \beta_{\min})$$

The constants β_{min} and β_{max} are set so that the integral matches the budget:

$$\int_0^1 \beta(t)dt = \frac{1}{2}\,(\beta_{\min} + \beta_{\max}) = B$$

Introducing a ratio parameter $r \in (0, 1)$, we define

$$\beta_{\min} = rB, \quad \beta_{\max} = (2-r)B,$$

which ensures the correct average while allowing flexibility in the temporal profile of noise injection. Smaller r concentrates noise near $t = 1$, while larger r distributes noise more evenly across time.

In the VP formulation, B is the logarithmic noise budget: it quantifies the total exponential damping of the signal. In the VE formulation, the corresponding budget is the variance scale σ^2. The two are linked by $\sigma^2 = e^B - 1$. Hence, the VP schedule can be constructed from a single intuitive parameter σ, which specifies the effective strength of the forward noise process, while B serves as its natural exponential coordinate.

F STANDARD GAUSSIAN DATA

For validation we include experiments where the data distribution p_{data} is a standard Gaussian. In this case, the score function is exactly linear:

$$\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}) = -\mathbf{x},$$

which can be fit by a single-layer neural network with linear weights. This setting provides a ground truth baseline where the score is known analytically, allowing us to verify the tightness of the lower bound and the accuracy of our numerical estimators.

Consider the case in which the data is normally distributed $\mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a driftless forward process with variance increment $v(t) = \int_0^t g(u)^2 du$, with $g^2(t) = \sigma^{2t}$ and $v(t) = \frac{\sigma^{2t}-1}{2\ln\sigma}$. Then $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + v(t)\boldsymbol{I})$ and

$$\mathbf{s}_{\text{true}}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -(\boldsymbol{\Sigma} + v(t)\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

which is exactly linear in $\mathbf{x}$ for every t. A tiny network (even a single linear layer conditioned on t) can represent this perfectly, so training can drive $\mathbf{s}_{\boldsymbol{\theta}} \to \mathbf{s}_{\text{true}}$.

Proof. Let

$$p_t(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C}(t)), \quad \mathbf{C}(t) = \boldsymbol{\Sigma} + v(t)\mathbf{I}_d,$$

so $\mathbf{C}(t)$ is symmetric positive-definite. The multivariate Gaussian probability density function (pdf) is

$$p_t(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\det(\mathbf{C})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

Take logs:

$$\log p_t(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}\log\det(2\pi\mathbf{C}).$$

Only the quadratic term depends on $\mathbf{x}$. With $h = (x-\mu)_j A_{jk}(x-\mu)_k$,

$$\frac{\partial h}{\partial x_i} = A_{ij}(x-\mu)_j + A_{ji}(x-\mu)_j = \left[\left(A + A^\top\right)(x-\mu)\right]_i$$

and we have

$$\nabla_{\mathbf{x}}\left[(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right] = \left(\mathbf{A} + \mathbf{A}^\top\right)(\mathbf{x} - \boldsymbol{\mu}) = 2\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) \quad \left(\mathbf{A} = \mathbf{A}^\top\right),$$

with $\mathbf{A} = \mathbf{C}^{-1}$, we get

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -\frac{1}{2} \cdot 2\mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Therefore the true score is

$$\mathbf{s}_{\text{true}}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -(\boldsymbol{\Sigma} + v(t)\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

exactly as claimed.

G EXACT SCORE OF THE UNIFORM AND NORMAL DISTRIBUTIONS

Pixels are independent under Uniform[0,1] and the Gaussian noise factorizes across coordinates, so we consider the 1D case and adopt scalar notation throughout. For the driftless diffusion,

$$dx_t = g(t)dW_t, \quad v(t) = \int_0^t g(u)^2 du, \quad s = \sqrt{v(t)}$$

Conditioned on x_0, we have

$$x_t \mid x_0 \sim \mathcal{N}\left(x_0, s^2\right)$$

If the data is Uniform on $[0, 1]$ (density $p_0(u) = \mathbf{1}_{[0,1]}(u)$), the marginal at time t is the convolution

$$p_t(x) = \int_0^1 \phi_s(x - u)du$$

where

$$\phi_s(z) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{z^2}{2s^2}\right)$$

is the $\mathcal{N}\left(0, s^2\right)$ pdf. With a change of variable $z = (x-u)/s$ and $du = -sdz$, we have

$$p_t(x) = \int_{(x-1)/s}^{x/s} \phi(z)dz = \Phi\left(\frac{x}{s}\right) - \Phi\left(\frac{x-1}{s}\right)$$

with ϕ and Φ the standard normal pdf and cumulative distribution function. Differentiate with respect to x :

$$\partial_x p_t(x) = \frac{1}{s}\left[\phi\left(\frac{x}{s}\right) - \phi\left(\frac{x-1}{s}\right)\right].$$

The score is the gradient of the log-density,

$$s(x,t) = \partial_x \log p_t(x) = \frac{\partial_x p_t(x)}{p_t(x)} = \frac{\frac{1}{s}\left[\phi\left(\frac{x}{s}\right) - \phi\left(\frac{x-1}{s}\right)\right]}{\Phi\left(\frac{x}{s}\right) - \Phi\left(\frac{x-1}{s}\right)}.$$

Part II Numerical Methods

H TOY DISTRIBUTION EXPERIMENTAL DETAILS

This appendix documents the synthetic experiments used only to support Figure 1 and the toy-data rows in Figure 3. The goal of these experiments is not to introduce additional benchmarks, but to provide controlled settings where the NLL, the entropy terms, and the slack in the lower bound can be checked against known distributions.

Distributions. For the Gaussian validation, data are drawn from a factorized standard normal distribution $p_0 = \mathcal{N}(0, I_d)$, so $S_{\text{data}} = \frac{d}{2}\log(2\pi e)$. For the Uniform validation, each coordinate is sampled independently from $\text{Unif}[0,1]$, so $S_{\text{data}} = 0$ nats in continuous coordinates. The synthetic image Uniform runs use the same independent-coordinate law, interpreted as independent grayscale pixel intensities in $[0,1]$. For the Swiss roll accuracy–dissipation and VP/VE comparison studies, the data generation procedure is the one in Appendix K.1.

Noising processes and ground truth scores. The Gaussian and Uniform validation experiments use the driftless variance-exploding (VE) process

$$d\mathbf{x}_t = g(t)\, d\mathbf{w}_t, \qquad v(t) = \int_0^t g(u)^2\, du, \tag{23}$$

with terminal noise scale swept over $\sigma \in \{10, 15, 20, 25, 30\}$. The Gaussian score is analytic after convolution with the Gaussian kernel. The Uniform score is the product of the one-dimensional truncated-Gaussian convolution scores derived in Appendix G. This gives a coordinate-wise exact score for the independent-pixel Uniform distribution. These exact scores are used only for validation and for computing the score-error term in the decomposition. All NLL and entropy rate quantities reported for learned models are computed from the trained score network itself.

Training and repeated runs. For each noise scale in Figure 1, we train ten independent models for the Gaussian and Uniform distributions using independent random seeds. Swiss roll points in the same figure are obtained by sweeping the model capacity described in Appendix K.3. All toy models are trained with denoising score matching against samples from the closed-form perturbed marginal $p_t(\mathbf{x}_t \mid \mathbf{x}_0)$, using the same time interval and probability flow evaluation convention as in the main text.

NLL and entropy rate evaluation. For each trained toy model, the reported NLL is computed by integrating the probability flow ODE

$$\frac{d\mathbf{x}_t}{dt} = f(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \tag{24}$$

with the corresponding instantaneous change of variables. The prior entropy is the analytic entropy of the terminal Gaussian p_T. For VP this is $\frac{d}{2}\log(2\pi e)$ and for VE it is $\frac{d}{2}\log(2\pi e \sigma_{\max}^2)$. Entropy rates are estimated by drawing $\mathbf{x}_t$ from the exact perturbed marginal and evaluating

$$\dot{S}_{\boldsymbol{\theta}}(t) = \mathbb{E}[\nabla \cdot f(\mathbf{x}_t, t)] + \frac{g(t)^2}{2}\, \mathbb{E}\big[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|^2\big], \tag{25}$$

with the intrinsic and exchange rates computed from the same batches. In the VE case $f = 0$, which yields the diagnostic identity $\dot{S}^e_{\boldsymbol{\theta}} = -2\dot{S}^i_{\boldsymbol{\theta}}$ and $\dot{S}_{\boldsymbol{\theta}} = -\dot{S}^i_{\boldsymbol{\theta}}$. The right panel of Figure 1 shows this check for the best Uniform VE run.

Slack diagnostic. For the Gaussian and Uniform cases, the exact data score is known. We therefore compute the score-error term of the decomposition as the integrated squared score error, weighted by $g(t)^2/2$ under the model's perturbed marginal. This is the score-error quantity plotted against NLL in the main text. It explains why the lower bound tightens as the learned score approaches the exact score.

I ISING EXPERIMENTAL DETAILS

This appendix supports only the Ising results in Figure 2 and Table 1. The main text Ising results use exact finite lattice sampling and exact finite lattice entropy targets.

Finite 2D Ising distribution. We consider an $L \times L$ square lattice with periodic boundary conditions ($L = 5$), spin variables $\sigma_{ij} \in \{-1, +1\}$, nearest neighbor coupling $J = 1$, and zero external field $h = 0$ throughout. The energy is

$$H(\boldsymbol{\sigma}) = -\sum_{i=1}^{L}\sum_{j=1}^{L} \sigma_{ij}\big(\sigma_{i+1,j} + \sigma_{i,j+1}\big), \tag{26}$$

where indices are taken modulo L. At temperature T, the discrete distribution is

$$p_T(\boldsymbol{\sigma}) = \frac{\exp[-H(\boldsymbol{\sigma})/T]}{Z_L(T)}, \qquad Z_L(T) = \sum_{\boldsymbol{\sigma}\in\{-1,+1\}^{L^2}} \exp[-H(\boldsymbol{\sigma})/T], \tag{27}$$

the partition function that normalizes p_T over the full 2^{L^2}-configuration state space (both sectors of the global spin flip).

Dequantization (fuzzing) and differential entropy. The diffusion model is trained in continuous space by adding independent uniform dequantization noise to each spin (we use $\boldsymbol{\eta}$ for dequantization noise to avoid clashing with the standard Gaussian diffusion noise $\boldsymbol{\epsilon}$ used elsewhere):

$$\mathbf{x}_0 = \boldsymbol{\sigma} + \boldsymbol{\eta}, \qquad \eta_k \sim \mathrm{Unif}[-b, b], \qquad b = 0.1. \tag{28}$$

The corresponding differential entropy target is the discrete Ising entropy plus the dequantization entropy,

$$S_{\mathrm{diff}}(T, L, b) = S_{\mathrm{disc}}(T, L) + L^2 \log(2b). \tag{29}$$

All values reported in the main text are converted to bits per dimension by dividing by $L^2 \log 2$.

Exact entropy target. For the zero-field finite lattice we evaluate $\log Z_L(T)$ in closed form (Onsager, 1944; Ferdinand & Fisher, 1969) via the spinor diagonalization of Kaufman (Kaufman, 1949), simplified by Kastening (2001). Defining the eigenvalue parameters γ_k ($k = 0, 1, \ldots, 2L-1$) implicitly by Kastening's Eq. (37), specialized to the isotropic case $J_a{=}J_b{=}1$ via the duality $\sinh(2\bar{a})\sinh(2a){=}1$ with $a{=}1/T$,

$$\cosh\gamma_k \;=\; \frac{\cosh^2(2/T)}{\sinh(2/T)} - \cos\frac{\pi k}{L}, \tag{30}$$

together with the signed-γ_0 convention of Kastening's Eq. (38), $\gamma_0 = 2(\bar{a} - a)$. Defining the four spectral products $Y_i^{c/s} := \prod_{k=1}^{L} 2\,(\cosh \text{ or } \sinh)(L\gamma_{2k-i}/2)$ for $i \in \{1, 2\}$, Kastening's Eq. (48) reads

$$Z_L(T) \;=\; \tfrac{1}{2}\,[2\sinh(2/T)]^{L^2/2}\,\big[\,Y_1^c \;+\; Y_1^s \;+\; Y_2^c \;-\; Y_2^s\,\big]. \tag{31}$$

The signed γ_0 is the only γ_k whose sign flips at the critical coupling: $\gamma_0 > 0$ for $T > T_c$ and $\gamma_0 < 0$ for $T < T_c$ (with $T_c{=}2/\ln(1{+}\sqrt{2}) \approx 2.269$). Honoring this sign is essential: without it, $Z_L(T)$ silently reduces to the single-sector partition function below T_c and undercounts the twofold $\mathbb{Z}_2$ ground state degeneracy. The discrete entropy follows by standard differentiation,

$$S_{\mathrm{disc}}(T, L) = \log Z_L(T) + T\,\frac{\partial}{\partial T}\log Z_L(T), \tag{32}$$

evaluated by central difference with step 10^{-4} and plotted as the analytic black curve in Figure 2. Because Eq. 31 sums over all 2^{L^2} states with the correct γ_0 sign, $S_{\mathrm{disc}}(T, L) \to \log 2$ as $T \to 0$ from the twofold ground state degeneracy. At $T = 3$ and $b = 0.1$, Eq. 29 gives $S_{\mathrm{diff}} = -1.586$ bpd, the exact row in Table 1. We verified Eq. 31 against direct enumeration of all 2^{L^2} configurations for $L = 5$ at every temperature in the sweep (the temperatures listed in Table 5); agreement is to floating-point precision.

Training pools. For each temperature in the sweep we draw a pool of $N_{\text{pool}}{=}10^5$ exact independent and identically distributed (i.i.d.) Boltzmann samples from p_T by autoregressive transfer matrix decomposition along lattice rows. Indexing each lattice row by a row state $\sigma \in \{-1,+1\}^L$, define the row-to-row transfer matrix $\mathbf{T} \in \mathbb{R}^{2^L \times 2^L}$ explicitly by

$$\mathbf{T}[\sigma,\sigma'] \;=\; \exp\Big(-\tfrac{1}{T}\big[\tfrac{1}{2}H_{\text{H}}(\sigma) + H_{\text{V}}(\sigma,\sigma') + \tfrac{1}{2}H_{\text{H}}(\sigma')\big]\Big), \tag{33}$$

where $H_{\text{H}}(\sigma) = -\sum_j \sigma_j\sigma_{j+1 \bmod L}$ is the intra-row coupling and $H_{\text{V}}(\sigma,\sigma') = -\sum_j \sigma_j\sigma'_j$ is the row-pair vertical coupling. The half-weighting on H_{H} ensures each horizontal bond is counted exactly once when $\mathbf{T}$ is iterated along the chain. Precomputing the powers $\mathbf{T}^0, \mathbf{T}^1, \ldots, \mathbf{T}^L$ once (each $32{\times}32$ for $L{=}5$), each sample is generated in $O(L \cdot 2^L)$ work as follows. The first row σ_1 is drawn from the row marginal

$$p(\sigma_1) \;\propto\; (\mathbf{T}^L)[\sigma_1,\sigma_1], \tag{34}$$

the diagonal of $\mathbf{T}^L$. Then for $k = 2, \ldots, L$, row σ_k is drawn from the conditional

$$p(\sigma_k \mid \sigma_{k-1},\sigma_1) \;\propto\; \mathbf{T}[\sigma_{k-1},\sigma_k] \cdot \mathbf{T}^{L-k+1}[\sigma_k,\sigma_1], \tag{35}$$

where the factor $\mathbf{T}^{L-k+1}[\sigma_k,\sigma_1]$ accounts for the remaining $L{-}k{+}1$ ring transitions $\sigma_k \to \sigma_{k+1} \to \cdots \to \sigma_L \to \sigma_1$ that close back to the anchor. For $k = L$ this reduces to $\mathbf{T}^1[\sigma_L,\sigma_1] = \mathbf{T}[\sigma_L,\sigma_1]$, recovering the closing edge. This samples p_T directly with no Markov chain mixing time, no autocorrelation, and no Monte Carlo bias. Independent uniform noise $\boldsymbol{\eta} \in [-b,b]^{L^2}$ with $b = 0.1$ is then added to produce the dequantized continuous samples used for training. Note that $\mathbf{T}$ is built explicitly only here, for the sampler; the entropy target above is computed from the closed form Eq. 31 without numerically forming $\mathbf{T}$ or its powers.

Sector balance and absence of a $\log 2$ *correction.* At $h = 0$ the full row-chain law is invariant under the global spin flip $\sigma \mapsto -\sigma$, so each configuration and its flipped partner have equal probability. With $L^2 = 25$ odd, the positive- and negative-magnetization sectors each carry probability $1/2$, reducing to the two pure ground states as $T \to 0$. The empirical pool therefore covers both sectors in proportion to their (equal) Boltzmann weights, matching the convention of $S_{\text{disc}}(T,L)$ in Eq. 32. The comparison between the model-based UB and the analytic curve in Figure 2b therefore requires no $\log 2$ correction at any temperature.

Score model and cosine VP process. The Ising score network is a periodic-convolutional network on the $5{\times}5$ lattice: an input $3{\times}3$ convolution on the spin field stacked with a constant channel carrying t, four residual $3{\times}3$ convolution blocks with 256 channels and circular padding, so that the receptive field wraps around the periodic boundary, and a $1{\times}1$ output convolution. It has 2.37M parameters in total. The forward process is the cosine VP process

$$\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + g(t)\boldsymbol{\epsilon}, \qquad \alpha(t) = \cos\frac{\pi t}{2}, \qquad g(t) = \sin\frac{\pi t}{2}, \tag{36}$$

with $t \in [0.0025, 0.995]$ in numerical evaluations. For this process the SDE diffusion coefficient is $D(t) = \frac{\pi}{2}\tan\frac{\pi t}{2}$, whose integral over $[0,1]$ diverges, so the Gaussian endpoint is a singular limit rather than an attained state. Every quantity reported for this system is computed on this truncated integration interval (the window), and the terminal prior mismatch K_1 of §C.3 is nonzero by construction. Truncation at t_{start} is safe in the direction that matters: the entropy bound evaluated on the window bounds $S[p_{t_{\text{start}}}]$. For this schedule the smoothing gain $S[p_{t_{\text{start}}}] - S[p_0]$ ($+3.5 \times 10^{-2}$ nats per dimension) exceeds the contraction loss $\log \alpha(t_{\text{start}}) = -7.7 \times 10^{-6}$ nats per dimension by more than three orders of magnitude, so the bound remains an upper bound on S_{data}. *Notation:* within this Ising appendix only, $g(t) = \sin(\pi t/2)$ denotes the cosine-VP marginal standard deviation of the perturbed sample, not the SDE diffusion coefficient. The corresponding SDE diffusion coefficient is denoted $g_{\text{SDE}}(t)$ where it appears (e.g. in the likelihood-weighted DSM weighting). Models are trained in v-prediction form,

$$\mathbf{v}_t = \alpha(t)\boldsymbol{\epsilon} - g(t)\mathbf{x}_0, \qquad \mathcal{L} = \mathbb{E}\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) - \mathbf{v}_t\|^2, \tag{37}$$

and the score used by the NLL and entropy rate estimators is recovered as

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) = -\frac{g(t)\mathbf{x}_t + \alpha(t)\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)}{g(t)}. \tag{38}$$

Training and selection. One recipe is used at every temperature: exact pool of 10^5 configurations, likelihood-weighted DSM with weight $g_{\mathrm{SDE}}(t)^2$, log-uniform time sampling, cosine learning rate decay from 10^{-4} to 10^{-5}, exponential moving average 0.999, and at most 3×10^4 steps. Ten seeds are trained per temperature. For each seed the reported checkpoint is the one with the smallest upper bound on a validation pool, among checkpoints whose validation slack (upper bound minus exact entropy) is nonnegative within noise. It is then evaluated on samples drawn fresh from the exact distribution. The exact entropy target is fixed and independent of the learned model.

NLL and entropy rate estimation. For each selected checkpoint, both likelihoods are evaluated on 4096 fresh exact samples. We compute $\mathrm{NLL}_{\mathrm{actual}}$ by integrating the probability flow ODE with fixed-step fourth-order Runge–Kutta on a shared grid of 400 nodes, and $\mathrm{NLL}_{\mathrm{plug}}$ by quadrature of the same divergence integrand under the closed-form forward marginals on the same grid, with the same noise draws and the same random probe vectors of the Hutchinson trace estimator of the divergence. The system entropy integral uses the cosine VP marginal, 1600 log-spaced quadrature nodes in t, and fresh samples from the exact dequantized distribution. The model-derived entropy upper bound plotted in Figure 2b is

$$\widehat{S}^{\mathrm{UB}}_{\mathrm{data,actual}} = 2\,\mathrm{NLL}_{\mathrm{actual}} - S_{\mathrm{noise}} + \int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\,dt \;=\; \widehat{S}^{\mathrm{UB}}_{\mathrm{data,plug}} + 2R_{\boldsymbol{\theta}}, \tag{39}$$

evaluated from the probability flow NLL. Its seed mean slack lies in $[+0.029, +0.061]$ bpd across the sweep, loosest at $T{=}3$. The selection pipeline's own test split estimate, with batch 256 and adaptive Runge–Kutta (RK45), gives $[+0.016, +0.055]$ bpd.

Residual and validity on the Ising sweep. Because $\mathbf{s}_{\mathrm{true}}$ is available in closed form by transfer matrix and S_{data} by exact enumeration of all 2^{25} configurations, every term of Theorem 1 is separately computable here. Table 5 reports, per temperature, the seed means of the score-error integral $\int D\,\mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2 dt$ measured directly against the exact score, the window term, the residual, and the reported slack, which is their sum. The window term is the plug-in slack minus the directly measured score error, 0.041–0.052 bpd on every checkpoint, close to the 0.05 bpd smoothing gain $S[p_{t_{\mathrm{start}}}] - S[p_0]$ computed above. On the truncated window the bound is on $S[p_{t_{\mathrm{start}}}]$, not on S_{data}. The score error rises from 0.047 bpd at $T{=}1$ to 0.102 bpd at $T{=}3$, and the residual is negative on all 60 checkpoints. Validity: against S_{data} the bound holds on all 60 checkpoints on both likelihood routes (plug-in and probability flow), with the one-sided lower confidence limit of the margin positive in every case. The condition stated with Eq. 11, a score error exceeding $2|R_{\boldsymbol{\theta}}|$, holds on only 38 of 60: on 1 of 10 seeds at $T{=}1.0$ and at $T{=}1.5$, and on 8 to 10 of 10 at $T \geq 2$. Where it fails, the bound evaluated with $\mathrm{NLL}_{\mathrm{actual}}$ undercuts $S[p_{t_{\mathrm{start}}}]$ and holds against S_{data} only because the window term exceeds the deficit.

Table 5: **Ising sweep: score error, residual, and reported slack (bpd, means over 10 seeds).** Periodic-convolutional score network. Score error is $\int D\,\mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2 dt$ measured against the exact score; the window term is the contribution of truncating the time interval; reported slack is $\widehat{S}^{\mathrm{UB}}_{\mathrm{actual}} - S_{\mathrm{data}}$, the sum of the three columns before it, with the 95% CI over seeds.

T	score error	window	$2R_{\boldsymbol{\theta}}$	reported slack
1.0	0.047	0.049	-0.067	$+0.029 \pm 0.009$
1.5	0.066	0.052	-0.085	$+0.033 \pm 0.010$
2.0	0.090	0.042	-0.076	$+0.056 \pm 0.011$
2.269	0.095	0.043	-0.078	$+0.060 \pm 0.017$
2.5	0.096	0.042	-0.086	$+0.052 \pm 0.011$
3.0	0.102	0.045	-0.086	$+0.061 \pm 0.017$

J IMAGE DATASET EXPERIMENTAL DETAILS

This appendix supports Figure 4 and Table 2. It records the dataset, checkpoint, dequantization, NLL, and entropy rate conventions used to produce the plotted image model points.

Common image likelihood convention. For an image with $d = CHW$ pixels/channels, we report 8-bit corrected bits per dimension as

$$\text{bpd} = \frac{\text{nats}}{d \log 2} + \text{offset}. \tag{40}$$

The offset follows the model's data scaling convention. Centered checkpoints use inputs in $[-1, 1]$ and offset 7.0. Uncentered checkpoints use inputs in $[0, 1]$ and offset 8.0. The CIFAR-10/ImageNet-32 VP/sub-VP and U-ViT checkpoints and the 256×256 VE NCSN++ checkpoints are all centered (offset 7.0) per their released wrappers. The only uncentered configurations are the toy and Ising experiments described above. Uniform dequantization is implemented as $(k + u)/256$ with $u \sim \text{Unif}[0, 1]$, followed by the checkpoint's expected scaling. Unless otherwise stated, the reported NLL values use this uniform dequantization convention.

Probability flow NLL. All image model NLLs are evaluated by the same probability flow ODE used in the toy experiments,

$$\frac{d\mathbf{x}_t}{dt} = f(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \qquad \text{NLL}(\mathbf{x}_0) = -\log p_T(\mathbf{x}_T) - \int_0^T \nabla \cdot \frac{d\mathbf{x}_t}{dt}\, dt. \tag{41}$$

The divergence term is estimated with a Rademacher Hutchinson–Skilling trace estimator, using a fixed probe along each ODE trajectory and independent seeds across batches/repeats. Adaptive RK45 uses relative and absolute tolerances 10^{-5} for the image sweeps, except for the tightened Ising runs described above.

Entropy rate quadrature. For each checkpoint, a held-out batch is dequantized and noised exactly from the model's forward marginal. We estimate $\dot{S}_{\boldsymbol{\theta}}(t)$ and $\dot{S}^i_{\boldsymbol{\theta}}(t)$ at $N_t = 50$ log-spaced times between the numerical lower endpoint and $1 - 10^{-5}$, then integrate on the log-time grid. The same score wrapper, SDE drift, diffusion coefficient, and convention for the divergence of the drift are used for NLL and entropy rate evaluation. The entropic lower bound and the plotted dissipation therefore share one evaluation pipeline. This interior-node rule in $\log t$ omits half a strip at each end of the grid and converges at first order. On the 256×256 checkpoints it is unconverged at N_t=50: refining to N_t=400 raises $\int \dot{S}_{\boldsymbol{\theta}}\, dt$ by $+1.4346$ bpd on FFHQ and $+1.4372$ bpd on LSUN Church, still increasing at fitted order ≈ 0.97. Every 256^2 upper bound in Table 2 is therefore biased low by at least 1.43 bpd, and every lower bound in the floor column is biased high by half of it. The CIFAR-10 and ImageNet-32 integrals move by 0.001–0.018 bpd over the same refinement and are converged for the purposes of Figure 4. Table 2 reports the N_t=50 values with this bias stated rather than corrected.

CIFAR-10 checkpoints in Figure 4. The CIFAR-10 panel uses the full test set with batch size 256, five NLL passes, and entropy rate batches of size 256. The NCSN++/DDPM++ VP, sub-VP, and VE checkpoints are the released continuous-time checkpoints from (Song et al., 2021b). The U-ViT-S/2 point uses the pixel space transformer baseline from (Bao et al., 2023) evaluated with the same probability flow and entropy rate code path after adapting the checkpoint to a VP score function. The dashed horizontal line marks $S_0 \approx 2.06$ bpd, the CIFAR-10 NLL reported by the Variational Polya Tree density estimator (Xu et al., 2025).

Rectified Flow in Figure 4. The Rectified Flow point is the released CIFAR-10 1-Rectified Flow checkpoint (Liu et al., 2023). Its network is the VP DDPM++ architecture above with a velocity output, trained on the path $\mathbf{x}_\tau = (1 - \tau)\mathbf{x}_0 + \tau\boldsymbol{\epsilon}$ in the diffusion time of this paper. On a Gaussian path the velocity determines the score exactly (Holderrieth & Erives, 2025): with $\mathbf{v}_{\boldsymbol{\theta}}$ the network output at flow time $1 - \tau$, $\mathbf{s}_{\boldsymbol{\theta}} = ((1 - \tau)\mathbf{v}_{\boldsymbol{\theta}} - \mathbf{x})/\tau$. The forward SDE sharing these marginals has $f(\tau) = -1/(1 - \tau)$ and $D(\tau) = \tau/(1 - \tau)$. Both diverge at $\tau = 1$, so the integrals run from $\tau = 10^{-3}$, the same forward noise standard deviation as the VP start point, to $1 - 10^{-3}$, the model's own training cutoff. The point comes from the residual evaluation code of Appendix C.4 on 2000 test images. On the VP checkpoint that code reproduces the plotted point to 0.02 bpd in both coordinates, and on this model its $\text{NLL}_{\text{actual}}$ agrees with the released likelihood code to 0.021 ± 0.009 bpd. Rectified Flow reaches the VP model's NLL (3.168 vs 3.156 bpd) at higher intrinsic dissipation. Because D differs between the two paths, we compared them at matched forward noise standard deviation at the start point: the excess is $+0.47$, $+0.95$, and $+1.38$ bpd at 10^{-2}, 3×10^{-3}, and 10^{-3}, with the NLLs equal to 0.011 bpd at each. The ordering therefore does not depend on the cutoff within both models' training ranges. On the system rate the two differ by 0.09 bpd.

ImageNet-32 checkpoints in Figure 4. The ImageNet-32 panel evaluates the released `score_flow` DDPM++ continuous checkpoints on the Oord downsampled ImageNet-32 validation set saved as `oord_val_data.npz`. The validation set contains 49,999 images after conversion to the channel-first format expected by the PyTorch evaluation code. We use centered VP/sub-VP inputs, offset 7.0, batch size 256, five NLL passes, and entropy rate batches of size 256. The panel restricts to the eight likelihood-weighted (LW) checkpoints from Song et al. (2021a) (four LW with uniform-t sampling and four LW with importance-weighted t-sampling, across VP/sub-VP and standard/deep) because these are the variants whose training objective matches the 3.76 bpd ImageNet-32 NLL reference (dashed line) reported by Song et al. (2021a). The four DSM-trained "standard" DDPM++ checkpoints from the same release achieve NLL 5.01–5.88 bpd on this set, consistent with the DSM-vs-LW gap originally reported by Song et al. (2021a). We omit them from the panel as a mismatched-objective comparison rather than as a result. For parity with the released likelihood code, the lower ODE endpoint is checkpoint-dependent: 10^{-5} for VP LW, 5×10^{-5} for VP LW with importance-weighted t-sampling, and 10^{-2} for sub-VP variants. The endpoint matters for the plotted dissipation. For sub-VP, $D(t) = \frac{1}{2}\beta(t)\big(1 - e^{-2\int_0^t \beta}\big) \to 0$ as $t \to 0$ while $\mathbf{f} = -\frac{1}{2}\beta\mathbf{x}$ stays finite, so $\dot{S}^i = \|\mathbf{f} - D\mathbf{s}\|^2/D \sim 1/t$ and $\int \dot{S}^i\, dt$ diverges logarithmically at the lower endpoint. With the exact score on the $\beta \in [0.1, 20]$ schedule, moving the sub-VP endpoint from 10^{-2} to 10^{-5} shifts it from 0.63 to 1.31 bpd, whereas VP moves by 0.003 bpd. Absolute sub-VP dissipation values are therefore cutoff-dependent and are compared only within family in Figure 4. Dequantization for the LW evaluations follows Song et al. (2021a)'s Flow++-style variational dequantizer, applied as released.

High-resolution VE checkpoints in Table 2. The 256×256 table evaluates continuous-time VE NCSN++ checkpoints from (Song et al., 2021b): FFHQ, CelebA-HQ, LSUN Bedroom, and LSUN Church. Each row uses a strictly held-out nontraining split, $n = 500$ images except LSUN Bedroom with $n = 300$, centered VE inputs in $[-1, 1]$, offset 7.0, and the released NCSN++ geometric VE schedule with $\sigma_{\max}$=380 (the LSUN default that the released wrappers also apply to FFHQ and CelebA-HQ). This schedule gives the per-dimension equilibrium entropy $S_{\text{noise}} = \frac{1}{2}\log_2(2\pi e\sigma_{\max}^2) + \text{offset} = 17.617$ bpd. NLL is reported with the standard error over the evaluated split. The model-derived data entropy upper bound in Table 2 is

$$\widehat{S}^{\text{UB}}_{\text{data,plug}} = 2\,\text{NLL}_{\text{plug}} - S_{\text{noise}} + \int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\, dt, \qquad \text{NLL}_{\text{plug}} = \text{NLL}_{\text{actual}} - R_{\boldsymbol{\theta}}, \tag{42}$$

with $R_{\boldsymbol{\theta}}$ measured as in Appendix C.4. The same bound evaluated from $\text{NLL}_{\text{actual}}$ differs by $2R_{\boldsymbol{\theta}}$, which is negative here, so it is smaller by 0.86–1.84 bpd on these four models. Where an external entropic proxy (another model's published likelihood on the same data) exists, the floor column substitutes it for S_{data} in Eq. 8. Because the proxy is a cross-entropy and hence an upper bound on S_{data}, and the lower bound increases with the entropy supplied, that column overstates the lower bound and is reported as a proxy quantity only. Where no proxy exists the floor column equals NLL_{plug} algebraically and is not an independent result. The 8-bit offset converts a continuous NLL into an upper bound on the discrete codelength by Jensen's inequality, $\text{NLL}_{\text{discrete}} \le \text{NLL}_{\text{continuous}} - d\log c$ with c the dequantization cell side. It does not carry a lower bound to the discrete side, so the floor column is a statement about the continuous-density likelihood and is not comparable term by term with the discrete Efficient-VDVAE codelengths beside it.

K SWISS ROLL EXPERIMENTAL DETAILS

We adapt our implementation from the open-source codebase `sdeflow-light` (https://github.com/CW-Huang/sdeflow-light).

K.1 SWISS ROLL DATASET SPECIFICATIONS

We generate a 2D Swiss roll dataset using `sklearn.datasets.make_swiss_roll`. For each draw, we sample n points from the standard 3D Swiss roll with additive isotropic Gaussian noise (default `noise`= 0.5), then retain the (x, z) coordinates and rescale by a factor of $1/5$, yielding $\mathbf{x}_0 \in \mathbb{R}^2$. Concretely, if the raw sample is $\tilde{\mathbf{x}}_0 \in \mathbb{R}^3$, we form

$$\mathbf{x}_0 \;=\; \frac{1}{5}\begin{bmatrix}\tilde{x}_0\\ \tilde{z}_0\end{bmatrix} \in \mathbb{R}^2. \tag{43}$$

During training, minibatches are generated on the fly by repeatedly sampling fresh i.i.d. points from this procedure (i.e., we do not materialize a fixed finite training set). For visualization, we plot 2D histograms over a fixed region, approximately $[-5, 5]^2$.

K.2 VP SCHEDULE SPECIFICATIONS

We use the variance-preserving (VP) forward diffusion in the continuous-time form

$$d\mathbf{x}_t \;=\; -\frac{1}{2}\beta(t)\,\mathbf{x}_t\,dt \;+\; \sqrt{\beta(t)}\,d\mathbf{w}_t, \qquad t \in [0, T], \tag{44}$$

where $\mathbf{w}_t$ is standard Brownian motion in $\mathbb{R}^2$. We take a linear noise schedule

$$\beta(t) \;=\; \beta_{\min} + t\big(\beta_{\max} - \beta_{\min}\big), \tag{45}$$

and set $T = 1$ in all experiments.

Closed-form marginal sampling. The conditional distribution $\mathbf{x}_t \mid \mathbf{x}_0$ is Gaussian:

$$\mathbf{x}_t \mid \mathbf{x}_0 \sim \mathcal{N}\big(m(t)\mathbf{x}_0,\ \mathrm{var}(t)\,\mathbf{I}\big), \tag{46}$$

with

$$m(t) \;=\; \exp\Big(-\tfrac{1}{4}t^2(\beta_{\max} - \beta_{\min}) - \tfrac{1}{2}t\beta_{\min}\Big), \tag{47}$$

$$\mathrm{var}(t) \;=\; 1 - \exp\Big(-\tfrac{1}{2}t^2(\beta_{\max} - \beta_{\min}) - t\beta_{\min}\Big), \tag{48}$$

and diffusion amplitude

$$g(t) \;=\; \sqrt{\beta(t)}. \tag{49}$$

In the reference implementation, we use $(\beta_{\min}, \beta_{\max}) = (0.1, 20.0)$.

K.3 MODEL SPECIFICATIONS

Time-conditioned MLP. We parameterize the learned score/drift network as a time-conditioned multilayer perceptron (MLP). The input is formed by concatenating time $t \in \mathbb{R}$ to the state $\mathbf{x} \in \mathbb{R}^2$:

$$[\mathbf{x}; t] \in \mathbb{R}^3, \tag{50}$$

and the network outputs a 2D vector:

$$\mathrm{MLP}_\theta : \mathbb{R}^3 \to \mathbb{R}^2, \qquad (\mathbf{x}, t) \mapsto \mathrm{MLP}_\theta(\mathbf{x}, t). \tag{51}$$

We use Swish activations and three hidden layers.

We sweep model capacity by varying the hidden width $H \in \{4, 8, 16, 32, 64, 128, 256\}$ (three hidden layers fixed), training separate models from scratch for each H. We run 5 independent replicates per configuration with different random seeds.

K.4 TRAINING SPECIFICATIONS

Denoising score matching (DSM) objective. At each iteration, we sample $t \sim \mathrm{Unif}[0, T]$, draw $\mathbf{x}_t \mid \mathbf{x}_0$ using the closed form in (46), and obtain the associated injected noise $\boldsymbol{\epsilon}$ such that

$$\mathbf{x}_t \;=\; m(t)\mathbf{x}_0 + \mathrm{std}(t)\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathrm{std}(t) = \sqrt{\mathrm{var}(t)}. \tag{52}$$

We minimize the DSM loss (as implemented) given by

$$\mathcal{L}_{\mathrm{DSM}}(\theta) \;=\; \frac{1}{2}\,\mathbb{E}\left[\left\|\mathrm{MLP}_\theta(\mathbf{x}_t, t) \cdot \frac{\mathrm{std}(t)}{g(t)} + \boldsymbol{\epsilon}\right\|_2^2\right], \tag{53}$$

where the expectation is approximated by a minibatch average. At inference time, the score is recovered from the MLP output by

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \;=\; \frac{\mathrm{MLP}_\theta(\mathbf{x}_t, t)}{g(t)}, \tag{54}$$

so that the minimizer of Eq. 53 over functions of $(\mathbf{x}_t, t)$ is the conditional expectation of the regression target,

$$\mathbb{E}[-\boldsymbol{\epsilon}/\mathrm{std}(t) \mid \mathbf{x}_t] \;=\; \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \;=\; \mathbf{s}_{\mathrm{true}}(\mathbf{x}_t, t),$$

the marginal score. It is not the realized conditional score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\boldsymbol{\epsilon}/\mathrm{std}(t)$, which depends on the unobserved $\boldsymbol{\epsilon}$ and is not a function of $\mathbf{x}_t$ alone. The distinction matters here because the slack of Theorem 1 is a marginal-score error. The two objectives share their minimizer and differ in value by the conditional variance $\mathbb{E}\|\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{x}_0) - \mathbf{s}_{\mathrm{true}}\|^2$. All NLL and entropy rate estimators evaluate the score through this recovery.

Optimization and runtime settings. We train using Adam with learning rate 10^{-3}, batch size 256, for 100,000 iterations. All experiments are run on an NVIDIA A100 graphics processing unit.

K.5 Capacity confound

The integrated quantities of Theorem 1 are invariant under a smooth increasing reparameterization of time with consistently transformed drift, diffusion, and score: under $t = \phi_\kappa(u) = (e^{\kappa u} - 1)/(e^\kappa - 1)$, $\kappa \in \{-2, \dots, 2\}$, the integrated dissipation, score error, and endpoint map of the VP Swiss roll sweep (35 checkpoints, seven hidden widths, five seeds) agree to 10^{-5} relative. The axes of Figure 3 therefore do not depend on the clock. On that sweep hidden width alone accounts for dissipation (Spearman $\rho = +0.93$), and at fixed width the dissipation standard deviation is 0.006–0.125 bpd against a 0.98 bpd between-width range. The design therefore has little power to detect a dissipation-specific effect on any downstream quantity: it cannot separate dissipation from capacity, which is not to say that dissipation is irrelevant. Sub-VP, VE, and CondOT checkpoints were not run through this control.

K.6 CondOT flow matching models

Figure 3b also shows 40 flow matching networks trained on the conditional optimal transport (CondOT) path (Lipman et al., 2022), with the same MLP, hidden widths, seeds, optimizer, batch size, and iteration count as the score networks. In the paper's clock ($t = 0$ data, $t = 1$ noise) the path is $\mathbf{x}_t = (1 - t)\,\mathbf{x}_0 + t\,\mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the network $\mathbf{v}_{\boldsymbol{\theta}}$ minimizes the conditional flow matching loss $\mathbb{E}\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - (\mathbf{z} - \mathbf{x}_0)\|^2$. The Gaussian conditional path with $m(t) = 1 - t$ and $\sigma(t) = t$ is the transition kernel of exactly one affine SDE $d\mathbf{x} = f(t)\,\mathbf{x}\,dt + g(t)\,d\mathbf{W}$, namely $f = \dot{m}/m = -1/(1 - t)$ and $g^2 = 2(\sigma\dot{\sigma} - f\sigma^2) = 2t/(1 - t)$, so $D(t) = g^2/2 = t/(1 - t)$. This is the diffusion whose entropy rates the figure reports. Identifying the learned velocity with the probability flow velocity $f\mathbf{x} - D\,\mathbf{s}$ of that SDE (Holderrieth & Erives, 2025) gives

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) = \frac{f(t)\,\mathbf{x} - \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, t)}{D(t)} = -\,\frac{\mathbf{x} + (1 - t)\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, t)}{t}, \tag{55}$$

which returns the exact score whenever the flow is exact. With this score every quantity is evaluated as for the score models: $\dot{S}^i_{\boldsymbol{\theta}} = \mathbb{E}_{p_t}\|f\mathbf{x} - D\mathbf{s}_{\boldsymbol{\theta}}\|^2/D = \mathbb{E}_{p_t}\|\mathbf{v}_{\boldsymbol{\theta}}\|^2/D$, so the intrinsic dissipation is the flow's kinetic energy weighted by $1/D$; $\dot{S}_{\boldsymbol{\theta}} = d\,\dot{f}(t) + D\,\mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}}\|^2$; and $\mathrm{NLL}_{\mathrm{plug}}$ by the instantaneous change of variables along $d\mathbf{x}/dt = \mathbf{v}_{\boldsymbol{\theta}}$. Because D vanishes at the data end and diverges at the noise end, all integrals are taken over the window $[0.0104, 0.9934]$, matched to the VP sweep's $[10^{-3}, 0.999]$ in signal-to-noise ratio $m^2 v_{\mathrm{data}}/\sigma^2$ rather than in t. Here S_{data} and S_{noise} in Corollary 1 are k-NN estimates of the forward marginals at those endpoints (1.400 and 2.037 bpd). Quadrature uses the union of a grid log-spaced in t, which resolves $D \to 0$, and a grid uniform in $\int D\,dt$, which resolves the pole at $t \to 1$. Since $\|\mathbf{v}_{\boldsymbol{\theta}}\| = O(1)$ while $D \sim t$ near the data, $\dot{S}^i_{\boldsymbol{\theta}} \sim v_{\mathrm{data}}/t$ there. The exact velocity already dissipates 7.7 bpd on this window: the size of CondOT's dissipation is set by its path, not by training.

L Reference lines of the accuracy–dissipation figure

The speed limit lower bound (intrinsic axis of Figure 3a; values tabulated below). For the forward SDE the intrinsic rate is the squared Wasserstein speed of the marginal path in the metric set by the diffusion coefficient: with the local velocity $\mathbf{v} = \mathbf{f} - D\,\mathbf{s}$ of Appendix B, $\dot{S}^i(t) =$

$\frac{1}{D(t)}\mathbb{E}_{p_t}\|\mathbf{v}\|^2 = |\dot{p}_t|^2_{W_2}/D(t)$, because for affine $\mathbf{f}$ the current velocity is a gradient and hence the minimal-norm velocity transporting $\{p_t\}$ (Benamou & Brenier, 2000). For a constant diffusion coefficient, Benamou–Brenier applied to the kinetic action gives the thermodynamic speed limit $\Delta S_{\mathrm{tot}} \geq W_2^2(\rho_0, \rho_*)/(D\, t_{\mathrm{tot}})$ (Aurell et al.; Dechant and Sakurai; Klinger and Rotskoff (Klinger & Rotskoff, 2025), Eq. 3, stated there for an overdamped Langevin system at constant $D = \beta^{-1}$; their Eq. 5 adds the factor $(1-\epsilon)^2$ when the final distribution misses the target by a relative W_2 error ϵ). The diffusion schedules here have time-dependent $D(t)$, so we extend the bound by the time change $d\vartheta = D(t)\, dt$: the action becomes $\int d\vartheta\, \mathbb{E}\|\mathbf{v}/D\|^2$ and Benamou–Brenier on the ϑ clock gives

$$\int_0^1 \dot{S}^{\,i}(t)\, dt \;\geq\; \frac{W_2^2\big(p_{\mathrm{data}}, p_1\big)}{\int_0^1 D(t)\, dt}, \tag{56}$$

which replaces $D\, t_{\mathrm{tot}}$ by $\int_0^1 D\, dt$ and reduces to the constant-D statement when D is constant. Equivalently, Eq. 56 is Cauchy–Schwarz applied to the instantaneous bound $\dot{S}_{\mathrm{tot}}(t) \geq v_2(t)^2/T_t$ of Ikeda et al. (Ikeda et al., 2025), with v_2 the W_2 speed and T_t the time-dependent noise temperature. Their integrated statement is written for constant T. The $\int D\, dt$-for-t_{tot} step appears in neither paper. We take $\epsilon \to 0$, the distribution at $t = 1$ being the prior. For VE the actual terminal marginal $p_{\mathrm{data}} * \mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$ misses $\mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$ by $\epsilon \approx 0.02$, a 4% loosening of the VE floor (the right-hand side of Eq. 56) that we do not apply. The bound holds for any velocity field along its own flow, so it applies to the exact process and to the model flow alike. It is a statement about the marginal path, not about the score. The W_2^2 between the data and the distribution at $t = 1$ is computed exactly, as the optimal assignment cost between equal-size samples ($n = 4000$, three seeds; finite-n bias is $O(\log n/n)$, about $+0.01$). The Gelbrich moment bound $W_2^2 \geq \|\mathbf{m}_0 - \mathbf{m}_1\|^2 + \mathrm{tr}\big(\mathbf{C}_0 + \mathbf{C}_1 - 2(\mathbf{C}_1^{1/2}\mathbf{C}_0\mathbf{C}_1^{1/2})^{1/2}\big)$ is exact for Gaussians but $1.6\times$ too small for the Swiss roll against a unit Gaussian (0.44 against 0.69), so it would place the VP, sub-VP and CondOT floors 37% too low. For VE the $\sigma^2(1)$ term dominates and the two agree to 0.2%.

The floors differ across families only through the two ingredients of Eq. 56. For the Swiss roll ($d = 2$, data variance $v_0 = 1.84$ per coordinate):

family	$\int_0^1 D\, dt$	p_1	$W_2^2(p_{\mathrm{data}}, p_1)$	floor (bpd/dim)
VP, $\beta \in [0.1, 20]$	$\frac{1}{2}\int_0^1 \beta\, dt = 5.02$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	0.687 ± 0.007	0.099
sub-VP	$\frac{1}{2}\int_0^1 \beta\,(1 - e^{-2\int_0^t \beta})\, dt = 4.78$	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	0.687 ± 0.007	0.104
VE, $\sigma \in [0.01, 50]$	$\frac{1}{2}(\sigma^2(1) - \sigma^2(0)) = 1250$	$\mathcal{N}(\mathbf{0}, \sigma^2(1)\mathbf{I})$	4742 ± 62	2.74
CondOT, window	$\int_{0.0104}^{0.9934} \frac{t}{1-t}\, dt = 4.03$	$\approx \mathcal{N}(\mathbf{0}, 0.99\,\mathbf{I})$	0.693 ± 0.007	0.124

VP's distribution at $t = 1$ is one W_2 step from the data and its $\int D$ is small, so its floor is close to the equilibrium value zero. VE's distribution at $t = 1$ is far wider, and the ratio grows. CondOT (Appendix K.6) has $D(t) = t/(1-t)$, so on the full interval $\int D\, dt$ diverges and the floor is zero. On the window $[0.0104, 0.9934]$, matched in signal-to-noise ratio and used for every CondOT quantity, $\int D\, dt = 4.03$ and the floor is 0.124 bpd. CondOT ends at the same unit Gaussian as VP, so its W_2^2 matches VP's to 1%. Its floor is higher than VP's only because its windowed $\int D\, dt$ is smaller. Its exact-score dissipation on that window is 7.67 bpd, $62\times$ the floor, because $D \to 0$ at the data end makes $\dot{S}^{\,i} \sim v_{\mathrm{data}}/t$ there. Measured against its floor, the least dissipative trained model runs $1.8\times$ (VE) and $3.6\times$ (VP) above the speed limit, and the exact VE process ($\Delta S_{\mathrm{true}} = 6.28$ bpd) $2.3\times$. Sub-VP, which shares VP's endpoints and appears in Figure 3b on the system axis, has an intrinsic $\int \dot{S}^{\,i} dt$ of 13.6 to 551 bpd that diverges logarithmically at $t \to 0$ ($D \to 0$ with finite drift). It therefore has no useful place on the intrinsic axis. The exact VE process cannot approach its floor under any VE clock: for a driftless path the W_2 speed in the diffusion clock $d\vartheta = D\, dt$ is $1/\mathrm{std}(t)$, so the path runs $37\times$ faster near the data than near the noise, while the floor assumes constant speed.

Full entropy restoration (dotted line, Figure 3b). On the system axis of Corollary 1 the natural reference is not a bound but the value the exact score takes, $\int_0^1 \dot{S}_{\mathrm{true}}\, dt = S_{\mathrm{noise}} - S_{\mathrm{data}} =: \Delta S_{\mathrm{true}}$ (Eq. 14); 0.633 bpd for VP and sub-VP, 6.28 for VE, 0.64 for CondOT on its window. $\Delta S_{\boldsymbol{\theta}} - \Delta S_{\mathrm{true}}$ is a state function difference and has either sign, so models sit on both sides of it.

Corollary 1 as a line (dashed line, Figure 3b). With $x = \int_0^1 \dot{S}_{\boldsymbol{\theta}}\, dt$ and $y = \mathrm{NLL}_{\mathrm{plug}} - S_{\mathrm{data}}$, Corollary 1 reads $y \geq \frac{1}{2}(\Delta S_{\mathrm{true}} - x)$: a line of slope $-\frac{1}{2}$ through $(\Delta S_{\mathrm{true}}, 0)$, attained only at $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\mathrm{true}}$. On the shifted axis $x - \Delta S_{\mathrm{true}}$ of Figure 3b it is the one line $y = -\frac{1}{2}(x - \Delta S_{\mathrm{true}})$ for every family. In terms of the entropy reduction ratio of Eq. 12 it reads $y \geq \frac{1}{2}\,\Delta S_{\mathrm{true}}\,\mathrm{ERR}$. It is a supporting tangent of the achievable region, not its boundary.

The Cauchy–Schwarz frontier. Writing $I_{\boldsymbol{\theta}} := \int_0^1 D\,\mathbb{E}_{p_t}\|\mathbf{s}_{\boldsymbol{\theta}}\|^2 dt = x - F$ and $I_{\mathrm{true}} := \int_0^1 D\,\mathbb{E}_{p_t}\|\mathbf{s}_{\mathrm{true}}\|^2 dt = \Delta S_{\mathrm{true}} - F$ with $F := \int_0^1 \mathbb{E}_{p_t}[\nabla \cdot \mathbf{f}]\, dt = d\int_0^1 f(t)\, dt$ (zero for VE, $-\frac{d}{2}\int_0^1 \beta\, dt$ for VP and sub-VP), the cross term of Theorem 1 obeys $\langle \mathbf{s}_{\boldsymbol{\theta}}, \mathbf{s}_{\mathrm{true}} \rangle \leq \sqrt{I_{\boldsymbol{\theta}} I_{\mathrm{true}}}$, hence

$$y \;\geq\; I_{\mathrm{true}} - \sqrt{I_{\boldsymbol{\theta}}\, I_{\mathrm{true}}} \;=\; I_{\mathrm{true}}\Big(1 - \sqrt{I_{\boldsymbol{\theta}}/I_{\mathrm{true}}}\Big). \tag{57}$$

This is the envelope of the one-parameter family of bounds obtained by weighting the two norms in the polarization identity by λ and $1/\lambda$ (Young's inequality). The member $\lambda = 1$ is Corollary 1, and Eq. 57 is its optimum $\lambda^\star = \sqrt{I_{\mathrm{true}}/I_{\boldsymbol{\theta}}}$. Its slack is the misalignment term $\sqrt{I_{\boldsymbol{\theta}} I_{\mathrm{true}}}\,(1 - \cos\phi)$ alone, with $\cos\phi$ the cosine between the learned and true scores in the D-weighted inner product. The frontier is therefore attained by every score proportional to the true one. On the Swiss roll the gain over the tangent is 0.0001–0.04 bpd, and the measured gap to the frontier is 93–100% misalignment. On the driftless VE, where $x = \int \dot{S}^{\,i}$, Eq. 57 evaluated at the floor forces at least 2.1 bpd of plug-in excess y: a model at the speed limit cannot be accurate.

M Computing the entropy reduction ratio

The entropy reduction ratio of Eq. 12 is computed from a trained model with forward passes of the score network alone: no divergences, no ODE or SDE solves, and no held-out likelihood. Algorithm 1 (Section 4) is what produced Figure 3c; this appendix gives its inputs and its reading in more detail.

Inputs. The schedule enters through $\mathbf{f}(\mathbf{x}, t) = f(t)\,\mathbf{x}$, $D(t) = g(t)^2/2$, and the marginal coefficients $m(t)$, $\sigma(t)$. For a flow model the score is the converted velocity of Appendix K.6. The data entropy reference S_{data} is a nonparametric estimate for low-dimensional data (Appendix N), and a published likelihood or the model-based upper bound $\widehat{S}^{\mathrm{UB}}_{\text{data}}$ of Eq. 11 for images. S_{noise} is the entropy of the Gaussian prior, of variance $\sigma^2(1) = 1$ per coordinate for VP and sub-VP and $\sigma^2_{\max}$ for VE. On a truncated window, both endpoint entropies are taken at the window edges, as for CondOT. The drift term $d\int f\, dt$ is analytic: $-\frac{d}{2}\int \beta\, dt$ for VP and sub-VP, 0 for VE, and $d\ln\big((1-t_1)/(1-t_0)\big)$ for CondOT. The quadrature grid should be log-spaced near $t \to 0$ wherever $D(t) \to 0$ there.

Reading. $\mathrm{ERR} = 0$ is full entropy restoration, attained at the exact score. A value $\mathrm{ERR} > 0$ means the model restores less entropy than the schedule removed, and Corollary 1 then bounds its plug-in excess by $\frac{1}{2}\Delta S_{\mathrm{true}}\,\mathrm{ERR}$. ERR is not sign-definite: a model may overproduce entropy ($\mathrm{ERR} < 0$), in which case the corollary is uninformative and the excess is governed by the misalignment term of Eq. 57. Across a capacity sweep, the width at which $|\mathrm{ERR}|$ stops decreasing is the width beyond which added capacity restores no further entropy.

Cost and scale. Cost is one score evaluation per $(t_j, \mathbf{x}_0^{(i)})$ pair. On the Swiss roll the sample sizes and grids are those of Appendix K (for CondOT, the two-grid union of Appendix K.6), with five seeds per width. On images the same estimator underlies the $\int \dot{S}_{\boldsymbol{\theta}}\, dt$ column of Table 2. The normalization by ΔS_{true} makes ERR a schedule-relative quantity: VE, whose ΔS_{true} is 6.28 bpd, reaches $\mathrm{ERR} \leq 0.03$ at $h = 16$ while its absolute entropy deficit (0.2 bpd) is still larger than VP's at $h = 64$. The comparison across schedules is of fractions restored, not of bits.

N Numerical Estimates of Exact NLL Terms

N.1 Equilibrium entropy S_{NOISE}

For the toy driftless VE process with diffusion coefficient $g(t) = \sigma^t$ used in our synthetic experiments (Gaussian, Uniform, and Swiss roll), the marginal at $t = 1$ has covariance $v(1)\,\mathbf{I}$ with

$$v(1) = \frac{\sigma^2 - 1}{2 \ln \sigma}.$$

Hence the equilibrium entropy in nats for a d-dimensional Gaussian is

$$S_{1,\text{nats}} = \frac{d}{2} \ln(2\pi e v(1)) = \frac{d}{2} \ln\left(2\pi e \frac{\sigma^2 - 1}{2 \ln \sigma}\right)$$

and in bits per dimension (bpd),

$$S_1 = \frac{S_{1,\text{nats}}}{d \ln 2}$$

This term is computed analytically. We compute the exact closed form and convert it to bpd. It is $S[p_{\text{prior}}]$, the entropy of the Gaussian the sampler starts from. The exact forward marginal p_1 is a convolution whose entropy exceeds it by $\Delta = I(\mathbf{x}_0; \mathbf{x}_1)$, 0.033 bpd at σ=10 and 0.005 at σ=30 for the Gaussian data (§C.3). *Scope:* this $v(1)$ formula is specific to the $g(t) = \sigma^t$ schedule of our toy experiments. The continuous-time NCSN++ VE schedule used in the 256×256 image evaluations of Table 2 is the geometric schedule $\sigma(t) = \sigma_{\min}(\sigma_{\max}/\sigma_{\min})^t$, for which $v(1) \approx \sigma_{\max}^2$ and $S_{\text{noise}} = \frac{1}{2}\log_2(2\pi e \sigma_{\max}^2) + \text{offset}$ as used in Appendix J.

N.2 Dataset entropy S_{DATA}

N.2.1 Uniform and Gaussian datasets

These constants are independent of the diffusion schedule (VE vs VP) and enter directly into the NLL lower bound formulas. For the standard Gaussian datasets considered, the data entropy is fixed by the closed-form expression

$$S_{\text{data}} = \frac{1}{2} d \log(2\pi e)$$

corresponding to the entropy of a d-dimensional standard normal.

For the Uniform $[0,1]^d$ datasets, the entropy vanishes, $S_{\text{data}} = 0$, since the density is constant on its support. Let $X \sim \text{Unif}([0,1]^d)$, so $p(x) = 1$ for $x \in [0,1]^d$ and $p(x) = 0$ otherwise. The differential entropy is

$$S_{\text{data}} \;=\; -\int_{\mathbb{R}^d} p(x)\, \log p(x)\, dx \;=\; -\int_{[0,1]^d} 1 \cdot \log 1 \, dx \;=\; 0.$$

By factorization across coordinates, $S_{\text{data}} = \sum_{i=1}^d h(X_i)$ with $X_i \sim \text{Unif}([0,1])$ and $h(X_i) = -\int_0^1 1 \cdot \log 1 \, dx_i = 0$, hence $S_{\text{data}} = 0$.

N.2.2 Estimating the entropy of the Swiss roll dataset

This section describes three complementary ways we obtain the differential entropy of the 2D Swiss roll distribution.

(1) Model-based upper bound on S_{data}. Rearranging the lower bound on negative log-likelihood gives the upper bound

$$S_{\text{data}} \;\le\; 2\,\text{NLL}(\theta) + \int_0^1 \dot{S}_{\boldsymbol{\theta}}(t)\, dt - S_{\text{noise}}.$$

We sweep model capacity by varying the hidden width $H \in \{4, 8, \ldots, 256\}$ (three hidden layers fixed). We train separate models from scratch for each H and run 5 independent replicates per configuration with different random seeds. We report the smallest upper bound observed across all configurations, attained by the model with $H = 256$.

(2) Differential entropy via the Kozachenko–Leonenko k-NN estimator. As an independent estimator of S_{data}, we use the Kozachenko–Leonenko k-nearest neighbors estimator Kozachenko & Leonenko (1987). Given i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, let ε_i be the Euclidean distance from $\mathbf{x}_i$ to its k-th nearest neighbor. The estimator of the differential entropy (in nats) is

$$\widehat{S}_{\text{data}}^{\text{kNN}}(k) = \psi(N) - \psi(k) + \log V_d + d\,\frac{1}{N}\sum_{i=1}^{N}\log(\varepsilon_i),$$

where $\psi(\cdot)$ is the digamma function and $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the unit ball in $\mathbb{R}^d$. We compute ε_i efficiently using a k-dimensional tree query and report $\widehat{S}_{\text{data}}^{\text{kNN}}(k)$ in bits per dimension.

(3) Differential entropy via a 2D histogram discretization. We also estimate S_{data} using a discretization approach. Fix a rectangular range $[x_{\min}, x_{\max}] \times [z_{\min}, z_{\max}]$ and a uniform $B \times B$ grid (bins). Let p_{ij} be the empirical mass in bin (i, j). The discrete entropy of the quantized variable X^Δ is

$$H(X^\Delta) = -\sum_{i,j} p_{ij}\log p_{ij}.$$

To convert this to a differential entropy estimate, we add the log cell area:

$$\widehat{S}_{\text{data}}^{\text{hist}}(B) = H(X^\Delta) + \log(\Delta x\,\Delta z), \qquad \Delta x = \frac{x_{\max}-x_{\min}}{B}, \quad \Delta z = \frac{z_{\max}-z_{\min}}{B}.$$

This is the standard Riemann sum correction for moving from probabilities (masses) to densities Cover & Thomas (2005).

Choosing the histogram range and resolution is crucial. We set the range using a pilot sample with a small padding fraction to reduce boundary artifacts. We then sweep over candidate B and find the B whose $\widehat{S}_{\text{data}}^{\text{hist}}(B)$ is consistent with $\widehat{S}_{\text{data}}^{\text{kNN}}(k)$ (e.g. $k = 10$).

N.3 Squared norm of the model score

We estimate

$$I_{\boldsymbol{\theta}} = \frac{1}{2}\int_0^1 g(t)^2 \mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right)\right\|^2\right] dt$$

Estimator (per time grid t_k, batch size B):

1. Draw $\mathbf{x}_0 \sim p_{\text{data}}, \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.
2. Form $\mathbf{x}_t = \mathbf{x}_0 + \sqrt{v\left(t_k\right)}\mathbf{z}$.
3. Evaluate the model score $\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t_k\right)$.
4. Compute the batch mean $\widehat{E}_k = \frac{1}{B}\sum_{i=1}^{B}\left\|\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t^{(i)}, t_k\right)\right\|^2$.

Finally, we integrate over t with a Riemann sum on the log-time grid:

$$\hat{I}_{\boldsymbol{\theta}} = \frac{1}{2}\sum_k w_k \widehat{E}_k, \quad w_k = g\left(t_k\right)^2 \Delta t_k,$$

where the leading $1/2$ comes from the $D(t) = g(t)^2/2$ weighting in the definition of $I_{\boldsymbol{\theta}}$ rather than from a trapezoidal rule. The result is converted to bpd by dividing by $d\ln 2$.

N.4 Squared difference term, I_{diff}

We write the score-error integral as

$$\begin{aligned} I_{\text{diff}} &:= \int_0^1 D(t)\,\mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{s}_{\text{true}}(\mathbf{x}_t, t)\right\|^2\right] dt \\ &= \tfrac{1}{2}\int_0^1 g(t)^2\,\mathbb{E}_{p_t}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\text{true}}\right\|^2\right]\,dt \;\geq\; 0, \end{aligned}$$

using the convention $D(t) = \frac{1}{2}g(t)^2$ from Section 2.2. The slack of Theorem 1 is $\frac{1}{2}I_{\mathrm{diff}}$, and the slack of the entropy bound Eq. 11 is I_{diff} itself. The score-error column of Table 5 is I_{diff} in this convention. Note that I_{diff} differs from the standard denoising score matching training loss $L_{\mathrm{DSM}} = \frac{1}{2}\int \mathrm{std}(t)^2\, \mathbb{E}_{p_t}[\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2]\, dt$ by the schedule-dependent ratio $\mathrm{std}(t)^2/D(t)$. The two coincide only on schedules with $\mathrm{std}(t)^2 = D(t)$.

We estimate I_{diff} in two ways: directly by computing $\|\mathbf{s}_{\boldsymbol{\theta}} - \mathbf{s}_{\mathrm{true}}\|^2$ per sample and averaging, and using the polarization identity as a sanity check:

$$\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2\langle a, b\rangle$$

to obtain the squared difference term from separate estimates of $\|\mathbf{s}_{\boldsymbol{\theta}}\|^2$, $\|\mathbf{s}_{\mathrm{true}}\|^2$, and $\langle \mathbf{s}_{\boldsymbol{\theta}}, \mathbf{s}_{\mathrm{true}}\rangle$. We use the agreement between the two as a useful consistency diagnostic.