Article

Not peer-reviewed version

The Information Dynamics of Generative Diffusion

Dejan Stančević * and Luca Ambrogioni *

Posted Date: 5 January 2026

doi: 10.20944/preprints202601.0131.v1

Keywords: generative diffusion models; stochastic thermodynamics; information theory; entropy production; symmetry breaking; phase transition

 doi:10.20944/preprints202601.0131.v1

Article

The Information Dynamics of Generative Diffusion

Dejan Stančević * and Luca Ambrogioni *

Donders Institute for Brain, Cognition and Behaviour, Radboud University
* Correspondence: dejan.stancevic@donders.ru.nl (D.S.); luca.ambrogioni@donders.ru.nl (L.A.)

Abstract

Generative diffusion models have emerged as a powerful class of models in machine learning, yet a unified theoretical understanding of their operation is still developing. This paper provides an integrated perspective on generative diffusion by connecting the information-theoretic, dynamical, and thermodynamic aspects. We demonstrate that the rate of conditional entropy production during generation (i.e. the generative bandwidth) is directly governed by the expected divergence of the score function's vector field. This divergence, in turn, is linked to the branching of trajectories and generative bifurcations, which we characterize as symmetry-breaking phase transitions in the energy landscape. Beyond ensemble averages, we demonstrate that symmetry-breaking decisions are revealed by peaks in the variance of pathwise conditional entropy, capturing heterogeneity in how individual trajectories resolve uncertainty. Together, these results establish generative diffusion as a process of controlled, noise-induced symmetry breaking, in which the score function acts as a dynamic nonlinear filter that regulates both the rate and variability of information flow from noise to data.

Keywords: generative diffusion models; stochastic thermodynamics; information theory; entropy production; symmetry breaking; phase transition

1. Introduction

Generative diffusion models have rapidly become one of the most successful frameworks for high-dimensional generative modeling. They were introduced in Sohl-Dickstein et al. [1] in analogy with stochastic thermodynamics. Several works elucidated the theoretical foundations of the method [2–4] and their practical implementation procedures [3,5]. Despite these efforts, a unified conceptual understanding of their behavior is still emerging. Several perspectives on information theory, stochastic thermodynamics, and the statistical physics of symmetry breaking have each shed light on different aspects of diffusion models, but their interrelations remain fragmented. The purpose of this perspective paper is to integrate these viewpoints into a single coherent theoretical picture.

Our central thesis is that *generation in diffusion models proceeds through a sequence of noise-driven symmetry-breaking transitions*. These transitions determine when and how the model commits to a specific generative outcome, structure the flow of information, regulate entropy production, and shape the geometry of trajectories in state space. We refer to this synthesis as the *information thermodynamics of generative diffusion*.

Information Theory and Entropy-Based Perspectives

A growing line of work has examined diffusion models from the standpoint of information theory, focusing especially on how information about the clean sample x_0 is progressively revealed as noise is removed. Recent works have proposed information-theoretic decompositions of diffusion dynamics [6,7] and have explored the role of conditional entropy in designing improved training and sampling schedules [8,9]. Furthermore, Franzese et al. [10] show how information-theoretic tools reveal the mechanisms by which latent abstractions guide generation. These approaches treat diffusion as a sequential information transfer process and highlight that the effectiveness of generation depends

 doi:10.20944/preprints202601.0131.v1

on how rapidly uncertainty about x_0 can be reduced. Central to these results is the observation that the conditional entropy rate is directly linked to geometric quantities such as the divergence of the score and the curvature of the log-density. This suggests that information flow is deeply connected to the underlying dynamical and geometric structure of the generative process.

Phase Transitions, Associative Memories, and Symmetry Breaking

Parallel developments in statistical physics have revealed that diffusion models exhibit noise-driven *symmetry-breaking events*, where the score field undergoes bifurcations and the generative trajectories split into distinct modes [11,12]. High-dimensional analyses have linked these transitions to mean-field phase transitions [13] and to dynamical behavior captured by stochastic localization [14–17]. These bifurcations correlate with sharp changes in the Hessian of the log-density, revealing a connection between symmetry breaking and information geometry. Similar mechanisms have been studied in hierarchical generative settings [18,19] and in analyses of memorization, mode formation, and semantic emergence [20–23]. Generative diffusion models have also been directly connected to modern Hopfield networks and other associative memory networks [24–27], where generalization has been associated with the emergence of spurious states [28]. Across these domains, the key unifying insight is that the Hessian (or score Jacobian) mediates both stability of generative trajectories and the structure of the data manifold.

Thermodynamics and the Role of Inferential Entropy

The connection between diffusion models and stochastic thermodynamics was first made explicit in Sohl-Dickstein et al. [1], motivating a thermodynamic view of generation. Furthermore, this connection was strengthened with a mathematical framework based on stochastic differential equations (SDEs) formulated in Song et al. [2], Rombach et al. [29] and is central to the modern understanding of diffusion models. However, the notion of entropy that is commonly used in stochastic thermodynamics [30,31] measures the irreversibility of the forward process. While such quantities yield elegant speed-accuracy tradeoffs [32], they characterize the evolution of the distribution of trajectories rather than the uncertainty relevant for generating a single sample. Instead, we argue that what matters during generation is the uncertainty about the clean sample x_0.

For this reason, it is more natural to study the conditional entropy $\mathbf{H}[\mathbf{x}_0|\mathbf{x}_t]$ and, at the trajectory level, the pathwise conditional entropy $h_t(\mathbf{x}_t)$, whose fluctuations capture the temporary multimodality experienced by individual generative paths. As in stochastic thermodynamics, such pathwise entropies can increase along single trajectories, even when the average conditional entropy decreases. We show that these fluctuations reveal symmetry-breaking events; the model becomes momentarily undecided among competing hypotheses for x_0, leading to spikes in conditional entropy variance and amplified sensitivity to noise.

Our Perspective

We unify these threads by showing that information flow, symmetry breaking, and dynamical instability are different manifestations of the same underlying mechanism governing diffusion models. In particular, we show that:

1. **Entropy as a detector of symmetry breaking:** Symmetry-breaking transitions are accompanied by pronounced changes in ensemble-level information measures. In particular, the conditional entropy rate $\dot{\mathbf{H}}[\mathbf{x}_0 \mid \mathbf{x}_t]$ exhibits sharp peaks around bifurcation points, reflecting an increased sensitivity of the generative process to noise. These peaks provide direct information-theoretic signatures of symmetry-breaking transitions.
2. **Noise-driven decisions:** These information-theoretic signatures arise when the score field becomes weak along a low-curvature direction, temporarily losing its ability to suppress stochastic fluctuations. In such regimes, noise plays an active role in selecting which generative branch the trajectory follows, effectively making the generative decision.

 doi:10.20944/preprints202601.0131.v1

3. **Path divergence via Lyapunov instability:** The same loss of curvature is reflected in the Jacobian of the score, whose spectrum develops positive eigenvalues along the unstable directions. As a result, nearby generative trajectories diverge exponentially, leading to macroscopic separation between paths that correspond to different generative outcomes.
4. **Non-monotonic pathwise entropy:** At the level of individual trajectories, this divergence manifests as heterogeneous resolution of uncertainty about x_0. Consequently, the pathwise conditional entropy need not evolve monotonically along single paths and may transiently increase, reflecting temporary ambiguity during the symmetry-breaking decision process. This results in the variance of pathwise conditional entropies peaking.

Together, these results establish a conceptual framework in which entropy production, posterior geometry, and dynamical stability are unified through the lens of noise-driven symmetry breaking. This perspective clarifies the mechanisms by which diffusion models transform noise into structured data and highlights the central role of symmetry in shaping generative dynamics.

2. Information Theory

We start by presenting an introduction to the information theory of sequential generative modeling, which will open the door to the analysis of generative diffusion.

Consider a game of *Twenty Questions* where an interrogator player may ask twenty binary questions concerning a set to an "oracle" player in order to gradually reveal the identity of a predetermined element $\mathbf{y}^*$ in a finite set $\Omega = \{\mathbf{y}_1, \dots, \mathbf{y}_{N_0}\}$ with N_0 elements. We denote the size of the possible set $\Omega_{j-1}(a_{1:j-1})$ after $j-1$ questions as $N_{j-1}(a_{1:j-1})$. The answer a_j to the j-th question q_j then divides the set into two possible subsets with sizes $N_j^1(a_{1:j-1}) = N_j(a_j = 1, a_{1:j-1})$ and $N_j^0(a_{1:j-1}) = N_{j-1}(a_{1:j-1}) - N_j(a_j = 1, a_{1:j-1})$. Assuming a fixed set of questions, the expected uncertainty experienced by the player after the j-th question can be quantified by the conditional entropy:

$$H(\mathbf{y}^* \mid a_{1:j}) = -\mathbb{E}_{\mathbf{y}^*, a_{1:j}}\left[\log_2 p(\mathbf{y} \mid a_{1:j})\right] = \mathbb{E}_{a_{1:j}}\left[\log_2 N_j(a_{1:j})\right] \tag{1}$$

where $\mathbf{y}^*$ is sampled uniformly from Ω. Under these conditions, the expected entropy reduction associated to a given question is given by

$$\Delta H_j = \mathbb{E}_{a_{1:j}}\left[\log_2 N_{j-1} - \frac{N_j^0}{N_{j-1}} \log_2 N_j^0 + \frac{N_j^1}{N_{j-1}} \log_2 N_j^1\right], \tag{2}$$

where we left the dependence on the set of answers implicit to unclutter the notation. It is easy to see that the maximum bit rate is 1, which is achieved when $N_j^0 = N_{j-1}/2$. Assuming that 20 questions are enough to fully identify the value of $\mathbf{y}^*$, we can encode each $\mathbf{y}$ in the string of binary values $a_{1:20}$, which makes clear that the question answering process consists of gradually filling in this string. Using the language of generative diffusion, we can re-frame this process in terms of a 'forward' process, where the string $a_{1:20}$ corresponding to an element of Ω is sampled in advance and then transmitted to the j-th 'time point' through the following non-injective forward process

$$R_j(a_{1:20}) = a_{1:j}\,, \tag{3}$$

which deterministically suppresses information by masking the values of the string. The solution of a Twenty Question game can then be seen as inverting this 'forward process'. Note that the forward process leads to a sequence of monotonically non-decreasing marginal conditional entropies $H(\mathbf{y}^* \mid a_{1:j}) < H(\mathbf{y}^* \mid a_{1:j-1})$, which is a fundamental feature of a forward process in diffusion models that captures the fact that information is lost by the forward transformation.

Now consider the case where a lazy oracle forgot to select a word in advance and decides instead to answer the questions at random under the probability determined by the sizes N_j^0 and N_j^1, which we assume to be fixed given the questions. Strikingly, this reformulation does not make any observable

 doi:10.20944/preprints202601.0131.v1

difference from the point of view of the interrogator as each (randomly sampled) answer equally reduces the space of possible words and it results in the same entropy reduction, until a final guess can be offered. Therefore, the game of Twenty Questions with a random oracle can be interpreted as a sequential generative process where the state at 'time' j is given by a binary string $a_{1:j}$ with Markov transition probabilities

$$p(a_{j+1} = 0 \mid a_{1:j}) = \frac{N^0_{j+1}(a_{1:j})}{N_j(a_{1:j})} \tag{4}$$

The conditional entropy rate ΔH_j determines how much information is transferred from 'time' j to the final generation.

As we shall see, the reverse diffusion process can be seen as analogous to this 'generative game' with the score function playing the part of the interrogator and the noise $\boldsymbol{\epsilon}_t$ playing the role of the oracle. Like in the interrogator in the generative Twenty Questions game, the score function can reduce the information transfer by tilting the probabilities of the stochastic increments out of uniformity, which reduces the impact of the noise. This phenomenon is related to the divergence of the vector field induced by the score function, which causes amplification of small perturbations during the generative dynamics. We will also see that the phenomenon is connected to the branching of paths of fixed-points of the score and consequently to the phenomenon of generative phase transitions and spontaneous symmetry breaking [11].

2.1. Score-Matching Generative Diffusion Models

The sequential generation example outlined above is analogous to the masked diffusion models [33,34]. On the other hand, score-matching generative diffusion models are continuous-time sequential generative models where the forward process is given by a diffusion process such as

$$\mathrm{d}\boldsymbol{x}_t = \nu(t)\, \mathrm{d}\boldsymbol{W}_t\,, \tag{5}$$

which is initialized with the data source $p(\boldsymbol{x}_0) = \rho(\mathbf{y})$. Generation in score-matching diffusion consists of integrating the 'reverse equation' [2]:

$$\mathrm{d}\boldsymbol{x}_t = -\nu^2(t)\nabla \log p_t(\boldsymbol{x}_t)dt + \nu(t)\, \mathrm{d}\boldsymbol{W}_t\,, \tag{6}$$

where, for notational simplicity, we restrict our attention to the forward process given in Eq. 5. Note that Eq. 6 must be integrated backward with initial condition determined by the stationary distribution of the forward process.

The fundamental mathematical object that determines the reverse dynamics is the score function, which in this case can be expressed as $\nabla \log p_t(\boldsymbol{x}_t) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_t}\left[\frac{\boldsymbol{y}-\boldsymbol{x}_t}{\sigma^2(t)}\right]$, where $\sigma^2(t) = \int_0^t \nu^2(\tau)\mathrm{d}\tau$ is the total variance of the noise at time t and the expectation is taken with respect of the conditional distribution $p(\mathbf{y} \mid \boldsymbol{x}_t) \propto p(\boldsymbol{x}_t \mid \mathbf{y})\rho(\mathbf{y})$. This expression can be further simplified by noticing that $\boldsymbol{x}_t = \mathbf{y} + \sigma(t)\mathbf{z}_t$:

$$\nabla \log p_t(\boldsymbol{x}_t) = -\mathbb{E}_{\mathbf{z}_t|\boldsymbol{x}_t}\left[\frac{\mathbf{z}_t}{\sigma(t)}\right] \tag{7}$$

where $\mathbf{z}$ is a standard normal vector. In other words, the score is the negative of the average (rescaled) noise and it therefore provides the optimal (infinitesimal) denoising direction.

In dynamical term, the score function determines the vector field (i.e. the drift) that guides the generative paths towards the distribution of the data.

In practice, a normalized score network $s(\mathbf{x}_t;\theta)$ should be trained to minimize the rescaled score-matching loss:

$$\mathcal{L}_{\mathrm{sm}}(\theta, t) = \mathbb{E}_{\boldsymbol{x}_t}\left[\|\sigma(t)\nabla \log p_t(\boldsymbol{x}_t) - s(\boldsymbol{x}_t;\theta)\|^2\right] \tag{8}$$

 doi:10.20944/preprints202601.0131.v1

This loss function cannot be computed directly because the true score is not available. However, Eq. 8 can be re-written using Eq. 7 and expanding the square:

$$\begin{aligned}\mathcal{L}_{\text{sm}}(\theta,t) &= \mathbb{E}_{\boldsymbol{x}_t}\left[\left\|\mathbb{E}_{\mathbf{z}_t|\boldsymbol{x}_t}[\mathbf{z}_t] \; + s(\boldsymbol{x}_t;\theta)\right\|^2\right] \\ &= \mathbb{E}_{\mathbf{z}_t,\mathbf{y}}\left[\|\mathbf{z}_t + s(\boldsymbol{y}+\sigma(t)\mathbf{z}_t;\theta)\|^2\right] - \mathbb{E}_{\mathbf{z}_t,\mathbf{y}}\left[\|\mathbf{z}_t + \sigma(t)\nabla \log p_t(\boldsymbol{y}+\sigma(t)\mathbf{z}_t)\|^2\right].\end{aligned} \tag{9}$$

Note that the second term is constant in θ, which means that the gradient solely depends on the denoising loss:

$$\mathcal{L}_{\text{d}}(\theta,t) = \mathbb{E}_{\mathbf{z}_t,\mathbf{y}}\left[\|\mathbf{z}_t + s(\boldsymbol{y}+\sigma(t)\mathbf{z}_t;\theta)\|^2\right]. \tag{10}$$

The constant term

$$C_t = \mathbb{E}_{\mathbf{z}_t,\mathbf{y}}\left[\|\mathbf{z}_t + \sigma(t)\nabla \log p_t(\boldsymbol{y}+\sigma(t)\mathbf{z}_t)\|^2\right] \tag{11}$$

is of high importance for our current purposes. It quantifies the loss of the denoiser obtained from the score function. This is therefore the unavoidable part of the denoising error that is still present given a perfectly trained network. With a few manipulations, it is possible to show that this term is in fact equal to the variance of the posterior denoising distribution:

$$C_t = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t}[\text{var}(\boldsymbol{y} \mid \boldsymbol{x}_t)], \tag{12}$$

which allows us to interpret this term as a measure of uncertainty at time t on the final outcome of the generative trajectory.

3. Generative Information Transfer in Score Matching Diffusion

To characterize the generative information transfer we need to compute the conditional entropy rate $\dot{\mathbf{H}}[\boldsymbol{y} \mid \boldsymbol{x}_t]$, which is the analogous of the discrete entropy reduction we gave in Eq. 2. The conditional entropy is defined as

$$\mathbf{H}[\boldsymbol{y} \mid \boldsymbol{x}_t] = -\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t}[\log p(\boldsymbol{y} \mid \boldsymbol{x}_t)] \tag{13}$$

To find the entropy rate, we can take the temporal derivative of Eq. 13 and use the Fokker-Planck equation, which in our case is just the heat equation:

$$\partial_t p_t(\boldsymbol{x}_t) = \frac{1}{2}\nu^2(t)\nabla^2 p_t(\boldsymbol{x}_t). \tag{14}$$

Using integration by parts, this results in

$$\begin{aligned}\dot{\mathbf{H}}[\boldsymbol{y}|\boldsymbol{x}_t] &= \frac{\nu^2(t)}{2}\left(\mathbb{E}_{p(\boldsymbol{x}_t,\boldsymbol{x}_0)}[\|\nabla \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2] - \mathbb{E}_{p_t(\boldsymbol{x}_t)}[\|\nabla \log p(\boldsymbol{x}_t)\|^2]\right) \\ &= \frac{\nu^2(t)}{2}\left(\frac{D}{\sigma^2(t)} - \mathbb{E}_{p_t(\boldsymbol{x}_t)}[\|\nabla \log p(\boldsymbol{x}_t)\|^2]\right),\end{aligned} \tag{15}$$

where D is the dimensionality of the ambient space. From this formula, we can see that the maximal bandwidth is reached when the Euclidean norm of the score function is minimized.

3.1. Score Norm and Posterior Concentration

To gain some insight on the significance of the square norm and the expression for the conditional entropy we will consider the following case. We assume a discrete data distribution $p_0(y) = \frac{1}{N}\sum_{i=1}^{N}\delta(y-y_i)$ with empirical mean equal to zero.

 doi:10.20944/preprints202601.0131.v1

At time t, the marginal density is given by a Gaussian smoothing of the data,

$$p_t(x) = \frac{1}{N}\sum_{i=1}^{N} \varphi_{\sigma(t)}(x - y_i), \tag{16}$$

where $\varphi_{\sigma(t)}$ denotes an isotropic Gaussian with variance $\sigma^2(t)$. The posterior distribution over datapoints is

$$p(y_i \mid x_t) = \frac{\varphi_{\sigma(t)}(x_t - y_i)}{\sum_{k=1}^{N} \varphi_{\sigma(t)}(x_t - y_k)}. \tag{17}$$

The score function can then be written as the posterior average

$$\nabla \log p_t(x_t) = \mathbb{E}_{y|x_t}\left[\frac{y - x_t}{\sigma^2(t)}\right] = \frac{1}{\sigma^2(t)}\Big(\mu(x_t) - x_t\Big), \qquad \mu(x_t) := \mathbb{E}[y \mid x_t]. \tag{18}$$

We now assume that the data vectors satisfy

$$y_i^\top y_j \approx 0 \quad (i \neq j), \qquad \|y_i\|^2 \approx R^2, \tag{19}$$

i.e. datapoints are approximately orthogonal and lie at a common distance R from the mean. Under this assumption, the squared norm of the posterior mean simplifies to

$$\|\mu(x_t)\|^2 = \Big\| \sum_{i=1}^{N} p(y_i \mid x_t)\, y_i \Big\|^2 \approx R^2 \sum_{i=1}^{N} p(y_i \mid x_t)^2. \tag{20}$$

Taking expectations with respect to $p_t(x_t)$, we obtain

$$\mathbb{E}_{x_t}\Big[\|\nabla \log p_t(x_t)\|^2\Big] = \frac{1}{\sigma^4(t)}\Big(\mathbb{E}_{x_t}\Big[\|\mu(x_t)\|^2\Big] - 2\,\mathbb{E}_{x_t}[x_t^\top \mu(x_t)] + \mathbb{E}_{x_t}[\|x_t\|^2]\Big). \tag{21}$$

The first term captures the data-dependent structure of the score and, using Eq. (20), can be written as

$$\mathbb{E}_{x_t}\Big[\|\mu(x_t)\|^2\Big] \approx R^2\, \mathbb{E}_{x_t}\left[\sum_{i=1}^{N} p(y_i \mid x_t)^2\right]. \tag{22}$$

The quantity $\sum_i p(y_i \mid x_t)^2$ measures the concentration of the posterior over datapoints. It satisfies $1/N \leq \sum_i p(y_i \mid x_t)^2 \leq 1$, interpolating between a fully diffuse posterior and complete concentration on a single datapoint.

The remaining two terms in Eq. (21) can be estimated explicitly under the forward model $x_t = y + \sigma(t)z$, where $z \sim \mathcal{N}(0, I)$ is independent of y. We have

$$\mathbb{E}_{x_t}[x_t^\top \mu(x_t)] = \mathbb{E}_{x_t}[x_t^\top \mathbb{E}[y \mid x_t]] = \mathbb{E}_{x_t,y}[x_t^\top y] = \mathbb{E}\|y\|^2 \approx R^2, \tag{23}$$

$$\mathbb{E}_{x_t}\Big[\|x_t\|^2\Big] = \mathbb{E}\|y\|^2 + \sigma^2(t)\,\mathbb{E}\|z\|^2 \approx R^2 + D\sigma^2(t), \tag{24}$$

where D denotes the ambient dimensionality.

Substituting Eqs. (22), (23), and (24) into Eq. (21) yields

$$\mathbb{E}_{x_t}\Big[\|\nabla \log p_t(x_t)\|^2\Big] \approx \frac{R^2}{\sigma^4(t)}\left(\mathbb{E}_{x_t}\left[\sum_{i=1}^{N} p(y_i \mid x_t)^2\right] - 1\right) + \frac{D}{\sigma^2(t)}. \tag{25}$$

The second term coincides with the expected squared norm of the score of the forward Gaussian kernel and therefore represents a data-independent baseline contribution. The first term encodes the deviation from pure diffusion induced by the structure of the dataset and depends solely on the posterior distribution over datapoints.

 doi:10.20944/preprints202601.0131.v1

Using the bound $1/N \leq \sum_{i=1}^{N} p(y_i \mid x_t)^2 \leq 1$, we obtain the inequality

$$-\frac{(N-1)R^2}{N\sigma^4(t)} \;\leq\; \frac{R^2}{\sigma^4(t)}\left(\mathbb{E}_{x_t}\left[\sum_{i=1}^{N} p(y_i \mid x_t)^2\right] - 1\right) \;\leq\; 0. \tag{26}$$

As a consequence, the expected squared norm of the score is always bounded above by the forward kernel contribution, ensuring that the marginal entropy remains a monotonically increasing function of time.

Further insight can be gained by rewriting the purity term as

$$\mathbb{E}_{x_t}\left[\sum_{i=1}^{N} p(y_i \mid x_t)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\int p(y_i \mid x_t)\, p(x_t \mid y_i)\, dx_t. \tag{27}$$

This expression makes explicit that the deviation from the diffusion baseline is controlled by the overlap of the forward kernels. If, at time t, the datapoints have effectively merged into m indistinguishable groups (with identical posteriors), the purity evaluates to m/N, yielding

$$\mathbb{E}_{x_t}\left[\sum_{i=1}^{N} p(y_i \mid x_t)^2\right] - 1 = \frac{m-N}{N}. \tag{28}$$

Therefore, increasing mixing among datapoints (smaller m) makes the data-dependent term more negative, reducing the expected score norm and increasing the conditional entropy rate.

This result allows us to interpret the magnitude of the score vector as a quantitative estimate of uncertainty in the denoising process: when multiple datapoints are compatible with the noisy state x_t, posterior averaging suppresses the score, leading to enhanced entropy production (equation 15). As we shall see in the rest of the paper, we can associate peaks in the entropy rates with symmetry-breaking bifurcations that correspond to noise-induced 'choices' between possible data points.

3.2. Conditional Entropy Production as Optimal Error

The conditional entropy rate quantifies the instantaneous generative information transfer at any given moment in time. It can be shown (see [8]) that this quantity is closely connected to the optimal denoising squared error, which is the variance of the denoising distribution:

$$\dot{\mathbf{H}}[\boldsymbol{y}|\boldsymbol{x}_t] = \frac{1}{2}\frac{\nu^2(t)}{\sigma^4(t)}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t}[\mathrm{var}(\boldsymbol{y} \mid \boldsymbol{x}_t)]\,. \tag{29}$$

Intuitively, this means that the information rate is directly related to the denoising uncertainty at a given time.

Using this relation, we can now re-express the denoising score matching formula in Eq. 9 in terms of the conditional entropy rate:

$$\mathbb{E}_{\boldsymbol{x}_t}\left[\left\|\mathbb{E}_{\mathbf{z}_t|\boldsymbol{x}_t}[\mathbf{z}] - s(\boldsymbol{x}_t;\theta)\right\|^2\right] + \frac{2\sigma^4(t)}{\nu^2(t)}\dot{\mathbf{H}}[\boldsymbol{y}|\boldsymbol{x}_t] = \mathbb{E}_{\mathbf{z}_t,\mathbf{y}}\left[\|\mathbf{z}_t - s(\boldsymbol{y}+\sigma(t)\mathbf{z}_t;\theta)\|^2\right], \tag{30}$$

which implies that the entropy rate can be estimated from the training loss if we assume that the network is well-trained.

3.3. Generative Bandwidth

It is insightful to investigate under what circumstances the score-matching diffusion model can achieve the maximum possible generative bandwidth. From equation 15, it is clear that this happens when $\mathbb{E}[\|\nabla \log p_t(\boldsymbol{x}_t)\|] = 0$, which in turn is obtained if the score vanishes almost everywhere.

 doi:10.20944/preprints202601.0131.v1

To realize this situation, we can consider a data distribution $p_h(\boldsymbol{y})$ to be a centered multivariate normal with variance h^2. In this case, the score function is just:

$$\nabla \log p_t(\boldsymbol{x}_t) = -\frac{\boldsymbol{x}_t}{\sigma^2(t) + h^2}\,, \tag{31}$$

which vanishes everywhere for $h \to \infty$, giving a maximum entropy rate:

$$\dot{\mathbf{H}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \frac{1}{2}\frac{D\nu^2(t)}{\sigma^2(t)}\,. \tag{32}$$

This corresponds to a setting where the particles are free to diffuse since every possible generation is equally likely. From this, we can conclude that the score function has the negative role of suppressing fluctuations along 'unwanted directions' to preserve the statistics of the data and that peaks in the information transfer comes from periods where noise fluctuations are not suppressed. Note that the maximum bandwidth scales with the dimensionality D.

Now consider the case where the distribution of the data is a centered Gaussian in a D_{data}-dimensional subspace with $D_{\text{data}} \leq D$. In this case, the expected norm of the score decomposes as follows

$$\mathbb{E}\left[\|\nabla \log p_t(\boldsymbol{x}_t)\|^2\right] = \frac{D_{\text{data}}}{\sigma^2(t) + h^2} + \frac{D - D_{\text{data}}}{\sigma^2(t)} \;\to\; \frac{D - D_{\text{data}}}{\sigma^2(t)} \tag{33}$$

which leads to the entropy rate

$$\dot{\mathbf{H}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \frac{1}{2}\frac{D_{\text{data}}\nu^2(t)}{\sigma^2(t)}\,. \tag{34}$$

In this case, the score function suppresses entropy reduction in the subspace orthogonal to the data and therefore acts as a linear analog filter. Note that the entropy rate is zero when D_{data} is equal to zero since all the distribution is in this case collapsed into a single point and no 'decision' needs to be made.

4. Statistical Physics, Order Parameters and Phase Transitions

In this section, we will connect the information theoretical concepts we outlined above with concepts from statistical physics such as order parameters, phase transitions and spontaneous symmetry breaking. We will start by studying the paths of fixed-points of the score function and use them to track 'generative decisions' (i.e. bifurcations) along the denoising trajectories. As we will see, the stability of these fixed-points paths is regulated by the Jacobian of the score and it is deeply connected with the conditional entropy production.

4.1. Branching Paths of Fixed-Points and Spontaneous Symmetry Breaking

The fixed-points of the score function are defined by the equation:

$$\nabla \log p_t(\boldsymbol{x}_t^*) = \mathbf{0}. \tag{35}$$

We denote the set of fixed-points at time t as Ψ_t. The solutions of this fixed-point equation can be organized in a set Ω of piecewise continuous paths $\gamma : \mathbb{R}^+ \to \mathbb{R}^d \in \Omega$. To remove ambiguities, we assume that, if $\gamma(\tau)$ is discontinuous at τ_0, then the one-sided limit exists and $\gamma(\tau_0)$ is equal to $\lim_{t\to\tau_0^+}\gamma(t) = \arg\min_{\boldsymbol{x}\in\Psi_{\tau_0}} \lim_{t\to\tau_0^+}[\|\boldsymbol{x} - \gamma(t)\|]$. We know that $\lim_{t\to\infty}\gamma(t) = \mathbf{0}$ for all paths since the zero vector is the only fixed-point of the score of the asymptotic Gaussian distribution. Any two paths $\gamma_1(t)$ and $\gamma_2(t)$ can be proven to overlap for a finite range of time, meaning that $\gamma_1(t) = \gamma_2(t)$ if $t \geq \tau_{1,2} \in \mathbb{R}^+$ (this follows from the results in [35–37] on the number of modes of mixture of normal distributions). We refer to $\tau_{1,2}$ as the *branching time* of the two paths. The branching time of two paths of fixed points can roughly be interpreted as a *decision time* in the generative process, where the sample will be 'pushed' by the noise in either one or the other path during the reverse dynamics. It is therefore insightful to study the behavior of the paths at the branching times. In general, this can happen if

 doi:10.20944/preprints202601.0131.v1

there is a discontinuous jump in a path $\gamma(t)$. Perhaps more interestingly, two paths can also branch continuously at a finite time. This can be studied by analyzing the Jacobian matrix of the score function:

$$J_t(\boldsymbol{x}_t^*) = \nabla^T \nabla \log p_t(\boldsymbol{x}_t^*). \tag{36}$$

We call a path point $\gamma(t)$ stable at time t if $J_t(\gamma(t))$ is negative-definite. We say that the path is stable if this is true for all $t \in \mathbb{R}^+$ except for a countable set of time points t_j where the Jacobian is negative semi-definite. Now consider two stable paths $\gamma_1(t)$ and $\gamma_2(t)$ that branch continuously at time $\tau_{1,2}$. Given the asymptotic separation vector

$$\boldsymbol{v}_{1,2} = \lim_{t \to \tau_{1,2}^-} \frac{(\gamma_2(t) - \gamma_1(t))}{\|\gamma_2(t) - \gamma_1(t)\|},$$

it can be shown that $\boldsymbol{v}^T J_t(\gamma(t))\boldsymbol{v} < 0$ in a finite interval $(\tau_{1,2}, \tau_{1,2} + \epsilon)$ and that

$$\lim_{t \to \tau_{1,2}^+} \boldsymbol{v}^T J_t(\gamma_1(t))\boldsymbol{v} = 0,$$

which implies that the second directional derivative of $D_{\boldsymbol{v}}^2 \log p_t(\boldsymbol{x}_t)$ along $\boldsymbol{v}$ vanishes at the branching point.

Consider now a generative diffusion with an initial distribution given as

$$p_0(\boldsymbol{y}) = \frac{1}{K} \sum_{j=1}^{K} \delta(\boldsymbol{y}^{(j)} - \boldsymbol{y}), \tag{37}$$

with K distinct data-points $\boldsymbol{y}^{(j)} \in \mathbb{R}^d$. In this case, there are exactly K distinct stable fixed-point paths $\gamma_j(t)$, with $\gamma_j(0) = \boldsymbol{y}^{(j)}$. Again, any two paths branch at a finite time $\tau_{j,k}$. For a given t, we can partition the set of data-points in equivalence classes, where two data-points $\boldsymbol{y}^{(j)}$ and $\boldsymbol{y}^{(k)}$ share the same class if their associated path coincide at t. Importantly. Each equivalence class corresponds to an individual fixed-point, which allows us to associate each fixed-point $\boldsymbol{x}^* \in \Psi_t$ to a sub-set of data-points that are, using colorful language, fused together. More precisely, we can express the fixed-points as weighted averages of data-points obtained by solving the self-consistency equation:

$$\boldsymbol{x}^* = \sum_{j=1}^{K} w_j(\boldsymbol{x}^*) \boldsymbol{y}^{(j)} \tag{38}$$

where

$$w_j(\boldsymbol{x}) = \frac{e^{\left(-\left\|\boldsymbol{y}^{(j)}\right\|^2/2 + \boldsymbol{x}^T \boldsymbol{y}^{(j)}\right)/\sigma^2(t)}}{\sum_{k=1}^{K} e^{\left(-\left\|\boldsymbol{y}^{(k)}\right\|^2/2 + \boldsymbol{x}^T \boldsymbol{y}^{(k)}\right)/\sigma^2(t)}}. \tag{39}$$

Note that this average has non-zero weight on all data points, which is why we cannot find the location of the fixed-point solely based on its equivalence class. However, usually the weights corresponding to data-points in the equivalence class will be substantially larger than the other weights and will therefore dominate the average. In summary, we can interpret the set of fixed-points as a decision tree where each branching point roughly coincides with a split between two sets of data points.

An example of spontaneous symmetry breaking happens when the generative path needs to 'decide' between two isolated data-points. Consider again the mixture of delta case and two neighboring data-points $\boldsymbol{y}_1 = \boldsymbol{v}$ and $\boldsymbol{y}_2 = -\boldsymbol{v}$. If the distance between the center of mass of these two points and the nearest external data-point is much larger than $\sigma(t)$, there will be a fixed point approximately

 doi:10.20944/preprints202601.0131.v1

located along the line segment connecting the two points. In these conditions, we can consider the fixed-point equation restricted to the projections on v:

$$x_v^* = \tanh\left(\frac{x_v^* + \phi(x_v^*, t)}{\sigma^2(t)}\right) \tag{40}$$

where $\phi(\boldsymbol{x}_v^*, t)$ encapsulates the interference due to all other data-points, which we, in this example, we assume to be small relative to the norm of the separation vector:

$$\phi(x_v^*, t) = \frac{\sigma^2(t)}{2}\left(\log\left(e^{x_v^*} + \sum_{j\neq 1,2}^{K} y_v^{(k)} e^{\left(-\|\boldsymbol{y}^{(k)}\|^2/2 + x_v y_v^{(k)}\right)/\sigma^2(t)}\right) - x_v^*\right).$$

If we approximate the interference function with constant ϕ using a zero-th order Taylor expansion, Eq. 40 becomes the self-consistency equation of a Curie-Weiss model of magnetism, with temperature $T = \sigma^2(t)$ and external magnetic field ϕ. The solutions of this equation can be visualized as intersection points between a straight line and a hyperbolic tangent (see [12] and [11] for a detail analysis). When ϕ is finite, the system transitions discontinuously from one to two fixed-points, which corresponds to a first-order phase transition in the magnetic system. However, the size of the discontinuity vanishes when $\phi = 0$, when there is an exact symmetry between the two data-points (see Figure 1). This gives rise to a so called *critical phase transition*, where a single fixed-point at $x^* = 0$ continuously splits into two paths $\boldsymbol{x}_1(t)$ and $x_2(t)$ with $\boldsymbol{x}_{1,2}(t - t_c) \sim \pm(t - t_c)^{1/2}$ for $t \to 0$. The loss of stability of the fixed-point at the origin corresponds to the vanishing of the quadratic well around the point:

$$\frac{\partial^2}{\partial x_v^2} \log p_{t_c}(x_{t_c}^*) = 0, \tag{41}$$

where, in this case, $\boldsymbol{x}_{t_c}^* = 0$ for $t < t_c$. The analysis we just carried out involves the breaking of the permutation symmetry between two isolated data-points. On the other hand, if the symmetry is broken along all directions like in the case where the data manifold is a sphere centered at $\boldsymbol{x}_t^*$, Eq. 41 implies that

$$\mathrm{Tr}\left[\nabla^T \nabla \log p_t(\boldsymbol{x}_{t_c}^*)\right] = \nabla \cdot \nabla \log p_t(\boldsymbol{x}_{t_c}^*) = \mathbf{0} \tag{42}$$

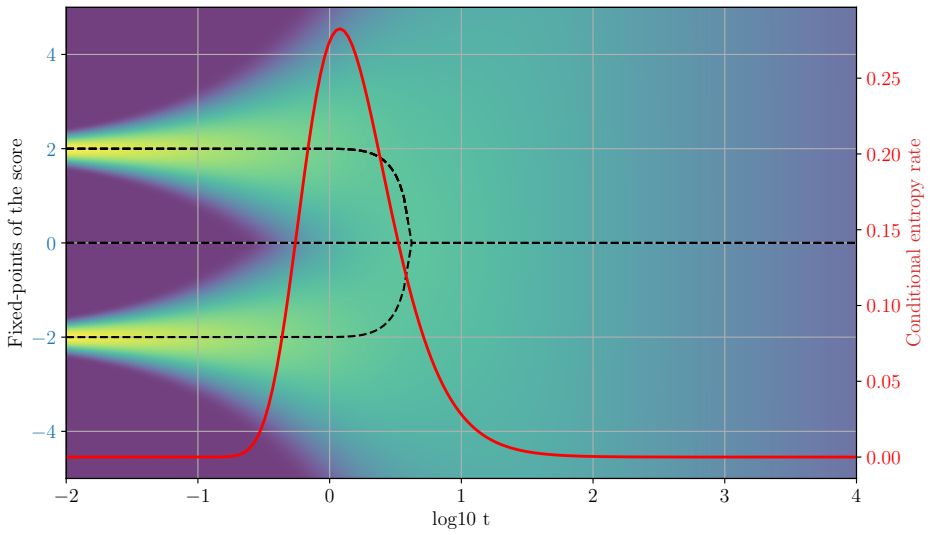

Figure 1. Conditional entropy profile (left) and paths of fixed-points (right) for a mixture of two delta distributions. The color in the background visualizes the (log)density of the process.

Therefore, the change in stability condition can be reformulated as the local vanishing (or suppression in a less symmetric case) of the divergence of the vector field that drives the generative dynamics. The transition from the super-critical ($t > t_c$) and the sub-critical ($t < t_c$) phases then corresponds to a

 doi:10.20944/preprints202601.0131.v1

sign change in the divergence of the vector field (i,e, the score) in the spherically-symmetric case, or a sign change of the divergence restricted to a sub-space in the general case, with the sub-critical regime being characterized by positive eigenvalues of the Jacobi matrix that lead to divergent local trajectories (see Figure 2).

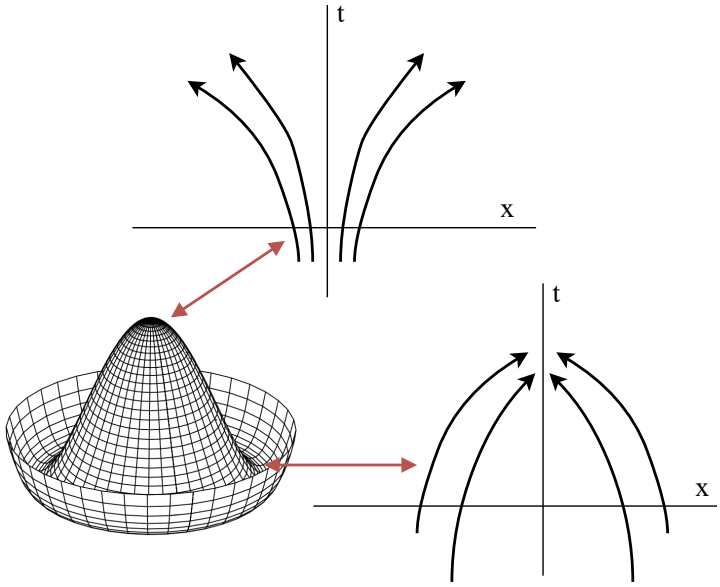

Figure 2. Stability and instability of trajectories in different parts of a symmetry-breaking potential. Generative branching is associated with divergent trajectories.

5. Dynamics of the Generative Trajectories

Around a point $\boldsymbol{x}^*$, the local behavior of the generative trajectories under the deterministic ODE flow dynamics can be characterized by its Lyapunov exponent, which quantifies the separation rate of infinitesimally close trajectories. In particular, the minimal local exponent for a perturbation along a unit vector $\boldsymbol{w}$, which in the reverse dynamics can be defined as

$$l_{\boldsymbol{w}}(t,\boldsymbol{x}) = \lim_{\tau\to\infty}\lim_{\boldsymbol{w}\to\boldsymbol{0}} \frac{1}{\tau}\log\frac{\|\boldsymbol{x}_{t-\tau}(\boldsymbol{x}_t+\boldsymbol{w}) - \boldsymbol{x}_{t-\tau}(\boldsymbol{x}_t)\|}{\|\boldsymbol{w}\|}\,, \tag{43}$$

where $\boldsymbol{x}_{t-\tau}(\boldsymbol{x}_t+\boldsymbol{w})$ denotes a deterministic generative trajectory with the perturbed initial condition. Note that in reality τ cannot tend to infinity in the generative dynamics since time is only defined up to 0. However, we will still consider this limit since we are only interested in the local asymptotic behavior of the linearized dynamics around a bifurcation point. Under the reverse dynamics, when the Lyapunov exponent along $\boldsymbol{w}$ is negative, infinitesimal perturbations are amplified exponentially (at least locally) by the generative dynamics.

We can use the notion of minimal local Lyapunov exponent to formalize the phenomenon of local divergence of trajectories after a spontaneous symmetry-breaking event at t_c. To study the local sub-critical behavior, we consider the linearization of the dynamics around the unstable fixed point for $t < t_c$ and $t_c - t = \epsilon$:

$$l_{\boldsymbol{w}}(t_c+\epsilon,\boldsymbol{x}) = \lim_{\tau\to\infty}\lim_{\boldsymbol{w}\to\boldsymbol{0}} \frac{1}{\tau}\log\frac{\left\|e^{-\tau J_{t_c-\epsilon}(\boldsymbol{x}^*_{t_c-\epsilon})}\boldsymbol{w}\right\|}{\|\boldsymbol{w}\|} = \lambda_{\min}(\boldsymbol{x}^*_{t_c-\epsilon}, t_c+\epsilon)\,, \tag{44}$$

where $\lambda_{\min}(\boldsymbol{x}^*_{t_c-\epsilon}, t_c+\epsilon)$ is the smallest of the eigenvalue of the Jacobi matrix whose eigenvectors overlap with $\boldsymbol{w}$. In the immediate sub-critical phase of a symmetry breaking phase transition, we know that there is a non-empty sub-space spanned by the eigenvector of the Jacobian corresponding

 doi:10.20944/preprints202601.0131.v1

to negative eigenvalues. Therefore, perturbations along this unstable eigen-space will be exponentially amplified by the generative dynamics. In the stochastic case, this can be seen as a critical 'macroscopic amplification' of the infinitesimal noise input, where the noise breaks the symmetry of the generative model. In the deterministic dynamics, the symmetry is instead broken by the amplification of small differences between the generative trajectories.

In general, we will refer to the spectrum of Jacobian eigenvalues $\lambda_j(\boldsymbol{x}_t, t)$ as the local Lyapunov spectrum. As we shall see, this spectrum can be directly related to the conditional entropy production.

5.1. The Global Perspective on Generative Bifurcations

In the previous sections, we characterized the generative dynamics of diffusion models by studying the associated paths of fixed-points in term of their stability and bifurcations, which led us to establish formal connections with the statistical physics of phase transitions and symmetry breaking. However, in high dimension, small volumes around a fixed-point have vanishingly low probability of being visited. In fact, due to the dispersive effect of the noise, the generative trajectories are concentrated on fixed-variance shells around the fixed points. More formally, these set of "typical" points form tubular neighborhoods of the set of fixed-points (see 1). It is therefore unclear how a bifurcation in a path of fixed-points affects the behavior of the generative trajectories, since the analysis we presented in the previous sections was purely local.

To gain insight into the global behavior of the typical generative trajectories, we can study the expected divergence of the vector field at time t

$$\mathrm{div}(t) = \mathbb{E}_{\boldsymbol{x}_t}[\nabla \cdot \nabla \log p_t(\boldsymbol{x}_t)] = \mathbb{E}_{\boldsymbol{x}_t}\left[\mathrm{Tr}\left[\nabla^T \nabla \log p_t(\boldsymbol{x}_t)\right]\right] . \tag{45}$$

If $\mathrm{div}(t)$ is negative, the separation between the generative trajectories will, on average, be contracted by the generative dynamics. The simplest example of this contractive behavior can be studied by considering a data distribution with a single point: $p_0(\boldsymbol{x}_t) = \delta(\boldsymbol{y} - \boldsymbol{c})$. In this case, all trajectories converge to $\boldsymbol{c}$ for $t \to 0$, and we have

$$\mathrm{div}_1(t) = -\frac{D}{\sigma^2(t)} . \tag{46}$$

where D is the dimensionality of the space. In the reverse dynamics, the negative sign implies that the forward process produces a stable dynamics where the particles 'fall' towards the data points.

In the general case, this quantity can be identified with the "trivial component" of the expected divergence since it does not depend on the data but only on the forward process. In the general case, it can be expressed as

$$\mathrm{div}_1(t) = \mathbb{E}_{\boldsymbol{x}_t}\left[\mathrm{Tr}\left[\nabla^T \nabla \log p_t(\boldsymbol{x}_t \mid \boldsymbol{y})\right]\right] . \tag{47}$$

We can therefore study the purely data-dependent part of the expected divergence by subtracting this "trivial component":

$$\Delta\mathrm{div}(t) = \mathrm{div}(t) - \mathrm{div}_1(t) . \tag{48}$$

Intuitively, $\Delta\mathrm{div}(\boldsymbol{x}_t)$ encodes the separation of the typical trajectories in the reverse process due to bifurcations in the generative process, which mirrors the local analysis we carried out in the previous sections at the level of the fixed-points.

Using integration by parts, it is straightforward to connect the expected divergence with the conditional entropy rate

$$\dot{\mathbf{H}}[\boldsymbol{y} \mid \mathbf{x}_t] = \frac{\nu^2(t)}{2} \Delta\mathrm{div}(t) \tag{49}$$

Therefore, the expected data-dependent divergence of the generative trajectories directly determines the conditional entropy rate. From this identity, we can immediately deduce that $\Delta\mathrm{div}(\boldsymbol{x}_t)$ is non-negative valued and consequently that $\mathrm{div}(t) \geq \mathrm{div}_1(t)$.

 doi:10.20944/preprints202601.0131.v1

We can also show that the marginal entropy is produced by the expected divergence

$$\dot{\mathbf{H}}[x_t] = -\frac{\nu^2(t)}{2}\mathrm{div}(t)\,, \tag{50}$$

which implies that $\mathrm{div}(t) \leq 0$ since the marginal entropy is a monotonically increasing function of t under our forward process. This reflects the fact that the forward process always lead to a dispersion of the trajectories, regardless to the nature of the initial distribution. From this, we can conclude that the maximum bandwidth is achieved when

$$\mathrm{div}(t) = \mathbb{E}_{\boldsymbol{x}_t}\Big[\mathrm{Tr}\Big[\nabla^T\nabla \log p_t(\boldsymbol{x}_t)\Big]\Big] \to 0\,. \tag{51}$$

This gives us a clear connection between the local vanishing of the Jacobian in spontaneous symmetry breaking (Eq. 42) with the expected vanishing that corresponds to saturation of the generative bandwidth.

5.2. Information Geometry

The derivation in the previous sub-section suggests a deep connection between the information production and the geometry of the data manifold. We can further analyze this connection by using concepts from information geometry [38]. The key connection is that conditional entropy rate is in fact just the expected value of the trace of the Fisher information matrix, which can be defined as follows:

$$\mathcal{I}_t(\boldsymbol{x}_t) = -\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_t}\Big[\nabla\nabla^T \log p(\boldsymbol{y} \mid \boldsymbol{x}_t)\Big]\,. \tag{52}$$

This quantity quantifies the sensitivity of the posterior distribution $p(\boldsymbol{y} \mid \boldsymbol{x}_t)$ to changes in $\boldsymbol{x}_t$ and can be interpreted as a natural metric tensor on the variable $\boldsymbol{x}_t$. Using Bayes theorem and our simplified forward process, the expression can be rewritten as

$$\mathcal{I}_t(\boldsymbol{x}_t) = \sigma^{-2}(t)\Big(I + \sigma^2(t)J(\boldsymbol{x}_t)\Big)\,, \tag{53}$$

Geometric information such as the manifold dimensionality is encoded in the spectrum of this matrix [22,39–41]. The Fisher information metric provides information on the (local) manifold structure of the data $\boldsymbol{y}$ as seen through the lenses of the noisy state $\boldsymbol{x}_t$. This is easy to see in the case where the data is Gaussian with covariance matrix Σ_0, which gives the formula

$$\mathcal{I}_t = \sigma^{-2}(t)I - \Big(\Sigma_0 + \sigma^2(t)I\Big)^{-1}\,. \tag{54}$$

When $\boldsymbol{y}$ is supported on a D_{data} manifold, the (degenerate) eigenvalue $\lambda_{||}$ corresponding to the orthogonal complement is equal to zero. On the other hand, in the flat limit, the tangent eigenvalues become equal to Σ_0^{-1}. This implies that the dimensionality of the manifold is given by the dimensionality of the eigenspace corresponding to the eigenvalue $\lambda_{\|} = \sigma^{-2}(t)$. In the general case, the eigen-decomposition of $\mathcal{I}(\boldsymbol{x}_t)$ characterizes the local tangent structure of the manifold [40,41].

We can now use these expressions to cast light on the geometry of entropy production. The conditional entropy rate is directly related to the trace of the Fisher information matrix:

$$\dot{\mathbf{H}}[\boldsymbol{y} \mid \boldsymbol{x}_t] = \frac{1}{2}\nu^2(t)\mathbb{E}_{\boldsymbol{x}_t}[\mathrm{Tr}[\mathcal{I}(\boldsymbol{x}_t)]]\,, \tag{55}$$

which reduces to Eq. 34 in the linear manifold case we just considered. From this perspective, it is clear that the reduction in bandwidth is the result of the suppression of the eigenvalues of $\mathcal{I}(\boldsymbol{x}_t)$. This can

 doi:10.20944/preprints202601.0131.v1

also be seen in the general case by re-expressing the entropy rate in terms of the expected eigenvalues of the Jacobi matrix:

$$\dot{\mathbf{H}}[\boldsymbol{y} \mid \boldsymbol{x}_t] = \frac{\nu^2(t)}{2\sigma^2(t)} \left(D + \sigma^2(t) \sum_j \mathbb{E}\left[\lambda_j(\mathbf{x}_t)\right] \right). \tag{56}$$

This equation shows that the entropy production is directly regulated by the spectrum of expected local Lyapunov exponents, as studied in our local analysis.

We can better understand this formula by rewriting it as follows:

$$\dot{\mathbf{H}}[\boldsymbol{y} \mid \boldsymbol{x}_t] = \frac{\nu^2(t)}{2} \sum_j \Big(1/\sigma^2(t) + \mathbb{E}\left[\lambda_j(\mathbf{x}_t)\right] \Big). \tag{57}$$

From this, we can see that conditional entropy production in an eigenspace is fully suppressed when $\mathbb{E}\left[\lambda_j(\mathbf{x}_t)\right] = -1/\sigma^2(t)$, which is the eigenvalue of the Jacobian of the conditional score under the isotropic forward process.

6. A Stochastic Thermodynamic Perspective

A central question in generative diffusion is how uncertainty about the clean sample x_0 is resolved as the model evolves from the noisy state x_t toward the data manifold. As argued throughout this paper, the appropriate notion of inferential uncertainty is the previously discussed conditional entropy $\mathbf{H}[\mathbf{x}_0 \mid \mathbf{x}_t]$ and, more fundamentally, its pathwise realization. The study of pathwise entropy is naturally motivated by ideas from stochastic thermodynamics. However, we believe that the commonly used entropy in stochastic thermodynamics [30,31] is not the correct quantity for understanding generative dynamics. It measures the irreversibility of the forward diffusion, not the uncertainty relevant to generating a single outcome.

For a given point on the trajectory $\mathbf{x}_t$, we define its path-dependent conditional entropy as

$$h_t(\mathbf{x}_t) = -\int p(x_0 \mid \mathbf{x}_t)\, \log p(x_0 \mid \mathbf{x}_t)\, dx_0. \tag{58}$$

This quantity measures the uncertainty experienced along a single generative path. Its expectation is the usual conditional entropy,

$$\mathbb{E}[h_t(\mathbf{x}_t)] = H[\mathbf{x}_0 \mid \mathbf{x}_t],$$

but its fluctuations encode a structure that is invisible to marginal entropies. In particular, as illustrated in Figure 3, the pathwise conditional entropy $h_t(\mathbf{x}_t)$ can locally increase along individual generative trajectories even as the mean conditional entropy decreases, a behavior reminiscent of entropy fluctuations in stochastic thermodynamics. Such effects do not arise in autoregressive models, where each generation step reduces uncertainty about the final sequence by revealing one token, since $\mathbf{H}[\mathbf{x}_{i+1:n} \mid \mathbf{x}_{1:i}] \leq \mathbf{H}[\mathbf{x}_{i+1:n} \mid \mathbf{x}_{1:i}] + \mathbf{H}[\mathbf{x}_i \mid \mathbf{x}_{1:i-1}] = \mathbf{H}[\mathbf{x}_{i:n} \mid \mathbf{x}_{1:i-1}]$. Whether these entropy fluctuations in diffusion-based generation have any practical advantage, however, remains an open question.

To expose this dynamical heterogeneity, we consider the variance of the pathwise conditional entropy,

$$\mathcal{V}_h(t) \;:=\; \mathrm{Var}\left[h_t(\mathbf{x}_t)\right]. \tag{59}$$

 doi:10.20944/preprints202601.0131.v1

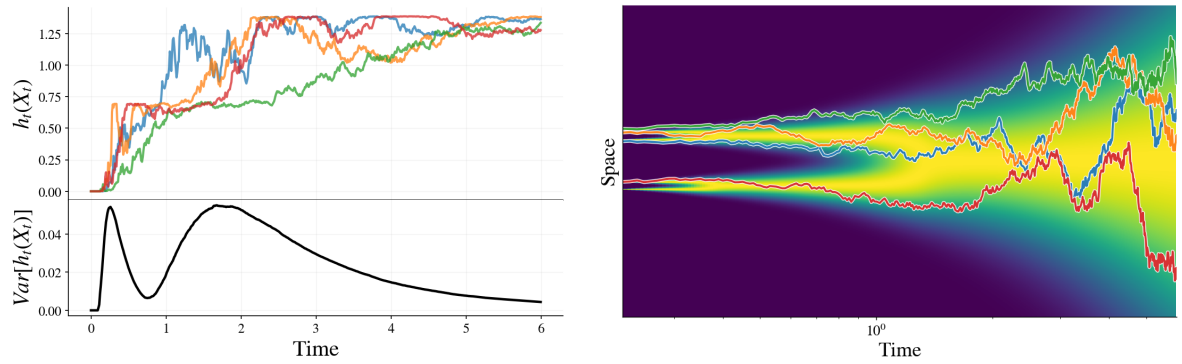

Figure 3. An example of conditional entropies and corresponding paths for a dataset of four data points.

6.1. Variance of Pathwise Conditional Entropy as a Signature of Symmetry Breaking

We find that the variance captures both symmetry-breaking transitions (see figure 3). To gain a better understanding, we explore the general behavior of the variance and demonstrate a connection with the speciation time [20].

Two limits are immediate. At very early times, the noise scale is negligible compared to the curvature of the data manifold. The posterior $p(x_0 \mid x_t)$ is effectively confined to the local tangent plane, behaving as an isotropic Gaussian whose shape is determined solely by the intrinsic dimension D_{data} (we assume that the dimension is uniform across the manifold). Because the entropy of this Gaussian depends on k and t but is insensitive to the specific location on the manifold,

$$h_t(x_t) \approx \text{const} \qquad \Rightarrow \qquad \mathcal{V}_h(t) \approx 0. \tag{60}$$

At very late times, the diffusion has effectively mixed the data distribution: x_t carries little discriminative information about the origin x_0 and the posterior becomes approximately independent of x_t, again implying

$$h_t(x_t) \approx \text{const} \qquad \Rightarrow \qquad \mathcal{V}_h(t) \approx 0. \tag{61}$$

Thus, nontrivial variance can only arise in an intermediate regime where different trajectories resolve uncertainty in different ways.

Furthermore, near a bifurcation/decision time t_c, the ensemble contains a substantial fraction of trajectories that are already decisively committed to a branch and a substantial fraction that remain ambiguous. In this regime, $h_t(x_t)$ becomes broadly distributed (some paths yield low entropy, others high entropy), and $\mathcal{V}_h(t)$ is therefore maximized.

6.1.1. Connection with the Speciation Time

As already hinted, the variance of the pathwise conditional entropy can be used to locate the speciation time for Gaussian-mixture data in the sense of Biroli et al. [20]. We provide a short argument for why $\mathrm{Var}[h_t(\mathbf{x}_t)]$ peaks at the speciation crossover by using a two-region picture of Biroli et al. [20]: points where the class is effectively determined versus maximally mixed. A fully rigorous derivation is possible, but we focus on the essential mechanism and keep the argument streamlined to make the discussion clear and self-contained.

Recall that Biroli et al. [20] define the speciation time t_S as the crossover at which (viewed in the forward noising process) the injected noise blurs the principal "class" direction so that class identity becomes hard to infer; equivalently, by time-reversal, it is the time in the backward process at which trajectories start to commit to one of the classes.

We start by noticing that a single Gaussian has $\mathcal{V}(t) = 0$. If $p_0(x_0) = \mathcal{N}(\mu, \Sigma_0)$ and the forward kernel $q_t(x_t \mid x_0)$ is Gaussian, then $p(x_0 \mid x_t)$ is Gaussian with covariance $\Sigma_{0|t}$ independent of x_t, so $h_t(x_t) = \frac{1}{2} \log\det(2\pi e\, \Sigma_{0|t})$ is constant and the variance vanish.

 doi:10.20944/preprints202601.0131.v1

Let $p_0(x_0) = \sum_{z=1}^{K} \pi_z \mathcal{N}(\mu_z, \Sigma_0)$ with latent index $\mathbf{z} \in \{1, \dots, K\}$, and let $q_t(x_t \mid x_0)$ be the (Gaussian) forward noising kernel. Define the pathwise conditional entropy

$$h_t(x_t) \equiv -\int p(x_0 \mid x_t) \log p(x_0 \mid x_t)\, dx_0, \qquad \mathcal{V}(t) \equiv \mathrm{Var}_{x_t \sim p_t}\big[h_t(x_t)\big].$$

Now, assume that the mixture is well-separated. For each fixed t, the posterior admits the standard separated-mixture approximation

$$p(x_0 \mid x_t) = \sum_{z=1}^{K} w_z(x_t)\, p(x_0 \mid x_t, z), \qquad w_z(x_t) = p(z \mid x_t),$$

where $p(x_0 \mid x_t, z)$ is Gaussian, hence $H\big(p(x_0 \mid x_t, z)\big) = h_G(t)$ for all z. Using the entropy decomposition for (nearly) disjoint mixtures then yields

$$h_t(x_t) \approx h_G(t) + H(\mathbf{z} \mid X_t = x_t), \tag{62}$$

so the fluctuations of $h_t(\mathbf{x}_t)$ are governed by those of the discrete uncertainty $H(\mathbf{z} \mid x_t)$.

Let $w_z(x_t) = p(z \mid x_t)$ be the posterior class weights and fix a small ε. For each t define the two subsets

$$\mathcal{A}_t \doteq \Big\{x_t : \max_z w_z(x_t) \geq 1 - \varepsilon\Big\}, \quad \mathcal{B}_t \doteq \Big\{x_t : \|w(\cdot \mid x_t) - \pi\|_1 \leq \varepsilon\Big\},$$

and write $\alpha(t) \doteq p_t(\mathcal{A}_t)$ and $\beta(t) \doteq p_t(\mathcal{B}_t)$. The dynamical-regimes setting precisely corresponds to the statement that, for the mixture-of-Gaussians class considered in Biroli et al. [20], the regions $\mathcal{A}_t$ and $\mathcal{B}_t$ carry the most mass for times outside a narrow window around the speciation time, i.e. $p_t(\mathcal{A}_t) \approx 1$ or $p_t(\mathcal{B}_t) \approx 1$ except near $t \simeq t_S$.

On $\mathcal{A}_t$, the posterior is nearly one-hot, while on $\mathcal{B}_t$, the posterior is close to π. Hence, neglecting the small boundary region $\mathcal{C}_t \doteq (\mathcal{A}_t \cup \mathcal{B}_t)^c$, the random variable $H(\mathbf{z} \mid \mathbf{x}_t)$ is approximately:

$$H(\mathbf{z} \mid \mathbf{x}_t) \approx \begin{cases} 0 & \mathbf{x}_t \in \mathcal{A}_t, \\ H(\pi) & \mathbf{x}_t \in \mathcal{B}_t. \end{cases}$$

As a consequence, the variance admits the sharp lower/upper control

$$\mathcal{V}(t) \approx \mathrm{Var}_{\mathbf{x}_t}\big(H(\mathbf{z} \mid \mathbf{x}_t)\big) \approx \alpha(t)\,\beta(t)\,\big(H(\pi) - 0\big)^2, \tag{63}$$

where we used that, when $\mathcal{C}_t$ is negligible, $\beta(t) \approx 1 - \alpha(t)$ and the variance of a two-point mixture is the product of the two masses times the squared gap. Equation (63) makes the mechanism transparent: the variance is small when essentially all points are class-diagnostic ($\alpha(t) \approx 1$) or essentially all points are well-mixed ($\alpha(t) \approx 0$), and it is largest when the population splits nontrivially.

To connect this directly to the dynamical-regimes diagnostics, note that the cloning "same-class" probability can be written as

$$\mathsf{P}(t) \doteq \mathbb{E}_{x_t \sim p_t}\Big[\sum_{z=1}^{K} w_z(x_t)^2\Big].$$

Pre-speciation, $w(\cdot \mid x_t)$ is almost one-hot for typical x_t, so $\mathsf{P}(t) \approx 1$ and thus $\alpha(t) \approx 1$, implying $\mathcal{V}(t) \approx 0$. Post-speciation (well-mixed), $w(\cdot \mid x_t) \approx \pi$ for typical x_t, so $\mathsf{P}(t) \approx \sum_z \pi_z^2$ and thus $\alpha(t) \approx 0$, again implying $\mathcal{V}(t) \approx 0$. Since $\mathsf{P}(t)$ varies continuously with t (it is an expectation of a bounded, smooth functional of the time-marginal p_t under the Gaussian kernel), it must interpolate continuously between these limiting values; correspondingly, the mass $\alpha(t)$ must continuously move from ≈ 1 to ≈ 0. Therefore the product $\alpha(t)(1 - \alpha(t))$ necessarily becomes $\Omega(1)$ in the crossover window, and by (63) $\mathcal{V}(t)$ must develop a peak there. Under the dynamical-regimes picture where the transition in $\mathsf{P}(t)$ sharpens with dimension/separation, this peak concentrates around the speciation time t_S.

 doi:10.20944/preprints202601.0131.v1

For more general distributions (e.g. strongly non-Gaussian components), $\mathrm{Var}[h_t(\mathbf{x}_t)]$ can also exhibit an additional early-time peak associated with rapid local "Gaussianization" of p_t under the forward kernel. This peak is absent for the ideal Gaussian case and is suppressed when each component is close to Gaussian.

7. Discussion & Conclusions

This paper has presented a unified framework that connects the dynamics, information theory, and statistical physics of generative diffusion. We have shown that the generative process is governed by the conditional entropy rate, which is directly tied to the expected divergence of the score function's vector field and, equivalently, to the expected squared norm of the score. This quantity captures how uncertainty about the clean sample is resolved during denoising and reveals when the score is suppressed, allowing noise to drive the dynamics. In this view, the branching of generative trajectories arises from noise-induced symmetry-breaking transitions that occur when multiple datapoints remain compatible with the noisy state, and the model is forced to commit to a specific outcome.

By analyzing the fixed points of the score function and their stability, we showed that these generative decisions are formalized as bifurcations of the score field, which can be mapped onto classical symmetry-breaking phase transitions such as those described by mean-field models like the Curie–Weiss magnet. Peaks in the conditional entropy rate coincide with these bifurcation points, marking moments of maximal posterior mixing and heightened sensitivity to noise, where small fluctuations determine the generative branch taken by the system.

Our results also clarify the relationship between generative diffusion and stochastic thermodynamics. While stochastic thermodynamic entropy characterizes the irreversibility of the process, the conditional entropy studied here captures the inferential uncertainty relevant to generating a single sample. At the trajectory level, the pathwise conditional entropy and its variance reveal heterogeneity in how different generative paths resolve uncertainty, with variance peaks emerging precisely during symmetry-breaking events. From this perspective, entropy fluctuations are not incidental but constitute an information-theoretic signature of generative decisions.

In conclusion, generative diffusion can be understood as a dynamical system that progressively breaks symmetries in the energy landscape while regulating the flow of information through posterior mixing. The score function acts as a dynamic filter that suppresses noise along resolved directions while leaving unresolved directions weakly constrained, thereby controlling the generative bandwidth. This perspective provides a coherent explanation of how diffusion models transform noise into structured data and connects the learning dynamics of modern generative models to fundamental principles of information theory and statistical physics.

Beyond conceptual unification, this framework suggests practical implications for model design and analysis. Because entropy production and posterior mixing are directly linked to the score norm, they offer principled signals for identifying critical periods of high information transfer, motivating adaptive training and sampling strategies that target generative decision points [8]. More broadly, the information-thermodynamic perspective developed here provides a natural language for studying memorization, mode formation, and generalization, and may guide the development of future generative models that explicitly leverage controlled symmetry breaking to represent hierarchical and semantic structure.

Author Contributions: Conceptualization, D.S. and L.A.; methodology, D.S. and L.A.; software, D.S. and L.A.; validation, D.S. and L.A.; formal analysis, D.S. and L.A.; investigation, D.S. and L.A.; data curation, D.S. and L.A.; writing—original draft preparation, D.S. and L.A.; writing—review and editing, D.S.; visualization, D.S. and L.A.; project administration, L.A.; funding acquisition, L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

 doi:10.20944/preprints202601.0131.v1

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sohl-Dickstein, J.; Weiss, E.A.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *arXiv preprint arXiv:1503.03585* **2015**.
2. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv preprint arXiv:2011.13456* **2021**.
3. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* **2020**, *33*, 6840–6851.
4. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. *arXiv preprint arXiv:2010.02502* **2022**.
5. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint arXiv:2105.05233* **2021**.
6. Kong, X.; Brekelmans, R.; Ver Steeg, G. Information-Theoretic Diffusion. *arXiv preprint arXiv:2302.03792* **2023**.
7. Kong, X.; Liu, O.; Li, H.; Yogatama, D.; Ver Steeg, G. Interpretable Diffusion via Information Decomposition. *arXiv preprint arXiv:2310.07972* **2023**.
8. Stancevic, D.; Handke, F.; Ambrogioni, L. Entropic Time Schedulers for Generative Diffusion Models. *arXiv preprint arXiv:2504.13612* **2025**.
9. Dieleman, S.; Sartran, L.; Roshannai, A.; Savinov, N.; Ganin, Y.; Richemond, P.H.; Doucet, A.; Strudel, R.; Dyer, C.; Durkan, C.; et al. Continuous Diffusion for Categorical Data. *arXiv preprint arXiv:2211.15089* **2022**.
10. Franzese, G.; Martini, M.; Corallo, G.; Papotti, P.; Michiardi, P. Latent Abstractions in Generative Diffusion Models. *Entropy* **2025**, *27*, 371.
11. Raya, G.; Ambrogioni, L. Spontaneous Symmetry Breaking in Generative Diffusion Models. *arXiv preprint arXiv:2305.19693* **2023**.
12. Ambrogioni, L. The Statistical Thermodynamics of Generative Diffusion Models: Phase Transitions, Symmetry Breaking and Critical Instability. *arXiv preprint arXiv:2310.17467* **2024**.
13. Biroli, G.; Mézard, M. Generative Diffusion in Very Large Dimensions. *Journal of Statistical Mechanics: Theory and Experiment* **2023**, *2023*, 093402.
14. Alaoui, A.E.; Montanari, A.; Sellke, M. Sampling from Mean-Field Gibbs Measures via Diffusion Processes. *arXiv preprint arXiv:2310.08912* **2023**.
15. Huang, B.; Montanari, A.; Pham, H.T. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization. *arXiv preprint arXiv:2404.15651* **2024**.
16. Montanari, A. Sampling, Diffusions, and Stochastic Localization. *arXiv preprint arXiv:2305.10690* **2023**.
17. Benton, J.; De Bortoli, V.; Doucet, A.; Deligiannidis, G. Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization. In Proceedings of the International Conference on Learning Representations, 2024.
18. Sclocchi, A.; Favero, A.; Wyart, M. A Phase Transition in Diffusion Models Reveals the Hierarchical Nature of Data. *Proceedings of the National Academy of Sciences* **2025**, *122*, e2408799121.
19. Sclocchi, A.; Favero, A.; Levi, N.I.; Wyart, M. Probing the Latent Hierarchical Structure of Data via Diffusion Models. *Journal of Statistical Mechanics: Theory and Experiment* **2025**, *2025*, 084005.
20. Biroli, G.; Bonnaire, T.; de Bortoli, V.; Mézard, M. Dynamical Regimes of Diffusion Models. *Nature Communications* **2024**, *15*. https://doi.org/10.1038/s41467-024-54281-3.
21. Bonnaire, T.; Urfin, R.; Biroli, G.; Mézard, M. Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training. *arXiv preprint arXiv:2505.17638* **2025**.
22. Achilli, B.; Ambrogioni, L.; Lucibello, C.; Mézard, M.; Ventura, E. Memorization and Generalization in Generative Diffusion under the Manifold Hypothesis. *Journal of Statistical Mechanics: Theory and Experiment* **2025**, *2025*, 073401.
23. Achilli, B.; Ventura, E.; Silvestri, G.; Pham, B.; Raya, G.; Krotov, D.; Lucibello, C.; Ambrogioni, L. Losing Dimensions: Geometric Memorization in Generative Diffusion. *arXiv preprint arXiv:2410.08727* **2024**.
24. Ambrogioni, L. In Search of Dispersed Memories: Generative Diffusion Models Are Associative Memory Networks. *Entropy* **2024**, *26*, 381.
25. Hoover, B.; Strobelt, H.; Krotov, D.; Hoffman, J.; Kira, Z.; Chau, D.H. Memory in Plain Sight: A Survey of the Uncanny Resemblances Between Diffusion Models and Associative Memories, 2023. Associative Memory & Hopfield Networks in 2023.

 doi:10.20944/preprints202601.0131.v1

26. Hess, J.; Morris, Q. Associative Memory and Generative Diffusion in the Zero-Noise Limit. *arXiv preprint arXiv:2506.05178* **2025**.
27. Jeon, D.; Kim, D.; No, A. Understanding Memorization in Generative Models via Sharpness in Probability Landscapes. *arXiv preprint arXiv:2412.04140* **2024**.
28. Pham, B.; Raya, G.; Negri, M.; Zaki, M.J.; Ambrogioni, L.; Krotov, D. Memorization to Generalization: Emergence of Diffusion Models from Associative Memory. *arXiv preprint arXiv:2505.21777* **2025**.
29. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752* **2022**.
30. Premkumar, A. Neural Entropy. *arXiv preprint arXiv:2409.03817* **2024**.
31. Seifert, U. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Physical review letters* **2005**, *95*, 040602.
32. Ikeda, K.; Uda, T.; Okanohara, D.; Ito, S. Speed-accuracy relations for diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport. *Physical Review X* **2025**, *15*, 031031.
33. Lou, A.; Meng, C.; Ermon, S. Discrete Diffusion Language Modeling by Estimating the Ratios of the Data Distribution. *arXiv preprint arXiv:2305.14627* **2023**.
34. Sahoo, S.; Arriola, M.; Schiff, Y.; Gokaslan, A.; Marroquin, E.; Chiu, J.; Rush, A.; Kuleshov, V. Simple and Effective Masked Diffusion Language Models. *Advances in Neural Information Processing Systems* **2024**, *37*, 130136–130184.
35. Carreira-Perpinan, M.A. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *22*, 1318–1323.
36. Carreira-Perpinán, M.A.; Williams, C.K. On the number of modes of a Gaussian mixture. In Proceedings of the International Conference on Scale-Space Theories in Computer Vision, 2003.
37. Améndola, C.; Engström, A.; Haase, C. Maximum number of modes of Gaussian mixtures. *Information and Inference: A Journal of the IMA* **2020**, *9*, 587–600.
38. Amari, S.i. *Information Geometry and Its Applications*; Vol. 194, Springer, 2016.
39. Kadkhodaie, Z.; Guth, F.; Simoncelli, E.P.; Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *arXiv preprint arXiv:2310.02557* **2023**.
40. Stanczuk, J.P.; Batzolis, G.; Deveney, T.; Schönlieb, C.B. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In Proceedings of the International Conference on Machine Learning, 2024.
41. Ventura, E.; Achilli, B.; Silvestri, G.; Lucibello, C.; Ambrogioni, L. Manifolds, Random Matrices and Spectral Gaps: The Geometric Phases of Generative Diffusion. *arXiv preprint arXiv:2410.05898* **2024**.