A New Look on Diffusion Times for Score-based Generative Models

Giulio Franzese, Simone Rossi, Lixuan Yang, Alessandro Finamore, Dario Rossi, Maurizio Filippone, Pietro Michiardi

A New Look on Diffusion Times for Score-based Generative Models

Giulio Franzese [1] **Simone Rossi** [1] **Lixuan Yang** [2] **Alessandro Finamore** [2] **Dario Rossi** [2] **Maurizio Filippone** [1]
Pietro Michiardi [1]

Abstract

Score-based diffusion models map noise into data using stochastic differential equations. While current practice advocates for a large T to ensure closeness to steady state, a smaller value of T should be preferred for a better approximation of the score-matching objective and computational efficiency. We conjecture, contrary to current belief and corroborated by numerical evidence, that the optimal diffusion times are smaller than current practice.

1. Introduction

Diffusion-based models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2021c; Vahdat et al., 2021; Kingma et al., 2021; Ho et al., 2020; Song et al., 2021a) generate samples from an unknown density p_{data} by reversing a *diffusion process* which injects noise into the data. This diffusion process is a forward Stochastic Differential Equation (SDE)

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t \quad \text{with} \quad \boldsymbol{x}_0 \sim p_{data}\,, \tag{1}$$

where $\boldsymbol{x}_t$ is a random variable at time t, $\boldsymbol{f}(\cdot, t)$ is the *drift term*, $g(\cdot)$ is the *diffusion term* and $\boldsymbol{w}_t$ is a *Wiener process*. We denote the time-varying probability density by $p(\boldsymbol{x}, t)$, by definition $p(\boldsymbol{x}, 0) = p_{data}(\boldsymbol{x})$, and the conditional on the initial condition $\boldsymbol{x}_0$ by $p(\boldsymbol{x}, t \,|\, \boldsymbol{x}_0)$. The forward SDE is usually considered for a sufficiently long *diffusion time* T as in principle, when $T \to \infty$, $p(\boldsymbol{x}, T)$ converges to Gaussian noise. Given initial condition $p(\boldsymbol{x}, T)$, the backward SDE (Anderson, 1982)

$$\mathrm{d}\boldsymbol{x}_t = \left[-\boldsymbol{f}(\boldsymbol{x}_t, t') + g^2(t')\boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t')\right]\mathrm{d}t + g(t')\mathrm{d}\boldsymbol{w}_t \quad t' \stackrel{\text{def}}{=} T - t, \tag{2}$$

after a *reverse diffusion time* T will be distributed as $p_{data}(\boldsymbol{x})$. The time varying density of Eq. (2) is denoted with $q(\boldsymbol{x}, t)$.

Practical considerations on diffusion time. In practice, diffusion models are challenging to work with (Song et al., 2021c). First, a direct access to the true *score* function $\boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t)$ is required in the reverse diffusion is unavailable. This can be solved by approximating it with a parametric function $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, which is trained using the following loss function,

$$\mathcal{L}(\boldsymbol{\theta}) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)}\mathbb{E}_{\sim(1)}g^2(t)\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\nabla}\log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)\|^2\,, \tag{3}$$

where the notation $\mathbb{E}_{\sim(1)}$ means that the expectation is taken with respect to the random process $\boldsymbol{x}_t$ in Eq. (1). Considering affine drift, the term $p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)$ is analytically known and normally distributed for all t (expression in Särkkä & Solin, 2019). Intuitively, the estimation of the *score* is akin to a denoising objective, which operates in a challenging regime. Second, the noise distribution $p(\boldsymbol{x}, T)$ is analytically known only when the diffusion time is $T \to \infty$. The common solution is to replace $p(\boldsymbol{x}, T)$ with a simple distribution $p_{noise}(\boldsymbol{x})$ which, for the classes of SDEs we consider in this work, is a Gaussian distribution. Indeed, in the infinite diffusion time regime, it is possible to derive $p(\boldsymbol{x}, T \to \infty) = p_{noise}(\boldsymbol{x})$ analytically.

[1]Department of Data Science, Eurecom, Biot, France [2]Huawei Technologies, Paris, France. Correspondence to: Giulio Franzese <giulio.franzese@eurecom.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

In the literature, the discrepancy between $p(\boldsymbol{x}, T)$ and $p_{noise}(\boldsymbol{x})$ has been neglected, under the informal assumption of a sufficiently large diffusion time. While this is a valid approach to simulate and generate samples, the reverse diffusion process starts from a different initial condition $q(\boldsymbol{x}, 0)$ and, as a consequence, it converges to a solution $q(\boldsymbol{x}, T)$ that is different from the true $p_{data}(\boldsymbol{x})$. Later, we will expand on this error, but for illustration purposes Fig. 1 shows quantitatively this behavior for a simple 1D toy example $p_{data}(\boldsymbol{x}) = \pi\mathcal{N}(1, 0.1^2) + (1-\pi)\mathcal{N}(3, 0.5^2)$, with $\pi = 0.3$: when T is small, the distribution $p_{noise}(\boldsymbol{x})$ is very different from $p(\boldsymbol{x}, T)$ and samples from $q(\boldsymbol{x}, T)$ exhibit very low likelihood of being generated from $p_{data}(\boldsymbol{x})$. Crucially, Fig. 1 (zoomed region) illustrates an unknown behavior of diffusion models, unveiled in our analysis. In practice, there exists an optimal diffusion time that strikes right the balance between efficient *score* estimation, and sampling quality.

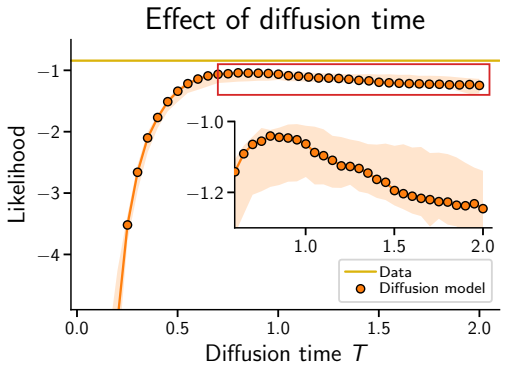

Figure 1: Effect of T on a toy model.

Contribution. In this work we provide a new characterization of score-based diffusion models to obtain a formal understanding of the impact of the diffusion time T. We consider a decomposition of the evidence lower bound (ELBO), which emphasizes the roles of (i) the discrepancy between the "ending" distribution of the diffusion and the "starting" distribution of the reverse diffusion, and of (ii) the *score* matching objective. This allows us to investigate the existence of an optimal diffusion time $< \infty$, differently from current best practice for selecting T.

2. A new look to the variational interpretation of diffusion models

The dynamics of a diffusion model can be studied through the lens of variational inference, bounding the (log-)likelihood using an evidence lower bound (ELBO) (Huang et al., 2021). Our interpretation emphasizes the two main factors affecting the quality of sample generation: an imperfect *score*, and a mismatch, measured in terms of the Kullback-Leibler (KL) divergence, between the noise distribution $p(\boldsymbol{x}, T)$ of the forward process and the distribution p_{noise} used to initialize the backward process.

By manipulating the $\mathcal{L}_{\text{ELBO}}$ derived in (Huang et al., 2021, Eq. (25)), we can write

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq \mathcal{L}_{\text{ELBO}}(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = \mathbb{E}_{\sim(1)} \log p_{noise}(\boldsymbol{x}_T) - I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) + R(T), \tag{4}$$

where

$$R(T) = \frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{\sim(1)} \left[g^2(t) \|\nabla \log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)\|^2 - 2\boldsymbol{f}^\top(\boldsymbol{x}_t, t) \nabla \log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0) \right] \mathrm{d}t \tag{5}$$

$$I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = \frac{1}{2}\int_{t=0}^{T} g^2(t) \mathbb{E}_{\sim(1)} \left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)\|^2 \right] \mathrm{d}t. \tag{6}$$

Note that $R(T)$ depends neither on $\boldsymbol{s}_{\boldsymbol{\theta}}$ nor on p_{noise}, while $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$, or an equivalent reparameterization (Huang et al., 2021; Song et al., 2021b, Eq. (1)), is used to learn the approximated *score*, by optimization of the parameters $\boldsymbol{\theta}$. It is then possible to show that

$$I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) \geq \underbrace{I(\nabla \log p, T)}_{\stackrel{\text{def}}{=} K(T)} = \frac{1}{2}\int_{t=0}^{T} g^2(t) \mathbb{E}_{\sim(1)} \left[\|\nabla \log p(\boldsymbol{x}_t, t) - \nabla \log p(\boldsymbol{x}_t, t \,|\, \boldsymbol{x}_0)\| \right]^2 \mathrm{d}t. \tag{7}$$

Consequently, we can rewrite $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T)) = K(T) + \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$, where $\mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)$ is a positive term that we call the *gap* term, accounting for the practical case of an imperfect *score*, i.e. $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \neq \nabla \log p(\boldsymbol{x}_t, t)$. It also holds that

$$\mathbb{E}_{\sim(1)} \log p_{noise}(\boldsymbol{x}_T) = \int \left[\frac{\log p_{noise}(\boldsymbol{x}) p(\boldsymbol{x}, T)}{p(\boldsymbol{x}, T)} \right] p(\boldsymbol{x}, T) \mathrm{d}\boldsymbol{x} = \mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - \text{KL}\left[\log p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x}) \right]. \tag{8}$$

Therefore, we can rewrite the ELBO in Eq. (4) as

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq -\text{KL}\left[p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x}) \right] + \mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - K(T) + R(T) - \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T). \tag{9}$$

Before concluding our derivation it is necessary to introduce an important observation.

Proposition 1. *Given the stochastic dynamics defined in Eq. (1), it holds that*

$$\mathbb{E}_{\sim(1)} \log p(\boldsymbol{x}_T, T) - K(T) + R(T) = \mathbb{E}_{p_{data}(\boldsymbol{x})} \log p_{data}(\boldsymbol{x}). \tag{10}$$

Intuitively, when $\boldsymbol{s}_{\boldsymbol{\theta}} = \nabla \log p$ and consequently $I(\boldsymbol{s}_{\boldsymbol{\theta}}, T) = K(T)$, Eq. (4) is attained with equality. Moreover when $p_{noise}(\boldsymbol{x}) = p(\boldsymbol{x}, T)$, then $q(\boldsymbol{x}, T) = p_{data}(\boldsymbol{x})$. The formal justification of Proposition 1 is obtained by manipulating the results of Huang et al. (2021) and the equality between $q(\boldsymbol{x}, t')$ and $p(\boldsymbol{x}, t)$ when the score estimation is exact and $q(\boldsymbol{x}, 0) = p(\boldsymbol{x}, T)$. Finally, we can now bound the value of $\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T)$ as

$$\mathbb{E}_{p_{data}(\boldsymbol{x})} \log q(\boldsymbol{x}, T) \geq \underbrace{\mathbb{E}_{p_{data}(\boldsymbol{x})} \log p_{data}(\boldsymbol{x}) - \mathcal{G}(\boldsymbol{s}_{\boldsymbol{\theta}}, T) - \mathrm{KL}\left[p(\boldsymbol{x}, T) \parallel p_{noise}(\boldsymbol{x})\right]}_{\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{s}_{\boldsymbol{\theta}}, T)}. \tag{11}$$

Eq. (11) clearly emphasizes the roles of an approximate score function, through the gap term $\mathcal{G}(\cdot)$, and the discrepancy between the noise distribution of the forward process, and the initial distribution of the reverse process, through the KL term. In the ideal case of perfect *score* matching, the ELBO in Eq. (11) is attained with equality. If, in addition, the initial conditions for the reverse process are ideal, i.e. $q(\boldsymbol{x}, 0) = p(\boldsymbol{x}, T)$, then the results of Anderson (1982) allow us to claim that $q(\boldsymbol{x}, T) = p_{data}(\boldsymbol{x})$.

3. Is there an optimal diffusion time?

While diffusion processes are generally studied for $T \to \infty$, for practical reasons, diffusion times in score-based models have been arbitrarily set to be "sufficiently large" in the literature. In principle, with an arbitrarily flexible model the gap term could be brought arbitrary close to 0 for any diffusion time T. In practice, fixing the model capacity, we expect the gap term to increase with T. Here we conjecture the existence of an optimal diffusion time, which strikes the right balance between the increasing gap $\mathcal{G}(\cdot)$ and the KL term of the ELBO in Eq. (11).

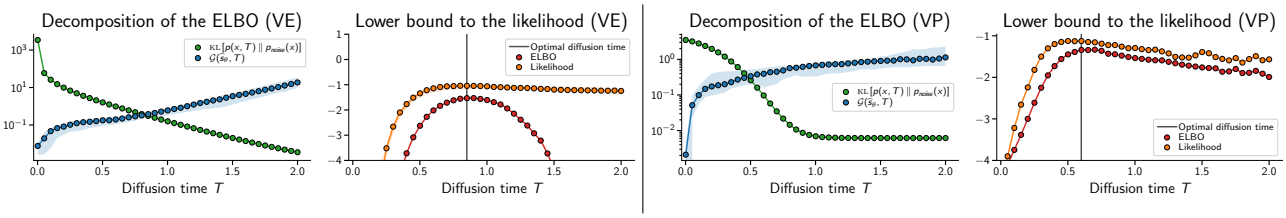

Figure 2: Decomposition of the ELBO and likelihood for a 1D toy model, as a function of diffusion time T.

Empirically, we use Fig. 2 to illustrate the existence of this tradeoff for both variance exploding (VP) and variance preserving (VE) SDEs. In this example, we use the same setup as in Fig. 1, where we have 8192 samples from a mixture of Gaussians with two components as target $p_{data}(\boldsymbol{x})$. The choice of a Gaussian mixture model allows us to write down explicitly the time-varying density $p(\boldsymbol{x}_t, t)$ and to compute the exact score $\nabla \log p(\boldsymbol{x}_t, t)$ (which is done via automatic differentiation, for convenience). On the left we show the ELBO decomposition, while on the right we show the values of the ELBO and of the likelihood as a function of T. We can observe that $\mathcal{G}(\widehat{\boldsymbol{s}}_{\boldsymbol{\theta}}, T)$, the *gap* term obtained with the optimal set of parameters $\boldsymbol{\theta}$ for each T, is an increasing function of T, whereas the KL term is a decreasing function of T. Even in the simple case of a toy example, the tension between small and large values of T is clear. We then verify the validity of our claims: the ELBO is neither maximized by an infinite diffusion time, nor by a "sufficiently large" value. Instead, there exists an optimal diffusion time ($T \approx 0.85$ for VE, $T \approx 0.6$ for VP) which is smaller than what is typically used in practical implementations, i.e. $T = 1.0$.

A comment on score model capacity. We can also investigate the role of the model capacity on the gap term $\mathcal{G}$ as a function of T. With the same toy example as above, we compare three different architectures for the score approximation network. This analysis is possible in this case because, as discussed above, the true score is known. Fig. 3 reports the results: as T grows, more complex models are required to keep the gap term small (i.e. the ELBO higher) and avoid training instabilities.

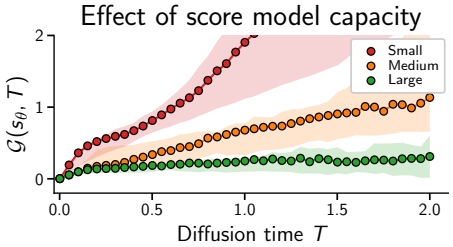

Figure 3: Analysis of the effect of the score model capacity on the gap term for the toy example.

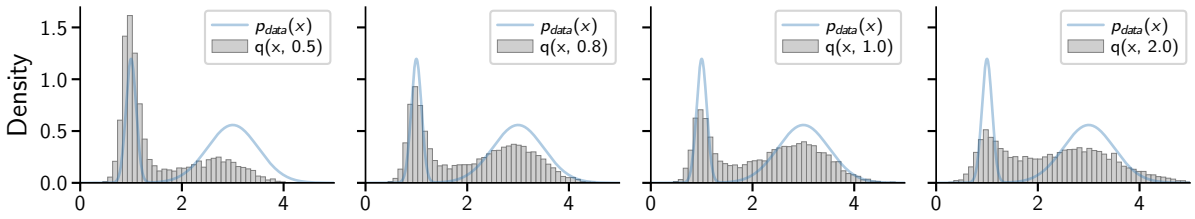

Figure 4: Histograms of generated samples for the toy example with models optimized with different $T \in \{0.5, 0.8, 1.0, 2.0\}$

3.1. Experiments on image datasets

We now present numerical results on the toy example, MNIST and CIFAR10 datasets, in order to support our conjecture on a more challenging benchmark. We follow a standard experimental setup (Song et al., 2021a,b; Huang et al., 2021; Kingma et al., 2021): we use a standard U-Net architecture with time embeddings (Ho et al., 2020) and we report the log-likelihood in terms of bit per dimension (BPD). Note that while the theoretical properties of the models we consider are obtained through the lens of ELBO maximization, the log-likelihood measured in terms of BPD should be considered with care (Theis et al., 2016). Finally, we also report the number of neural function evaluations (NFE) for computing the relevant metrics. Training and evaluation is performed on a small cluster with 16 NVIDIA V100 GPUs. We considered Variance Preserving (VP) SDE with default β_0, β_1 parameter settings. When experimenting on CIFAR10 we considered the NCSN++ architecture as implemented in (Song et al., 2021c). Training of the score matching network has been carried out with the default set of optimizers and schedulers of (Song et al., 2021c), independently of the selected T. For the MNIST dataset we reduced the architecture by considering 64 features, ch_mult $= (1, 2)$ and attention resolutions equal to 8. The optimizer has been selected as the one in the CIFAR10 experiment but the warmup has been reduced to 1000 and the total number of iterations to 65000.

Exploring different diffusion times. We look for further empirical evidence of the existence of an optimal time. The histograms of toy example samples, depicted in Fig. 4, clearly indicates that it does not hold that the larger the T, the better the quality. We shall focus on the baseline model (Song et al., 2021c); results are reported in Table 1. For MNIST, we observe how times $T = 0.6$ and $T = 1.0$ have comparable performance in terms of BPD, implying that any $T \geq 0.6$ is at best unnecessary and generally detrimental. Similarly, for CIFAR10, it is possible to notice that the best value of BPD is achieved for $T = 0.6$, outperforming all other values.

Table 1: Optimal T

Dataset	Time T	BPD (↓)	NFE (↓)
MNIST	1.0	1.16	300
	0.6	**1.16**	258
	0.4	1.25	235
	0.2	1.75	191
CIFAR10	1.0	3.09	221
	0.6	**3.07**	200
	0.4	3.09	187
	0.2	3.38	176

Training and sampling efficiency Reducing T has the benefits of reducing training and sampling cost. For training, smaller times T allows to use simpler parametric score models. Sampling speed benefits are evident from Table 1. Notice that when considering the SDE version of the methods the number of sampling steps can decrease linearly with T, in accordance with known theory (Kloeden & Platen, 1995). Consequently, if our conjecture holds, reducing the diffusion time T can be beneficial for **both** log-likelihood performance and sampling/training efficiency.

4. Conclusion

Diffusion-based generative models emerged as competitive approaches. The standard understanding of diffusion-based models is that smaller values of T are preferable for efficiency reasons, but sufficiently large T are required to reduce approximation errors of the forward dynamics. Starting from a variational interpretation, we proposed to explore the key idea of considering the diffusion time T as a free variable, which should be chosen appropriately. A first empirical validation corroborates our conjecture indicating that reducing T to smaller than usual values can improve performance and efficiency. Further theoretical analyses and thorough experimental validations are required to fully validate our discussion.

Acknowledgements MF gratefully acknowledges support from the AXA Research Fund and from the Agence Nationale de la Recherche (grant ANR-18-CE46-0002 and ANR-19-P3IA-0002).

References

Anderson, B. D. Reverse-Time Diffusion Equation Models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Huang, C.-W., Lim, J. H., and Courville, A. C. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22863–22876. Curran Associates, Inc., 2021.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707. Curran Associates, Inc., 2021.

Kloeden, P. E. and Platen, E. Numerical solution of Stochastic Differential Equations. 1995.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021a.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021c.

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

Theis, L., van den Oord, A., and Bethge, M. A Note on the Evaluation of Generative Models. In Bengio, Y. and LeCun, Y. (eds.), *International Conference on Learning Representations*, 2016.

Vahdat, A., Kreis, K., and Kautz, J. Score-based Generative Modeling in Latent Space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302. Curran Associates, Inc., 2021.