Demystifying Variational Diffusion Models

Other titles in Foundations and Trends® in Computer Graphics and Vision

Step-by-Step Diffusion: An Elementary Tutorial
Preetum Nakkiran, Arwen Bradley, Hattie Zhou and Madhu Advani
ISBN: 978-1-63828-534-2

Tutorial on Diffusion Models for Imaging and Vision
Stanley Chan
ISBN: 978-1-63828-432-1

Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research
Timnit Gebru and Remi Denton
ISBN: 978-1-63828-354-6

Multimodal Foundation Models: From Specialists to General-Purpose Assistants
Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang and Jianfeng Gao
ISBN: 978-1-63828-336-2

Computational Imaging Through Atmospheric Turbulence
Stanley H. Chan and Nicholas Chimitt
ISBN: 978-1-63828-999-9

Towards Better User Studies in Computer Graphics and Vision
Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka and Yile Zhang
ISBN: 978-1-63828-172-6

Demystifying Variational Diffusion Models

Fabio De Sousa Ribeiro
Imperial College London
f.de-sousa-ribeiro@imperial.ac.uk

Ben Glocker
Imperial College London
b.glocker@imperial.ac.uk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

F. De Sousa Ribeiro and B. Glocker. *Demystifying Variational Diffusion Models*. Foundations and Trends® in Computer Graphics and Vision, vol. 17, no. 2, pp. 76–170, 2025.

ISBN: 978-1-63828-561-8

Foundations and Trends® in Computer Graphics and Vision
Volume 17, Issue 2, 2025

Editorial Board

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2025, Volume 17, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Contents

Demystifying Variational Diffusion Models

Fabio De Sousa Ribeiro and Ben Glocker

Imperial College London, UK; f.de-sousa-ribeiro@imperial.ac.uk, b.glocker@imperial.ac.uk

ABSTRACT

Despite the growing interest in diffusion models, gaining a deep understanding of the model class remains an elusive endeavour, particularly for the uninitiated in non-equilibrium statistical physics. Thanks to the rapid rate of progress in the field, most existing work on diffusion models focuses on either applications or theoretical contributions. Unfortunately, the theoretical material is often inaccessible to practitioners and new researchers, leading to a risk of superficial understanding in ongoing research. Given that diffusion models are now an indispensable tool, a clear and consolidating perspective on the model class is needed to properly contextualize recent advances in generative modelling and lower the barrier to entry for new researchers. To that end, we revisit predecessors to diffusion models, such as hierarchical latent variable models, and synthesize a holistic perspective using only directed graphical modelling and variational inference principles. The resulting narrative is easier to follow as it imposes fewer prerequisites on the average reader relative to the view from non-equilibrium thermodynamics or stochastic differential equations.

Fabio De Sousa Ribeiro and Ben Glocker (2025), "Demystifying Variational Diffusion Models", Foundations and Trends® in Computer Graphics and Vision: Vol. 17, No. 2, pp 76–170. DOI: 10.1561/0600000113.

1

Introduction

A generative model is a simulation of a data-generating process. Understanding the true generative process of data is valuable as it naturally reveals the causal relationships in the world. These causal relationships are advantageous as they tend to generalize more effectively to new situations than mere correlations, which may be spurious and unreliable. Generative modelling typically consists of using data from observations of $\mathbf{x}$ to estimate the marginal distribution $p(\mathbf{x})$. Knowing $p(\mathbf{x})$ facilitates many useful tasks, such as: (i) sample generation, (ii) density estimation, (iii) compression, (iv) data imputation, (v) model selection, etc. As $p(\mathbf{x})$ is typically unknown and/or intractable, we often have to approximate it with a model $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p(\mathbf{x})$, by optimizing some parameters $\boldsymbol{\theta}$. Although various generative modelling strategies exist, diffusion models [14, 60] have emerged as the latest dominant paradigm. With that said, gaining a deep understanding of the model class remains an elusive endeavour, particularly for the uninitiated in non-equilibrium statistical physics.

Thanks to the rapid rate of progress in the field, existing work on diffusion models focuses on either applications or theoretical contributions. However, research material on diffusion is often inaccessible to practitioners and new researchers. Given that diffusion models are now

an indispensable tool, we argue that a clear, consolidating perspective on the model class is needed to properly contextualize recent advances in generative modelling and lower the barrier to entry. To that end, we revisit predecessors to diffusion models like hierarchical latent variable models (HLVMs) [57, 61, 70], and synthesize a holistic perspective using *only* directed graphical modelling and variational inference principles. The resulting narrative is easier to follow as it imposes fewer prerequisites on the reader relative to the view from non-equilibrium thermodynamics [60] or stochastic differential equations (SDEs) [62, 66]. Other variational perspectives on diffusion have been studied [21, 28, 69], but their expositions are optimized for technical and empirical contributions to the model class rather than accessibility. A notable exception is the technical review by Luo [36]; however, our account is far more comprehensive, covers a lot more recent material, and is more mathematically consistent with the seminal works in the field [28, 29].

We begin our exposition by revisiting deep latent variable models [30, 48] and their hierarchical counterparts (Section 2). We then highlight the difficulties with bottom-up inference procedures for even modestly deep hierarchies and present a compelling argument in favour of the *top-down* hierarchical model using a concept called *generative feedback* (Section 2.3). This contrasts with prior work [5, 36], which offers an incomplete efficiency-based view. We then show that the top-down hierarchy is ubiquitous in both classical HLVMs [32, 61, 70] and diffusion models. We explain how both model classes share optimization objectives and offer an intuitive understanding of diffusion models as a specific instantiation of HLVMs with top-down inference. In Section 2.5, we reproduce the *hole problem* in LVMs, explain how diffusion models overcome it by construction, and stress its importance for sample quality. In Section 3, we provide a comprehensive account of modern diffusion models from the top-down hierarchy perspective, and in Section 5, we conclude with a forward-looking discussion.

Appendix

A

Notation and Extras

A.1 Notation

Symbol	Description	Section
$\mathbf{x}$	Observed datapoint, e.g. input image	§1
t	Time index variable $t \in \{1, 2, \ldots, T\}$, or $t \in [0, 1]$ for continuous-time	§2.2
$\mathbf{z}_t$	Latent variable at time t	§2.2
$\mathbf{z}_{1:T}$	Finite set of latent variables representing $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$	§2.2
$\mathbf{z}_{0:1}$	Set of latent variables in continuous-time from $t = 0$ to $t = 1$	§3.1
α_t	Noise schedule coefficient $\alpha_t \in (0, 1)$	§3.1
σ_t^2	Noise schedule variance $\sigma_t^2 \in (0, 1)$	§3.1
$\boldsymbol{\epsilon}_t$	Isotropic random noise, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$	§1

$\mathrm{SNR}(t)$	Signal-to-noise ratio (SNR) function at time t, defined as α_t^2/σ_t^2	§A.2
$q(\mathbf{z}_t \mid \mathbf{x})$	Latent variable distribution given $\mathbf{x}$	§3.1
$q(\mathbf{z}_t \mid \mathbf{z}_s)$	Transition distribution from time s to time t, where $s < t$	§3.2
$\alpha_{t\|s}$	Transition coefficient from time s to t	§3.2
$\sigma_{t\|s}^2$	Variance of transition distribution	§3.2
$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x})$	Top-down posterior distribution at time s	§2.4
$\boldsymbol{\mu}_Q(\mathbf{z}_t, \mathbf{x}; s, t)$	Mean of top-down posterior distribution at time s; $\boldsymbol{\mu}_Q$ for short	§3.3
$\sigma_Q^2(s, t)$	Variance of top-down posterior distribution; σ_Q^2 for short	§3.3
$p(\mathbf{z}_s \mid \mathbf{z}_t)$	Generative transition distribution defined as $q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x} = \hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t))$	§3.4
$p(\mathbf{x} \mid \mathbf{z}_0)$	Observation likelihood (e.g. input image), analogous to $p(\mathbf{x} \mid \mathbf{z}_1)$ in discrete-time	§3.4
$\boldsymbol{\phi}$	Variational parameters related to $q_{\boldsymbol{\phi}}$	§1
$\boldsymbol{\theta}$	Model parameters pertaining to $p_{\boldsymbol{\theta}}$	§1
$\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$	Denoising model mapping any $\mathbf{z}_t$ to $\mathbf{x}$	§3.5
$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$	Noise prediction model, which approximates $\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$	§3.5
$\hat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$	Score prediction model, equivalent to $-\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t, t)/\sigma_t$	§3.5
$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t)$	Predicted posterior mean at time $s < t$	§3.5
$\mathrm{VLB}(\mathbf{x})$	Single-datapoint variational lower bound; also denoted as $\mathrm{ELBO}(\mathbf{x})$	§1
$\mathcal{L}_T(\mathbf{x})$	Discrete-time diffusion loss	§3.6
$\mathcal{L}_\infty(\mathbf{x})$	Continuous-time diffusion loss	§3.9

$\mathcal{L}_w(\mathbf{x})$	Weighted diffusion loss; also $\mathcal{L}_\infty(\mathbf{x}, w)$	§4.4
$\boldsymbol{\gamma}_{\boldsymbol{\eta}}(t)$	Neural network with parameters $\boldsymbol{\eta}$ for learning the noise schedule	§A.2
$w(\cdot)$	Noise level weighting function	§4.4
λ	Logarithm of $\mathrm{SNR}(t)$; also λ_t	§4.5
$f_\lambda(t)$	Noise schedule function, mapping t to λ	§4.5
$\lambda_{\min}$	Lowest log SNR given by $f_\lambda(t=1)$	§4.5
$\lambda_{\max}$	Highest log SNR given by $f_\lambda(t=0)$	§4.5
$p(\lambda)$	Density over noise levels	§4.5
$\mathcal{L}(t;\mathbf{x})$	Joint KL divergence up to time t	§4.7
$p_w(t)$	Augmentation kernel specified by $w(\cdot)$	§4.7

A.2 Learning the Noise Schedule

Perturbing data with multiple noise scales and choosing an appropriate *noise schedule* is instrumental to the success of diffusion models. The noise schedule of the forward process is typically pre-specified and has no learnable parameters, however, VDMs learn the noise schedule via the parameterization:

$$\sigma_t^2 = \mathrm{sigmoid}\left(\gamma_{\boldsymbol{\eta}}(t)\right), \tag{A.1}$$

where $\gamma_{\boldsymbol{\eta}}(t)$ is a *monotonic* neural network comprised of linear layers with weights $\boldsymbol{\eta}$ restricted to be positive. A monotonic function is a function defined on a subset of the real numbers which is either entirely non-increasing or entirely non-decreasing. As explained later, the noise schedule can be conveniently parameterized in terms of the signal-to-noise ratio. The signal-to-noise ratio (SNR) is defined as $\mathrm{SNR}(t) = \alpha_t^2/\sigma_t^2$, and since $\mathbf{z}_t$ grow noisier over time we have that: $\mathrm{SNR}(t) < \mathrm{SNR}(s)$ for any $t > s$.

To remain consistent with prior work and avoid notational clutter, we may use the same symbols to denote random variables and their outcomes whenever our intentions can be clearly understood from context.

For now, we provide some straightforward derivations of the expressions for α_t^2 and $\mathrm{SNR}(t)$ as a function of $\gamma_{\boldsymbol{\eta}}(t)$. Recall that in a variance-preserving diffusion process $\alpha_t^2 = 1 - \sigma_t^2$, therefore:

$$\alpha_t^2 = 1 - \sigma_t^2 \tag{A.2}$$

$$= 1 - \mathrm{sigmoid}\left(\gamma_{\boldsymbol{\eta}}(t)\right) \tag{A.3}$$

$$\implies \alpha_t^2 = \mathrm{sigmoid}\left(-\gamma_{\boldsymbol{\eta}}(t)\right), \tag{A.4}$$

as for an input $x \in \mathbb{R}$ the following holds

$$1 - \mathrm{sigmoid}\left(x\right) = 1 - \frac{1}{1+e^{-x}} \tag{A.5}$$

$$= \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}} \tag{A.6}$$

$$= \frac{e^{-x}}{1+e^{-x}} \cdot \frac{e^{x}}{e^{x}} \tag{A.7}$$

$$= \mathrm{sigmoid}\left(-x\right). \tag{A.8}$$

To derive $\mathrm{SNR}(t)$ as a function of $\gamma_{\boldsymbol{\eta}}(t)$, we simply substitute in the above equations and simplify:

$$\mathrm{SNR}(t) = \frac{\alpha_t^2}{\sigma_t^2} = \frac{\mathrm{sigmoid}\left(-\gamma_{\boldsymbol{\eta}}(t)\right)}{\mathrm{sigmoid}\left(\gamma_{\boldsymbol{\eta}}(t)\right)} \qquad \text{(by definition)} \tag{A.9}$$

$$= \frac{(1+e^{\gamma_{\boldsymbol{\eta}}(t)})^{-1}}{(1+e^{-\gamma_{\boldsymbol{\eta}}(t)})^{-1}} \tag{A.10}$$

$$= \frac{1+e^{-\gamma_{\boldsymbol{\eta}}(t)}}{1+e^{\gamma_{\boldsymbol{\eta}}(t)}} \tag{A.11}$$

$$= \frac{\frac{e^{\gamma_{\boldsymbol{\eta}}(t)}}{e^{\gamma_{\boldsymbol{\eta}}(t)}} + \frac{1}{e^{\gamma_{\boldsymbol{\eta}}(t)}}}{1+e^{\gamma_{\boldsymbol{\eta}}(t)}} \cdot \frac{e^{\gamma_{\boldsymbol{\eta}}(t)}}{e^{\gamma_{\boldsymbol{\eta}}(t)}} \tag{A.12}$$

$$= \frac{e^{\gamma_{\boldsymbol{\eta}}(t)}+1}{e^{\gamma_{\boldsymbol{\eta}}(t)}(1+e^{\gamma_{\boldsymbol{\eta}}(t)})} \tag{A.13}$$

$$= \frac{1}{e^{\gamma_{\boldsymbol{\eta}}(t)}}, \tag{A.14}$$

which is equivalently expressed as $\mathrm{SNR}(t) = \exp(-\gamma_{\boldsymbol{\eta}}(t))$.

A.3 Numerically Stable Primitives

The closed-form expressions for the mean and variance of $p(\mathbf{z}_s \mid \mathbf{z}_t)$ can be further simplified to include more numerically stable functions like $\text{expm1}(\cdot) = \exp(\cdot) - 1$, which are available in standard numerical packages. The resulting simplified expressions – which we derive in detail next – enable more numerically stable implementations as highlighted by [28].

Recall from Appendix A.2 that the noise schedule parameters are given by: $\sigma_t^2 = \text{sigmoid}(\gamma_{\boldsymbol{\eta}}(t))$, and $\alpha_t^2 = \text{sigmoid}(-\gamma_{\boldsymbol{\eta}}(t))$, for any t. For brevity, let s and t be shorthand notation for $\gamma_{\boldsymbol{\eta}}(s)$ and $\gamma_{\boldsymbol{\eta}}(t)$ respectively. The posterior variance simplifies to:

$$\sigma_Q^2(s,t) = \frac{\sigma_{t|s}^2 \sigma_s^2}{\sigma_t^2} = \frac{\sigma_s^2 \left(\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2\right)}{\sigma_t^2} \tag{A.15}$$

$$= \frac{\frac{1}{1+e^{-s}} \cdot \left(\frac{1}{1+e^{-t}} - \frac{(1+e^t)^{-1}}{(1+e^s)^{-1}} \cdot \frac{1}{1+e^{-s}}\right)}{\frac{1}{1+e^{-t}}}$$

(cancel denominator) (A.16)

$$= \left(1+e^{-t}\right) \cdot \frac{1}{1+e^{-s}} \cdot \left(\frac{1}{1+e^{-t}} - \frac{1+e^s}{1+e^t} \cdot \frac{1}{1+e^{-s}}\right)$$

(distribute $1+e^{-t}$) (A.17)

$$= \frac{1}{1+e^{-s}} \cdot \left(1 - \frac{1+e^s}{1+e^t} \cdot \frac{1+e^{-t}}{1+e^{-s}}\right) \tag{A.18}$$

$$= \frac{1}{1+e^{-s}} \cdot \left(1 - \frac{e^s\left(1+e^{-s}\right)}{1+e^t} \cdot \frac{e^{-t}\left(1+e^t\right)}{1+e^{-s}}\right)$$

(cancel common factors) (A.19)

$$= \frac{1}{1+e^{-s}} \cdot \left(1 - e^{s-t}\right) \tag{A.20}$$

$$= \sigma_s^2 \cdot \left(-\text{expm1}\left(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t)\right)\right).$$

($\text{expm1}(\cdot) = \exp(\cdot) - 1$) (A.21)

The posterior mean – under a noise-prediction model $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)$ – simplifies in a similar fashion to:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t; s, t) = \frac{1}{\alpha_{t|s}}\mathbf{z}_t - \frac{\sigma^2_{t|s}}{\alpha_{t|s}\sigma_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t) \tag{A.22}$$

$$= \frac{\alpha_s}{\alpha_t}\left(\mathbf{z}_t - \frac{\sigma^2_{t|s}}{\sigma_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right) \tag{A.23}$$

$$= \frac{\alpha_s}{\alpha_t}\left(\mathbf{z}_t - \frac{\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2}{\sigma_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right)$$
$$\text{(substituting } \sigma^2_{t|s} = \sigma_t^2 - \alpha^2_{t|s}\sigma_s^2\text{)} \tag{A.24}$$

$$= \frac{\alpha_s}{\alpha_t}\left(\mathbf{z}_t - \frac{\frac{1}{1+e^{-t}} - \frac{1+e^s}{1+e^t}\cdot\frac{1}{1+e^{-s}}}{\sqrt{\frac{1}{1+e^{-t}}}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right) \tag{A.25}$$

$$= \frac{\alpha_s}{\alpha_t}\Bigg(\mathbf{z}_t - (1+e^{-t})\cdot\sqrt{\frac{1}{1+e^{-t}}} \tag{A.26}$$

$$\cdot\left(\frac{1}{1+e^{-t}} - \frac{1+e^s}{1+e^t}\cdot\frac{1}{1+e^{-s}}\right)\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\Bigg) \tag{A.27}$$

$$= \frac{\alpha_s}{\alpha_t}\left(\mathbf{z}_t - \sigma_t\left(1 - e^{s-t}\right)\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right) \tag{A.28}$$

$$= \frac{\alpha_s}{\alpha_t}\left(\mathbf{z}_t + \sigma_t \text{expm1}\left(\gamma_{\boldsymbol{\eta}}(s) - \gamma_{\boldsymbol{\eta}}(t)\right)\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\right), \tag{A.29}$$

where Equation (A.27) simplifies significantly via the same logical steps in Equations (A.17)-(A.20) above.

A.3.1 Numerically Stable Loss Estimator

The estimator of the discrete-time diffusion loss can be made more numerically stable in practice by re-expressing the constant term inside the expectation using more numerically stable primitives. Specifically:

$$\frac{\text{SNR}(s)}{\text{SNR}(t)} - 1 = \frac{\alpha_s^2}{\sigma_s^2} \div \frac{\alpha_t^2}{\sigma_t^2} - 1 \tag{A.30}$$

$$= \frac{\alpha_s^2 \sigma_t^2}{\alpha_t^2 \sigma_s^2} - 1 \tag{A.31}$$

$$= \frac{\text{sigmoid}(-\gamma_{\boldsymbol{\eta}}(s)) \cdot \text{sigmoid}(\gamma_{\boldsymbol{\eta}}(t))}{\text{sigmoid}(-\gamma_{\boldsymbol{\eta}}(t)) \cdot \text{sigmoid}(\gamma_{\boldsymbol{\eta}}(s))} - 1, \tag{A.32}$$

letting s and t denote $\gamma_{\boldsymbol{\eta}}(s)$ and $\gamma_{\boldsymbol{\eta}}(t)$ for brevity we have:

$$\frac{\frac{1}{1+e^s} \cdot \frac{1}{1+e^{-t}}}{\frac{1}{1+e^t} \cdot \frac{1}{1+e^{-s}}} - 1 = \frac{(1+e^t)\,(1+e^{-s})}{(1+e^s)\,(1+e^{-t})} - 1 \tag{A.33}$$

$$= \frac{e^t\,(1+e^{-t})\,e^{-s}\,(1+e^s)}{(1+e^s)\,(1+e^{-t})} - 1 \tag{A.34}$$

$$= e^t e^{-s} - 1 \tag{A.35}$$

$$= \text{expm1}\left(\gamma_{\boldsymbol{\eta}}(t) - \gamma_{\boldsymbol{\eta}}(s)\right). \tag{A.36}$$

Substituting the above back into the (noise-prediction-based) diffusion loss estimator gives:

$$\mathcal{L}_T(\mathbf{x}) = \frac{T}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),i\sim U\{1,T\}}\Big[\tag{A.37}$$

$$\text{expm1}\left(\gamma_{\boldsymbol{\eta}}(t) - \gamma_{\boldsymbol{\eta}}(s)\right) \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_t; t)\|_2^2\Big], \tag{A.38}$$

which is the final form of the objective we wanted to show.

A.3.2 Dealing with Edge Effects

There is an edge effect at diffusion time $t = 0$, possibly causing numerical issues [60, 62], which we can avoid by setting the likelihood term to:

$$p(\mathbf{x} \mid \mathbf{z}_1) = \frac{q(\mathbf{z}_1 \mid \mathbf{x})p(\mathbf{x})}{p(\mathbf{z}_1)}, \tag{A.39}$$

and removing it from the variational lower bound. In discrete-time, this looks like:

$$\text{VLB} = \mathbb{E}_{q(\mathbf{z}_{1:T},\mathbf{x})}\left[\log \frac{p(\mathbf{x} \mid \mathbf{z}_1)}{q(\mathbf{z}_1 \mid \mathbf{x})} + \log p(\mathbf{z}_T) + \sum_{t=2}^{T} \log \frac{p(\mathbf{z}_{t-1} \mid \mathbf{z}_t)}{q(\mathbf{z}_t \mid \mathbf{z}_{t-1})}\right] \tag{A.40}$$

$$= \mathbb{E}_{q(\mathbf{z}_1,\mathbf{x})}\left[\log \frac{q(\mathbf{z}_1 \mid \mathbf{x})p(\mathbf{x})}{q(\mathbf{z}_1 \mid \mathbf{x})p(\mathbf{z}_1)}\right] \tag{A.41}$$

$$+ \mathbb{E}_{q(\mathbf{z}_{1:T}, \mathbf{x})} \left[\log p(\mathbf{z}_T) + \sum_{t=2}^{T} \log \frac{p(\mathbf{z}_{t-1} \mid \mathbf{z}_t)}{q(\mathbf{z}_t \mid \mathbf{z}_{t-1})} \right]. \quad \text{(A.42)}$$

The left-hand side (LHS) term above cancels out as the SNR $\to \infty$ (i.e. $\alpha_1 \to 1$ and $\sigma_1 \to 0$) since the least noisy latent variable $\mathbf{z}_1 = \alpha_1 \mathbf{x} + \sigma_1 \boldsymbol{\epsilon}$ approaches $\mathbf{x}$, meaning $p(\mathbf{z}_1) \approx p(\mathbf{x})$. Note that p in $p(\mathbf{z}_1)$ and $p(\mathbf{x})$ above refers to the (tractable) prior distribution of choice.

For continuous-time diffusion where $T \to \infty$ and $t \in [0, 1]$, we have that learning a model $p(\mathbf{z}_0)$ is practically equivalent to learning a model $p(\mathbf{x})$ since $\mathbf{z}_0$ (the least noisy latent variable) is almost identical to $\mathbf{x}$ in the limit given large enough log-SNR $\lambda_{\max} = \log(\alpha_0^2/\sigma_0^2)$. However, if one chooses to learn the noise schedule rather than fixing it, the $p(\mathbf{x} \mid \mathbf{z}_0)$ term may need to be incorporated back into the VLB objective, representing a final discrete step from latent space to image space. This manifests as some variation of a decoding step in both VDMs [28] and score-based diffusion models [62].

A.4 Equivalence of Diffusion Specifications

Kingma *et al.* [28] elaborate on the equivalence of diffusion noise-schedule specifications using the following straightforward example. Firstly, the change of variables we used implies that σ_v is given by:

$$v = \frac{\alpha_v^2}{\sigma_v^2} \implies \sqrt{v} = \frac{\alpha_v}{\sigma_v} \implies \sigma_v = \frac{\alpha_v}{\sqrt{v}}, \quad \text{(A.43)}$$

therefore, $\mathbf{z}_v$ can be equivalently expressed as

$$\mathbf{z}_v = \alpha_v \mathbf{x} + \sigma_v \boldsymbol{\epsilon} = \alpha_v \mathbf{x} + \frac{\alpha_v}{\sqrt{v}} \boldsymbol{\epsilon} = \alpha_v \left(\mathbf{x} + \frac{\boldsymbol{\epsilon}}{\sqrt{v}} \right), \quad \text{(A.44)}$$

which holds for any diffusion specification (forward process) by definition. Now, consider two distinct diffusion specifications denoted as $\left\{ \alpha_v^A, \sigma_v^A, \tilde{\mathbf{x}}_{\boldsymbol{\theta}}^A \right\}$ and $\left\{ \alpha_v^B, \sigma_v^B, \tilde{\mathbf{x}}_{\boldsymbol{\theta}}^B \right\}$. Due to Equation (A.44), any two diffusion specifications produce equivalent latents, up to element-wise rescaling:

$$\mathbf{z}_v^A = \frac{\alpha_v^A}{\alpha_v^B} \mathbf{z}_v^B \quad \text{(A.45)}$$

$$\alpha_v^A\left(\mathbf{x}+\frac{\boldsymbol{\epsilon}}{\sqrt{v}}\right)=\frac{\alpha_v^A}{\alpha_v^B}\alpha_v^B\left(\mathbf{x}+\frac{\boldsymbol{\epsilon}}{\sqrt{v}}\right). \tag{A.46}$$

This implies that we can denoise from any latent $\mathbf{z}_v^B$ using a model $\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^A$ trained under a different noise specification, by trivially rescaling the latent $\mathbf{z}_v^B$ such that it'd be equivalent to denoising from $\mathbf{z}_v^A$:

$$\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^B\left(\mathbf{z}_v^B,v\right)\equiv\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^A\left(\frac{\alpha_v^A}{\alpha_v^B}\mathbf{z}_v^B,v\right). \tag{A.47}$$

Furthermore, when two diffusion specifications have equal $\mathrm{SNR}_{\min}$ and $\mathrm{SNR}_{\max}$, then the marginal distributions $p^A(\mathbf{x})$ and $p^B(\mathbf{x})$ defined by the two generative models are equal:

$$\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^B\left(\mathbf{z}_v^B,v\right)\equiv\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^A\left(\frac{\alpha_v^A}{\alpha_v^B}\mathbf{z}_v^B,v\right)\implies p^A(\mathbf{x})=p^B(\mathbf{x}), \tag{A.48}$$

and both specifications yield identical diffusion loss in continuous time: $\mathcal{L}_\infty^A(\mathbf{x})=\mathcal{L}_\infty^B(\mathbf{x})$, due to Equation (4.80). Importantly, this does *not* mean that training under different noise specifications will result in the same model. To be clear, the $\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^B$ model is fully determined by the $\tilde{\mathbf{x}}_{\boldsymbol{\theta}}^A$ model and the rescaling operation α_v^A/α_v^B. Furthermore, this invariance to the noise schedule does not hold for the Monte Carlo estimator of the diffusion loss, as the noise schedule affects the *variance* of the estimator and therefore affects optimization efficiency.

References

[1] M. S. Albergo and E. Vanden-Eijnden, "Building normalizing flows with stochastic interpolants," in *The Eleventh International Conference on Learning Representations*, 2023.

[2] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, 1982, pp. 313–326.

[3] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*. Springer, 2023.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, 2017, pp. 859–877.

[5] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[6] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," in *International Conference on Learning Representations*, 2020.

[7] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 472–11 481, 2022.

[8] F. De Sousa Ribeiro, T. Xia, M. Monteiro, N. Pawlowski, and B. Glocker, "High fidelity image counterfactuals with probabilistic causal models," in *Proceedings of the 40th International Conference on Machine Learning*, pp. 7390–7425, PMLR, 2023.

[9] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, 2021, pp. 8780–8794.

[10] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan, *et al.*, "Continuous diffusion for categorical data," *arXiv preprint arXiv:2211.15089*, 2022.

[11] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. S. Grathwohl, "Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc," in *International conference on machine learning*, PMLR, pp. 8489–8510, 2023.

[12] T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, X. Geng, and B. Guo, "Efficient diffusion training via min-snr weighting strategy," *arXiv preprint arXiv:2303.09556*, 2023.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, 2020, pp. 6840–6851.

[15] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, vol. 23, no. 1, 2022, pp. 2249–2281.

[16] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.

[17] M. D. Hoffman and M. J. Johnson, "Elbo surgery: Yet another way to carve up the variational evidence lower bound," in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, vol. 1, 2016.

[18] E. Hoogeboom, J. Heek, and T. Salimans, "Simple diffusion: End-to-end diffusion for high resolution images," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, pp. 13 213–13 232, PMLR, 2023.

[19] E. Hoogeboom, T. Mensink, J. Heek, K. Lamerigts, R. Gao, and T. Salimans, "Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion," *arXiv preprint arXiv:2410.19324*, 2024.

[20] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3d," in *International conference on machine learning*, PMLR, pp. 8867–8887, 2022.

[21] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 22 863–22 876.

[22] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[23] A. Hyvärinen, I. Khemakhem, and R. Monti, "Identifiability of latent-variable and structural-equation models: From linear to nonlinear," *Annals of the Institute of Statistical Mathematics*, vol. 76, no. 1, 2024, pp. 1–33.

[24] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, 1999, pp. 183–233.

[25] H. Jun, R. Child, M. Chen, J. Schulman, A. Ramesh, A. Radford, and I. Sutskever, "Distribution augmentation for generative modeling," in *International Conference on Machine Learning*, PMLR, pp. 5006–5019, 2020.

[26] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 26 565–26 577.

[27] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24 174–24 184, 2024.

[28] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, 2021, pp. 21 696–21 707.

[29] D. P. Kingma and R. Gao, "Understanding the diffusion objective as a weighted integral of elbos," *arXiv preprint arXiv:2303.00848*, 2023.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[31] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, 2019, pp. 307–392.

[32] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," *Advances in neural information processing systems*, vol. 29, 2016.

[33] K. Kreis, R. Gao, and A. Vahdat, *Tutorial on denoising diffusion-based generative modeling: Foundations and applications*, 2022.

[34] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[35] X. Liu, C. Gong, and qiang liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *The Eleventh International Conference on Learning Representations*, 2023.

[36] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[37] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther, "Biva: A very deep hierarchy of latent variables for generative modeling," *Advances in neural information processing systems*, vol. 32, 2019.

[38] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," in *International conference on machine learning*, PMLR, pp. 1445–1453, 2016.

[39] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[40] C. Meng, J. Song, Y. Song, S. Zhao, and S. Ermon, "Improved autoregressive modeling with distribution smoothing," in *International Conference on Learning Representations*, 2020.

[41] M. Monteiro, F. D. S. Ribeiro, N. Pawlowski, D. C. Castro, and B. Glocker, "Measuring axiomatic soundness of counterfactual image models," in *The Eleventh International Conference on Learning Representations*, 2022.

[42] M. Mueller, K. Gruber, and D. Fok, "Continuous diffusion for mixed-type tabular data," *arXiv preprint arXiv:2312.10431*, 2023.

[43] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, PMLR, pp. 8162–8171, 2021.

[44] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*, PMLR, pp. 16 784–16 804, 2022.

[45] R. Ranganath, D. Tran, and D. Blei, "Hierarchical variational models," in *International conference on machine learning*, PMLR, pp. 324–333, 2016.

[46] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," *Advances in neural information processing systems*, vol. 28, 2015.

[47] D. J. Rezende, *Short notes on divergence measures*, 2018. URL: https://danilorezende.com/wp-content/uploads/2018/07/divergences.pdf.

[48] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*, PMLR, pp. 1278–1286, 2014.

[49] D. J. Rezende and F. Viola, "Taming vaes," *arXiv preprint arXiv:1810.00597*, 2018.

[50] S. Rissanen, M. Heinonen, and A. Solin, "Generative modelling with inverse heat dissipation," in *The Eleventh International Conference on Learning Representations*, 2023.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 684–10 695, 2022.

[52] A. Sabour, S. Fidler, and K. Kreis, "Align your steps: Optimizing sampling schedules in diffusion models," *arXiv preprint arXiv:2404.14507*, 2024.

[53] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 36 479–36 494.

[54] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[55] T. Salimans and J. Ho, "Should EBMs model the energy or the score?" In *Energy Based Models Workshop - ICLR 2021*, 2021.

[56] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*, 2022.

[57] T. Salimans, D. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *International conference on machine learning*, PMLR, pp. 1218–1226, 2015.

[58] J. E. Santos and Y. T. Lin, "Using ornstein-uhlenbeck process to understand denoising diffusion probabilistic model and its noise schedules," *arXiv preprint arXiv:2311.17673*, 2023.

[59] R. Shu and S. Ermon, "Bit prioritization in variational autoencoders via progressive coding," in *International Conference on Machine Learning*, PMLR, pp. 20 141–20 155, 2022.

[60] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, PMLR, pp. 2256–2265, 2015.

[61] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[62] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 1415–1428.

[63] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[64] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, 2020, pp. 12 438–12 448.

[65] Y. Song and D. P. Kingma, "How to train your energy-based models," *arXiv preprint arXiv:2101.03288*, 2021.

[66] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[67] J. Tomczak and M. Welling, "Vae with a vampprior," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1214–1223, 2018.

[68] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 19 667–19 679.

[69] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 11 287–11 302.

[70] H. Valpola, "From neural pca to deep unsupervised learning," in *Advances in independent component analysis and learning machines*, Elsevier, 2015, pp. 143–171.

[71] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, 2011, pp. 1661–1674.

[72] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[73] L. Wasserman, *All of statistics: a concise course in statistical inference*, vol. 26. Springer, 2004.