# Denoising Multi-$\beta$ VAE: Representation Learning for Disentanglement and Generation

Anshuk Uppal*
Technical University of Denmark
Copenhagen, Denmark
ansup@dtu.dk

Yuhta Takida
Sony AI
Tokyo, Japan

Chieh-Hsin Lai
Sony AI
Tokyo, Japan

Yuki Mitsufuji
Sony AI & Sony Group Corporation
New York, USA

## Abstract

*Disentangled and interpretable latent representations in generative models are often achieved at the expense of generation quality. The $\beta$-VAE framework introduces a hyperparameter $\beta$ to balance disentanglement and reconstruction quality, where setting $\beta > 1$ introduces an information bottleneck that favors disentanglement over sharp, accurate reconstructions. To address this trade-off, we propose a novel generative framework that leverages a range of $\beta$ values to learn multiple latent representations. First, we train these representations within a single variational autoencoder (VAE), with a new loss function that controls the information retained in each latent representation. We then, introduce a non-linear diffusion model that links latent spaces corresponding to different $\beta$ values. This model denoises latent variables toward less disentangled representations, ultimately leading to (almost) lossless representations, enabling sharp reconstructions. Furthermore, our model supports sample generation without input images, functioning as a standalone generative model. We evaluate our framework on both disentanglement and generation quality, showing competitive performance against $\beta$-VAE baselines and achieving high-quality image generation comparable to state-of-the-art models. Additionally, we observe smooth transitions in the latent spaces with respect to $\beta$ changes, facilitating consistent manipulation of generated outputs.*

## 1. Introduction

Today, numerous advanced latent generative models are capable of producing hyperrealistic images, providing end users with a broad array of options. Current advancements in state-of-the-art generative models focus primarily on qualitative improvements in generated outputs, with recent research emphasizing the study and analysis of training dynamics to enhance generation quality [21, 22, 25, 26]. Consequently, progress in generative modeling has largely shifted focus from learning and evaluating a model's latent representations to refining the generation process itself. However, research on deep latent generative models [30, 38] and unsupervised representation learning has shown that purposefully learned representations not only enhance generative performance but also offer practical advantages [54].

Previous generative modeling approaches aimed at learning disentangled and interpretable latent representations have often trailed behind in generation quality. $\beta$-VAE is a fundamental method for learning such representations, based on the variational autoencoder (VAE) framework [30]. Higgins et al. [18] modified the VAE objective by introducing a hyperparameter $\beta$, where setting $\beta = 1$ recovers the original objective function. This $\beta$ parameter governs the degree of disentanglement, balancing it against reconstruction and generation quality. A larger $\beta$ value imposes stronger regularization on the latent space, empirically shown to promote disentanglement, while a smaller $\beta$ prioritizes reconstruction accuracy but does not encourage disentanglement. Although $\beta$-VAE has been extensively studied [4, 8, 29, 46], overcoming this challenging trade-off remains difficult. To the best of our knowledge, no existing work has achieved satisfactory generation quality on widely-used, practical image datasets.

Inspired by research following $\beta$-VAEs, we aim to promote disentangled representation learning within modern generative models. To this end, we propose a novel generative modeling framework. Our model consists of two main components. First, we train a *single* VAE that learns a spectrum of disentangled representations by varying the parameter of $\beta$, which controls the degree of disentangle-

*Work done during an internship at Sony AI

ment. This VAE comprises an encoder and a decoder, each conditioned by $\beta$. However, this VAE still faces the trade-off issue: latent representations disentangled by larger $\beta$ values lose some information from the original input, resulting in blurred outputs similar to those of a standard $\beta$-VAE. To address this, we introduce a novel non-linear diffusion model that denoises the latent variable at a given $\beta$ back to an (ideally) non-lossy latent space corresponding to $\beta = 0$. This allows us to generate sharp, non-blurred images by decoding the denoised latent variable through the $\beta = 0$ decoder. Notably, our model can also generate new samples by starting from the largest value of $\beta$ without any input image, thanks to the carefully designed configuration of the VAE.

In our experiments, we evaluate our model in terms of both disentanglement and generation quality. For disentanglement, we benchmark on well-known toy datasets [27, 33]. Furthermore, we test our model's capability as a standalone generative model on widely-used image datasets both qualitatively and quantitatively. Additionally, we show that the set of learned latent spaces is smooth both with respect to $\beta$ and within each individual space. In particular, smoothness with respect to $\beta$ is essential for consistent manipulation.

Our contributions are briefly listed as follows.

- We propose a generative modeling framework that leverages multiple levels of latent representations ranging from fully-informed to disentangled representations. To obtain these latent representations, we extend $\beta$-VAE using a range of $\beta$, along with a novel model design and objective.
- We propose a novel non-linear diffusion model that connects latent spaces induced by different $\beta$ values. By integrating the VAE with the diffusion, our model enables both disentanglement and high-quality generation in principle.
- We empirically show that our proposed model achieves disentanglement performance competitive with $\beta$-VAE-based baselines, while also generating high-quality images comparable to those produced by state-of-the-art generative models.

## 2. Overview of $\beta$-VAE

This section provides an overview of the $\beta$-VAE while establishing the notations used throughout the rest of the paper.

We begin with the formulation of a vanilla VAE [30]. Suppose we have a training dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^M$, where $\boldsymbol{x}_i \in \mathbb{R}^D$ for $i = \{1, \dots, M\}$, drawn from an unknown underlying distribution. We denote the empirical distribution defined by $\mathcal{D}$ as $p_{\mathcal{D}}(\boldsymbol{x})$. A VAE aims to uncover a reduced set of latent factors that give rise to this dataset.

Specifically, a latent variable $\boldsymbol{z} \in \mathbb{R}^d$ ($d < D$) is introduced, with its prior distribution set as $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Data samples are generated by first sampling $\boldsymbol{z} \sim p(\boldsymbol{z})$ and then decoding it using a probabilistic decoder, denoted as $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. The decoder is commonly parameterized by a conditional isotropic Gaussian as $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|g_\theta(\boldsymbol{z}), s^2\boldsymbol{I}_D)$ with a function $g_\theta : \mathbb{R}^d \to \mathbb{R}^D$. We then wish to maximize the marginal log-likelihood $\log p_\theta(\boldsymbol{x})$, where $p_\theta(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{z})}[p_\theta(\boldsymbol{x}|\boldsymbol{z})]$. However, this maximization is generally not tractable. Therefore, in the VAE framework, a surrogate objective function called the evidence lower bound (ELBO) is maximized instead, formulated as $\log p_\theta(\boldsymbol{x}) \geq$

$$\underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{\mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))}_{\text{regularization term}}, \quad (1)$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is a variational distribution used to approximate the posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, a.k.a., the encoder. A common way to model the variational distribution is using a conditional Gaussian as $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|f_\phi(\boldsymbol{x}), \mathrm{diag}(\sigma_\phi(\boldsymbol{x})))$, with functions $f_\phi : \mathbb{R}^D \to \mathbb{R}^d$ and $\sigma_\phi : \mathbb{R}^D \to \mathbb{R}^d_{\geq 0}$.

Through the maximization of Eq. (1), the encoder learns to recover latent generative factors from the dataset, while the decoder attempts to reconstruct $\boldsymbol{x}$ from $\boldsymbol{z}$ as accurately as possible. In other words, the encoder and decoder are trained to compress the data without information loss, effectively becoming stochastic inverses of each other. The $\beta$-VAE [18] is a variant of the above model that employs a slightly different objective function:

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta\,\mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})). \quad (2)$$

In this objective function, the *regularization term* in Eq. (1) is scaled by a hyperparameter $\beta$.

The choice of $\beta$ creates a trade-off between the reconstruction quality and disentanglement of the latent representation. Some previous works have suggested that increasing the contribution of the *regularization term*, i.e., setting $\beta > 1$, not only promotes independence among latent dimensions but also facilitates the learning of interpretable generative factors (please refer to Appendix A for this literature review). On the other hand, a downside to increasing regularization is the loss of information. When the variational approximation approaches the prior according to the KL term, all encodings $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ begin to collapse to the prior, resulting in a lack of distinct information about individual data points, which hampers accurate reconstruction. This indicates that setting the value of $\beta$ is non-trivial due to the precarious balance between desirable disentanglement and undesirable loss of information. Even with a suitable parameter $\beta$ for disentangled representation, reconstructed samples may still be blurred due to the loss of information.

Alternatives to the $\beta$-VAE, such as Chen et al. [4], Kim and Mnih [29], address the issue of poor reconstructions but do not evaluate sample quality and reconstructions on practical image datasets. In the following section, we introduce our novel conditional $\beta$-VAE architecture, which incorporates the strengths of these previously proposed enhancements.

In the following sections, we propose our solutions to two critical issues:

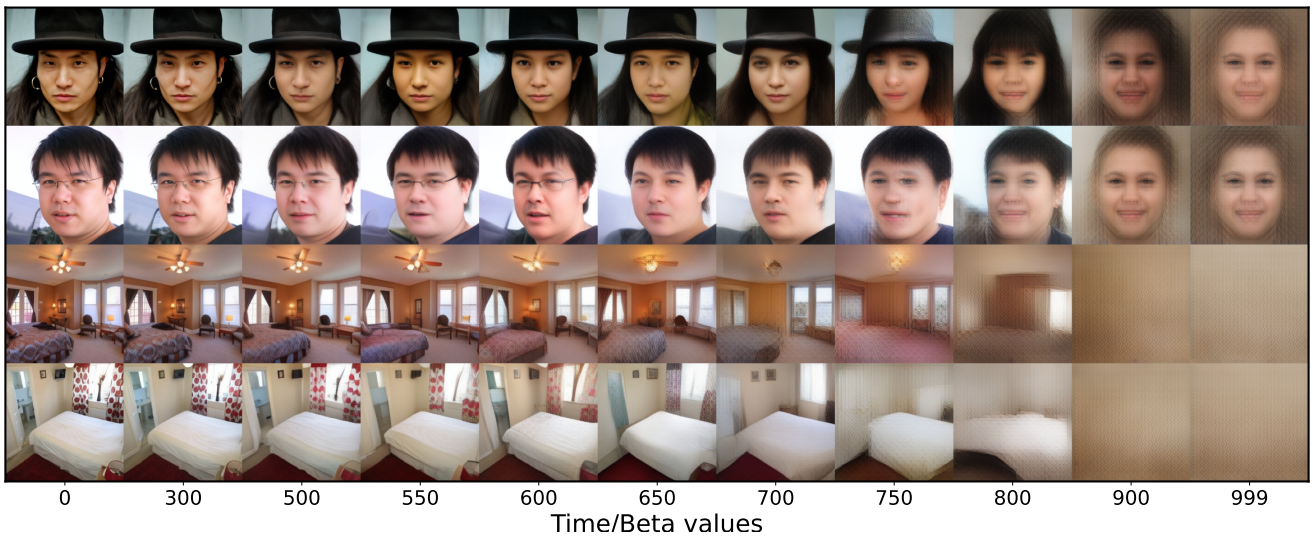


Figure 1. We visualize reconstructions from a trained conditional multi-level $\beta$-VAE on FFHQ [24] and LSUN Bedrooms [56], using $\beta$ values between $[0, 1]$ over 1,000 steps. Without the non-linear diffusion model proposed in Sec. 4, our multi-level $\beta$-VAE produces blurry reconstructions at higher $\beta$ values, where greater disentanglement is observed.

1. **Problem 1:** Choosing the optimal regularization coefficient ($\beta$) is nontrivial and requires multiple training runs. Our solution to this issue is detailed in Sec. 3.
2. **Problem 2:** Regularization leveraged for learning disentangled representations deteriorates reconstruction and sample quality, we propose inclusion of a novel diffusion model in the latent space in Sec. 4 to resolve this.

## 3. Multi-$\beta$ Representation Learning

To overcome the severe trade-off between reconstruction accuracy and the disentanglement of latent representation in existing VAE variants, we propose multi-$\beta$ latent representation learning. First, we extend $\beta$-VAE by treating $\beta$ as a variable rather than a hyperparameter in Section 3.1. Using a monotonic property of multi-$\beta$ latent space presented in Section 3.2, a subsequently learned diffusion model allows us to move across latent spaces corresponding to different $\beta$ (see Section 4).

### 3.1. Conditional Multi-level $\beta$-VAE

Here we extend the $\beta$-VAE to incorporate $\beta$ as a variable within a range of values. In our setup, unlike the typical $\beta$-VAE, $\beta$ lies in $[0, B]$ instead of being fixed, and it scales the reconstruction and regularization terms with weights of $(B - \beta)$ and $\beta$, respectively (see Eq. (5)). This approach allows us to achieve a full range of weighting with finite values. We expect that larger values of $\beta$ result in more disentangled latent representations, while smaller values will yield higher fidelity in the reconstructed samples. We assume that each value of $\beta$ has its latent space, which is denoted as $Z_\beta$. In our VAE, the decoder and encoder for a given $\beta$ are designed as follows:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}_\beta;\beta) = \mathcal{N}(\boldsymbol{x}|g_\theta(\boldsymbol{z}_\beta,\beta), s_\beta^2\boldsymbol{I}) \tag{3}$$

$$q_\phi(\boldsymbol{z}_\beta|\boldsymbol{x};\beta) = \mathcal{N}(\boldsymbol{z}_\beta|f_\phi(\boldsymbol{x},\beta), \sigma_\beta^2\boldsymbol{I}), \tag{4}$$

where $s_\beta^2,\ \sigma_\beta^2 \in \mathbb{R}^d_{\geq 0}$, $\theta$ and $\phi$ represent the parameters for the decoder and encoder, respectively. We model both the encoder and decoder to depend on $\beta$, which induces different latent spaces. Additionally, the conditional covariance matrices in both the data and latent spaces are modeled as learnable isotropic matrices that depend solely on $\beta$.

Under this model setup, we propose a novel objective function based on a rescaled ELBO as $\mathcal{L} = \mathbb{E}_\beta \mathcal{L}_\beta$, where

$$\mathcal{L}_\beta = \mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})}\Big[(B-\beta)\mathbb{E}_{q_\phi(\boldsymbol{z}_\beta|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_\beta)\right] - \beta\, \mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_\beta|\boldsymbol{x})||p(\boldsymbol{z}))\Big]. \tag{5}$$

Here, we sample $\beta$ from a prior distribution to train the VAE across multiple $\beta$ values. For this work, we set it as a uniform distribution, with $B = 1$. Notably, $\mathcal{L}_\beta$ with $\beta = 0$ and $0.5$ corresponds to the objective functions for plain autoencoder and a VAE, respectively, without considering the scaling factors*. Algorithm 1 contains the training algorithm for this upgraded $\beta$-VAE.

### 3.2. Controlling Information Loss with $\beta$

As described in the previous section, our objective function (5) smoothly interpolates and extrapolates between those

*Our $\beta$ is different from that of the typical $\beta$-VAE in the relationship between $\beta$ values and their respective models.

**Algorithm 1:** Training of multi-level $\beta$-VAE

**Input:** Dataset $\mathcal{D}$, $\beta$-schedule $\{\beta_i\}_{i=1}^N$, learning rate $\eta$, number of training steps $S$

**Output:** Trained networks, $\phi$, $\theta$, and $\boldsymbol{\sigma} = \{\sigma_\beta\}_{\beta \in \{\beta_i\}_{i=1}^N}$

**for** $s = 1, 2, \ldots, S$ **do**

1. Sample: $\boldsymbol{x} \sim p_{\mathcal{D}}(\boldsymbol{x}),\ \beta \sim \mathcal{U}([0,1]),\ \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
2. Generate noisy encoding: $\boldsymbol{z}_\beta = f_\phi(\boldsymbol{x}, \beta) + \sigma_\beta \epsilon$
3. Compute the objective $\mathcal{L}$ based on Eq. (5)
   **for** $\omega = \{\theta, \phi, \boldsymbol{\sigma}\}$ **do**
4. $\omega \leftarrow \omega - \eta \nabla_\omega \mathcal{L}$

**return** $f_\phi, g_\theta, \sigma$

for autoencoder and VAE. Specifically, $\beta$ controls the information loss in the latent spaces. We expect that $\beta$ larger than 0.5 promotes disentanglement in the latent space $\mathcal{Z}_\beta$, albeit at the cost of information essential for reconstruction. Conversely, $\beta$ less than 0.5 results in higher fidelity of reconstruction, sacrificing disentanglement of representations.

A smaller value of $\beta$ places larger weight on the reconstruction term, leading to a reduced latent variance $\sigma_\beta^2$. In the extreme case, setting $\beta = 0$ theoretically results in perfect reconstruction with $\sigma_0^2 = 0$, as demonstrated in the following proposition.

*Proposition* 1. Under certain regularity conditions, the global optimum of $\mathcal{L}_0$ is achieved when $\sigma_0^2 = 0$.

*Proof.* Please refer to Appendix B □

In contrast, increasing $\beta$ towards 1 enhances the degree of latent regularization, which encourages disentanglement of the latent representation. In the extreme case, when $\beta = 1$, the objective function (5) reduces to the KL regularization term, causing $q_\phi(\boldsymbol{z}|\boldsymbol{x}; \beta = 1)$ collapse to the prior distribution $p(\boldsymbol{z})$. Consequently, the latent space, i.e., $Z_1$, no longer retains any information about the input $\boldsymbol{x}$, as demonstrated in the following proposition.

*Proposition* 2. The mutual information between the input and the reconstructed samples produced by the VAE becomes zero as $\mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}; \beta = 1) \parallel p(\boldsymbol{z}))$ converges to zero. It holds for any decoder function $g_\theta(\cdot, \beta = 1)$.

*Proof.* Please refer to Appendix B □

This phenomenon causes a gradual loss of information in the latent space, characterised by a gradual increase in the variance of latent representations such that $\sigma_\beta < \sigma_{\beta'}$ for $0 \leq \beta < \beta' \leq 1$, which is also observed in previous studies [49].

In summary, our multi-level $\beta$-VAE is equipped to learn a slew of latent representations which due to Eq. (5), capture major axes of variation present across the dataset by down-weighting accurate reconstructions. This does not mitigate Problem 2 (see Fig. 1). To this end, we purposely combine the model Eq. (4) and Eq. (5) so that learnt $\sigma_\beta$ parallel a typical noising process in diffusion models ($\beta \equiv t$). When trained well, diffusion models can capture the target distribution and sample realisitic data by repeated denoising. However, due to the involvement of an encoder that specifies the mean $f_\phi(\boldsymbol{x}, \beta)$ at all $\beta \in [0, 1]$, our noising process diverges from the commonly used linear inference/noising process. We detail our formulation of non-linear denoising diffusion in the next section.

## 4. Reversing the Information Loss

Increasing regularization enhances representation learning but negatively affects sample and reconstruction quality. This effect is shown in Fig. 1, where images are decoded from latent spaces learned by our VAE on the FFHQ [24] and LSUN Bedrooms [56] dataset. At higher $\beta$ values, the reconstructions tend to collapse into an "averaged" image, a phenomenon also noted by Collins et al. [6]. To address Problem 2, we propose reversing information loss by training a denoising model based on a diffusion process. First, we review the standard diffusion model in Sec. 4.1 and its nonlinear extension in Sec. 4.2, which can be viewed as a specific instance of a Hierarchical VAE. In this approach, the time-varying mean is governed by the encoder ($f_\phi$), with noise conditioning parameterized by $\beta$ or equivalently by time $t$.

### 4.1. Primer on Diffusion Models

We start with a brief primer on the vanilla diffusion models with a linear diffusion process. Diffusion models consist of a fixed hierarchical encoding process, known as the forward or noising process, and a decoding process for generation. In the encoding stage, incremental noise is gradually added to the data, transforming it into a Gaussian noise:

$$q(\boldsymbol{z}_t|\boldsymbol{x}; t) := \mathcal{N}(\boldsymbol{z}_t|\boldsymbol{x}, \sigma_t^2 \boldsymbol{I}_d), \tag{6}$$

where $t \in [0, T]$, $\tau > 0$ is a small constant, and $\sigma_t > 0$ is a predefined noise schedule that increases with $t$. Next, the Markovian forward distributions are derived as

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-\tau}) = \mathcal{N}(\boldsymbol{z}_t|\boldsymbol{z}_{t-\tau}, \sigma_{t|t-\tau}^2 \boldsymbol{I}_d), \text{ where} \tag{7}$$

$$\sigma_{t|t-\tau}^2 := \sigma_t^2 - \sigma_{t-\tau}^2. \tag{8}$$

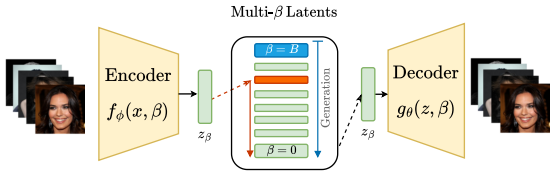


Figure 2. Architecture of our model. Our approach embeds $\beta$s to control multiple levels of disentanglement as time conditioning in our newly designed nonlinear diffusion model, enabling both effective disentanglement and high-quality generation.

A tractable sequential reverse decoding process is obtained via Bayes' rule, leading to:

$$q(\boldsymbol{z}_{t-\tau}|\boldsymbol{z}_t,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_{t-\tau}|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{z}_t,\boldsymbol{x}),\tilde{\sigma}_t^2\boldsymbol{I}_d), \text{ where} \quad (9)$$

$$\tilde{\sigma}_t^2 = \frac{\sigma_{t|t-\tau}^2\sigma_{t-\tau}^2}{\sigma_t^2}, \quad \tilde{\boldsymbol{\mu}}_t(\boldsymbol{z}_t,\boldsymbol{x}) = \frac{\sigma_{t-\tau}^2}{\sigma_t^2}\boldsymbol{z}_t + \frac{\sigma_{t|t-\tau}^2}{\sigma_t^2}\boldsymbol{x}. \quad (10)$$

Diffusion models are trained to match the generative reverse conditional distributions in Eq. (9), and are generally parameterized as

$$p_\psi(\boldsymbol{z}_{t-\tau}|\boldsymbol{z}_t) := \mathcal{N}(\boldsymbol{z}_{t-\tau}|\boldsymbol{\mu}_\psi(\boldsymbol{z}_t,t),\sigma_r^2(t)\boldsymbol{I}_d), \quad (11)$$

$$\text{where } \boldsymbol{\mu}_\psi(\boldsymbol{z}_t,t) := \frac{\sigma_{t-\tau}^2}{\sigma_t^2}\boldsymbol{z}_t + \frac{\sigma_{t|t-\tau}^2}{\sigma_t^2}\hat{\boldsymbol{x}}_\psi(\boldsymbol{z}_t,t). \quad (12)$$

$\hat{\boldsymbol{x}}_\psi$ is known as the *denoiser*. Equivalently, we parametrize it as a noise prediction model $\hat{\boldsymbol{\epsilon}}_\psi$, where $\hat{\boldsymbol{x}}_\psi(\boldsymbol{z}_t,t) = \big(\boldsymbol{z}_t - \sigma_t\hat{\boldsymbol{\epsilon}}_\psi(\boldsymbol{z}_t,t)\big)/\alpha_t$. The loss function of a diffusion model is derived using an ELBO (by extending Eq. (1) with time hierarchy), aiming to match the conditional distributions in Eq. (9) and Eq. (11) over all $t \in [0,T]$ using the KL divergence. This loss boils down to a simple regression loss:

$$\min_\psi \mathbb{E}_{t,\boldsymbol{x}}\mathbb{E}_{\boldsymbol{z}_t|\boldsymbol{x}}\left[\frac{1}{2\tilde{\sigma}_t^2}\|\boldsymbol{\mu}_\psi(\boldsymbol{z}_t,t) - \tilde{\boldsymbol{\mu}}_t(\boldsymbol{z}_t,\boldsymbol{x})\|_2^2\right]. \quad (13)$$

By discretizing the time such that $t \in \{iT/N\}_{i=0}^N$ for the iterative decoding, we obtain a hierarchical generator:

$$p_\psi(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z}_0)p(\boldsymbol{z}_B)\prod_{i=1}^N p_\psi(\boldsymbol{z}_{\frac{(i-1)T}{N}}|\boldsymbol{z}_{\frac{iT}{N}}), \quad (14)$$

where $p(\boldsymbol{z}_T) = \mathcal{N}(\boldsymbol{z}_T|0,\boldsymbol{I}_d)$.

### 4.2. Non-linear Diffusion in Latent Space

We propose a non-linear (in $\boldsymbol{x}$) denoising diffusion for use in our model. In this subsection, time variables $t$ (and $T$) are interchangeable with $\beta$ (and $B$), as they represent the same concept. We hinted in Sec. 3.2 that our non-linear diffusion process is prescribed by the $\beta$- or time-dependent encoder. Formally, the distribution of $\boldsymbol{z}_t$ for given $\boldsymbol{x}$ is

$$q_\phi(\boldsymbol{z}_t|\boldsymbol{x};t) = \mathcal{N}(\boldsymbol{z}_t|f_\phi(\boldsymbol{x},t),\sigma_t^2\boldsymbol{I}_d). \quad (15)$$

This expression is just an adaptation of Eq. (4) with $\beta = t$, and is more general than Eq. (6). We propose the nonlinear Markovian encoding process as

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-\tau},\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t|\boldsymbol{z}_{t-\tau}+f_\phi(\boldsymbol{x},t) - f_\phi(\boldsymbol{x},t-\tau),\sigma_{t|t-\tau}^2). \quad (16)$$

This form closely resembles Eq. (7). Following the development in Sec. 4.1, we now define the reverse of Eq. (16) as:

$$q(\boldsymbol{z}_{t-\tau}|\boldsymbol{z}_t,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t|\tilde{\boldsymbol{\mu}}_t(\boldsymbol{z}_t,\boldsymbol{x}),\tilde{\sigma}_t^2\boldsymbol{I}_d), \text{ where} \quad (17)$$

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{z}_t,\boldsymbol{x}) = \frac{\sigma_{t-\tau}^2}{\sigma_t^2}\boldsymbol{z}_t + \frac{\sigma_{t|t-\tau}^2}{\sigma_t^2}\boldsymbol{x} + f_\phi(\boldsymbol{x},t-\tau) - f_\phi(\boldsymbol{x},t), \quad (18)$$

with $\tilde{\sigma}_t^2$ remaining the same as Eq. (10). Due to the additional $f_\phi$ terms in this flavour of the diffusion model, $\boldsymbol{\mu}_\psi$ cannot follow the same parameterization as in Eq. (12). Instead we introduce a new approach to express the mean prediction in this case, as follows:

$$\boldsymbol{\mu}_\psi(\boldsymbol{z}_t,t) := \frac{\sigma_{t-\tau}^2}{\sigma_t^2}\boldsymbol{z}_t + \frac{\sigma_{t|t-\tau}^2}{\sigma_t^2}\hat{\boldsymbol{x}}_\psi(\boldsymbol{z}_t,t) + \hat{\boldsymbol{\Delta}}_\psi(\boldsymbol{z}_t,t). \quad (19)$$

Eq. (19) introduces an extra predictor $\hat{\boldsymbol{\Delta}}_\psi$ for learning the evolution of encodings with time. In practice, we train noise prediction, reparameterizing $\hat{\boldsymbol{x}}_\psi(\boldsymbol{z}_t,t)$ with $\hat{\boldsymbol{\epsilon}}_\psi(\boldsymbol{z}_t,t)$, along with an encoding difference predictor $\hat{\boldsymbol{\Delta}}_\psi(\boldsymbol{z}_t,t)$, which is novel to the best of our knowledge.

Following the parameterization of $\boldsymbol{\mu}_\psi$, our model training differs from standard practice in a few key ways. First, we do not train the noise prediction network to predict the noise added to $\boldsymbol{z}_0$ given a sample $\boldsymbol{z}_t$. The transition from $\boldsymbol{z}_0$ to $\boldsymbol{z}_t$ is non-linear due to the encoder and depends on its Jacobian, $\frac{\mathrm{d}f_\phi(\boldsymbol{x},t)}{\mathrm{d}t}$. Rather than learning an inversion of this time-varying encoding, we aim to learn the direction of the noise ($\epsilon_t$) at each time step $t \in [0,T]$ using $\hat{\boldsymbol{\epsilon}}_\psi(\boldsymbol{z}_t,t)$, i.e.,

$$\boldsymbol{z}_t = f_\phi(\boldsymbol{x},t) + \sigma_t\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d) \quad (20)$$

$$\min_\psi \mathbb{E}_{t,\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\epsilon}_t}\mathbb{E}_{\boldsymbol{z}_t|\boldsymbol{x},\boldsymbol{\epsilon}_t}\left[\frac{1}{w(t)}\|\boldsymbol{\epsilon}_\psi(\boldsymbol{z}_t,t) - \boldsymbol{\epsilon}_t\|_2^2\right], \quad (21)$$

where $w(t)$ is a weighting function. This approach trains the model to denoise at each time step.Additionally, $\hat{\boldsymbol{\Delta}}_\psi$ is necessary for sampling and is trained to predict the change in $f_\phi$ over a small time interval $\tau$[*]. Assuming that the encoder

[*]In implementation, we set $\tau = T/N$ in the discrete time setup, learning the single time-step encoding difference for all times.

**Algorithm 2:** Sampling (Noise Prediction model)

**Input:** Trained model $\boldsymbol{\epsilon}_\psi$, total time steps $N$, largest time $T$, noise schedule $\boldsymbol{\sigma} = \{\sigma_t\}_{t\in\{0,T/N,\dots,(N-1)T/N,T\}}$

**Output:** Generated sample $\boldsymbol{z}_0$

**Initialize:** $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$

**for** $t = T, (N-1)T/N \dots, T/N$ **do**

1. Predict noise and diff:
   $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}_\psi(\boldsymbol{z}_t, t),\ \hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}_\psi(\boldsymbol{z}_t, t)$
2. Compute mean for $z_t$:
   $\boldsymbol{\mu}_t = \boldsymbol{z}_t - \sigma_t\hat{\boldsymbol{\epsilon}}$
3. Predict previous mean:
   $\boldsymbol{\mu}_{t-T/N} = \boldsymbol{\mu}_t - \hat{\boldsymbol{\Delta}}$
4. Update $\boldsymbol{z}_{t-T/N}$:
   $\epsilon \sim \mathcal{N}(0, I),\ \boldsymbol{z}_{t-T/N} = \mu_{t-T/N} + \sigma_{t-T/N}\boldsymbol{\epsilon}$

**return** $\boldsymbol{z}_0$

is a smooth function, this design is based on the intuition that learning encoding differences over one time step is easier than over arbitrarily large steps.

We adjust the diffusion model's U-Net to produce two outputs, $\hat{\boldsymbol{\epsilon}}(\boldsymbol{z}_t, t),\ \hat{\boldsymbol{\Delta}}_\psi(\boldsymbol{z}_t, t)$. Using these predictions, we build our DDPM [19]-inspired sampling algorithm, as shown in Algorithm 2. The actual loss function used to train this new diffusion model is defined as

$$\min_\psi \mathbb{E}_{t,\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\epsilon}_t}\mathbb{E}_{\boldsymbol{z}_t|\boldsymbol{x},\boldsymbol{\epsilon}_t}\Big[\frac{1}{w(t)}\|\boldsymbol{\epsilon}_\psi(\boldsymbol{z}_t, t) - \boldsymbol{\epsilon}_t(\boldsymbol{x}, \boldsymbol{\epsilon}_t)\|_2^2 + \|f_\phi(\boldsymbol{x}, t-\tau) - f_\phi(\boldsymbol{x}, t) - \hat{\boldsymbol{\Delta}}_\psi(\boldsymbol{z}_t, t)\|_2^2\Big]. \quad (22)$$

Integrating the conditional multilevel $\beta$-VAE, as introduced in Sec. 3.1, with the diffusion model described in Sec. 4.2 is key to mitigating the disentanglement-reconstruction trade-off in our framework.

We now combine the two proposed modules from Sec. 3 and Sec. 4 into a single model, trained in two phases. In the first phase, we train the conditional multilevel $\beta$-VAE using the loss defined in Eq. (5) (Algorithm 1). After this, we train the non-linear diffusion model from Sec. 4.2, keeping the autoencoder parameters fixed. Additionally, depending on the dataset, we fine-tune the decoder with an adversarial loss, following Rombach et al. [43], to enhance generation quality. Further training details are provided in Appendix D.

## 5. Related Works

**$\beta$-VAE** and its variants are extensively studied for their distinct capabilities [3] and wide-ranging applications in domains such as images [18], text [46], and molecular generation [42]. These models are especially valued for their interpretable latent representations, achieved through $\beta$-controlled regularization.

Kim and Mnih [29] enhanced disentanglement while preserving reconstruction quality by combining the ELBO with a total correlation term, balanced by a hyperparameter. Similarly, Chen et al. [4] improved mutual information between latent variables and observed data to support independence among latent factors. Shao et al. [46] introduced an adaptive feedback mechanism that adjusts $\beta$ during training based on KL divergence, using a non-linear PI controller. Dewangan and Maurya [8] employed a deep convolutional $\beta$-VAE for feature extraction in fault diagnosis of industrial machines, with a variable $\beta$ training protocol that regularizes the model but does not condition the encoder and decoder on $\beta$.

Both Collins et al. [6] and Bae et al. [1] explored training VAEs with multiple $\beta$ values. Bae et al. [1] aimed to capture the full rate-distortion curve using a hypernetwork conditioned on $\beta$, while Collins et al. [6] trained over multiple $\beta$ values using conditional variational and generative distributions similar to ours, focusing on particle physics applications. Unlike these works, our framework supports high-fidelity generation from disentangled representations (with larger $\beta$) by linking latent spaces through a novel non-linear diffusion process. Additionally, we introduced specific adjustments to the $\beta$-conditioned VAE, tailoring the latent space for diffusion.

**Hierarchical VAEs** are relevant to our model framework. Sønderby et al. [48] introduced a graphical model with top-down and bottom-up paths to build hierarchical latent structures in both generative and inference processes. This design improves network efficiency by enabling feature sharing between the generative process $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ and the inference process $q_\phi(\boldsymbol{x}, \boldsymbol{z})$. Subsequent studies have leveraged this graphical approach to enhance generation quality and log-likelihood estimation [5, 50].

Razavi et al. [39] extended vector quantized-VAE (VQ-VAE) by incorporating this graphical modeling [52]. Alternatively, Dhariwal et al. [10] trained separate VQ-VAEs to learn distinct levels of discrete latent representations. To integrate the multiple latent spaces, they introduced a prior and upsamplers, modeled by separate unconditional and conditional autoregressive transformers, respectively. Unlike these approaches, our model uses a single VAE and prior model to manage hierarchical levels, supporting continuous (infinite) levels and providing efficient memory usage and reduced training costs.

Non-linear diffusion models represent a special case of hierarchical VAEs, offering learnable forward and denoising processes for improved modeling flexibility. Singhal et al. [47] and Bartosh et al. [2] focus on algorithmic strategies for training non-linear diffusion models. The former addresses intractable score expressions in non-linear kernels

by linearizing the non-linear drift, while Nielsen et al. [36] improve generation quality by introducing a non-linear drift term through a time-dependent encoder. Unlike Bartosh et al. [2], their method employs fixed noise schedules and skips input-space compression, avoiding decoders in the latent-to-data mapping. Our model shares similarities with Nielsen et al. [36] but with notable differences: our model outputs two heads for noise and difference prediction (see Eq. (19)), learns noise schedules independently of data, and encodes to a lower-dimensional latent space.While Wang et al. [53] try to solve this issue with linear diffusion processes with an auxiliary variable that's trained with addtional losses.

## 6. Experiments

In this section, we first quantitatively evaluate our proposed approach with two types of assessments. In Sec. 6.1, we examine the disentanglement of the learned latent representations, using toy datasets for quantitative analysis. In Sec. 6.2, we evaluate our model's unconditional generation performance and report standard metrics on commonly-used image datasets. Additionally, we demonstrate the effectiveness of representations learned by our method qualitatively using common image datasets in Sec. 6.3.

### 6.1. Evaluating Disentanglement

We adopt the evaluation protocols of Locatello et al. [33] and Khrulkov et al. [27] using toy datasets: Cars3d [40] with 3 ground truth factors, Shapes3D [29] with 6 ground truth factors, and MPI3D [15] with 7 ground truth factors. For assessment, we use the Mutual Information Gap (MIG)[4] and the Disentanglement metric[11]. The MIG measures how well latent dimensions respond to changes in individual generative factors, while DCI (Disentanglement, Completeness, and Informativeness) evaluates the extent to which factors are distinctly represented by individual latent dimensions, the exclusivity of each factor to a specific dimension, and the comprehensiveness of the overall representation.

We compare our approach with other methods that address the disentanglement-generation trade-off: FactorVAE [29], $\beta$-TCVAE [4], and InfoGAN-CR [32]. Although other unsupervised representation learning methods have been proposed [55], they primarily focus on disentangled representation learning without paying attention to generation quality. We also compare with Ren et al. [41] that uses contrastive learning to find the right directions in the latent space of a pre-trained generative model. Our results are shown in Tab. 1. For all toy datasets, we train with 500 values of $\beta$ equally spaced within $[0, 1]$. We perform a sweep over all $\beta$ values in our model to identify the optimal latent representation for each metric. Specifically, for Cars3d [40] we achieve the highest score at $\beta = 285/500$, for Shapes3D [29] at $\beta = 210/500$, and for MPI3D [15] at $\beta = 280/500$. Our experiments indicate that our model achieves disentanglement performance, comparable to or surpassing existing methods. Moreover, *finding* based approaches like Ren et al. [41] are complementary to our multi-$\beta$ VAE and offer marginal performance gains when combined with our model.

Table 1. **Disentanglement Metrics:** We evaluate our multi-$\beta$ VAE representations by benchmarking on well-known toy datasets and comparing them to baselines that address the disentanglement-generation trade-off. In each row, the best results are highlighted in bold, and the second-best results are marked with an asterisk ($^*$).

| | Metrics | FactorVAE | $\beta$-TCVAE | InfoGAN-CR | Ours |
|---|---|---|---|---|---|
| Cars | MIG(↑) | **0.128 ± 0.036** | 0.080 ± 0.024 | 0.011 ± 0.009 | 0.114 ± 0.009$^*$ |
| | DCI(↑) | **0.160 ± 0.020** | 0.140 ± 0.020 | 0.020 ± 0.011 | 0.157 ± 0.010$^*$ |
| Shapes | MIG(↑) | 0.411 ± 0.163 | 0.406 ± 0.190 | 0.297 ± 0.124 | **0.422 ± 0.090** |
| | DCI(↑) | 0.611 ± 0.127 | 0.613 ± 0.151 | 0.478 ± 0.055 | **0.621 ± 0.090** |
| MPI | MIG(↑) | 0.098 ± 0.027 | 0.108 ± 0.053 | **0.161 ± 0.077** | 0.147 ± 0.035$^*$ |
| | DCI(↑) | 0.246 ± 0.066 | 0.239 ± 0.062 | 0.242 ± 0.076 | **0.253 ± 0.043** |

To further investigate the properties of latent representations learned at different $\beta$ values, we plot the evolution of the MIG score as a function of $\beta$ using the Cars3d dataset. These results are presented in Fig. 3. The plot reveals that MIG scores fluctuate in alignment with the loss function (5) associated with each $\beta$ value. For models like $\beta$-TCVAE, $\beta$-VAE, and FactorVAE, determining an optimal coefficient for regularization can be challenging, often requiring numerous training runs as the ideal coefficient depends on both the dataset and model architecture. In our approach, this issue is mitigated by training a conditional $\beta$ model across a wide range of $\beta$ values (Sec. 3.1). Building on this analysis, we highlight an additional insight gained from using multiple $\beta$ values. By leveraging the distinct generative factors in the toy datasets, we plotted the highest mutual information for two factors across all $\beta$ values, showing that while mutual information (MI) increases for 'Pitch' as $\beta$ increases, it decreases for 'Identity' (see Fig. 3). This observation supports the idea of training a model across multiple $\beta$ values, as it enables targeted manipulation of image factors by selecting specific $\beta$ values. Interestingly, in Fig. 1, we observe that sharper facial and hair features gradually fade until around the 500th step, with features like 'glasses' and 'gender' diminishing between steps 600 and 700, while 'facial angle' persists until approximately the 800th step. We provide more observations related to our multi $\beta$-VAE in Appendix C.1..

### 6.2. Evaluating Generation Quality

In this subsection, we evaluate the generation quality of our model on practical image datasets, including CelebA-HQ [23], FFHQ [24], and LSUN-Bedrooms [56], at a resolution of $256 \times 256$. While our model is designed to learn disentangled latent representations, it is crucial that this capability does not compromise generation quality. For the experiments in this section, we train and sample from the non-linear diffusion model unconditionally to assess perfor-

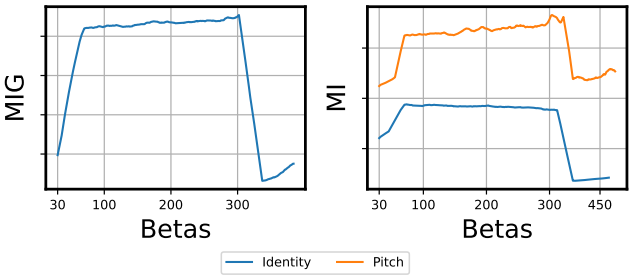


Figure 3. **Variation of MIG and MI:** We analyze the variation in MIG and MI scores of generative factors across different $\beta$ values on the Cars3D dataset [40] to quantitatively track disentanglement in the latent space. Notably, the highest MIG scores for each factor are achieved at different $\beta$ values.

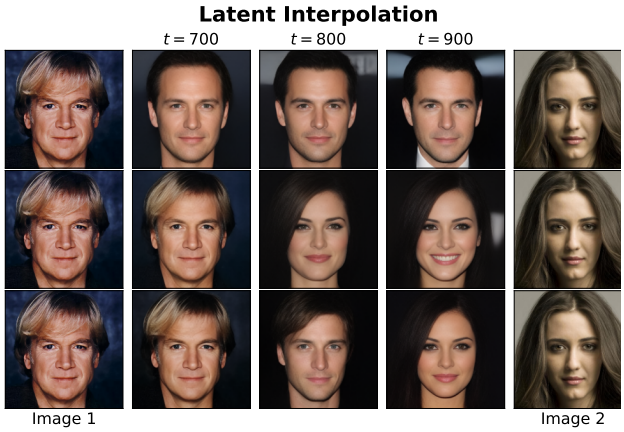


Figure 4. **Smoothness of latent space:** We provide evidence for the smoothness of the learned representations by partially interpolating between two images (leftmost and rightmost columns). In each row, we vary the timestep at which interpolation occurs, and in each column, we change the latent variables being interpolated.

mance. Both in this subsection and the next, we use the same trained generative models for consistency. For all the three image datasets, we train the diffusion model with one thousand time steps or $\beta$ values in the range $[0, 1]$.

For these image datasets, we evaluate our model by using FID [17] to assess image quality and precision-recall [31] to gauge data distribution coverage. Based on our comparative performance with generation-focused baselines in Tab. 2, we conclude that our model effectively generates high-quality images at this resolution across various datasets. Details of our architecture and training times are provided in Appendix D, and generated sample images are shown in Appendix C.2.

### 6.3. Exploring Learned Latents

First, we test the smoothness of the learned latent space by interpolating between two images. In this experiment, we use a trained model $\epsilon_\psi$ and a pair of images. Latent representations are obtained for both images by encoding them with a chosen $t$ (or $\beta$) value, then partially interpolating these representations by averaging $15\%$ of the latent dimensionality for each row. Next, we denoise these representations back to $t = 0$ using the learned diffusion model and decode them to image space, as shown in Fig. 4. The random seed is manually controlled throughout the process. This analysis shows that our representation space maintains smoothness along the time (or $\beta$) axis.


Figure 5. **Editing images:** As demonstrated by the example above, our model can also edit specific attributes of a given image. In the top row, we change the hair style, and in the middle row, we modify the hair colour and in the bottom row we remove glasses. The last two edits are performed between the 200th and 300th time steps while the first one is performed at 650th step.

Furthermore, achieving specific facial attributes requires adjustments along the time axis and selecting the appropriate set of latent variables, as different facial attributes disentangle at various time steps. In Fig. 5, we provide examples showcasing our model's capability to edit specific attributes within an image. Attribute editing is performed by manipulating the latent encodings of an image at certain time steps. This example highlights an additional benefit of the disentangled representations learned by our model. Empirically, we observe that attributes such as glasses, hairstyle, and hair color can be modified between the 200th–300th time steps, whereas attributes like face angle and age require adjustments between the 700th–800th time steps. Further results in this category are provided in Appendix C.1.

## 7. Conclusion

We propose a new generative modeling framework that leverages a range of $\beta$ values to learn disentangled representations and sharp generation quality, including unconditional generation. Our framework introduces two key components: (1) a multi-$\beta$ VAE, producing a spectrum of latent representations that can be refined via a denoising diffusion process, and (2)

Table 2. **Generation quality:** We evaluate our model for unconditional image synthesis and report standard metrics, comparing them against baselines specifically designed for generation.

| CelebA-HQ 256 × 256 | | | | FFHQ 256 × 256 | | | | LSUN-Bedrooms 256 × 256 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **FID ↓** | **Prec. ↑** | **Recall ↑** | **Method** | **FID ↓** | **Prec. ↑** | **Recall ↑** | **Method** | **FID ↓** | **Prec. ↑** | **Recall ↑** |
| DC-VAE[37] | 15.8 | - | - | ImageBART[12] | 9.57 | - | - | ImageBART[12] | 5.51 | - | - |
| VQGAN+T[13] (k=400) | 10.2 | - | - | U-Net GAN[45] (+aug) | 10.9 (7.6) | - | - | DDPM[19] | 4.9 | - | - |
| PGGAN[23] | 8.0 | - | - | UDM[28] | 5.54 | - | - | UDM[28] | 4.57 | - | - |
| LSGM[51] | 7.22 | - | - | StyleGAN[24] | 4.16 | 0.71 | 0.46 | StyleGAN[24] | 2.35 | 0.59 | 0.48 |
| UDM[28] | 7.16 | - | - | ProjectedGAN[44] | 3.08 | 0.65 | 0.46 | ADM[9] | 1.90 | 0.66 | 0.51 |
| *ours* | 6.81 | 0.71 | 0.48 | *ours* | 5.65 | 0.72 | 0.48 | *ours* | 3.2 | 0.65 | 0.48 |

a non-linear diffusion model that links latent representations for different $\beta$ values. Our method achieves comparable disentanglement performance to dedicated baselines while maintaining high decoding quality and achieving generation quality on par with state-of-the-art models. To our knowledge, this is the first model that accomplishes both effective disentanglement and satisfactory generation quality.

## References


[1] Juhan Bae, Michael R. Zhang, Michael Ruan, Eric Wang, So Hasegawa, Jimmy Ba, and Roger Baker Grosse. Multi-rate VAE: Train once, get the full rate-distortion curve. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[2] Grigory Bartosh, Dmitry Vetrov, and Christian A Naesseth. Neural flow diffusion models: Learnable forward process for improved diffusion modelling. *Advances in Neural Information Processing Systems*, 37, 2024. 6, 7

[3] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $\beta$-vae. In *Neural Information Processing Systems*, 2017. 6, 1

[4] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 6, 7

[5] Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *Proc. International Conference on Learning Representation (ICLR)*, 2021. 6

[6] Jack H. Collins, Yifeng Huang, Simon Knapen, Benjamin Nachman, and Daniel Whiteson. Machine-Learning Compression for Particle Physics Discoveries. 2022. 4, 6

[7] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *Proc. International Conference on Learning Representation (ICLR)*, 2019. 2

[8] Gaurav Dewangan and Seetaram Maurya. Fault diagnosis of machines using deep convolutional beta-variational autoencoder. *IEEE Transactions on Artificial Intelligence*, 3(2): 287–296, 2022. 1, 6

[9] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 9

[10] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 6

[11] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. 7

[12] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 9

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 9

[14] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[15] Muhammad Waleed Gondal, Manuel Wüthrich, Đorđe Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Bessekon Akpo, Olivier Bachem, Bernhard Scholkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Neural Information Processing Systems*, 2019. 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 8

[18] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016. 1, 2, 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 6, 9

[20] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016. 1

[21] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. 1

[22] Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion, 2024. 1

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 7, 9, 4, 5, 6, 8

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 3, 4, 7, 9, 5, 6, 8

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 1

[26] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proc. CVPR*, 2024. 1

[27] Valentin Khrulkov, Leyla Mirvakhabova, I. Oseledets, and Artem Babenko. Disentangled representations from non-disentangled models. *ArXiv*, abs/2102.06204, 2021. 2, 7

[28] Dongjun Kim, Seung-Jae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, 2021. 9

[29] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018. 1, 2, 6, 7

[30] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 1, 2

[31] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Neural Information Processing Systems*, 2019. 8

[32] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. InfoGAN-CR and ModelCentrality: Self-supervised model training and selection for disentangling GANs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6127–6139. PMLR, 2020. 7

[33] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019. 2, 7

[34] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *ArXiv*, abs/1511.05644, 2015. 1

[35] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, 2018. 1

[36] Beatrix Miranda Ginn Nielsen, Anders Christensen, Andrea Dittadi, and Ole Winther. Diffenc: Variational diffusion with a learned encoder. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[37] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 823–832, 2021. 9

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 1

[39] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 14866–14876, 2019. 6

[40] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 7, 8

[41] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *ICLR*, 2022. 7

[42] Ryan J Richards and Austen M Groener. Conditional $\beta$-vae for de novo molecular generation. *arXiv preprint arXiv:2205.01592*, 2022. 6

[43] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 6, 3, 7

[44] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems*, pages 17480–17492. Curran Associates, Inc., 2021. 9

[45] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 9

[46] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. ControlVAE: Controllable variational autoencoder. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8655–8664. PMLR, 2020. 1, 6

[47] Raghav Singhal, Mark Goldstein, and Rajesh Ranganath. What's the score? automated denoising score matching for nonlinear diffusions. In *ICML*, 2024. 6

[48] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Neural Information Processing Systems*, 2016. 6

[49] Yuhta Takida, Wei-Hsiang Liao, Chieh-Hsin Lai, Toshimitsu Uesaka, Shusuke Takahashi, and Yuki Mitsufuji. Preventing oversmoothing in vae via generalized variance parameterization. *Neurocomputing*, 509:137–156, 2022. 4, 1, 2

[50] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 19667–19679, 2020. 6

[51] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, 2021. 9

[52] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 6306–6315, 2017. 6

[53] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: representation learning using information maximizing diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 7

[54] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1900–1910, 2023. 1

[55] Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 7

[56] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 3, 4, 7, 5, 8