f-DM: A MULTI-STAGE DIFFUSION MODEL VIA PROGRESSIVE SIGNAL TRANSFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models (DMs) have recently emerged as the SoTA tools for generative modeling in various domains. Standard DMs can be viewed as an instantiation of hierarchical variational autoencoders (VAEs) where the latent variables are inferred from input-centered Gaussian distributions with fixed scale and variance. Unlike VAEs, this formulation constrains DMs from changing the latent spaces and learning abstract representations. In this work, we propose f-DM, a generalized family of DMs, which allows progressive latent space transformation. More precisely, we extend DMs to incorporate a set of degradation functions, where the transformed input is the mean of each diffusion step. We then derive the corresponding de-noising objective together with a modified sampling algorithm. As a demonstration, we apply f-DM in image generation tasks with a range of functions including down-sampling, blurring and learned transformations based on VAEs. In addition, we identify the importance of rescaling the noise levels whenever the signal is sub-sampled. Our models can produce high-quality samples on standard image generation benchmarks like FFHQ, AFHQ, LSUN and ImageNet with better efficiency and semantic interpretation.

1 INTRODUCTION

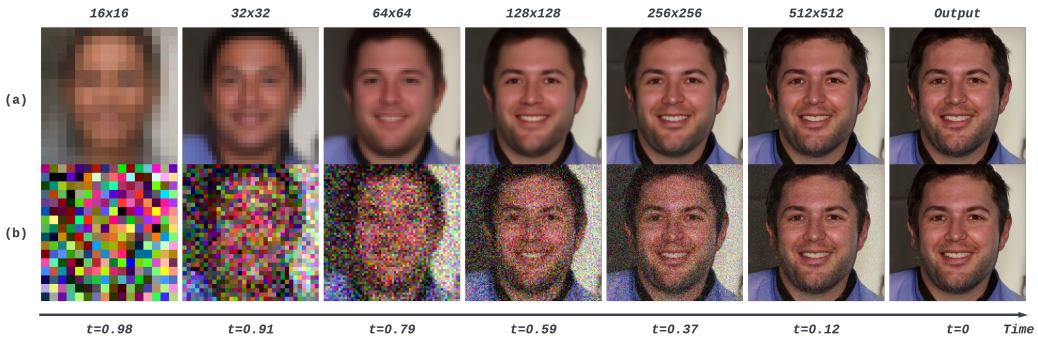

Figure 1: An example of image generation process from f-DM trained with *down-sampling*. (a) is the predicted target $\hat{\boldsymbol{x}}$ at each generation step, and (b) is the input $\hat{\boldsymbol{z}}_s$ for the next diffusion step.

Diffusion probabilistic models (DMs, Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021) and score-based (Song et al., 2021b) generative models have become increasingly popular as the tools for high-quality image (Dhariwal & Nichol, 2021), video (Ho et al., 2022b), text-to-speech (Popov et al., 2021) and text-to-image (Rombach et al., 2021; Ramesh et al., 2022; Saharia et al., 2022) synthesis. Despite the power of learning arbitrary distributions, conventional DMs are restricted to operate in the same space throughout the Gaussian noising process. On the other hand, common generative models such as VAEs (Kingma & Welling, 2013; Child, 2021) or GANs (Goodfellow et al., 2014; Karras et al., 2021) do not have such limitations, and typically employ a coarse-to-fine process that hierarchically generates high-resolution outputs.

In this paper, we propose f-DM, a generalization of DMs to incorporate progressive transformations to the inputs. By doing so, we achieve better efficiency and/or better sampling quality. We show

that signal transformation can easily be incorporated to a multi-stage diffusion framework, which makes it a strict generalization of standard DMs. This makes our work drastically different from other works (Ho et al., 2022a) where transformations (e.g., downsample) are introduced as a pre-processing step, after which DMs are applied as a black-box algorithm.

We consider general types of transformations including down-sampling, blurring and neural-based transformations. While some of these transformations are separately considered in prior works, we point out that it is critical to diffuse with a continuously transformed input based on interpolation, and rescale the noise levels when the stage changes. We further give principled guidelines by proposing a *resolution-agnostic* SNR.

Extensive experiments are performed on image generation benchmarks, including FFHQ, AFHQ, LSUN Bed/Church and ImageNet. We show that f-DM consistently match or outperform the baseline performance, while requiring significantly less compute. We also perform ablations on our design choices and demonstrate that our proposed noise schedule adjustment is effective.

2 Background

Diffusion Models Standard diffusion models (DMs) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) are deep hierarchical generative models defined by a Markovian Gaussian process. In this paper, we consider diffusion in continuous time similar to Song et al. (2021b); Kingma et al. (2021). Given a continuous datapoint $\boldsymbol{x} \in \mathbb{R}^N$, a DM models continuous time latent variables $\boldsymbol{z} = \{\boldsymbol{z}_t | t \in [0, 1], \boldsymbol{z}_0 = \boldsymbol{x}\}$ based on a pre-defined signal-noise schedule $\{\alpha_t, \sigma_t\}$:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_s) = \mathcal{N}(\boldsymbol{z}_t; \alpha_{t|s}\boldsymbol{z}_s, \sigma^2_{t|s}I), \quad 0 \leq s < t \leq 1, \tag{1}$$

where $\alpha_{t|s} = \alpha_t/\alpha_s, \sigma^2_{t|s} = \sigma^2_t - \alpha^2_{t|s}\sigma^2_s$. It also defines the marginal distribution $q(\boldsymbol{z}_t|\boldsymbol{x})$ as:

$$q(\boldsymbol{z}_t|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t; \alpha_t\boldsymbol{x}, \sigma^2_t I), \quad t \in [0, 1], \tag{2}$$

By default, we assume the variance preserving form (Ho et al., 2020). That is, $\alpha^2_t + \sigma^2_t = 1, \alpha_0 = \sigma_1 = 1$, and the signal-to-noise-ratio (SNR, α^2_t/σ^2_t) decreases monotonically with t. To synthesize data from the latents, a parametric function θ is optimized to reverse the diffusion process by denoising $\boldsymbol{z}_t = \alpha_t\boldsymbol{x} + \sigma_t\boldsymbol{\epsilon}$ to the clean input $\boldsymbol{x}$, with a weighted reconstruction loss $\mathcal{L}_\theta$. For example, the "simple loss" proposed in Ho et al. (2020) is equivalent to $\omega_t = \alpha^2_t/\sigma^2_t$:

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{z}_t \sim q(\boldsymbol{z}_t|\boldsymbol{x}), t\sim[0,1]} \left[\omega_t \cdot \|\boldsymbol{x}_\theta(\boldsymbol{z}_t, t) - \boldsymbol{x}\|^2_2\right]. \tag{3}$$

In practice, θ is parameterized as a U-Net (Ronneberger et al., 2015). As suggested in Ho et al. (2020), predicting the noise $\boldsymbol{\epsilon}_\theta$ empirically achieves better performance than predicting $\boldsymbol{x}_\theta$, where $\boldsymbol{x}_\theta(\boldsymbol{z}_t, t) = (\boldsymbol{z}_t - \sigma_t\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t))/\alpha_t$. Sampling from such learned model can be performed from ancestral sampling (DDPM, Ho et al., 2020), or a deterministic DDIM sampler (Song et al., 2021a). Starting from $\boldsymbol{z}_1 \sim \mathcal{N}(\mathbf{0}, I)$, a sequence of timesteps $1 = t_0 > \ldots > t_N = 0$ are sampled for iterative generation, and we can readily summarize both methods for each step as follows:

$$\boldsymbol{z}_s = \alpha_s \cdot \boldsymbol{x}_\theta(\boldsymbol{z}_t) + \sqrt{\sigma^2_s - \eta^2\bar{\sigma}^2} \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t) + \eta\bar{\sigma} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I), \quad s < t, \tag{4}$$

where $\bar{\sigma} = \sigma_s\sigma_{t|s}/\sigma_t$, and η controls how much noise is added (DDPM $\eta = 1$, DDIM $\eta = 0$).

Due to the element-wise summation and scaling in Equation 2, it is clear that all the latent variables $\boldsymbol{z}$ are forced to be the same shape as the input data $\boldsymbol{x}$ ($\mathbb{R}^N$). This not only leads to inefficient training, especially for steps with high noise level (Jing et al., 2022), but also makes DMs hard to learn abstract and semantically meaningful latent space as pointed out by Preechakul et al. (2022).

Hierarchical VAEs Variational autoencoders (VAEs, Kingma & Welling, 2013) are a broader class of generative models with latent variables, and DMs can be seen as a special case with a fixed encoder and "infinite" depth. Instead, the encoder ($q_\phi(\boldsymbol{z}|\boldsymbol{x})$) and decoder ($p_\theta(\boldsymbol{x}, \boldsymbol{z})$) of VAEs are optimized jointly for the variational lower bound similar to DMs. Unlike DMs, there are no restrictions for the dimensions of the latent space. Therefore, VAEs are widely known as the tools for learning disentangled semantics in various domains. Figure 2 shows a comparison between DMs and a "bottown-up style" hierarchical VAEs. Recently, deep VAEs (Vahdat & Kautz, 2020; Child, 2021) have made progress on high-quality generation with top-down architectures, however, the visual outputs still have gaps with other generative models such as GANs or DMs. In this work, our goal is to take inspiration from hierarchical VAEs, enabling diffusion in the transformed space.

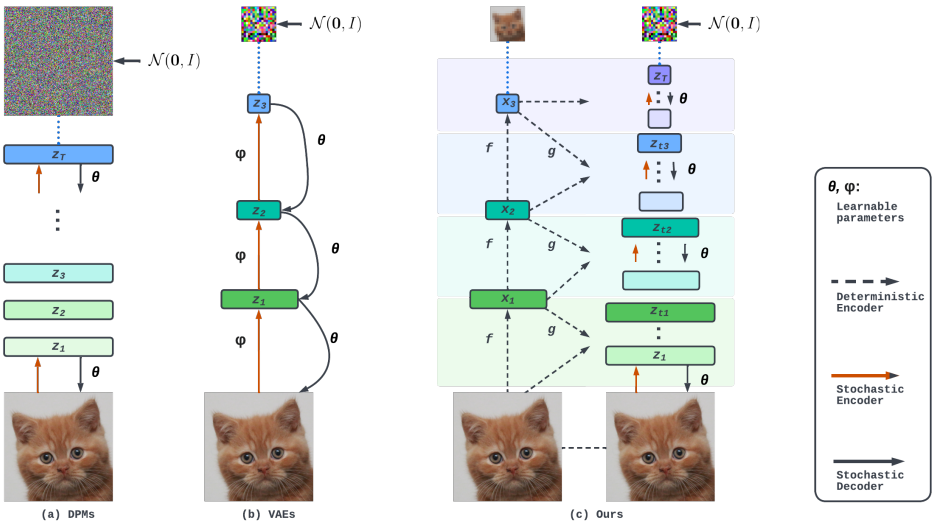

Figure 2: Motivation (a) the standard DMs; (b) a bottom-up instantiation of hierarchical VAEs; (c) our proposed f-DM which is a DM with transformed latent space.

DMs beyond Gaussian Noise Recently, several concurrent research (Rissanen et al., 2022; Bansal et al., 2022; Daras et al., 2022) have looked into DM alternatives to combine transformations (e.g., blur, pixelate) other than Gaussian noise. Although sharing similarities, our work starts from a different view on DMs and signal transformation. Furthermore, our experiments will show that Gaussian noise acts a critical role for high-fidelity image synthesis.

3 Method

In this section, we first dive into f-DM, an extended family of DMs, which enables diffusion over transformed signals with the modified training and sampling algorithms. Then, we will specifically present f-DM in several existing transformations.

3.1 Multi-stage Diffusion with Progressive Transformations

Degradation Functions We consider a sequence of deterministic functions $\boldsymbol{f} = \{f_0, f_1, \ldots, f_K\}$, where $f_0 \ldots f_k$ progressively transforms the input signal $\boldsymbol{x} \in \mathbb{R}^N$ into $\boldsymbol{x}_k = f_{0:k}(\boldsymbol{x}) \in \mathbb{R}^{M_k}$. We assume $\boldsymbol{x}_0 = f_0(\boldsymbol{x}) = \boldsymbol{x}$. In principle, $\boldsymbol{f}$ can be any functions. In this work, we focus on transformations that gradually destroy the information contained in $\boldsymbol{x}$ (e.g., down-sampling), leading towards more abstract and lower-dimensional representations. Without loss of generality, we assume $M_0 \geq M_1 \geq \ldots \geq M_K$. Then, a sequence of *inverse* mappings $\boldsymbol{g} = \{g_0, \ldots, g_{K-1}\}$ is used to connect each two consecutive spaces. Specifically, we define $\hat{\boldsymbol{x}}_k$ as:

$$\hat{\boldsymbol{x}}_k := \begin{cases} g_k\left(f_{k+1}(\boldsymbol{x}_k)\right) \approx \boldsymbol{x}_k, & \text{if } k < K, \\ \boldsymbol{x}_k, & \text{if } k = K. \end{cases} \tag{5}$$

In our formulation, the approximation of Equation 5 ($k < K$) is not necessarily (and sometimes impossibly) accurate. For instance, f_k downsamples an input image $\boldsymbol{x}$ from 128^2 into 64^2 with average pooling, and g_k can be bilinear interpolation that upsample back to 128^2.

The definition of $\boldsymbol{f}$ and $\boldsymbol{g}$ can be seen as a direct analogy of the encoder (ϕ) and decoder (θ) in hierarchical VAEs (see Figure 2 (b)). However, there are still major differences: (1) the VAE encoder/decoder are stochastic and the encoder's outputs are regularized by the prior. In contrast, $\boldsymbol{f}$ and $\boldsymbol{g}$ are deterministic mappings, and the encoder output $\boldsymbol{x}_K$ does not necessarily follow a simple prior; (2) VAEs directly use the decoder path for generation, while we will show later that $\boldsymbol{f}, \boldsymbol{g}$ are fused in the diffusion process which enables us to generate higher quality samples than VAEs.

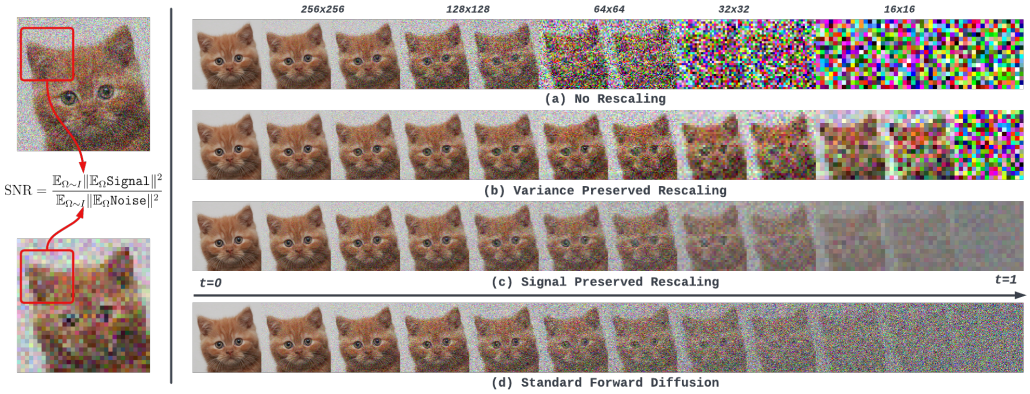

Figure 4: Left: an illustration of the proposed SNR computation for different sampling rates; Right: the comparison of rescaling the noise level for progressive down-sampling. Without noise rescaling, the diffused images in low-resolution quickly become too noisy to distinguish the underline signal.

Multi-stage Diffusion We extend the continuous-time DMs for transformed signals. We split the timesteps t into $K+1$ stages. For each stage, a partial diffusion process is performed in the corresponding space. More specifically, we define a set of time boundaries $0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = 1$, and for $t \in [0, 1]$, the latent $\boldsymbol{z}_t$ has the following marginal probability:

$$q(\boldsymbol{z}_t|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t; \alpha_t \hat{\boldsymbol{x}}_k^t, \sigma_t^2 I), \quad \text{where } \hat{\boldsymbol{x}}_k^t = \frac{(t-\tau_k)\hat{\boldsymbol{x}}_k + (\tau_{k+1}-t)\boldsymbol{x}_k}{\tau_{k+1}-\tau_k}, \quad \tau_k \le t < \tau_{k+1}. \tag{6}$$

The standard DMs is the special case of our f-DM when $K = 0$. As listed above, $\hat{\boldsymbol{x}}_k^t$ is the interpolation of $\boldsymbol{x}_k$ and its approximation $\hat{\boldsymbol{x}}_k$. We argue that such interpolation is crucial for synthesizing high-fidelity samples, as it creates a *continuous* signal transformation to slowly remove information inside each stage. In this way, such change can be easily reversed by our model.

It is also worth a mention that Equation 6 does not guarantee a Markovian transition. Nevertheless, in our formulation, only $q(\boldsymbol{z}_t|\boldsymbol{z}_s, \boldsymbol{x})$ is needed which has the following simple form:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_s, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}_t; \alpha_{t|s}\boldsymbol{z}_s + \alpha_t \cdot \left(\hat{\boldsymbol{x}}_k^t - \hat{\boldsymbol{x}}_k^s\right), \sigma_{t|s}^2 I), \quad \tau_k \le s < t < \tau_{k+1}. \tag{7}$$

From the above equation, we further re-write $\hat{\boldsymbol{x}}_k^t - \hat{\boldsymbol{x}}_k^s = \boldsymbol{\delta}_k^t \cdot (t-s)/(t-\tau_k)$, where $\boldsymbol{\delta}_k^t = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^t$ is the signal degradation. Equation 7 also indicates that the reverse diffusion distribution $p(\boldsymbol{z}_s|\boldsymbol{z}_t, \boldsymbol{x}) \propto q(\boldsymbol{z}_t|\boldsymbol{z}_s, \boldsymbol{x})q(\boldsymbol{z}_s|\boldsymbol{x})$ can be written the function of $\boldsymbol{z}_t$, $\hat{\boldsymbol{x}}_k^t$ and $\boldsymbol{\delta}_k^t$, and we will show later to use neural networks to predict $\hat{\boldsymbol{x}}_k^t$ and $\boldsymbol{\delta}_k^t$ from $\boldsymbol{z}_t$.

Boundary Condition Equation 7 defines only the diffusion steps ($s \to t$) in the same stage. To enable generation across stages, we also need to model the transition at each boundary $\tau_1, \ldots, \tau_K$. More specifically, when the diffusion step approaches the boundary $\tau - \tau^- \to 0^+$, the transition $q(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau^-}, \boldsymbol{x})$ as deterministic & smooth as possible.[1].

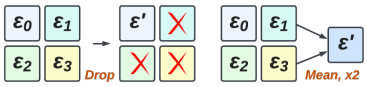

Figure 3: Two naïve ways of $\zeta(\boldsymbol{\epsilon})$ for down-sampling degradation.

For the first requirement, $q(\boldsymbol{z}_\tau|\boldsymbol{z}_{\tau^-}, \boldsymbol{x})$ becomes deterministic if we can find $\zeta(.)$ that:

$$\boldsymbol{z}_\tau = \alpha_\tau \cdot \boldsymbol{x}_k + \sigma_\tau \cdot \zeta(\boldsymbol{\epsilon}) \in \mathbb{R}^{M_k}, \quad \text{s.t. } p(\zeta(\boldsymbol{\epsilon})) = \mathcal{N}(\mathbf{0}, I), \tag{8}$$

where based on Equation 6, $\boldsymbol{\epsilon} = (\boldsymbol{z}_{\tau^-} - \alpha_{\tau^-} \cdot \hat{\boldsymbol{x}}_{k-1}))/\sigma_{\tau^-} \in \mathbb{R}^{M_{k-1}}$ and $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, I)$. Under the initial assumption of $M_k \le M_{k-1}$, this can be achieved easily by dropping or averaging elements from $\boldsymbol{\epsilon}^-$. Take down-sampling ($M_{k-1} = 4M_k$) as an example (see Figure 3). We can either drop 3 out of 4 values from $\boldsymbol{\epsilon}$, or average every 2×2 noises with $\times 2$ magnitude.

However, the second requirement of smooth transition is not as straightforward as it looks, which asks the "noisiness" of latents $\boldsymbol{z}$ unchanged when the stage shifts ($\mathbb{R}^{M_{k-1}} \to \mathbb{R}^{M_k}$). We argue that the conventional measure – the signal-to-noise-ratio (SNR) – in DM literature is not compatible with

[1]For simplicity, we omit the subscript k for τ_k in the following paragraphs.

resolution change as it averages the signal/noise power directly for each element. In this work, we propose a *resolution-agnostic* SNR by assuming the data as continuous domain signals:

$$\texttt{SNR}(\hat{\boldsymbol{o}}) = \frac{\mathbb{E}_{\Omega\sim I}\|\mathbb{E}_{i\sim\Omega}\tilde{\boldsymbol{o}}_i\|^2}{\mathbb{E}_{\Omega\sim I}\|\mathbb{E}_{i\sim\Omega}\boldsymbol{o}_i - \tilde{\boldsymbol{o}}_i\|^2}, \tag{9}$$

where I is the data range, and Ω is the minimal interested patch relative to I. $\boldsymbol{o}$ and $\tilde{\boldsymbol{o}}$ are the clean and corrupted signals, respectively. As shown in Figure 4 (left), we can obtain a reliable measure of noisiness for data with different sizes by averaging the signal/noise inside patches.

We can derive α_τ, σ_τ from $\alpha_{\tau^-}, \sigma_{\tau^-}$ for any transformations by forcing $\texttt{SNR}(\boldsymbol{z}_\tau) = \texttt{SNR}(\boldsymbol{z}_{\tau^-})$ under the new definition. Specifically, if the size difference is solely caused by change of sampling rate (e.g., down-sampling, average RGB channels, deconvolution), the following stands:

$$\alpha_\tau^2/\sigma_\tau^2 = d_k \cdot \gamma_k \cdot \alpha_{\tau^-}^2/\sigma_{\tau^-}^2\ , \tag{10}$$

where $d_k = M_{k-1}/M_k$ is the total dimension reduction, and $\gamma_k = \mathbb{E}||\hat{\boldsymbol{x}}_{k-1}||^2/\mathbb{E}||\boldsymbol{x}_k||^2$ is the change of signal power. For example, we have $d_k = 4, \gamma_k \approx 1$ for typical down-sampling with interpolation. We leave the details of derivation in Appendix TODO. Equation 10 gives us only the guideline of adjusting α, σ jointly. Here, we propose two possible implementations:

- **Variance preserved rescaling (VP)**: $\alpha_\tau = \frac{\sqrt{d_k}\alpha_{\tau^-}}{\sqrt{1+(d_k-1)\alpha_{\tau^-}^2}}, \sigma_\tau = \frac{\sigma_{\tau^-}}{\sqrt{d_k-(d_k-1)\sigma_{\tau^-}^2}}$. By doing so, we keep $\sigma_\tau^2 + \sigma_\tau^2 = 1$ and the generation starts from $\mathcal{N}(\mathbf{0}, I)$ as standard DM.
- **Signal preserved rescaling (SP)**: $\alpha_\tau = \alpha_{\tau^-}, \sigma_\tau = \frac{1}{\sqrt{d_k}}\sigma_{\tau^-}$. In this way, $\sigma_\tau^2 + \sigma_\tau^2 < 1$ but the resulting image looks closer to the standard DM.

We also show a comparison of the forward process for both choices in Figure 4.

Training Following Equation 3, we learn a neural network θ to "denoise" the input. In our case, the "noise" are caused from two sources (1) the stochastic perturbation $\boldsymbol{\epsilon}$ from noise injection; (2) the degradation $\boldsymbol{\delta}$ due to signal degradation. Thus, we propose to predict two outputs – $\hat{\boldsymbol{x}}_\theta$ and $\boldsymbol{\delta}_\theta$, which simultaneously remove both noises from $\boldsymbol{z}_t$ with the following "double reconstruction" loss:

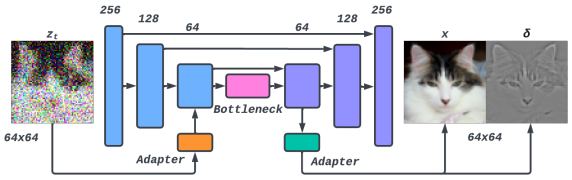

Figure 5: An illustration of the proposed architecture. Time conditioning is omitted for simplicity.

$$\mathcal{L}_\theta = \mathbb{E}_{\boldsymbol{z}_t\sim q(\boldsymbol{z}_t|\boldsymbol{x}),t\sim[0,1]}\left[\omega_t \cdot \left(\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_t,t) - \hat{\boldsymbol{x}}_k^t\|_2^2 + \|\boldsymbol{\delta}_\theta(\boldsymbol{z}_t,t) - \boldsymbol{\delta}_k^t\|_2^2\right)\right], \tag{11}$$

where we denote the denoised output as $\boldsymbol{x}_\theta(\boldsymbol{z}_t,t) = \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_t,t) + \boldsymbol{\delta}_\theta(\boldsymbol{z}_t,t)$. Unlike standard DMs, the f-DM objectives are the transformed signals of each stage, which is generally a simpler target to recover especially for high noise levels.

Architecture Figure 5 shows an illustration of our architecture, where the same U-Net is shared across all stages. Input $\boldsymbol{z}_t$ will be directed to the corresponding inner layer based on spatial resolutions, and a stage-specific adapter is adopted to transform the channel dimension. Such architecture enables memory-efficient batching for different stages (Appendix TODO). Optionally, the model predicts $\hat{\boldsymbol{\epsilon}}_\theta$ instead of $\hat{\boldsymbol{x}}_\theta$ similar to the standard DMs.

Generation Similar to standard DMs, we present the reverse diffusion steps in Algorithm 1, where $\hat{x}_k^t$ and δ_k^t are replaced by model's predictions $\hat{x}_\theta, \delta_\theta$. Thanks to the linear formulation (Equation 6), the resulting algorithm is independent of the transformations $\boldsymbol{f}$, and only $\boldsymbol{g}$ is needed at the stage boundaries. This brings flexibility & efficiency of learning complex or even test-time inaccessible transformations, which drastically differs our methods from related approaches (Bansal et al., 2022).

The noise schedulers are rescaled for stage changes before generation starts. In addition, the algorithm includes a "noise resampling step" at the stage boundary, which is the reverse process for $\zeta(\boldsymbol{\epsilon})$ in Equation 8. Although $\zeta(.)$ is deterministic in our assumption, the reverse $\zeta^{-1}(.)$ has additional degrees of freedom and depends on the choice of ζ which we refer as "noise resampling". For example (see Figure 3 left), if the ζ drops 3 out of 4 noise values ($\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}_0$), then the reverse step will inject standard Gaussian noise back to the dropped locations ($\boldsymbol{\epsilon}_0 = \boldsymbol{\epsilon}', \boldsymbol{\epsilon}_{1:3} \sim \mathcal{N}(\mathbf{0}, I)$). Similarly, we present the resampling method for averaging in Appendix todo.

Algorithm 1: Reverse diffusion for image generation using f-DM

Input: model θ, $\boldsymbol{f}, \boldsymbol{g}$, stage schedule $\{\tau_0, \ldots, \tau_K\}$, **rescaled** noise schedule functions $\alpha(.), \sigma(.)$, step-size Δt, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma(1) \cdot I)$, DDPM ratio η

1 **for** $(k = K; k \geq 0; k = k - 1)$ **do**
2 **for** $(t = \tau_{k+1}; t > \tau_k; t = t - \Delta t)$ **do**
3   $s = \max(t - \Delta t, \tau_k)$;
4   $\hat{\boldsymbol{\epsilon}}_\theta, \boldsymbol{\delta}_\theta = \theta(\boldsymbol{z}, t)$; $\hat{\boldsymbol{x}}_\theta = (\boldsymbol{z} - \sigma(t) \cdot \hat{\boldsymbol{\epsilon}}_\theta)/\alpha(t)$; $\boldsymbol{x}_\theta = \hat{\boldsymbol{x}}_\theta + \boldsymbol{\delta}_\theta$
5   $\boldsymbol{z} = \alpha(s) \cdot (\hat{\boldsymbol{x}}_\theta + \boldsymbol{\delta}_\theta \cdot (t - s)/(t - \tau_k)) + \sqrt{\sigma^2(s) - \eta^2 \bar{\sigma}^2} \cdot \hat{\boldsymbol{\epsilon}}_\theta + \eta \bar{\sigma} \cdot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$
6 **if** $k > 0$ **then**
7   Resample $\hat{\boldsymbol{\epsilon}}'_\theta = \zeta^{-1}(\hat{\boldsymbol{\epsilon}}_\theta)$; $\boldsymbol{z} = \alpha(s) \cdot g_k(\boldsymbol{x}_\theta) + \sigma(s) \cdot \hat{\boldsymbol{\epsilon}}'_\theta$
8 **return** $\boldsymbol{x}_\theta$

3.2 Application with Various Signal Transformations

With the definition in Section 3.1, we can show that when applying different transformations, the proposed f-DM are not only able to sample more efficiently, but enable various applications by "inverting" $\boldsymbol{f}$ during generation. In this paper, we consider the following transformations:

Progressive Super-Resolution As the major example we use to explain Section 3.1, we consider $\boldsymbol{f}$ a sequence of `downsampe` operations that transforms a given image (e.g., 256^2) progressively down to 16^2, where each $f_k(.)$ reduces the length by 2, and correspondingly $g_k(.)$ upsamples by 2. Thus, the generation starts from a low-resolution noise and progressively perform super-resolution. As discussed, $d_k = 4, \gamma_k = 1$ in Equation 10 and $K = 4$ for 256^2 images. In our early experiments, we found that implementing both f_k and g_k as bilinear interpolation achieves best performance.

De-blurring f-DM also supports general `blur` transformations. Unlike recent works (Rissanen et al., 2022; Hoogeboom & Salimans, 2022) that focuses on continuous-time blur (heat dissipation), Equation 6 can be seen as an instantiation of progressive blur function if we treat $\hat{\boldsymbol{x}}_k$ as a blurred version of $\boldsymbol{x}_k$. This design brings more flexibility of choosing any kind of blurring functions, and using the blurred versions as stages for f-DM to recover progressively. In this paper, we experiment two types of blurring functions: (1) `DownUpBlur` which utilizes the same operators in super-resolution cases, but always upsample the images to the original sizes; (2) `GaussianBlur` which applies the standard Gaussian blurring kernel following Rissanen et al. (2022). In both cases, we use $g_k(\boldsymbol{x}) = \boldsymbol{x}$. As the dimension is not changed, no rescaling and noise resampling are required.

Latent Diffusion We further consider diffusion with learned non-linear transformations like VAEs (Figure 2 (b)). Similar to Rombach et al. (LDM, 2021), we define *latent diffison* generation that starts from the low-dimensional latent space and progressively synthesizes images. As a major difference, LDM operates only on the latent variables, while we learn f-DM for the latent and image spaces jointly. Because of this, the performance will not be bounded by the quality of VAEs. In this paper, we adopt VQVAE (Van Den Oord et al., 2017) together with its variant (VQGAN, Esser et al., 2021) which is known to generate high fidelity samples thanks to the perceptual & adversarial objectives. For both cases, we transform $256^2 \times 3$ images into $32^2 \times 4$ (i.e., $d_k = 48$) latent space. We train our own VQVAE on ImageNet (Deng et al., 2009), and fix its weights for all the rest of experiments. For VQGAN experiments, we directly take the checkpoint [2] provided by Rombach et al. (2021). Since the signal magnitude also changes, it is necessary to tune γ_k separately for each encoder.

4 Experiments

In this section, we empirically validate f-DM on a set of image generation tasks. We show that f-DM is able to synthesize comparable or even better visual generation than the baselines, while enjoys appealing properties (e.g., efficient inference) when training on different signal transformation.

[2] `https://ommer-lab.com/files/latent-diffusion/vq-f8-n256.zip`

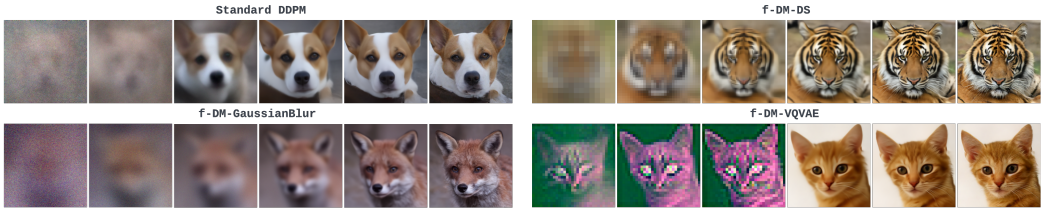

Figure 6: Visualization of the diffusion process (predicted signal at each step)

Table 1: (**placeholder**) Quantitative comparisons against baselines on various datasets. We calculate FID, Precsion, Recall, and the average rendering time (batch size $= 1$). For fair comparison, DDPM baseline (Ho et al., 2020) are implemented and evaluated under the same settings as f-DM.

Models	**FID**↓	**P**↑	**R**↑	**FID**↓	**P**↑	**R**↑	**Speed**
	FFHQ 256×256			**AFHQ** 256×256			
DDPM	10.8	0.76	0.53				
Cascaded DM							
Cold Diffusion							
f-DM-DS							
IHDM	64.9			43.4			
Cold Diffusion				54.7			
f-DM-Blur-G							
f-DM-Blur-U							
LDM (VAE)							
f-DM-VAE							
f-DM-GAN							

Models	**FID**↓
LSUN-Church 256×256	
DDPM	
f-DM-DS	
f-DM-VAE	
LSUN-Bedroom 256×256	
DDPM	
f-DM-DS	
f-DM-VAE	
ImageNet 256×256	
DDPM	
f-DM-DS	
f-DM-VAE	

4.1 Experimental Settings

Datasets We evaluate f-DM on five commonly used image generation benchmarks: FFHQ (Karras et al., 2019), AFHQ (Choi et al., 2020), LSUN Church& Bedroom (Yu et al., 2015), and ImageNet (Deng et al., 2009). All images are center-cropped and resized to 256×256. For more dataset and preprocessing details, please refer to Appendix todo.

Model Configurations As discussed in Section 3.2, we implement three types of transformations, where we denote f-DM-DS for downsampling, f-DM-Blur-G/U for Gaussian-based and up/down-based blurring, and f-DM-VAE/GAN for VQVAE/VQGAN encoders, respectively. In Figure 6, we show examples of the reverse diffusion from f-DM of three types, together with the standard DDPM (Ho et al., 2020). For all cases, we share the same architecture and hyper-parameters except for the transformation-specific adapters discussed in Figure 5.

Baselines Baselines: standard DM, cold diffusion (?), Upsample cascade?

4.2 Results

- Main results of using DS/AE on FFHQ, AFHQ, LSUN
- Conditional Generation for AFHQ and ImageNet (classifier-free guidance)
- Cold-diffusion 128 AFHQ of ours
- HD directly use number and figures, and show problems in prior
- VQVAE reconstruction FID? latent diff on VQVAE?
- Hierarchical VQVAE?
- Cascaded models?
- Interpolation latent space

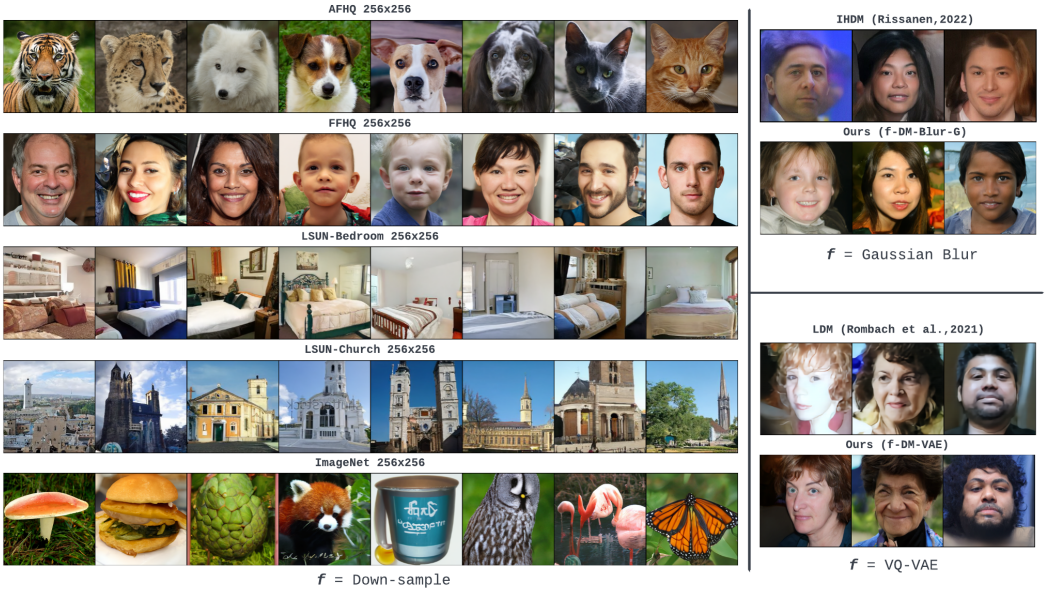

Figure 7: **Placeholder** Uncurated samples from models trained in 256×256. All faces presented are synthesized by the models without real identities.

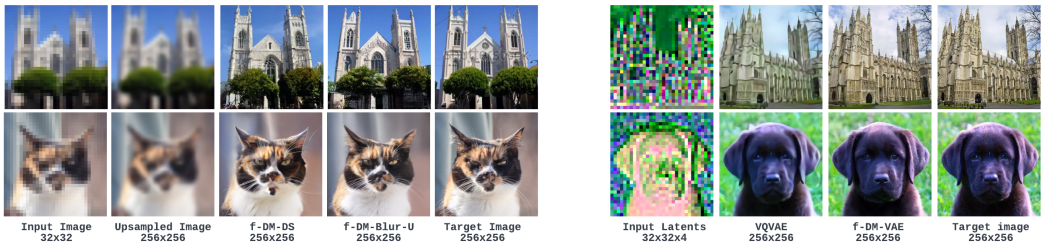

Figure 8: **Placeholder** Conditional generation (Alert: real human faces, need to change)

- blur variant hard to be fair comparison

4.3 Ablation Study

- Do we need interpolation?
- How noise rescaling affects?
- How stage schedule change the results
- Other possible ablation

5 Related Work

Do we need this section?

- score based generative modeling, discrete and continuous
- casecaded models, pyramid generation (connecting to DS)
- latent diffusion (connecting to autoencoder)

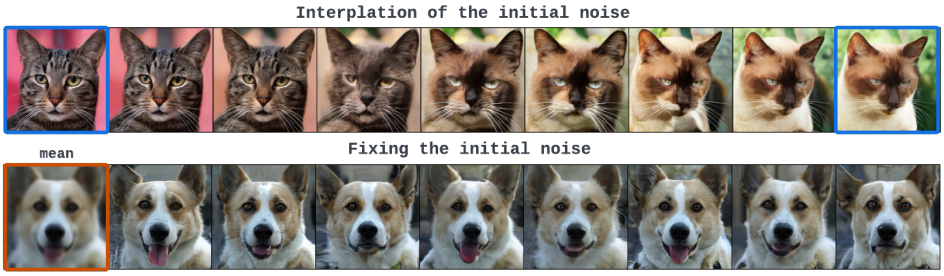

Figure 9: **Placeholder** Learning of the latent space

6 Conclusions

References

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G. Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions, 2022. URL `https://arxiv.org/abs/2209.05442`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

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*, pp. 12873–12883, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23: 47–1, 2022a.

Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022b.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models, 2022. URL https://arxiv.org/abs/2209.05557.

Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. *arXiv preprint arXiv:2205.01490*, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.