

# Exploiting the Signal-Leak Bias in Diffusion Models

Martin Nicolas Everaert<sup>1</sup>    Athanasios Fitsios<sup>1,2</sup>    Marco Bocchio<sup>2</sup>  
 Sami Arpa<sup>2</sup>    Sabine Süsstrunk<sup>1</sup>    Radhakrishna Achanta<sup>1</sup>

<sup>1</sup>School of Computer and Communication Sciences, EPFL, Switzerland    <sup>2</sup>Largo.ai, Lausanne, Switzerland

Project page: <https://ivrl.github.io/signal-leak-bias/>

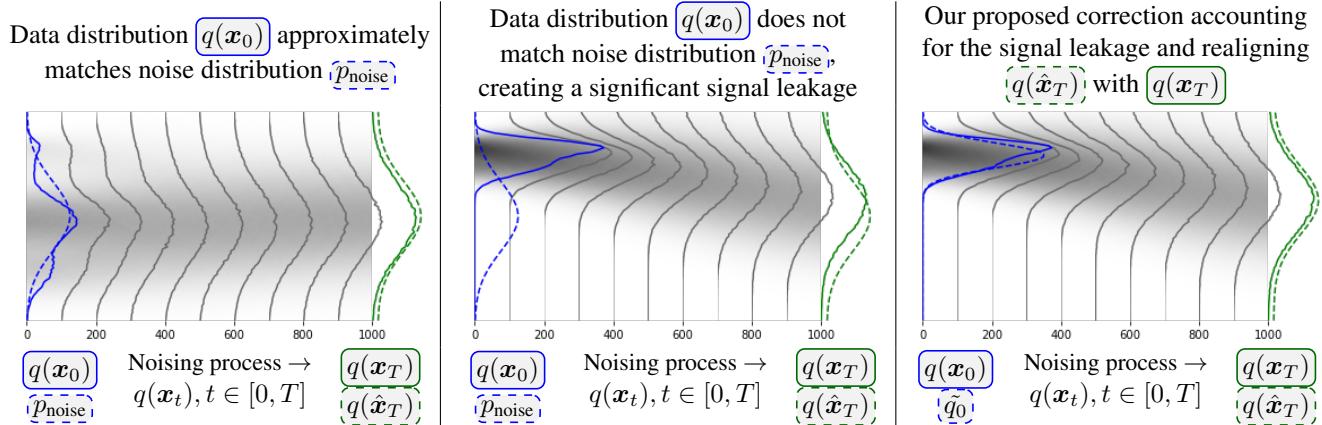


Figure 1. Current diffusion models contain a *signal leakage*, creating a mismatch between training distribution  $q(\mathbf{x}_T)$  and inference distribution  $q(\hat{\mathbf{x}}_T)$ . This leakage is pronounced when the diffusion model is tuned on a specific subset of images (middle column), but also exists in the original model (left column). We propose (right column) to realign the inference distribution  $q(\hat{\mathbf{x}}_T)$  with the training distribution  $q(\mathbf{x}_T)$ , by modeling the distribution of the signal leak via  $\tilde{q}_0 \approx q(\mathbf{x}_0)$ . (For visualization purposes, the plots consider images as 1D data points, their data distributions  $q(\mathbf{x}_0)$  are chosen arbitrarily, and the noise schedule used here accentuates the discrepancy.)

## Abstract

There is a bias in the inference pipeline of most diffusion models. This bias arises from a signal leak whose distribution deviates from the noise distribution, creating a discrepancy between training and inference processes. We demonstrate that this signal-leak bias is particularly significant when models are tuned to a specific style, causing sub-optimal style matching. Recent research tries to avoid the signal leakage during training. We instead show how we can exploit this signal-leak bias in existing diffusion models to allow more control over the generated images. This enables us to generate images with more varied brightness, and images that better match a desired style or color. By modeling the distribution of the signal leak in the spatial frequency and pixel domains, and including a signal leak in the initial latent, we generate images that better match expected results without any additional training.

## 1. Introduction

Denoising diffusion models [9] employ a sequential denoising process to generate visually appealing images from noise. During training, real images are corrupted with white noise, and the diffusion model is tasked to denoise the corrupted images back to their uncorrupted versions. During inference, the trained diffusion model is given white noise, which it progressively denoises to generate realistic images.

Interestingly, during training, images are corrupted to various degrees, but in the case of most currently available models, if not all, images are never corrupted down to *complete* noise [7, 15]. Even at the last timestep, when the noise level is maximal, corrupted images contain a *signal leak*, *i.e.* they are not composed only of noise but still contain a part of the original real images [15]. Only a limited number of studies examine whether the inference process actually matches the training process [14, 15, 22, 33], and only a few [7, 14, 15] explicitly mention this issue. Following

this observation, we argue that starting denoising from only noise at inference time is not aligned with the training process and often results in a *signal-leak bias*.

For instance, images generated with Stable Diffusion [28] tend to always have a medium brightness [7, 15]. This bias occurs because the model learns to utilize the brightness of the signal leak to infer the brightness of the real image. At inference time, *starting denoising from white noise is biased* toward generating images with medium brightness, because white noise, which the model interprets as the signal leak, has a medium brightness. Likewise, sampling the initial latent from white noise also biases the generated images to have medium low-frequency components in general, *i.e.* colors and brightness tend to be similar in different areas of the image. More importantly, we notice that the signal-leak bias *prevents models tuned on specific styles from faithfully reproducing the desired styles*.

Aware of the existence of this bias, recent research [7, 15] proposes to fine-tune the diffusion models to reduce or remove the signal leak during training, hence generating images with more varied brightness. We, on the contrary, propose to *exploit the signal-leak bias to our advantage*. Instead of fine-tuning or retraining models to eliminate the bias, our approach consists in estimating the distribution of the signal leak with a simple distribution  $\tilde{q}_0 \approx q(\mathbf{x}_0)$ . This is done by computing statistics on a small set of target images, for instance, the mean and covariance of their low-frequency content, or the mean and element-wise variance of the pixel values. During inference, rather than denoising from a latent made only of white noise, we *start denoising from a latent composed of both white noise and signal leak*, exactly like during the training of the diffusion model.

Figure 1 visually depicts the source of the signal-leak bias. The middle column shows intuitively the discrepancy between training and inference caused by the signal leakage when the model is tuned for a specific style. Our approach shown in the right column, mitigates this discrepancy, by introducing a signal leak  $\sqrt{\bar{\alpha}_T} \tilde{\mathbf{x}}, \tilde{\mathbf{x}} \sim \tilde{q}_0$  in the initial latents  $\hat{\mathbf{x}}_T$  during inference, hence mirroring the training distribution.

Our contributions are thus as follows:

- We provide an analysis of the signal-leak bias, with new insights on its origin and its implications. (Section 3)
- We propose a novel approach to include a signal leak in the sampling of the initial latents, instead of sampling them from noise only, biasing the generated images towards generating specific features. (Section 4)

We show in this paper how to use this approach and leverage the signal-leak bias to our advantage to:

- significantly enhance the outcome of models tuned on images of a specific style, without any additional fine-tuning. (Section 5.1)
- generate images in a particular style with Stable Diffusion [28] without any fine-tuning. (Section 5.2)

- obtain more diverse images with Stable Diffusion [28], fixing the issue of generated images having a medium brightness, and this without any training. (Section 5.3)
- provide greater control over the generated images, allowing to generate images with specific mean color, without any training. (Section 5.4)

## 2. Background and related work

### 2.1. Denoising Diffusion Probabilistic Models

Diffusion models [9] learn to denoise corrupted versions  $\mathbf{x}_t$  of images  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ . The noising diffusion process comprises  $T$  timesteps, typically  $T = 1000$ . At the first timestep  $t = 1$ , the image  $\mathbf{x}_1$  is a slightly noisy version of  $\mathbf{x}_0$ . At the last timestep  $t = T$ , the image  $\mathbf{x}_T$  is almost indistinguishable from noise. Transitions from  $\mathbf{x}_0$  to  $\mathbf{x}_T$  are parameterized by a noise schedule, *i.e.* a function  $\alpha_t$  of the timestep  $t$ . For any timestep  $t \in [1, T]$ , the noisier version  $\mathbf{x}_t$  of  $\mathbf{x}_{t-1}$  is obtained from the conditional distribution:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \quad (1)$$

$$\text{i.e. } \mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim p_{\text{noise}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

This describes a first-order Markov chain. Using the notation  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ , we have by the chain rule:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (2)$$

$$\text{i.e. } \mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim p_{\text{noise}}$$

Diffusion models are trained to reverse the forward process described in Equations 1 and 2. Namely, a neural network  $q_{\theta, t}(\mathbf{x}_{t-1} | \mathbf{x}_t)$  with learnable parameters  $\theta$  is trained to predict (the distribution of)  $\mathbf{x}_{t-1}$  from a sample  $\mathbf{x}_t$ . With some reparametrizations, the neural network can be trained to predict (the distribution of)  $\boldsymbol{\varepsilon}$  knowing  $\mathbf{x}_t$ ,  $\mathbf{x}_0$  knowing  $\mathbf{x}_t$ , or  $\sqrt{\bar{\alpha}_t} \boldsymbol{\varepsilon} - \sqrt{1 - \bar{\alpha}_t} \mathbf{x}_0$  knowing  $\mathbf{x}_t$ . These correspond to epsilon-prediction [9], sample-prediction [33], and velocity-prediction [33], respectively.

Assuming epsilon-prediction, one training iteration of the neural network  $\boldsymbol{\varepsilon}_{\theta, t}(\mathbf{x}_t)$  is as follows. An image  $\mathbf{x}_0$  of the dataset, a random timestep  $t \sim \mathcal{U}([1, T])$  and a noise  $\boldsymbol{\varepsilon} \sim p_{\text{noise}}$  are sampled. A noise-contaminated image  $\mathbf{x}_t$  is built according to Equation 2. The neural network is given  $\mathbf{x}_t$  and  $t$ , and outputs a predicted noise  $\boldsymbol{\varepsilon}_{\theta, t}(\mathbf{x}_t)$ . The loss for epsilon-prediction is typically set as the mean square error  $\|\boldsymbol{\varepsilon}_{\theta, t}(\mathbf{x}_t) - \boldsymbol{\varepsilon}\|_2^2$ .

At inference time, an initial latent  $\hat{\mathbf{x}}_T$  is sampled from:

$$q(\hat{\mathbf{x}}_T) = p_{\text{noise}} = \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (3)$$

$$\text{i.e. } \hat{\mathbf{x}}_T = \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim p_{\text{noise}}$$

and iteratively denoised. For any timestep  $t \in [1, T]$ :

$$q_{\theta}(\hat{\mathbf{x}}_{t-1}) = \int q(\hat{\mathbf{x}}_T) \prod_{s=t}^T q_{\theta, s}(\hat{\mathbf{x}}_{s-1} | \hat{\mathbf{x}}_s) d\hat{\mathbf{x}}_{t:T} \quad (4)$$

thus, to sample each  $\hat{\mathbf{x}}_{t-1}$ , we use  $\hat{\mathbf{x}}_t$  and  $q_{\theta, t}(\hat{\mathbf{x}}_{t-1} | \hat{\mathbf{x}}_t)$

Instead of denoising through all the timesteps of the model to generate an image  $\hat{x}_0$ , accelerated sampling algorithms have been proposed [16, 37], reducing the number of forward passes through the neural network  $q_{\theta,t}$  by orders of magnitude, *e.g.* from 1000 to 50 [16]. In such cases, the first denoising iteration does not always start denoising from the highest timestep  $t = 1000$ , but for instance  $t = 981$  (see details in Section 3.3. of [15]). Without loss of generality, we can assume in such a case that the model was only trained with  $T = 981$  timesteps and that the inference process always starts denoising from the last timestep  $T$ .

A commonly used diffusion model is Stable Diffusion, which generates high-quality images conditionally on a textual prompt. Stable Diffusion is a Latent Diffusion Model (LDM, [28]), meaning the images are represented in a latent space instead of the pixel space. The equations above still hold if we consider  $\mathbf{x}_0$  to be an image represented by a latent code. In particular, latent codes in LDMs still have channels and two spatial dimensions. We thus keep the terminology *pixel* to refer to an element of the latent code.

## 2.2. Fixing the training of diffusion models

Sampling the initial latents from noise only (Equation 3) is not totally aligned with the training process (Equation 2), where images are not corrupted up to complete noise [7], but always contain a signal leak  $\sqrt{\bar{\alpha}_T} \mathbf{x}_0$  from a real image  $\mathbf{x}_0$  [15]. In the case of Stable Diffusion, for which the signal leakage is particularly important [15], this leads for instance to generated images with medium brightness [7, 15]. To eliminate this issue, recent research [15, 33] trains diffusion models enforcing  $\alpha_T = 0$ , effectively training the last timestep from white noise, *i.e.* without signal leakage. Guttenberg [7] proposes to modify the noise distribution, such that the brightness of the real image cannot be deduced from the signal leak anymore. The signal leakage also leads to difficulties in generating style-specific images. To overcome this, Everaert *et al.* [4] propose to finetune Stable Diffusion on a new noise distribution that approximates the distribution of the style images. While not focusing on the signal leakage, Ning *et al.* [22] propose adding an extra perturbation term during training to make the model more robust to training and inference distribution changes. All of these require retraining/finetuning the models to remove or reduce any signal leakage.

Our approach, on the contrary, can be used directly with any existing diffusion model. It leverages the signal leakage instead of retraining or fine-tuning models. Rather than finetuning to realign the training distribution with the inference distribution, *i.e.* the distribution of initial latents, we focus on realigning the inference distribution with the training distribution, by adding a signal leak  $\sqrt{\bar{\alpha}_T} \tilde{\mathbf{x}}$  to the noise  $\mathbf{\epsilon}$  in the initial latents  $\hat{\mathbf{x}}_T$ .

## 3. Signal-leak bias

### 3.1. Discrepancies between training and inference distributions in diffusion models

The reverse diffusion (*i.e.* denoising) process described in Equation 4 inputs the neural network  $q_{\theta,t}$  with some data  $\hat{\mathbf{x}}_t$  to obtain  $\hat{\mathbf{x}}_{t-1}$ . However,  $\hat{\mathbf{x}}_t$  is obtained either from previous predictions of the model when  $t < T$  or, if  $t = T$ , from white noise. This differs from training, where  $\mathbf{x}_t$  is a corrupted version of a real image  $\mathbf{x}_0$ . In both cases,  $t < T$  and  $t = T$ , the inference distribution differs from the training distribution.

**Exposure bias:** The diffusion model  $q_{\theta,t}$  is trained using noise-corrupted versions of images (Equation 2), rather than with the predictions of the latter timesteps as done during inference (Equation 4). This creates a discrepancy between training and inference, which can cause error accumulation during the iterative denoising process to generate images [14, 22], similarly to the *exposure bias* [27, 35] in text-generation models.

Due to the signal leakage, a significant “error” often already exists at the last timestep ( $T$ ), as we explained below. This error will be accumulated forward through the image generation process, and hence should not be ignored.

**Signal-leak bias:** At inference time, the model is given white noise as initial latent  $\hat{\mathbf{x}}_T$  (Equation 3). However, we can deduce from Equation 2 that the model was trained at the last timestep with samples  $\mathbf{x}_T$  from  $q(\mathbf{x}_T)$ :

$$\mathbf{x}_T \sim q(\mathbf{x}_T) \Leftrightarrow \mathbf{x}_T = \sqrt{\bar{\alpha}_T} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_T} \mathbf{\epsilon}, \quad (5)$$

$$\mathbf{x}_0 \sim q(\mathbf{x}_0), \quad \mathbf{\epsilon} \sim p_{\text{noise}}$$

The two distributions  $q(\hat{\mathbf{x}}_T)$  and  $q(\mathbf{x}_T)$  differ, creating a discrepancy between training and inference distributions. Following Lin *et al.* [15], we can quantify the importance of the signal leakage by a signal-to-noise ratio (SNR):

$$\text{SNR} = \bar{\alpha}_T / (1 - \bar{\alpha}_T) \quad (6)$$

This SNR depends on  $\bar{\alpha}_T$ , and hence on the choice of the function  $\alpha_t$  [15]. The function  $\alpha_t$  is defined according to a  $\beta$ -schedule. The function  $\beta_t = 1 - \alpha_t$  is typically chosen to be a linear schedule in  $\beta$ -space [9], a squared capped cosine schedule in  $\beta$ -space [19, 20], a linear schedule in  $\sqrt{\beta}$ -space [28], or a sigmoid schedule in  $\beta$ -space [42]. The SNR is particularly high with Stable Diffusion [15], which uses the linear schedule in  $\sqrt{\beta}$ -space. Lin *et al.* [15] also link the signal leakage to the fact that Stable Diffusion always generates images with medium brightness.

We notice that this linear schedule in  $\sqrt{\beta}$ -space is not only used in Stable Diffusion, but is commonly employed in many variants of LDM [28] as well, *e.g.* [2, 38, 39].

### 3.2. Mismatch between noise and signal leak distributions

**The source of the signal-leak bias:** The strength of the signal leakage depends on  $\bar{\alpha}_T$ . If  $\bar{\alpha}_T = 0$ , then there is no signal leakage at the last timestep  $T$ , as can be deduced from Equation 2. This is rarely the case [15, 33].

When  $\bar{\alpha}_T \neq 0$ , the signal leakage exists, but does not necessarily imply a bias at inference time. As seen by the high quality of the images generated by LDMs [2, 28, 38, 39], the existence of the signal leak can have little implication in practice. In LDMs,  $\bar{\alpha}_T$  is high, leading to  $\mathbf{x}_T = 0.068265 \cdot \mathbf{x}_0 + 0.997667 \cdot \mathbf{\epsilon}$  [15]. Even though  $\bar{\alpha}_T$  is high, the training distribution  $q(\mathbf{x}_T)$  and inference distribution  $q(\hat{\mathbf{x}}_T)$  are relatively aligned because the noise distribution  $p_{\text{noise}}$  and the image distribution  $q(\mathbf{x}_0)$  are similar. Indeed, in LDMs [28], the diffusion happens in a normalized VAE latent space [13], where images  $\mathbf{x}_0$  are rep-

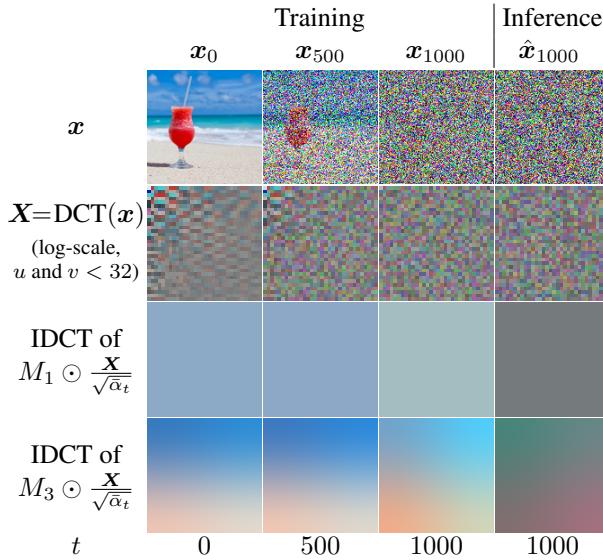


Figure 2. The first row shows values for  $\mathbf{x}_0$ ,  $\mathbf{x}_{500}$ ,  $\mathbf{x}_{1000}$ , and  $\hat{\mathbf{x}}_{1000}$ . The second row contains their 2D-DCT components, showing that natural images mostly contain low-frequency components, unlike white noise, which is equally spread across all frequencies. As shown in the third and fourth rows, we can recover some low-frequencies components of the original image  $\mathbf{x}_0$  from  $\mathbf{x}_{500}$  or partially from  $\mathbf{x}_{1000}$ . By eliminating all frequencies except the lowest one (third row) or the 3 lowest ones (fourth row), we successfully recover the low-frequency information of  $\mathbf{x}_0$  shown in the first column, third and fourth rows. The noise introduced during the diffusion process does not affect these low-frequency components; thus allowing the model to learn not to alter them during denoising, with the result of generating images with similar low-frequency content as white noise, *e.g.* medium brightness.  $M_N$  refers to the mask of the  $N$  lowest-frequencies and  $\odot$  symbolizes element-wise multiplication. IDCT refers to the Inverse Discrete Cosine Transform. We used the same  $\beta$ -schedule as LDM [28].

resented by their VAE latent codes. The VAE makes the distribution of images  $q(\mathbf{x}_0)$  relatively similar to the noise distribution  $p_{\text{noise}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The signal leak  $\sqrt{\bar{\alpha}_T} \mathbf{x}_0$  is then almost indistinguishable from noise, leading to almost no bias when the initial latents are sampled from noise only.

However, whenever the signal leak has a distribution that differs from the noise distribution, sampling the initial latent from only noise creates a bias. At inference time, the model expects to find a signal leak  $\sqrt{\bar{\alpha}_T} \mathbf{x}_0$  in the initial latent  $\hat{\mathbf{x}}_T$  to deduce information about the real image  $\mathbf{x}_0$ . Sampling the initial latents from noise only biases the generated images, because the model interprets noise as being the signal leak. This is especially noticeable when trying to generate images in a specific style: the model expects to find an initial latent from a distribution that is different from white noise (see the middle column of Figure 1).

**Why Stable Diffusion always generates images with medium brightness:** In Stable Diffusion, the distribution of images  $q(\mathbf{x}_0)$  does not *exactly* match  $p_{\text{noise}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ . For example, the mean of the pixels of a sample from  $p_{\text{noise}}$  always has a medium value, but the mean of pixels of a real image will be more varied, depending on the brightness of the image. This causes images generated with Stable Diffusion to always have a medium brightness [7, 15]. We discuss here the mismatch between  $q(\mathbf{x}_0)$  and  $p_{\text{noise}}$  from a *frequency domain* point of view. Natural images tend to be smooth and exhibit an average power spectrum that declines with a  $f^{-2}$  relationship [1, 5, 29, 41], indicating a concentration of signal power at the lowest spatial frequencies. On the other hand, the white noise  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  is equally distributed across all frequencies, meaning noising mostly affects high-frequency components [3]. Rewriting the Equation 6 for a specific spatial frequency  $(u, v)$ , we obtain:

$$\text{SNR}(u, v) = \frac{\bar{\alpha}_T \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}((\mathbf{X}_0^{u,v})^2)}{(1 - \bar{\alpha}_T) \mathbb{E}_{\mathbf{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}((\mathbf{\epsilon}^{u,v})^2)} = \frac{\bar{\alpha}_T}{1 - \bar{\alpha}_T} \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)}((\mathbf{X}_0^{u,v})^2) \quad (7)$$

where  $\mathbf{X}_0^{u,v}$  denotes the  $(u, v)$ -th term of the 2D-DCT of an image  $\mathbf{x}_0$ . Note that  $\mathbf{\epsilon}^{u,v}$ , the  $(u, v)$ -th term of the 2D-DCT of a noise sample  $\mathbf{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , also follows  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . Because of the prevalence of low spatial frequency content,  $\mathbb{E}_{\mathbf{x}_0}((\mathbf{X}_0^{u,v})^2)$  is high when  $u$  and  $v$  are small, and negligible for high frequencies. The SNR is then high for the lowest frequencies and almost 0 for the remaining frequencies. These observations can be visualized as in Figure 2. Note that the mean color of the image  $\mathbf{x}_0$ , *i.e.* the signal  $\mathbf{X}_0^{0,0}$ , is thus the least affected by the noise. The diffusion model then learns to recover the mean color  $\mathbf{X}_0^{0,0}$  of the image  $\mathbf{x}_0$  from the one  $\mathbf{X}_T^{0,0}$  of  $\mathbf{x}_T$ . When  $\hat{\mathbf{x}}_T$  is sampled from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , then  $\hat{\mathbf{X}}_T^{0,0} \approx 0$ , thus generated images  $\hat{\mathbf{x}}_0$  always result in  $\hat{\mathbf{X}}_0^{0,0} \approx 0$ , *i.e.*, a medium brightness.

**Limitation of Stable Diffusion after tuning on a style:** Fine-tuning Stable Diffusion to a specific style usually does not work as intended. Generated images do not match the colors or backgrounds of the style, as illustrated in the first rows of Figures 3a, 3b, and 3c. Even when fine-tuned on a single solid black image, Stable Diffusion is unable to produce a black image [7]. Everaert *et al.* [4] show that training with style-specific noise instead of white noise leads to better style adaptation. We then argue that, when fine-tuning for a specific style *with white noise*, there is a significant mismatch between noise and image distributions *in the pixel domain*. The distribution of images of a specific style is located in a specific part of the image space. Hence  $q(\mathbf{x}_0)$  cannot be considered similar to  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  anymore. Because of the signal leakage, the training distribution  $q(\mathbf{x}_T)$  is far from the inference distribution  $q(\hat{\mathbf{x}}_T)$ . Images generated by a diffusion model tuned to a specific style thus do not look as good as they potentially could.

## 4. Method

### 4.1. Exploiting the signal-leak bias

As discussed in Section 2.2, previously proposed solutions mainly focus on eliminating the signal-leak bias by setting  $\alpha_T = 0$  [15, 33], on adding noise perturbations [22], or on modifying noise distribution [4, 7]. Essentially, these methods attempt to realign the training distribution with the inference distribution. This comes at the cost of re-training or fine-tuning a model. To our knowledge, only Li *et al.* [14] propose a solution to realign the distributions at inference time, without re-training. At each denoising iteration, they propose to find the best timestep  $t'$  at which to denoise the current  $\hat{\mathbf{x}}_t$ . However, while efficient for the exposure bias, this cannot work for the signal-leak bias - there is simply no timestep  $t'$  trained with samples from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ .

Our solution to *exploit the signal-leak bias in diffusion models* is much simpler than these previous solutions. We focus on *realigning the distribution of initial latent  $q(\hat{\mathbf{x}}_T)$  with the training distribution  $q(\mathbf{x}_T)$* . This has the advantage of not requiring any additional training of the diffusion model. The key idea of our solution is to simply sample the initial latents  $\hat{\mathbf{x}}_T$  from the training distribution (Equation 5) instead of from only white noise. Although the distribution  $q(\mathbf{x}_0)$  is unknown, we can approximate it by computing statistics from a set of target images. Our approach then consists of obtaining an approximate distribution  $\tilde{q}_0$  of  $q(\mathbf{x}_0)$ . At inference time, we simply sample the initial latents in the same way as during training (Equation 2), *i.e.* with a random signal leak  $\sqrt{\bar{\alpha}_T} \tilde{\mathbf{x}}$ :

$$\begin{aligned} \hat{\mathbf{x}}_T &= \sqrt{\bar{\alpha}_T} \tilde{\mathbf{x}} + \sqrt{1 - \bar{\alpha}_T} \boldsymbol{\varepsilon} \\ \tilde{\mathbf{x}} &\sim \tilde{q}_0, \quad \boldsymbol{\varepsilon} \sim p_{\text{noise}} = \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \quad (8)$$

Note that no other operations are needed and we then

just follow the usual process of generating images with diffusion models. Equation 8 has similarities with the image-editing work SDEdit [18], which samples intermediate latents  $\hat{\mathbf{x}}_{t_0}$  as  $\sqrt{\bar{\alpha}_{t_0}} \mathbf{x}^{(g)} + \sqrt{1 - \bar{\alpha}_{t_0}} \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim p_{\text{noise}}$  (*i.e.* Equation 2), where  $\mathbf{x}^{(g)}$  is an image to be edited. SDEdit [18] focuses only on image editing and uses  $t_0 \approx 0.3T$  to  $0.6T$ . Note that in our work, unlike SDEdit, we generate images starting from the timestep  $T$ .

### 4.2. Modeling the distribution of the signal leak

The current image generation process of diffusion models, which samples  $\hat{\mathbf{x}}_T$  from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , is equivalent to using Equation 8 with  $\tilde{q}_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ . We now discuss better choices for  $\tilde{q}_0$ . Following our previous insights, we conclude that the signal leak mismatches the noise distribution either in the *pixel domain* or in the *frequency domain*.

We provide here two models for the distribution of the signal leak. The first one estimates the distribution of the signal leak in the pixel domain. We use it in Sections 5.1 and 5.2. The second one, used in Sections 5.3 and 5.4, estimates the distribution of the signal leak in the frequency domain *and* in the pixel domain, for the low-frequency (LF) *and* high-frequency (HF) contents, respectively. To be more specific, we provide the dimensions of the elements for Stable Diffusion V1, where latent codes of images have 4 channels and  $64 \times 64$  pixels, *i.e.*  $\mathbf{x}_0 \in \mathbb{R}^{64 \times 64 \times 4}$ . These values are to be adapted to the model being used and do not imply that our approach requires specific architectural changes.

#### 4.2.1 Pixel-domain model

We first model the approximate distribution  $\tilde{q}_0$  as  $\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$ , a Gaussian distribution with diagonal covariance. The location  $\boldsymbol{\mu}$  and covariance  $\text{diag}(\boldsymbol{\sigma}^2)$  are obtained from statistics of the target images.

$$(\boldsymbol{\mu})^{i,j,k} = \text{Mean}_{\mathbf{x}_0} \mathbf{x}_0^{i,j,k} \quad (9)$$

$$(\boldsymbol{\sigma})^{i,j,k} = \text{Std}_{\mathbf{x}_0} \mathbf{x}_0^{i,j,k} \quad (10)$$

$$\boldsymbol{\mu} \in \mathbb{R}^{64 \times 64 \times 4}, \boldsymbol{\sigma} \in \mathbb{R}^{64 \times 64 \times 4}$$

These equations have similarities with prior research [4, 43]. However, we only use this distribution to model the signal leak  $\sqrt{\bar{\alpha}_T} \mathbf{x}_0$ , instead of training the model with it. As we show in Sections 5.1 and 5.2, this *pixel-domain model* of the signal leak is effective for style adaptation of diffusion models. Yet, because the distribution of natural images in LDMs [28] is already approximately  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , this pixel-domain model does not help to generate images with more varied brightness.

#### 4.2.2 Frequency and pixel domain model

As mentioned before, the training distribution of diffusion models on natural images differs from the inference distri-

bution mostly in the lowest frequencies. We can thus explicitly model the  $N$  lowest frequencies of the signal leak by computing the mean and covariance of the low-frequency components from a small set of natural images. The remaining, *i.e.* the components with higher frequencies, are modeled in the pixel domain as in the previous paragraph. To model the  $N$  lowest frequencies, we compute the DCT  $\mathbf{X}_0$  of each target image  $\mathbf{x}_0$ . We obtain a multivariate Gaussian distribution  $\tilde{q}_{0,\text{LF}}$  with location  $\boldsymbol{\mu}_{\text{LF}}$  and covariance  $\boldsymbol{\Sigma}_{\text{LF}}$  by computing statistics from the DCTs  $\mathbf{X}_0$ . With the notation  $\mathbf{X}_{0,\text{LF}} = M_N \odot \mathbf{X}_0$ , the location and covariance are estimated as follows:

$$\begin{aligned} \tilde{q}_{0,\text{LF}} &= \mathcal{N}(\boldsymbol{\mu}_{\text{LF}}, \boldsymbol{\Sigma}_{\text{LF}}) & (11) \\ (\boldsymbol{\mu}_{\text{LF}})^{u,v,k} &= \text{Mean}_{\mathbf{x}_0} \mathbf{X}_{0,\text{LF}}^{u,v,k} \\ (\boldsymbol{\Sigma}_{\text{LF}})^{u_1,v_1,k_1,u_2,v_2,k_2} &= \text{Cov}_{\mathbf{x}_0}(\mathbf{X}_{0,\text{LF}}^{u_1,v_1,k_1}, \mathbf{X}_{0,\text{LF}}^{u_2,v_2,k_2}) \\ \boldsymbol{\mu}_{\text{LF}} &\in \mathbb{R}^{4N}, \quad \boldsymbol{\Sigma}_{\text{LF}} \in \mathbb{R}^{4N \times 4N} \end{aligned}$$

The high-frequency components  $\mathbf{x}_{0,\text{HF}} = \text{IDCT}(\mathbf{X}_0 - \mathbf{X}_{0,\text{LF}})$  are modeled with a pixel-domain distribution  $\tilde{q}_{0,\text{HF}}$  as in Section 4.2.1. The signal leak  $\sqrt{\alpha_T} \tilde{\mathbf{x}}$  that we add in the initial latent at inference time is sampled such that:

$$\begin{aligned} \tilde{\mathbf{x}} \sim \tilde{q}_0 \Leftrightarrow \tilde{\mathbf{x}} &= \text{IDCT}(\tilde{\mathbf{X}}_{\text{LF}}) + \tilde{\mathbf{x}}_{\text{HF}}, & (12) \\ \text{with } \tilde{\mathbf{X}}_{\text{LF}} &\sim \tilde{q}_{0,\text{LF}} \text{ and } \tilde{\mathbf{x}}_{\text{HF}} \sim \tilde{q}_{0,\text{HF}} \end{aligned}$$

By combining the two components LF and HF, we create a distribution  $\tilde{q}_0$  that encompasses a broader range of colour and brightness variations than  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . Sampling initial latents as in Equation 8 with this estimation  $\tilde{q}_0$  enables the generation of images with more diverse brightness and colours than with  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , as we show in Section 5.3. The value of  $N$  is chosen empirically, for instance,  $N = 3$ . While we present results obtained with DCT, note that different approaches to model frequency components could also be used, such as PCA or Fourier Transform.

## 5. Results

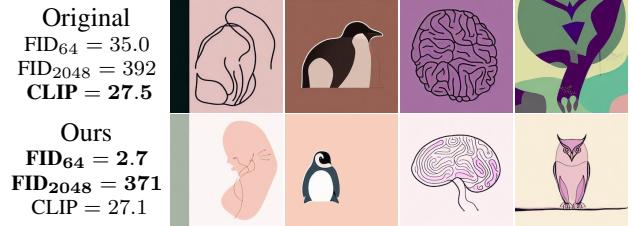
We experiment with Stable Diffusion, which has a significant signal leakage [15]. Following current evaluations of diffusion models [24, 28, 32], we compute FID (lower is better, [8]) and CLIP (higher is better, [26]) scores from the TorchMetrics library [21]. Metrics are computed using 200 generated images. Wherever a CLIP score is reported (Figures 3a, 3b, 4a, 4b, and 5), the 200 images are generated from the textual prompts of the DrawBench benchmark [32], with a guidance scale of 7.5 [10]. In the other cases (Figures 3c, 4c), the 200 images are generated without classifier-free guidance [10]. We compute two versions of the FID,  $\text{FID}_{64}$  and  $\text{FID}_{2048}$ , using the 64-th or 2048-th InceptionV3 [40] feature layers, respectively. All images are generated with 50 PNDM denoising steps [16].

### 5.1. Improved style for style-specific models

We apply our pixel-domain approach from Section 4.2.1 to different existing fine-tuned versions of Stable Diffusion, covering different styles and fine-tuning strategies. Note that the fine-tuning has already been done: we do not do any additional fine-tuning with our approach. Our results for three such models are shown in Figure 3. Figure 3a shows the results of the Pokemon-LoRA model [11, 31, 34]. For this style, we use the first 50 images of the Pokemon BLIP captions dataset [25] to obtain our signal leak distribution  $\sqrt{\alpha_T} \tilde{\mathbf{x}}$ ,  $\tilde{\mathbf{x}} \sim \tilde{q}_0$ , but use all 833 images to compute the FID metrics. The results obtained with the current inference process (first row) do not correctly match the expected style, as seen qualitatively and with the FID scores. In particular, the generated images do not have a white background, as



(a) Pokemon-LoRA model [34] fine-tuned from Stable Diffusion v1.4 on the Pokemon BLIP captions dataset [25] with LoRA fine-tuning [11, 31].



(b) Stable Diffusion v1.4 model. The concept of “line-art style” [12] was learned with Textual Inversion [6], from 7 line-art images [12].



(c) NASA-space model [17] fine-tuned with DreamBooth [30] from Stable Diffusion v2 on 24 images of astronomical phenomena [17].

Figure 3. When using white noise as initial latent, existing fine-tuning strategies lead to sub-optimal style-matching, as illustrated in the first row of each subfigure 3a, 3b and 3c. In each of the second rows, images are generated with the same prompts and the same models, which we did not additionally fine-tune. In this second row, images are generated with our proposed approach, sampling a signal leak  $\sqrt{\alpha_T} \tilde{\mathbf{x}}$  from our pixel-domain estimation, to generate images that better match the target style. Please check the style of the target images in references [12, 17, 25].

opposed to those used in training the model [25, 34]. Sampling the initial latents with a signal leak  $\sqrt{\bar{\alpha}_T} \tilde{x}$  generates images matching the expected style (second row). Similar observations are made in Figure 3b for the “line-art” style tuned with Textual Inversion [6], and in Figure 3c for the “Nasa-space” style tuned with DreamBooth [30]. For these two styles, we use all the target images (respectively 7 and 24 images) to obtain our estimated distribution of the signal leak and to compute the FID scores. The FID<sub>64</sub> score is improved significantly, indicating, accordingly to the qualitative assessment, that the images generated with our approach reproduce more faithfully the style of the target images. Note that the CLIP score remains high, which implies that our approach does not affect how well the generated images match their textual prompt.

## 5.2. Improving style for non-style-specific models

We additionally experiment with exploiting the signal-leak bias directly in the original diffusion model, without using a tuned version of the model. Simply describing a desired style in the textual prompt is often insufficient to generate images in the desired style [4]. However, when we combine the style description with our approach of exploiting the signal-leak bias, the generated images seem to match well the desired style. This suggests that fine-tuning the model for a specific style may not always be necessary, as shown in Figure 4. It is possible to generate images in a particular style by exploiting the signal-leak bias without any fine-tuning. By putting a signal leak  $\sqrt{\bar{\alpha}_T} \tilde{x}$ ,  $\tilde{x} \sim \tilde{q}_0$  into the initial latent, we bias the denoising process toward generating images that look like  $\tilde{x}$ . Our strategy here takes only a few seconds to estimate the distribution  $\tilde{q}_0$ , all without any compromise on the inference time to generate an image, for instance as opposed to guidance [23]. **Limitation:** Some specific styles may not be easily described in words and may correspond to characteristics not captured by our pixel-domain model. We design such an example in Section 2.4 of the supplementary material. For such styles, fine-tuning or a different model for  $\tilde{q}_0$  would be required.

## 5.3. Generating more varied images

As mentioned earlier, images currently generated with Stable Diffusion tend to have medium low-frequency components, *e.g.* medium brightness and little variation of colors between different areas of an image. This observation is noticeable in the top rows of Figure 5. To generate images with more varied low-frequency components, we apply our approach from Section 4.2.2 by estimating the distribution of the signal leak in both frequency and pixel domains. Especially, we use 323 images from the LAION-6+ dataset [36] to model the 3 lowest-frequency components, *i.e.* a value  $N = 3$  following the notation in Section 4.2.2. We use the same 323 images to compute the FID scores.



(a) Targeting the style of the Pokemon BLIP captions dataset [25], generating images with the prompt “[text] In the style of Satoshi Tajiri, white background.”



(b) Targeting the style of the 7 line-art images [12], generating images with the prompt “[text] In the style of line art, pastel colors, white background.”



(c) Targeting the style of the 24 images of astronomical phenomena [17], generating images with the prompt “A photo of the sky made by the NASA.”

Figure 4. When using white noise as initial latent, describing the style in the textual prompt is insufficient to generate images that match the desired style, as illustrated in the first row of each subfigure 4a, 4b and 4c. In each of the second rows, images are generated with the same prompts and model using our proposed approach, sampling a signal leak  $\sqrt{\bar{\alpha}_T} \tilde{x}$  from our pixel-domain estimation. We used the same approach as in Section 5.1 to estimate the distribution in the pixel domain for the three styles. All images here are generated with standard Stable Diffusion 2.1, without fine-tuning for specific styles.

We visualize in Figure 5 the effect of using our method to sample the initial latents instead of sampling them from white noise. The effect is slight, but noticeable on the 8 randomly-picked images of this figure. Images generated by sampling from a distribution containing a signal leak with more varied low-frequency components also have more varied low-frequency components. This not only solves the issue of generating “only” medium-brightness images but, also results in more natural variances of colors inside each image; this all without extra training, as opposed to previous solutions [7, 15].

The FID scores are slightly improved, suggesting, according to our visual assessment, that images better match

the distribution of low-level features of natural images. The CLIP score worsens only very slightly, suggesting our approach has almost no impact on the content and high-frequency alignment of the generated images. **Limitation:** The signal leak  $\sqrt{\alpha_T} \tilde{x}$ ,  $\tilde{x} \sim \tilde{q}_0$  is sampled randomly with our approach. One advantage of prior work based on re-training [7, 15] is that the brightness of the generated image matches the textual prompt instead of being random.

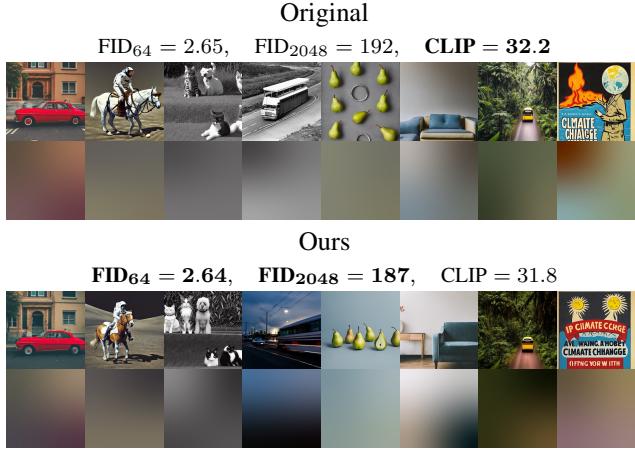


Figure 5. We generated 8 images from Stable Diffusion 2.1, at the default resolution  $768 \times 768$  with 8 prompts from DrawBench [32], in the top starting from white noise (*i.e.* the default behavior), and in the bottom from our distribution realigned with the distribution of natural images, as explained in Sections 4.2.2 and 5.3. Below each generated image, we show a low-pass filtered version of it. We observe, as expected from our analysis, that images generated by taking into account the signal leak have more varied low-frequency components, as in natural images. Images in the top tend to be greyish, with medium brightness and little variation of colors inside each image. Images in the bottom have more varied colors and brightness across images. They tend to be less greyish and to have more variation of colors inside each image. This all comes without any additional training of Stable Diffusion. Quantitative comparisons of average pixel values, contrast, and luminance are provided in the supplementary Section 3.2.

#### 5.4. Explicit influence on low-frequency attributes

Instead of randomly sampling the signal leak  $\sqrt{\alpha_T} \tilde{x}$ , its low-frequencies components can be manually selected by the user. This provides explicit control over the generated image atop the textual prompt, without needing any target images. Following the notations of Section 4.2.2, we set  $N = 1$  and use 323 images from the LAION-6+ dataset [36] to obtain  $\tilde{q}_{0\text{HF}}$ . Instead of sampling  $\tilde{x}$  as in Equation 12, we manually select a value for  $\tilde{X}_{\text{LF}}$  and sample  $\tilde{x}$  as  $\text{IDCT}(\tilde{X}_{\text{LF}}) + \tilde{x}_{\text{HF}}$ , with  $\tilde{x}_{\text{HF}} \sim \tilde{q}_{0\text{HF}}$ . As we show in Figure 6, it is easy to interpret the effect of the different values of  $\tilde{X}_{\text{LF}}$ . We can consistently bias the generation of images towards a specific brightness or desired colors.



Figure 6. Explicit control on the mean color of the generated images. Instead of sampling the signal leak  $\sqrt{\alpha_T} \tilde{x}$  from the computed distribution, we can manually set its low-frequency components  $\tilde{X}_{\text{LF}}$  and randomly sample only the high-frequency components  $\tilde{x}_{\text{HF}} \sim \tilde{q}_{0\text{HF}}$ . This provides a specific bias towards generating images with desired low-frequency components, here, the mean color of the image. In this figure, images are generated from the prompt “A glass on the beach” with Stable Diffusion 2.1 by setting  $\tilde{X}_{\text{LF}} \in \mathbb{R}^4$  to  $0$ , except for one of the four channels, for which we set the value indicated at the bottom of each column. The four rows correspond to the 4 channels in Stable Diffusion’s latent space.

## 6. Conclusion

In this paper, we show that the signal-leak bias in diffusion models is not only caused by a non-zero SNR during the training of the last timestep, but also a discrepancy between the noise and the data distributions. When generating natural images, the discrepancy between the noise and the data distributions lies in the frequency domain, explaining why generated images always tend to have medium low-frequency values, including medium brightness. When diffusion models are tuned to a specific style, the discrepancy between the noise and the data distributions lies in the pixel domain, explaining the unsatisfactory outcomes of style adaptation of diffusion models.

We propose a simple way to exploit this signal-leak bias to our advantage to solve these issues. By injecting a signal leak in the initial latent at inference time, we can bias the image generation toward a desired specific color distribution or a specific style. This simple step does not require any fine-tuning making it much simpler than existing approaches for style or color-specific image generation.

We encourage future research to account for training and inference distribution gap when training or fine-tuning diffusion models, and to include a signal leak in the initial latents at inference time as well, in order to mirror the training process and achieve visually more pleasing results.

**Acknowledgements:** This work was supported by Innosuisse grant 48552.1 IP-ICT.

## References

[1] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987. 4

[2] Jonathan Chang. ttj/flex-diffusion-2-1 — huggingface.co, 2023. <https://huggingface.co/ttj/flex-diffusion-2-1>. 3, 4

[3] Majed El Helou, Ruofan Zhou, and Sabine Süsstrunk. Stochastic frequency masking to improve super-resolution and denoising networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 749–766. Springer, 2020. 4

[4] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in Style. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3, 5, 7

[5] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. 4

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*, 2022. 6, 7

[7] Nicholas Guttenberg. Diffusion with offset noise, 2023. <https://www.crosslabs.org/blog/diffusion-with-offset-noise>. 1, 2, 3, 4, 5, 7, 8

[8] 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. *Advances in neural information processing systems*, 30, 2017. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[10] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2021. 6

[12] Dhruv Karan. sd-concepts-library/line-art — huggingface.co, 2022. <https://huggingface.co/sd-concepts-library/line-art>. 6, 7

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[14] Mingxiao Li, Tingyu Qu, Wei Sun, and Marie-Francine Moens. Alleviating Exposure Bias in Diffusion Models through Sampling with Shifted Time Steps. *arXiv preprint arXiv:2305.15583*, 2023. 1, 3, 5

[15] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common Diffusion Noise Schedules and Sample Steps are Flawed. *arXiv preprint arXiv:2305.08891*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[16] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *International Conference on Learning Representations*, 2022. 3, 6

[17] MatAIart. sd-dreambooth-library/nasa-space-v2-768 — huggingface.co, 2022. <https://huggingface.co/sd-dreambooth-library/nasa-space-v2-768>. 6, 7

[18] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. 5

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[20] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2021. 3

[21] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stanci, Changsheng Quan, Maxim Gorchkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch, 2022. <https://github.com/Lightning-AI/torchmetrics>. 6

[22] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input Perturbation Reduces Exposure Bias in Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 26245–26265, 2023. 1, 3, 5

[23] Zhihong Pan, Xin Zhou, and Hao Tian. Arbitrary style guidance for enhanced diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4461–4471, 2022. 7

[24] Patrick von Platen, Kashif Rasul, Pedro Cuenca, Will Berman, Will Berman, and YiYi Xu. Evaluating Diffusion Models — huggingface.co, 2023. <https://huggingface.co/docs/diffusers/main/en/conceptual/evaluation>. 6

[25] Justin N. M. Pinkney. Pokemon BLIP captions, 2022. <https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions>. 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[27] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, 2015. 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2021. 2, 3, 4, 5, 6

[29] Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical review letters*, 73(6):814, 1994. 4

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2022. 6, 7

[31] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. <https://github.com/cloneofsimo/lora>. 6

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 6, 8

[33] Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 5

[34] Sayak Paul. sayakpaul/sd-model-finetuned-lora-t4 — huggingface.co, 2023. <https://huggingface.co/sayakpaul/sd-model-finetuned-lora-t4>. 6, 7

[35] Florian Schmidt. Generalization in Generation: A closer look at Exposure Bias. *EMNLP-IJCNLP 2019*, page 157, 2019. 3

[36] Christoph Schuhmann and Romain Beaumont. LAION-Aesthetics, 2022. [https://huggingface.co/datasets/ChristophSchuhmann/improved\\_aesthetics\\_6plus](https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_6plus) and <https://laion.ai/blog/laion-aesthetics/>. 7, 8

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021. 3

[38] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. LDM3D: Latent Diffusion Model for 3D. *arXiv preprint arXiv:2305.10853*, 2023. 3, 4

[39] Spencer Sterling. cerspense/zeroscope\_v2\_XL — huggingface.co, 2023. [https://huggingface.co/cerspense/zeroscope\\_v2\\_XL](https://huggingface.co/cerspense/zeroscope_v2_XL). 3, 4

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2015. 6

[41] David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 4

[42] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*, 2022. 3

[43] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted Diffusion for Text-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10157–10166, 2023. 5