CharDiff: Improving Sampling Convergence via Characteristic Function Consistency in Diffusion Models

Abhishek Kumar Sinha S. Manthira Moorthi
Signal and Image Processing Area,
Space Applications Centre (Indian Space Research Organization), Ahmedabad
aks, smmoorthi@sac.isro.gov.in

Abstract

Diffusion models have demonstrated extensive capabilities for generative modelling in both conditional and conditional image synthesis tasks. Reverse sampling has been the centre of interest to improve the overall image quality without retraining the model from scratch. In this work, we propose a plug-and-play module by utilizing the characteristic function of the distributions to minimize sampling drift. We experiment with existing diffusion solvers with our module during the denoising step to provide additional performance gain in image synthesis, linear inverse problem tasks and text-conditioned image synthesis. Moreover, We theoretically establish the method's effectiveness in terms of improved Fréchet Inception Distance (FID) and second-order Tweedie moment for reduced trajectory deviation.

1. Introduction

Diffusion model [13, 24, 26] has emerged as a powerful tool in image synthesis and solving inverse problems. Due to their unprecedented success in high-fidelity image generation, they are widely used in various low-level vision tasks, including super-resolution [11, 12, 16, 27], deblurring [4, 19], computational tomography [25], and so on.

These generative models first add noise to the data, as a part of the forward diffusion process, using Itô Stochastic Differential Equation (SDE) $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$, and then learns to predict the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ from the forward process at every timestep t. Multiple works have tried to address various concerns in training [23, 28] and sampling strategy [29, 30] to improve the performance of the model Despite these efforts, inaccurate estimates of score function may lead to sampling drift resulting in convergence to sub-optimal to data distribution. Though some efforts have been made in Dara *et al.* [8], it requires retraining the model from scratch with an improved training objective, which may not be an ideal choice for large-scale models. To address this problem of sub-optimal sampling without necessarily retraining from scratch, our work focuses on generalized moment matching of underlying probability density in score function as a plug-and-play module in the existing methods. For this, we propose characteristic function (ChF) matching to minimize the sampling drift due to imperfect score functions. We also theoretically prove that ChF matching upper bounds the FID score. Moreover, we also show that ChF matching implicitly incorporates higher-order Tweedie moments in the sampling to help reduce sampling drift. **We enumerate the contribution of this work as follows**:

1. We propose a plug-and-play characteristic function consistency in the training and sampling stage to improve the fidelity score of the generated score in both random image generation and linear inverse problems.
2. We provide theoretical analysis to show that the proposed approach is equivalent to second-order Tweedie correction. We also prove that the proposed correction also reduces FID score and therefore, improves the overall image fidelity.
3. We experimentally demonstrate that by introducing this consistency, the overall performance of the existing diffusion models can be significantly improved in terms of FID and LPIPS scores.

2. Preliminary Background

In this section, we briefly discuss the concept of characteristic function and principles of diffusion models before describing our methodology.

2.1. Characteristic Function

For any random variable X that admits a probability distribution $p(\mathbf{x})$, its characteristic function is the Fourier Transform of the probability density function. the characteristic function $\phi(\mathbf{u}) : \mathbb{R}^d \rightarrow \mathbb{R}$ is estimated by

$$\phi(\mathbf{u}) = \mathbb{E}(e^{i\mathbf{u}^T\mathbf{X}}) = \int_{\mathbb{R}^d} e^{i\mathbf{u}^T\mathbf{x}} p(\mathbf{x}) dx \qquad (1)$$

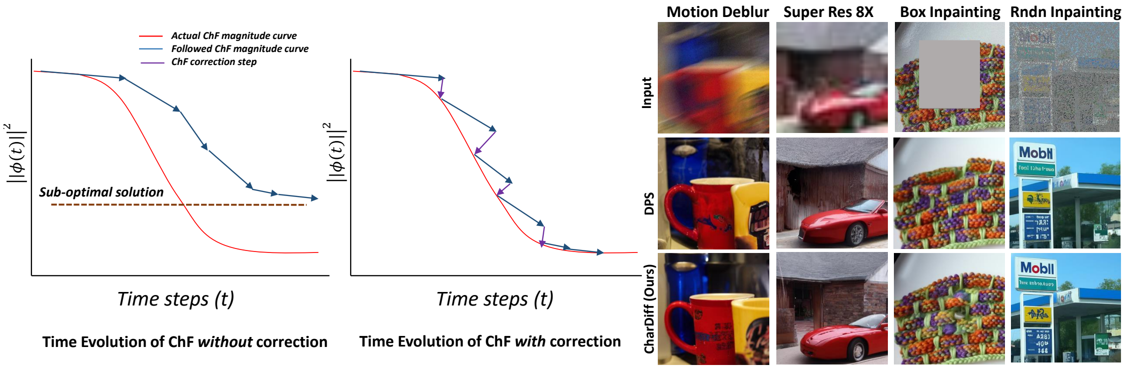


Figure 1. **Geometric Illustration (Left):** Due to imperfect score, the sampling drift may induce convergence of samples to a sub-optimal distribution, which is depicted in terms of ChF curve. The ChF correction forces the distribution to stay closer to the expected distribution at every time step. **Image inversion (Right):** CharDiff (*with DPS*) provides meaning reconstruction compared to DPS [6] alone. The jresults are compared here for three inversion taks: motion deblur, 8 × super-resolution, box inpainting and random inpainting.

For a normal distribution with mean μ and variance σ, the corresponding ChF is given by,

$$\phi(u) = e^{i\mu u - \frac{1}{2}\sigma^2 u} \tag{2}$$

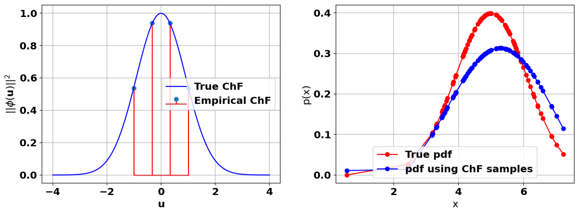


Figure 2. **Left:** Estimation of four ChF coefficients with 5 data samples. **Right:** Comparison of true probability density and density computed using these four ChF coefficients

For a given finite number of samples $\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ from distribution $\mathcal{X}$, the corresponding empirical characteristic function $\hat{\phi}(\mathbf{u})$ is computed as,

$$\hat{\phi}(\mathbf{u}) = \frac{1}{n}\sum_{j=1}^{n} e^{i\mathbf{u}^T\mathbf{x_j}} \tag{3}$$

The motivation to utilize the ChF is that it is always defined, since it is an integral of a bounded continuous function over finite probability measure. Moreover, it is unique for a given distribution. The probability density can therefore be estimated from ChF by,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^d}\int \phi(\mathbf{u})e^{-\mathbf{u}^T\mathbf{x}}d\mathbf{u}, \tag{4}$$

Subsequently, the squared characteristic function distance [5, 7] is between two distributions is given by,

$$d^2(\mathbb{X},\mathbb{Y}) = \mathbb{E}_{\mathbf{u}\sim\omega(\mathbf{u};\eta)}[||\phi_X(\mathbf{u}) - \phi_Y(\mathbf{u})||^2], \tag{5}$$

where $\omega(\mathbf{u};\eta)$ is the weighing function parameterized by η The idea of empirical characteristic function can be extended to measure the Squared Empirical Characteristic Function Distance to measure the distribution shift. For two given distributions $\mathbb{X}$ and $\mathbb{Y}$ with empirical ChF $\hat{\phi}_X(\mathbf{u})$ and $\hat{\phi}_Y(\mathbf{u})$ respectively, the corresponding distance is given by,

$$\hat{d}^2(\mathbb{X},\mathbb{Y}) = \frac{1}{N}\sum_{i=1}^{N}||\hat{\phi}_X(\mathbf{u}_i) - \hat{\phi}_Y(\mathbf{u}_i)||^2 \tag{6}$$

The probability density in 4 requires the knowledge of $\phi(\mathbf{u})$ over the full range of $\mathbf{u}$, which is practically intractable. Fortunately, [1] shows that the magnitude of $\phi(\mathbf{u})$ decays exponentially with the increase in the magnitude of $\mathbf{u}$ such that $||\phi(\mathbf{u})|| = \mathcal{O}(\frac{1}{1+|\mathbf{u}|_2^{\frac{d}{2}}})$. Consequently, we sample a finite set of $\mathbf{u}$ close to 0 to estimate Eq. 6. Figure 2 shows that that if very few ChF coefficients are sampled close to 0, the underlying density function estimated using those few coefficients remains closer to the true density function.

2.2. Diffusion Models

Diffusion models are primarily split in two categories, score based model (SGM) and denoising diffusion models (DDPM). In both of these approaches, target is to find $p_0 = p_{data}$ in inversion or denoising step. Moreover, DDPM is a discrete version of SGM, where time steps vary discretely in [0,T]. To this, Gaussian noise is artificially added to the target data at every time step t to get a new probability density p_t such $p_t \rightarrow \mathcal{N}(0, I)$ as $t \rightarrow \infty$. In SGM, time-rescaled Ornstein-Uhlenbeck process is commonly used Stochastic Differential Equation (SDE) used for noise addition.

$$d\mathbf{x_t} = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)}d\mathbf{w}_t, \quad \mathbf{x}_0 \sim \mathbf{x}_{data} \tag{7}$$

The corresponding reverse sampling-SDE for the above process is given by,

$$dx = \Big[\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)]dt + \sqrt{\beta(t)}d\mathbf{w} \quad (8)$$

The forward process in DDPM [13] gradually adds Gaussian noise to the data such that,

$$p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}, (1-\bar{\alpha}_t)\mathbf{I}), \quad (9)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_t = 1 - \beta_t$, where β_t is the variance. For the reverse process, the state$\mathbf{x}_t$ can be predicted with $\hat{\mu}_t$ and $\hat{\sigma}_t$ given by,

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \hat{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \hat{\sigma}^2\mathbf{I}), \quad (10)$$

where $\hat{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{1-\hat{\alpha}_t}\epsilon\Big)$ and $\hat{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$. Here, $\epsilon \sim (\prime, \mathbf{I}$.

For posterior sampling, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ in 8 is replaced by $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|y)$, which can further be written as,

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t|y) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}} \log p_t(y|\mathbf{x}_t) \quad (11)$$

Equation 8 involves two approximation to solve the reverse sampling equation. First, p_T is unknown since it is the noisy version of $x \sim p_{data}$. It can be solved by assuming that there are sufficient number of steps in the diffusion process such that $p_t \sim \mathcal{N}(0, I)$. Second, we also have no knowledge about $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$. It can be solved by score matching technique [14], where we use a model to approximate $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ using score function such that $s_\theta(\mathbf{x}_t, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$.

3. Related Works

Characteristic Function: Being a powerful generalization of probability measure, the characteristic function has been a promising approach in the generative models. Ansari *et al*. proposed OCFGAN [10] which formulated the learning of Implicit Generative Models in terms of characteristic functions. It is particularly based on the framework of MMD-GAN [2] by replacing the critic discrepancy measure using characteristic function loss. While this approach marginally improves the performance, it extensively undermines the physical interpretation of ChF. Li *et al*. proposed RCF-GAN [17] which interprets the meaning of real and complex parts of ChF as a way to strike a balance between diversity and accuracy. Though it experimentally validates this idea, it does not provide any theoretical interpretation of the claims. The application of ChF has not been studied for the diffusion models yet. In addition to experimental validation, we provide a theoretical guarantee of the observations so that it always holds for any class of datasets and models.

Inverse Problems: Diffusion models have set new benchmarks among the generative models and are now used to solve most of the inverse problems through methods like diffusion posterior sampling. They operate in either pixel space [13, 22] or latent space [20]. In pixel space diffusion models, DPS [6] utilizes the approximated posterior sampling in reverse sampling to solve noisy inverse problems. While $\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}_t)$ is intractable, Chung *et al*. in DPS provides an useful approximation which assumes that $p(\mathbf{y}|\mathbf{x}_t) \approx p(\mathbf{y}|\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])$. Rout *et al*. [21] proposed PSLD, which extends DPS to solve general noisy inverse problems by using orthogonal projection on the subspace of the Transformation matrix in between encoding and decoding steps to enforce fidelity.

Higher order Moments: The gradient of log-likelihood is key to converging at data distribution. Many methods [3, 18] explicitly estimate the Jacobian of score function or second-order moment to reduce the bias in the imperfectly estimated score function, which is a computationally expensive procedure. Instead of directly solving for higher-order moments, we prove that updating the sampling step with characteristic function Distance is equivalent to second-order correction.

4. Methodology

4.1. Unconditional Image Synthesis

In unconditional image synthesis, the objective is to randomly generate a sample from target data distribution from Gaussian noise as the only input. For this, we study the improvement using ChF in the context of both sampling and training. Moreover, since ChF in diffusion model is time dependent, we exchangably use $\phi(\mathbf{u}, t)$ and $\phi(\mathbf{u})$.

Training: To train the model, we add additional ChF matching loss using Eq. 6 along with score matching loss as shown in Algorithm 1. Given the finite numbers of $\mathbf{u}$, we can show that the upper bound on the MSE between true

Algorithm 1 Training DDPM with ChF correction

```
Require: u for sampling $\phi(\mathbf{u})$
Find mean of the target distribution $\mu = \mathbb{E}_{p_{data}}[x]$
for i in range (N) do
    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
    $t \sim \mathcal{U}(0, T)$
    $\mathbf{x}_t \sim \mathcal{N}(\sqrt{\overline{\alpha}_i}\mathbf{x}_0, \sqrt{1-\overline{\alpha}_i\mathbf{I}})$
    Take gradient descent step on,
    $\nabla_\theta ||s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)||^2$ ▷ Score Matching Loss
    Choose $t_i|_{i=1}^{K}$ such that $t_i \sim \mathcal{U}(0, T)$
    Reverse-sample $\hat{\mathbf{x}}_t$ for $t_i \in \mathbf{t}$ using $s_\theta$
    Take the gradient step on ChF loss,
    $\nabla_\theta \sum_{\mathbf{u}} ||\frac{1}{K}\sum_{i=1}^{K} e^{j\mathbf{u}^T\hat{\mathbf{x}}_{t_i}} - e^{j\mathbf{u}^T\overline{\alpha}_i\mu - 0.5\mathbf{u}^T(1-\overline{\alpha}_i)\mathbf{I}\mathbf{u}}||^2$
    ▷ ChF Matching Loss
end
```

Algorithm 2 Sampling DDPM with ChF correction

```
Require: Batch size $B$, Precalculated mean of the target distribution $\mu = \mathbb{E}_{p_{data}}[x]$
$\mathbf{x}_T \sim \mathcal{N}(0, I)$
for $t$ in $T-1$ to 0, do
    $\mathbf{z}_t \sim \mathcal{N}(0, I)$ if $t > 0$ else $\mathbf{z} = 0$
    $\mathcal{L} = \sum_{\mathbf{u}} || \frac{1}{B} \sum_{i=1}^{B} e^{j\mathbf{u}^T \hat{\mathbf{x}}_{t_i}} - e^{j\mathbf{u}^T \mu - 0.5\mathbf{u}^T \sigma_t \mathbf{u}} ||^2$
    $\hat{\mathbf{x}}_t = \mathbf{x}_t - \gamma \nabla_{\mathbf{x}_t} \mathcal{L}$    ▷ ChF Correction
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \hat{\mathbf{x}}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} s_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end
```

score function and ChF based score function s_θ is given by,

$$\sqrt{\int ||s(\mathbf{x}, t) - s_\theta(\mathbf{x}, t)||_2^2 p(\mathbf{x}) d\mathbf{x}} = \max_{|\mathbf{u}|} \mathcal{O}\Big(\frac{1}{1 + |\mathbf{u}|^d} \Big) \tag{12}$$

For a given normal distribution with mean and variance μ and Σ, respectively, the corresponding ChF function is given by $\phi(\mathbf{u}) = e^{j\mathbf{u}^T \mu - 0.5\mathbf{u}^T \Sigma \mathbf{u}}$. The ChF can serve two objectives under certain assumptions. Firstly, it always minimizes the distribution shift for the derived at every timestep and therefore, helps converge to target data distribution. Second, if we choose covariance matrix to be a diagonal matrix, as shown in Proposition 1, then ChF loss is equivalent to minimizing the shift between the time evolution of true ChF and the given predicted ChF.

Proposition 1 *Let $\frac{\partial \phi(\boldsymbol{u})}{\partial t}$ be the rate of change of the characteristic function $\phi(\boldsymbol{u})$ with time. If we choose the diffusion term to be a scaled identity matrix, then for a given timestep t, $\frac{\partial \phi(\boldsymbol{u})}{\partial t} \propto \phi(\mathbf{u})$. Consequently, $||\frac{\partial \phi(\boldsymbol{u})}{\partial t} - \frac{\partial \hat{\phi}(\boldsymbol{u})}{\partial t}||_2^2 \propto ||\phi(\mathbf{u}) - \hat{\phi}(\mathbf{u})||_2^2$.*

Sampling: Another alternative to generate sampling quality is to suitably modify the sampling strategy without retraining the model from scratch. Algorithm 2 illustrates the steps to directly implement ChF correction just before denoising step. If the resultant ChF the estimated $\mathbf{x}_t$ deviates from the true ChF, the gradient projection corrects it before denoising step to minimize drift. It is to be observed that we need a certain batch size to estimate empirical ChF. Experimentally, we found that batch size of at least 12 is sufficient to accurately estimate the ChF. Figure 1 illustrates the correction aided by ChF consistency to ensure that samples remains closer to the target distribution.

4.2. Inverse Problems

Given linear inverse problem $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$, where the measurement $\mathbf{y}$ and linear operator $\mathbf{A}$ are known, the aim is to find an estimate for $\mathbf{x}$. Unlike retraining baseline diffusion model from scratch for such complex tasks, we use ChF correction as a plug-and-play module in the sampling stage

Algorithm 3 DPS sampling with ChF correction

```
Require: $\mathbf{y}$, Number of steps $T$
$\mathbf{x}_T \sim \mathcal{N}(0, I)$
for $t$ in $T-1$ to 0, do
    $\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\overline{\alpha}_t}} \Big( \mathbf{x}_t + (1 - \overline{\alpha}_t) s_\theta(\mathbf{x}, t) \Big)$
    $\mathbf{z} \sim \mathcal{N}(0, I)$
    Generate $K$ no. of paired patches from $\mathbf{x}_t$ and $\hat{\mathbf{x}}_0$ given by $(\mathbf{x}_0^p, \mathbf{x}_t^p)$
    $\mathcal{L} = \sum_{\mathbf{u}} || \frac{1}{K} \sum_{i=1}^{K} e^{j\mathbf{u}^T \hat{\mathbf{x}}_{t_i}^p} - e^{j\mathbf{u}^T \sqrt{\overline{\alpha}_i} \hat{\mathbf{x}}_0 - 0.5\mathbf{u}^T \sigma_t \mathbf{u}} ||^2$ ▷ ChF correction
    $\mathbf{x}'_t = \mathbf{x}_t - \eta \nabla_{\mathbf{x}_t} \mathcal{L}$
    $\mathbf{x}'_{t-1} = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_t)}{1-\overline{\alpha}_t} \mathbf{x}'_t + \frac{\sqrt{\alpha_{t-1}}}{1-\overline{\alpha}_t} \mathbf{x}_0 + \sigma_t \mathbf{z}$
    $\mathbf{x}_{t-1} = \mathbf{x}'_{t-1} - \gamma \nabla_{\mathbf{x}_t} ||\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_0||^2$
end
```

of existing posterior sampling methods to further improve the quality of resultant estimates. However, there are subtle differences while using ChF correction for inverse problems compared to unconditional image synthesis. Firstly, we do not want to impose any batch size (to estimate empirical ChF) during sampling. To solve this problem, we can divide the sampled $\mathbf{x}_t$ into small patches and treat them as the independent samples to find $\hat{\phi}(\mathbf{u})$. We can do this since denoising step for each pixel in the image is independent of other pixels. Secondly, we do not have access to posterior mean to compute true ChF. Therefore, we directly use the approximate posterior mean evaluated at every time step.

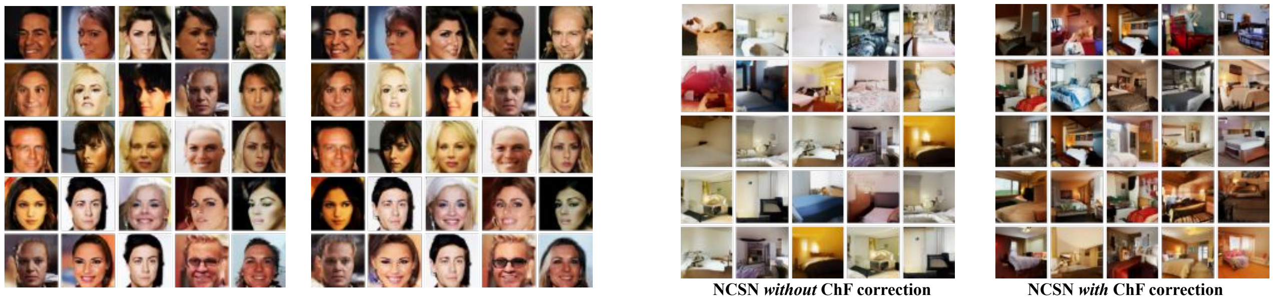
Remark 1 *To compute true characteristic function at every time step t, we use the approximate posterior mean given by $\hat{\mathbf{x}}_0 \approx \frac{1}{\sqrt{\overline{\alpha}_t}} \Big(\mathbf{x}_t + (1 - \overline{\alpha}_t) s_\theta(\mathbf{x}, t) \Big)$.*

Proposition 2 *Let $\mathbf{x}_t$ and $\mathbf{x}'_t$ are the corresponding samples in the forward diffusion and reverse sampling processes, respectively such that $||\mathbf{x}_t - \mathbf{x}'_t||_2^2 \leq \rho$. The upper bound on MSE loss between the ChF $\phi(\mathbf{u}, \mathbf{x}_0)$ with true posterior mean $\mathbf{x}_0$ and ChF $\phi(\mathbf{u}, \hat{\mathbf{x}}_0)$ using $\hat{\mathbf{x}}_0$ as the posterior mean is given by,*

$$||\phi(\mathbf{u}, \mathbf{x}_0) - \phi(\mathbf{u}, \hat{\mathbf{x}}_0)||^2 \leq \rho + \kappa ||s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{x}_0)||_2^2, \tag{13}$$

with probability of at least $1 - 2e^{-\frac{\rho^2}{4(1-\overline{\alpha}_t)}}$. Here, $\kappa = (1 - \overline{\alpha}_t)^2 ||\mathbf{u}||_2^2$.

Proposition 2 justifies our choice of approximating the true posterior mean in Eq. 2 by $\hat{\mathbf{x}}_0$. On the right hand side, ρ is the squared L2 distance between forward and reverse samples at time t, that is relatively small in suitable trained model. The term κ involves squared L2 distance of $\mathbf{u}$, $1 - \overline{\alpha}_t < 1$, and score matching loss. Since the sampled $\mathbf{u}$ is close to 0, the overall second term is also bound to remain closer to zero.



NCSN *without* ChF correction

NCSN *with* ChF correction

NCSN *without* ChF correction

NCSN *with* ChF correction

Figure 3. **Left:** Illustration of unconditional synthesis of CelebA faces with and without ChF correction. **Right:** Illustration of unconditional synthesis of LSUN bedroom images with and without ChF correction.

Sampling Method	Fashion MNIST	CIFAR-10		CelebA		LSUN-bedroom
	FID	**FID**	**IS**	**FID**	**NLL**	**FID**
DDPM	20.24	3.17	7.86	3.29	2.86	5.21
DDPM+ChF	19.25	**2.56**	8.21	**3.01**	**2.76**	4.89
DDIM	20.56	4.31	-	3.58	-	3.89
DDIM+ChF	19.42	4.14	**-**	3.16	-	**3.26**
NCSN	19.86*	25.32	8.87	8.21	-	38.92
NCSN+ChF	**19.21**	24.11	9.49	7.44	-	38.86

Table 1. Performance of diffusion model with and without ChF correction for three benchmark datasets on unconditional image synthesis with NFE=1000. * indicates that these numbers have not been reported in the previous works and have been computed at our end.

5. Theoretical Analysis

Theorem 1 *For two distributions $\mathbb{P}$ and $\mathbb{Q}$, let D be the diameter of the space supported by these two distributions. Then with probability of at least $1 - 2e^{-\frac{\delta^2 \mu_t^2}{2\sigma_t^2}}$, The Frechet Inception Distance between the samples at time step t from these two distributions is upper bounded by,*

$$\mathcal{FID}(p, q) \leq k\mathbb{E}_{\mathbf{u}}[\|\phi_P(\mathbf{u}) - \phi_Q(\mathbf{u})\|], \tag{14}$$

where $k = \frac{32L^2\mu_t^2\delta^2}{(2\pi)^d}$ and L is the Lipschitz constant of Inception Network.

Based on Theorem 1, we can infer that at every sampling step, the ChF correction ensures that the underlying distribution in the learned score function remains close to the true distribution. Since every sample estimated at a given time step t depends on the $t + 1$, ChF correction minimizes the propagation of sampling drift due to imperfect score and therefore, the final sample at $t = 0$ converges closer to p_{data}.

Theorem 2 *(Tweedie Sampler from ChF correction for Inverse Problems)* Let $\mathcal{L} = \mathbb{E}_{\mathbf{u}}[\|\phi(\mathbf{u}) - \hat{\phi}(\mathbf{u})\|_2^2]$. Let $\delta_{discrete}$ be the discretization error between the forward and reverse samples at every time step t. The gradient of the loss function is given by,

$$\nabla_{\mathbf{x}_t}\mathbb{E}_{\mathbf{u}}[\mathcal{L}] = \mathcal{A} + \mathcal{B}(\mathbb{I} - \nabla^2_{\mathbf{x}_t}\log p_\theta), \tag{15}$$

where $\mathcal{A} = \mathbb{E}\Big[\mathbf{u}^T\delta_{discrete}\mathbf{u}\Big(cos(\frac{\mathbf{u}^T\mathbf{x}_t}{2}) - sin(\frac{\mathbf{u}^T\mathbf{x}_t}{2})\Big)^2\Big]$ and $\mathcal{B} = \mathbb{E}\Big[e^{-\sigma_t^2||\mathbf{u}||_2^2}\mathbf{u}^T\delta_{discrete}\mathbf{u}\Big(cos(\frac{\mathbf{u}^T\mu_t}{2}) - sin(\frac{\mathbf{u}^T\mu_t}{2})\Big)^2\Big]$.

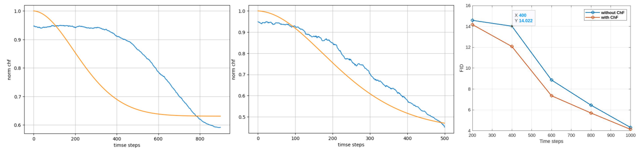


Figure 4. **Left:** Time Evolution of magnitude of Characteristic function with and without ChF correction. **Orange** curve denotes the magnitude of Original function, and **Blue** curve represents the observed one. **Right:** Variation of FID score with Time steps.

Theorem 2 shows that gradient of ChF correction can be expressed using the Hessian of log likelihood or gradient of score function. This second order correction in Tweedie estimator helps reduce bias towards $\mathbb{E}_{x_T \sim P(x_T|x_t)}[x_T]$ and tends to generate samples $x_T \sim p_{T-t}(x_T|x_t)$. This correction, in turn, improves the meaningful details in the reconstructed samples.

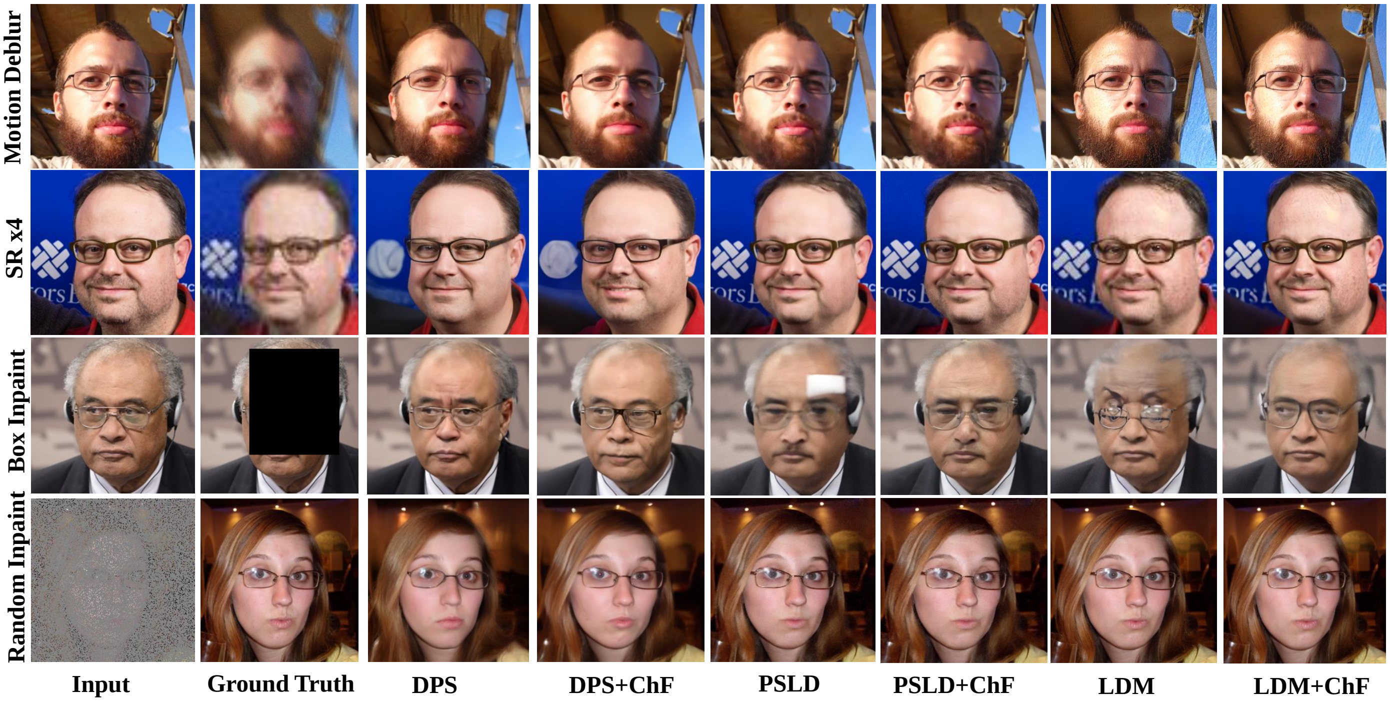


Figure 5. Qualitative comparison of diffusion models with different sampling methods.

Methods	Inpaint (Random)		SR x4		Gaussian Deblur		Inpaint (box)	
	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS
DDRM	69.71	0.587	62.15	0.294	74.92	0.332	42.93	0.204
DPS	33.48	0.212	39.35	0.214	44.05	0.257	35.14	0.216
PSLD	21.34	0.096	34.28	0.201	41.53	0.221	43.11	0.167
Score-SDE	76.54	0.612	96.72	0.563	109.0	0.403	60.06	0.331
DPS+ChF	31.89	**0.201**	38.77	0.199	42.87	0.236	**30.07**	0.176
PSLD+ChF	**20.67**	0.204	**32.87**	**0.146**	**40.26**	**0.216**	41.97	**0.159**

Table 2. Quantitative comparison of different methods for various linear inverse problem tasks on FFHQ 256×256 dataset.

6. Experiments

6.1. Experimental Setup

We evaluate our methodology for three different tasks, including unconditional image synthesis, inverse problems (super-resolution, inpainting, deblurring), and text-to-image generation. For all these tasks, we present and compare the results both quantitatively and qualitatively.

Quantitative metrics: Firstly, We use FID to evaluate the sample quality for all the tasks. For unconditional synthesis, we additionally use the Inception score and Negative log-likelihood. For inverse problem tasks, we evaluate the sample quality in terms of LPIPS due to access to the paired ground truth.

6.2. Unconditional Image Synthesis

To evaluate our methodology, we experiment with Fashion MNIST, CIFAR-10, CelebA and LSUN bedroom datasets. We implement the ChF correction in the sampling stage of DDPM, DDIM, and NCSN-v2. Unless specified, the value of γ is always set to $\frac{0.1}{t+1}$ for the time step t. Table 1 presents the quantitative analyses for these datasets in terms of Inception Score (**IS**), Frechet Inception Distance (**FID**) and Negative Log-Likelihood Score (**NLL**). It is evident that the inclusion of ChF correction in the sampling step drastically improves these numbers in most of the cases. Additionally, Figure 1 depicts the sample quality for CelebA and LSUN bedroom datasets for the NCSN model. We run the sampling stage under a fixed global setting due to which the given nosie converges to a particular sample in the distribution. For the CelebA dataset, we particularly observed that ChF correction helps reduce the overall distortion in the facial structure, whereas in the case of LSUN bedroom, it enhances the diversity in the composition of the bedroom images without introducing any significant distortion. Figure 4 further shows that in the absence of ChF correction, the reverse sampling steps remain flattened in the beginning indicating that there is no effective denoising taking place. ChF correction reduces the effective flattening that results in faster convergence towards data samples for CIFAR-10 datasets in DDIM.

6.3. Conditional Image Synthesis

Table 2 quantitatively compares the performance of ChF correction with other methods, inlcuding DDRM [15], PSLD [21], DPS [6] and Score-SDE [26], for FFHQ 256 × 256 dataset. It can be observed that ChF correction, when added to DPS and PSLD, significantly improves the performance in multiple linear inverse problem tasks. Figure 5 depicts visual comparison of the samples, which shows that artifacts in PSLD gets corrected through our approach. Similarly, the quality of generated image also improves significantly for both DPS and LDM.

Image Super-Resolution: To study the effectiveness of Chf correction, we add the proposed module to DPS [6], PSLD [21], and LDM [20]. It is sufficient to measure the LPIPS score at different NFE to observe the convergence of data sample. Table 3 shows that the LPIPS score for our method is consistently lower than that of DDRM and DPS alone. Larger steps sizes may introduce larger drift in the characteristic function leading to poor sample quality. This equivalently reflects in the FID score.

Method	1000	600	400	200	50
DDRM x4 (DDRM)	0.304	0.306	0.296	0.298	0.330
SR x4 (DPS)	0.214	0.218	0.234	0.262	0.371
SR x4 (DPS+ChF)	**0.199**	**0.206**	**0.213**	**0.238**	**0.317**

Table 3. LPIPS score vs NFE for 4× super-resolution task on FFHQ dataset.

Image deblurring: We compared the results for Gaussian deblur in Table 2 to show that ChF correction surpassed the image quality scores with respect to other methods. Additionally, we present the quality metrics in Table 4 for motion deblur with kernel size of 61 and intensity set to 0.5. We note that the ChF correction surpasses the performance of DPS by almost 23 % in terms of FID and and 32 % in terms of LPIPS. Similarly, we observe drastic improvement of 19 % and 9% in terms of FID and LPIPS scores of PSLD respectively.

Random Image Inpainting: Table 2 compares the results for (20 %, 80 %) uniform random drop in pixels in which we observe performance over both PSLD and DPS. Following this experiment, we further investigate the variation in these scores with varying pixel drop rates. Figure 7 presents such variation in which we observe that ChF correction, when augmented with DPS and PSLD, consistently maintains better LPIPS score for various extent of pixel drops.

Methods	Motion Deblur	
	FID	**LPIPS**
DPS	56.08	0.389
PSLD	51.02	0.292
DPS+ChF	42.92	0.306
PSLD+ChF	**40.86**	**0.263**

Table 4. Results comparison for Motion deblur problem.

Task	**DPS**	**PSLD**	**DPS+ChF**	**PSLD+ChF**
Box-96	25.62	32.19	**22.36**	31.69
Box-128	35.14	43.11	**30.07**	41.97
Box-160	51.06	53.18	**48.26**	50.91

Table 5. FID score for different methods under variable box sizes, where N in Box-N is the size of box.

Box inpainting: We further study the approach in line by experimenting with variable box sizes. Table 5 compares the methods for various box sizes in terms of FID score. IT can be clearly seen that ChF augmented sampling consistently preforms better than DPS and PSLD.

6.4. Text-to-Image Generation

Apart from conditional and unconditional image generation, we also evaluate our method for text-to-image generation task. Following [20], we experiment with MS-COCO captions dataset to generate 256 × 256 sized image using the captions. We use LDM as the baseline with BERT as the encoder network to generate text embeddings. The ChF correction is added to the latent space sampling which is sandwiched between an encoder and a decoder network. Table 6 compares the performance for various text conditional image synthesis among which ChF correction improves the performance of the latent diffusion model in terms of FID. However, we also observed a marginal decrease in the Inception score compared to LDM alone. The left-hand side of Figure 7 illustrates the visual and semantic quality of the text conditioned generated images. It can be observed that ChF correction incorporates more meaningful features in the image based on the given textual input. Similarly, on the right-hand side, we additionally compare the image-to-image translation tasks conditioned on user given prompt.

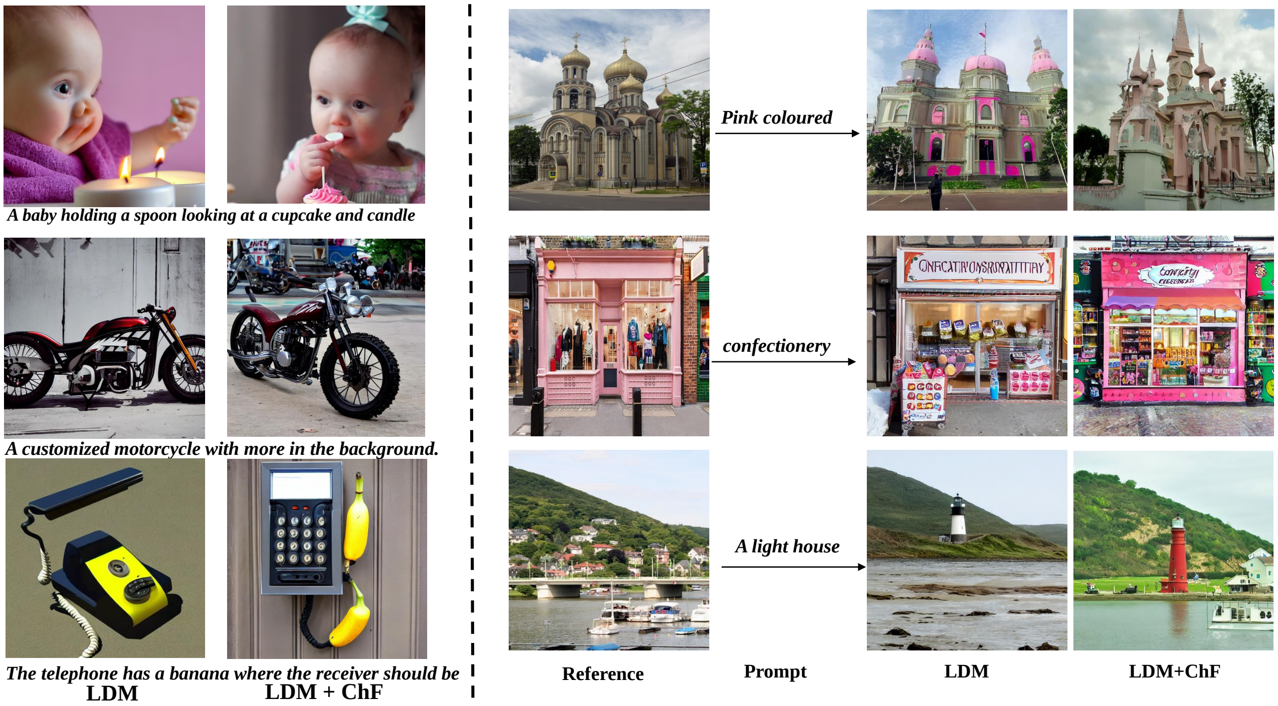

A baby holding a spoon looking at a cupcake and candle
A customized motorcycle with more in the background.
The telephone has a banana where the receiver should be
LDM
LDM + ChF
Pink coloured
confectionery
A light house
Reference
Prompt
LDM
LDM+ChF

Figure 6. **Left:** Illustration of Text-to-Image generation with and without ChF correction using Latent Diffusion Model. **Right:** Image-to-Image Translation conditioned on the prompt.

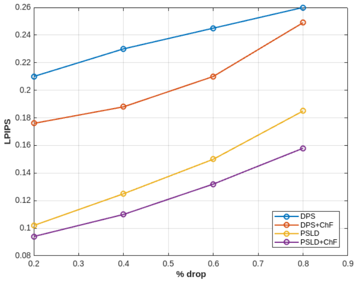


Figure 7. LPIPS vs % pixel drop in random inpainting task.

Method	FID	IS	Params
CogView [9]	27.10	18.20	4B
LAFITE [31]	26.94	26.02	75M
LDM-KL-8	23.31	20.03	1.45B
LDM-KL-8+ChF	22.16	19.20	1.45B

Table 6. Text conditional image synthesis results for MS-COCO dataset for 250 DDIM steps.

7. Discussions and Conclusion

7.1. Limitations

We observed that ChF correction improved the sample quality in different image generation tasks, including unconditional image synthesis, conditional image synthesis and text-to-image generation. We noted that though our approach worked extremely well in pixel space sampling, it appeared to lag in terms of performance in latent space sampling. Moreover, we also observed that in certain cases, for example changing initial random samples, it seemed to introduce color imbalance in the generated during latent space sampling.

7.2. Conclusions

In this work, we proposed a plug-and-play method for improving sampling convergence in diffusion models. We further showed its effectiveness by experimenting with various image-generation tasks. Moreover, we also provided theoretical justification to the assumption and choice we made in this work. As part of future work, it will be interesting to further explore the aforementioned limitations in addition to studying the adversarial robustness.

References

[1] Magda Amiridi, Nikos Kargas, and Nicholas D. Sidiropoulos. Low-rank characteristic tensor density estimation part i: Foundations. *IEEE Transactions on Signal Processing*, 70:2654–2668, 2022. 2

[2] Michael Arbel, Danica J. Sutherland, Mikoł aj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans.

In *Advances in Neural Information Processing Systems*, volume 31, 2018. 3

[3] Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O. Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems, 2023. 3

[4] Zheng Chen, Yulun Zhang, Liu Ding, Xia Bin, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *NeurIPS*, 2023. 1

[5] Jamal-Dine Chergui. The integrated squared error estimation of parameters. *Extracta Mathematicae*, 11(3):435–442, 1996. 2

[6] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 7

[7] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, 2015. 2

[8] Giannis Daras, Yuval Dagan, Alex Dimakis, and Constantinos Costis Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 8

[10] Abdul Fatir Ansari, Jonathan Scarlett, and Harold Soh. A characteristic function approach to deep implicit generative modeling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7476–7484, 2020. 3

[11] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9935–9946, 2023. 1

[12] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10021–10030, 2023. 1

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 3

[14] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. 3

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 7

[16] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1

[17] Shengxi Li, Zeyang Yu, Min Xiang, and Danilo Mandic. Reciprocal adversarial learning via characteristic functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 217–228, 2020. 3

[18] Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[19] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *2023 IEEE/CVF Conference on Computer Vision*, 2023. 1

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

[21] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 7

[22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, 2015. 3

[23] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[24] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1

[25] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. 1

[26] 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*, 2021. 1, 7

[27] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 1

[28] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang "Atlas" Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 72137–72154, 2023. 1

[29] Suttisak Wizadwongsa and Supasorn Suwajanakorn. Accelerating guided diffusion sampling with splitting numerical methods. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[30] Guoqiang Zhang, Kenta Niwa, and W. Bastiaan Kleijn. On accelerating diffusion-based sampling processes via improved integration approximation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[31] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, June 2022. 8