Improving Diffusion Models for Inverse Problems Using Optimal Posterior Covariance

Xinyu Peng [1] **Ziyang Zheng** [1] **Wenrui Dai** [1] **Nuoqian Xiao** [1] **Chenglin Li** [1] **Junni Zou** [1] **Hongkai Xiong** [1]

Abstract

Recent diffusion models provide a promising zero-shot solution to noisy linear inverse problems without retraining for specific inverse problems. In this paper, we reveal that recent methods can be uniformly interpreted as employing a Gaussian approximation with hand-crafted isotropic covariance for the intractable denoising posterior to approximate the conditional posterior mean. Inspired by this finding, we propose to improve recent methods by using more principled covariance determined by maximum likelihood estimation. To achieve posterior covariance optimization without retraining, we provide general plug-and-play solutions based on two approaches specifically designed for leveraging pre-trained models with and without reverse covariance. We further propose a scalable method for learning posterior covariance prediction based on representation with orthonormal basis. Experimental results demonstrate that the proposed methods significantly enhance reconstruction performance without requiring hyperparameter tuning.

1. Introduction

Noisy linear inverse problems are widely studied for a variety of image processing tasks, including denoising, inpainting, deblurring, super-resolution, among others. The noisy linear inverse problems are formulated to accommodate the widely adopted degradation model, where images are measured with a linear projection under noise corruption.

Recently, diffusion models have been emerging as promising methods for solving inverse problems. According to the training strategies, these diffusion-based solvers can be categorized into two groups: 1) *supervised methods* that aim to learn a conditional diffusion model using datasets consisting of pairs of degraded and clean images (Saharia et al., 2022; Whang et al., 2022; Luo et al., 2023; Chan et al., 2023), and 2) *zero-shot methods* that leverage pre-trained unconditional diffusion models for conditional sampling in various scenarios of inverse problems without the requirement of retraining. In this paper, we focus on the zero-shot methods that can accommodate various tasks without retraining. To ensure the consistency of optimizing fidelity under conditional sampling, existing zero-shot methods adopt projection onto the measurement subspace (Choi et al., 2021; Lugmayr et al., 2022; Song et al., 2022; Wang et al., 2023; Zhu et al., 2023), leverage the similar idea as classifier guidance (Song et al., 2021b; Dhariwal & Nichol, 2021) to modify the sampling process with the likelihood score (Song et al., 2023; Chung et al., 2023a), or resort to variational inference (Feng et al., 2023; Mardani et al., 2024).

Diffusion models initially establish a forward process that introduces noise to the original data $\mathbf{x}_0$ to generate noisy data $\mathbf{x}_t$ at time t. Subsequently, a reverse process generates $\mathbf{x}_0$ following the distribution of the original data. The key to realizing the reverse process lies in the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$, which represents the optimal estimate of $\mathbf{x}_0$ given $\mathbf{x}_t$ in the sense of minimum mean square error (MMSE). In inverse problems, achieving conditional sampling that ensures samples consistent with the measurement $\mathbf{y}$ requires considering the conditional posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$, which plays a similar role to $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ in unconditional sampling.

However, differing from the methods originated from approximating the conditional score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ that can be regarded as explicitly approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ (Song et al., 2023; Chung et al., 2023a), the underlying motivations of recent methods can differ significantly from the objective of approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$. These methods achieve conditional sampling by refining the null space during the reverse diffusion process (Wang et al., 2023), interpreting denoising diffusion models as plug-and-play image priors (Zhu et al., 2023), or optimizing the data consistency in the tangent space (Chung et al., 2023b). There remain two important problems unresolved, *i.e.*, *i) unified*

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Ziyang Zheng <zhengziyang@sjtu.edu.cn>, Wenrui Dai <daiwenrui@sjtu.edu.cn>, Junni Zou <zoujunni@sjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

interpretation in the sense of approximating the conditional posterior mean, and *ii) better approximation for the conditional posterior mean*.

In this paper, we address these two problems. We reveal that recent zero-shot methods (Song et al., 2023; Chung et al., 2023a; Wang et al., 2023; Zhu et al., 2023) can be uniformly interpreted as employing an isotropic Gaussian approximation for the intractable denoising posterior $p_t(\mathbf{x}_0|\mathbf{x}_t)$ to approximate $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$. This perspective allows us to not only reveal a unified interpretation for these methods but also extend the design space of diffusion-based solvers. Specifically, we propose a generalized method to improve recent methods by optimizing posterior covariance based on maximum likelihood estimation (MLE). Numerically, we achieve plug-and-play posterior covariance optimization using pre-trained unconditional diffusion models by converting (available) reverse covariance or via Monte Carlo estimation without reverse covariance. To overcome the quadratic complexity of covariance prediction, we further propose a scalable method for learning posterior covariance prediction by leveraging widely-used orthonormal basis for image processing (*e.g.*, DCT and DWT basis). Experimental results demonstrate that the proposed method significantly outperforms existing methods in a wide range of tasks, including inpainting, deblurring, and super-resolution, and eliminates the need for hyperparameter tuning.

2. Background

2.1. Bayesian Framework for Solving Inverse Problems

We consider the linear inverse problems for infering $\mathbf{x}_0 \in \mathbb{R}^d$ from noisy measurements $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times d}$ is known and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is an i.i.d. additive Gaussian noise with a known standard deviation of σ. This gives a likelihood function $p(\mathbf{y}|\mathbf{x}_0) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}_0, \sigma^2\mathbf{I})$. Under the Bayesian framework, we assume $\mathbf{x}_0$ obeys an unknown prior distribution $p(\mathbf{x}_0)$, and the inverse problems are solved by formulating the posterior distribution over $\mathbf{x}_0$ given $\mathbf{y}$ by Bayes' theorem: $p(\mathbf{x}_0|\mathbf{y}) = p(\mathbf{x}_0)p(\mathbf{y}|\mathbf{x}_0)/\int p(\mathbf{x}_0)p(\mathbf{y}|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0$.

2.2. Diffusion Models and Conditioning

We are interested in using diffusion models to model the complex posterior distribution $p(\mathbf{x}_0|\mathbf{y})$ for inverse problems. Let us define a family of Gaussian perturbation kernels $p_t(\mathbf{x}_t|\mathbf{x}_0)$ of $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ by injecting *i.i.d.* Gaussian noise of standard deviation σ_t to $\mathbf{x}_0$ and scaling by the factor of s_t, *i.e.*, $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$, where σ_t is monotonically increasing with respect to time $t = 0, 1, \cdots, T$. We start from $\sigma_0 = 0$ and reach a value of σ_T that is much larger than the standard deviation of $p(\mathbf{x}_0)$, to ensure that samples from $\mathbf{x}_T \sim p(\mathbf{x}_T)$ are indistinguishable to samples from $\mathcal{N}(\mathbf{0}, s_T^2\sigma_T^2\mathbf{I})$. Since $\mathbf{x}_t$ is independent of $\mathbf{y}$ once $\mathbf{x}_0$ is known, we characterize the joint distribution between $\mathbf{x}_0, \mathbf{y}$, and $\mathbf{x}_t$ as $p_t(\mathbf{x}_0, \mathbf{y}, \mathbf{x}_t) = p(\mathbf{x}_0)p(\mathbf{y}|\mathbf{x}_0)p_t(\mathbf{x}_t|\mathbf{x}_0)$, which can also be represented by a probabilistic graphical model $\mathbf{y} \leftarrow \mathbf{x}_0 \rightarrow \mathbf{x}_t$. There exist multiple formulations of diffusion models in the literature (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021a; Kingma et al., 2021; Song et al., 2021b; Karras et al., 2022). Here, we use the ordinary differential equation (ODE) formulation and select $s_t = 1, \sigma_t = t$ as suggested in (Karras et al., 2022) for simplicity. Let us consider the following ODE:

$$\mathrm{d}\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]}{t}\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T). \tag{2}$$

In Eq. (2), the only source of randomness is the initial sample $\mathbf{x}_T \sim p_T(\mathbf{x}_T)$. The ODE possesses an important property that $\mathbf{x}_t$ generated by the ODE maintains the exact same marginals to $\mathbf{x}_t$ obtained by injecting Gaussian noise to $\mathbf{x}_0$, *i.e.*, $p_t(\mathbf{x}_t)$. Generally, diffusion models approximate $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ with a time-dependent denoiser $D_t(\mathbf{x}_t)$ (referred to as the unconditional diffusion model throughout the paper) that is trained by minimizing the L_2 loss for all $t \in [0, T]$:

$$\min_{D_t} \mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)} \left[\|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_2^2\right]. \tag{3}$$

With sufficient data and model capacity, the optimal $D_t(\mathbf{x}_t)$ is the MMSE estimator of $\mathbf{x}_0$ given $\mathbf{x}_t$ and equals to $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$. Thus, samples from $p(\mathbf{x}_0)$ can be obtained by first sampling $\mathbf{x}_T$ from $\mathcal{N}(\mathbf{0}, s_T^2\sigma_T^2\mathbf{I})$ and then simulating Eq. (2) from $t = T$ to $t = 0$ using black box ODE solver that replaces $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ with a well-trained $D_t(\mathbf{x}_t)$.

To solve inverse problems, we are interested in $p(\mathbf{x}_0|\mathbf{y})$, which requires constructing an ODE whose marginals are $p_t(\mathbf{x}_t|\mathbf{y})$. The desired ODE is

$$\mathrm{d}\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]}{t}\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T|\mathbf{y}). \tag{4}$$

For sufficiently large σ_T, samples $\mathbf{x}_T \sim p(\mathbf{x}_T|\mathbf{y})$ are indistinguishable to samples from $\mathcal{N}(\mathbf{0}, s_T^2\sigma_T^2\mathbf{I})$ (see Appendix D.2 in (Dhariwal & Nichol, 2021)). The unconditioned sampling procedure in Eq. (2) can be used to achieve conditioned sampling in Eq. (4), by substituting $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ with the conditional posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$[1].

3. Unified Interpretation of Diffusion-based Solvers to Inverse Problems

To simulate Eq. (4) for solving inverse problems, a standard approach is training a conditional diffusion model for es-

[1]Note that this conclusion can be extended to DDPM, DDIM, and SDE-formalized diffusion models.

timating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ using supervised learning. However, this approach can be computationally demanding by training separate models for different scenarios. In this paper, we focus on zero-shot methods, which leverage pre-trained unconditional diffusion models for conditional sampling in inverse problems to avoid additional training. We provide a construction of zero-shot methods based on conditional posterior mean approximation as below:

1) **Denoising** Obtain an estimation of $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ through the unconditional diffusion model $D_t(\mathbf{x}_t)$;

2) **Modification** Obtain an approximation $\hat{\mathbf{x}}_0^{(t)}$ for the conditional posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ based on $\mathbf{y}$ and $D_t(\mathbf{x}_t)$;

3) **Substitution** Substitute $\hat{\mathbf{x}}_0^{(t)}$ for $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ in the unconditional sampling process Eq. (2).

In this section, we unify existing methods with the interpretation of approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ with isotropic Gaussian approximations $\mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), r_t^2\mathbf{I})$ to the intractable denoising posterior $p_t(\mathbf{x}_0|\mathbf{x}_t)$ with different r_t, as summarized in Table 1. We first review the Type I guidance methods that explicitly approximate $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$, including DPS (Chung et al., 2023a) and ΠGDM (Song et al., 2023). Subsequently, we reveal that the proximal-based methods like DDNM (Wang et al., 2023) and DiffPIR (Zhu et al., 2023) can also be interpreted as approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ and categorize them into the Type II guidance methods.

3.1. Type I Guidance: Approximating the Likelihood Score Function

We classify DPS and ΠGDM into the category of Type I guidance that originates from approximating the conditional score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$. We represent these methods with an equivalent form of conditional posterior mean approximation based on Proposition 3.1.

Proposition 3.1. *The conditional posterior mean is equal to the posterior mean drifted by scaled likelihood score function. Formally,*

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] + s_t\sigma_t^2\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t). \quad (5)$$

Proof. Please refer to Appendix A.2. □

To obtain $\hat{\mathbf{x}}_0^{(t)}$ according to Proposition 3.1[2], note that $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ can be estimated using $D_t(\mathbf{x}_t)$; however, the likelihood score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ is computationally intractable. In inverse problems, only the likelihood $p(\mathbf{y}|\mathbf{x}_0)$ at $t = 0$ is known, and $p_t(\mathbf{y}|\mathbf{x}_t)$ for any $t > 0$ is obtained by an intractable integral over all possible $\mathbf{x}_0$ as

$$p_t(\mathbf{y}|\mathbf{x}_t) = \int p(\mathbf{y}|\mathbf{x}_0)p_t(\mathbf{x}_0|\mathbf{x}_t)\mathrm{d}\mathbf{x}_0. \quad (6)$$

Thus, one possible way to obtain $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ is to consider approximating $p_t(\mathbf{y}|\mathbf{x}_t)$. DPS and ΠGDM achieve this by uniformly employing isotropic Gaussian approximation for $p_t(\mathbf{x}_0|\mathbf{x}_t)$, as elaborated below.

DPS (Chung et al., 2023a) DPS can be viewed as approximating $p_t(\mathbf{x}_0|\mathbf{x}_t)$ using a delta distribution $\delta(\mathbf{x}_0 - D_t(\mathbf{x}_t))$ centered at the posterior mean estimate $D_t(\mathbf{x}_t)$, which can be regarded as the limit of the Gaussian $\mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), r_t^2\mathbf{I})$ when the variance r_t^2 approaches zero. The likelihood $p_t(\mathbf{y}|\mathbf{x}_t)$ is approximated by

$$\begin{aligned} p_t(\mathbf{y}|\mathbf{x}_t) &\approx \int p(\mathbf{y}|\mathbf{x}_0)\delta(\mathbf{x}_0 - D_t(\mathbf{x}_t))\mathrm{d}\mathbf{x}_0 \\ &= p(\mathbf{y}|\mathbf{x}_0 = D_t(\mathbf{x}_t)). \end{aligned} \quad (7)$$

However, directly using Eq. (7) does not perform well in practice, and Chung et al. (2023a) empirically adjusts the strength of guidance by approximating the likelihood score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ with $-\zeta_t\nabla_{\mathbf{x}_t}\|\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)\|_2^2$, where $\zeta_t = \zeta/\|\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)\|_2$ with a hyper-parameter ζ.

ΠGDM (Song et al., 2023) The delta distribution used in DPS is a very rough approximation to $p_t(\mathbf{x}_0|\mathbf{x}_t)$ as it completely ignores the uncertainty of $\mathbf{x}_0$ given $\mathbf{x}_t$. As t increases, the uncertainty in $p_t(\mathbf{x}_0|\mathbf{x}_t)$ becomes larger and is closed to the original data distribution $p(\mathbf{x}_0)$. Thus, it is more reasonable to choose a positive r_t. In ΠGDM, r_t is heuristically selected as $\sqrt{\sigma_t^2/(1+\sigma_t^2)}$ under the assumption that $p(\mathbf{x}_0)$ is the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In such case, the likelihood $p_t(\mathbf{y}|\mathbf{x}_t)$ is approximated by

$$\begin{aligned} p_t(\mathbf{y}|\mathbf{x}_t) &\approx \int \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}_0, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), r_t^2\mathbf{I})\mathrm{d}\mathbf{x}_0 \\ &= \mathcal{N}(\mathbf{y}|\mathbf{A}D_t(\mathbf{x}_t), \sigma^2\mathbf{I} + r_t^2\mathbf{A}\mathbf{A}^T), \end{aligned} \quad (8)$$

where the gradient of the log-likelihood of Eq. (3) can be computed via Jacobian-vector product using back-propagation for approximating $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$.

Methods	Guidance	r_t
DPS (Chung et al., 2023a)	I	approach 0
ΠGDM (Song et al., 2023)	I	$\sqrt{\sigma_t^2/(\sigma_t^2+1)}$
DDNM (Wang et al., 2023)	II	any fixed value
DiffPIR (Zhu et al., 2023)	II	$\sigma_t/\sqrt{\lambda}$

Table 1: **Unified interpretation of diffusion-based solvers to inverse problems.** Recent methods can be regarded as making isotropic Gaussian approximations to the denoising posterior. $p_t(\mathbf{x}_0|\mathbf{x}_t)$

[2]In addition to realizing guidance for diffusion models, Proposition 3.1 can be used for refining the estimation of $\mathbf{x}_0$ for learning sampling patterns in MRI (Ravula et al., 2023) and realizing guidance for flow-based generative models to solve inverse problems (Pokle et al., 2023).

3.2. Type II Guidance: Approximating the Conditional Posterior Mean Using Proximal Solution

We classify DiffPIR and DDNM into the category of Type II guidance, which replaces $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ in the unconditional sampling process using proximal solutions. Although Type II guidance methods have been developed from perspectives that differ from approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$, we show that it can also be regarded as approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ by employing isotropic Gaussian approximation to the denoising posterior.

DiffPIR (Zhu et al., 2023) The core step in DiffPIR is replacing $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ in the unconditional sampling process with the solution to the following proximal problem:

$$\hat{\mathbf{x}}_0^{(t)} = \arg\min_{\mathbf{x}_0} \|\mathbf{y} - \mathbf{A}\mathbf{x}_0\|^2 + \rho_t \|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_2^2, \quad (9)$$

where $\rho_t = \lambda\sigma^2/\sigma_t^2$ and λ is a hyper-parameter. Our key insight is that Eq. (9) for DiffPIR can be interpreted as approximating $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ with the mean of an approximate distribution $q_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y})$ derived from the isotropic Gaussian approximation $q_t(\mathbf{x}_0|\mathbf{x}_t)$ to $p_t(\mathbf{x}_0|\mathbf{x}_t)$. Let us define the isotropic Gaussian approximation $q_t(\mathbf{x}_0|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), r_t^2\mathbf{I})$. We can obtain that the approximate distribution $q_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}_0)q_t(\mathbf{x}_0|\mathbf{x}_t)$ for $p_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}_0)p_t(\mathbf{x}_0|\mathbf{x}_t)$[3] is also a Gaussian and its mean $\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ can be obtained by solving the optimization problem:

$$\begin{aligned}\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] &= \arg\max_{\mathbf{x}_0} \log q_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) \\ &= \arg\max_{\mathbf{x}_0} [\log p(\mathbf{y}|\mathbf{x}_0) + \log q_t(\mathbf{x}_0|\mathbf{x}_t)] \\ &= \arg\min_{\mathbf{x}_0} \left[\|\mathbf{y} - \mathbf{A}\mathbf{x}_0\|^2 + \frac{\sigma^2}{r_t^2}\|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_2^2\right]. \quad (10)\end{aligned}$$

Note that Eq. (9) used in DiffPIR is a special case of Eq. (10) with $r_t = \sigma_t/\sqrt{\lambda}$. Therefore, DiffPIR can be viewed as using isotropic Gaussian approximation $q_t(\mathbf{x}_0|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), (\sigma_t^2/\lambda)\mathbf{I})$ for $p_t(\mathbf{x}_0|\mathbf{x}_t)$.

DDNM (Wang et al., 2023) The core step in DDNM resorts to the range-null space decomposition of $D_t(\mathbf{x}_t)$. Specifically, to ensure the measurement consistency $\mathbf{y} = \mathbf{A}\hat{\mathbf{x}}_0^{(t)}$ of the solution $\hat{\mathbf{x}}_0^{(t)}$, DDNM formulates $\hat{\mathbf{x}}_0^{(t)} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})D_t(\mathbf{x}_t)$ by replacing the range space component of $D_t(\mathbf{x}_t)$ with $\mathbf{A}^\dagger\mathbf{y}$ but keeping the null space component unchanged. We find that, when the measurement noise σ vanishes, replacing the range space component $\mathbf{A}^\dagger\mathbf{A}D_t(\mathbf{x}_t)$ with $\mathbf{A}^\dagger\mathbf{y}$ is equivalent to replacing $D_t(\mathbf{x}_t)$ with $\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$. Therefore, DDNM is equivalent to approximating $p_t(\mathbf{x}_0|\mathbf{x}_t)$ using an isotropic Gaussian, as formalized in Proposition 3.2.

[3] $p_t(\mathbf{y}|\mathbf{x}_0, \mathbf{x}_t) = p(\mathbf{y}|\mathbf{x}_0)$ since we leverage the conditional independence between $\mathbf{y}$ and $\mathbf{x}_t$ given $\mathbf{x}_0$.

Proposition 3.2. *For any $r_t > 0$, $\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ approaches $\hat{\mathbf{x}}_0^{(t)}$ used in DDNM as the variance of measurement noise approaches zero. Formally, we have*

$$\lim_{\sigma\to 0} \mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})D_t(\mathbf{x}_t). \quad (11)$$

Proof. Please refer to Appendix A.3. □

3.3. Solving Inverse Problems with Optimal Posterior Covariance

In Sections 3.1 and 3.2, we unify existing zero-shot diffusion-based solvers with the interpretation of isotropic Gaussian approximations to the intractable denoising posterior $p_t(\mathbf{x}_0|\mathbf{x}_t)$. This unified interpretation motivates us to further optimize the posterior covariance in the Gaussian approximation for enhancing existing diffusion-based solvers. Specifically, we consider to approximate $p_t(\mathbf{x}_0|\mathbf{x}_t)$ using variational Gaussian $q_t(\mathbf{x}_0|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_0|D_t(\mathbf{x}_t), \Sigma_t(\mathbf{x}_t))$ with learnable covariance $\Sigma_t(\mathbf{x}_t)$. In the pre-training stage, $q_t(\mathbf{x}_0|\mathbf{x}_t)$ is learned by minimizing the weighted integral of expected forward KL divergence between $p_t(\mathbf{x}_0|\mathbf{x}_t)$ and $q_t(\mathbf{x}_0|\mathbf{x}_t)$ as

$$\min_q \int \omega_t \mathbb{E}_{p_t(\mathbf{x}_t)}[D_{KL}(p_t(\mathbf{x}_0|\mathbf{x}_t)\|q_t(\mathbf{x}_0|\mathbf{x}_t))]\mathrm{d}t. \quad (12)$$

Eq. (12) can be achieved in a tractable way by maximizing the log-likelihood of $q_t(\mathbf{x}_0|\mathbf{x}_t)$. When $\Sigma_t(\mathbf{x}_t)$ is full rank, the optimal $D_t(\mathbf{x}_t)$ is exactly the MMSE estimator $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$. Once $q_t(\mathbf{x}_0|\mathbf{x}_t)$ is learned, we can leverage D_t and Σ_t to solve inverse problems and accommodate existing methods by plugging $D_t(\mathbf{x}_t)$ in Eq. (2) to achieve unconditional sampling. The solutions to Type I and Type II guidance are summarized as below.

Type I guidance. The likelihood is approximated in a similar way to Eq. (8) using D_t and Σ_t:

$$p_t(\mathbf{y}|\mathbf{x}_t) \approx \mathcal{N}(\mathbf{y}|\mathbf{A}D_t(\mathbf{x}_t), \sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T). \quad (13)$$

Type II guidance. $\hat{\mathbf{x}}_0^{(t)}$ is solved from the *auto-weighted* proximal problem associated with D_t and Σ_t:

$$\hat{\mathbf{x}}_0^{(t)} = \arg\min_{\mathbf{x}_0} \|\mathbf{y} - \mathbf{A}\mathbf{x}_0\|^2 + \sigma^2\|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_{\Sigma_t^{-1}}^2, \quad (14)$$

where $\|\mathbf{x}\|_\Lambda^2 = \mathbf{x}^T\Lambda\mathbf{x}$ with positive definite matrix Λ.

To avoid the inversion of high-dimensional matrices for realizing guidances in practice, we develop efficient closed-form solutions under the isotropic posterior covariance $\Sigma_t(\mathbf{x}_t) = r_t^2(\mathbf{x}_t)\mathbf{I}$, as well as numerical solutions based on the conjugate gradient method (CG) for more general covariance. Please refer to Appendix B for details.

4. Posterior Covariance Optimization

In this section, we discuss practical methods for posterior covariance optimization. We propose two plug-and-play methods which can be directly applied to recent methods for two common cases: 1) reverse covariance prediction is available from the given unconditional diffusion model (Section 4.1), and 2) reverse covariance prediction is not available (Section 4.2). Furthermore, to model the ubiquitous pixel-correlations in natural images, we propose to learn the variance in the transform space (Section 4.3), which addresses the quadratic complexity in covariance prediction.

4.1. Converting Optimal Reverse Variances

Recent pre-trained diffusion models often predict the optimal *reverse variances*[4] (Nichol & Dhariwal, 2021) for improving performance when using ancestral sampling of denoising diffusion probabilistic models (DDPM, (Ho et al., 2020; Sohl-Dickstein et al., 2015)). In this section, we reveal that the reverse variance prediction can be leveraged for posterior variance prediction. To achieve this, We first develop the fixed-point solutions to Eq. (12) in Proposition 4.1 and then establish the relation between the fixed-point solutions to Eq. (12) and DDPM in Eq. (19) in Theorem 4.2.

Proposition 4.1 (Fixed-point solutions of variational Gaussian posterior)**.** *The optimal mean $D_t^*(\mathbf{x}_t)$ and the optimal diagonal posterior covariance $\Sigma_t^*(\mathbf{x}_t) = \mathrm{diag}[\mathbf{r}_t^{*2}(\mathbf{x}_t)]$ to Eq. (12) are obtained by*

$$D_t^*(\mathbf{x}_t) = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t], \tag{15}$$

$$\mathbf{r}_t^{*2}(\mathbf{x}_t) = \mathbb{E}_{p_t(\mathbf{x}_0|\mathbf{x}_t)}[(\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])^2], \tag{16}$$

Proof. Please refer to Appendix A.5. □

where $(\cdot)^2$ denotes element-wise square. Subsequently, we consider the fixed-point solutions of DDPM. Unlike continuous ODE formulation introduced in Section 2.2, diffusion models pre-trained under the DDPM framework are latent variable models defined by $p(\mathbf{x}_0) = \int p(\mathbf{x}_{0:T})\mathrm{d}\mathbf{x}_{1:T}$. The joint distribution $p(\mathbf{x}_{0:T})$ is referred to as the *reverse process* defined as a Markov chain of learnable Gaussian transitions starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, which are characterized by the mean $\mathbf{m}_t$ and covariance $\mathbf{C}_t$:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{17}$$

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}|\mathbf{m}_t(\mathbf{x}_t), \mathbf{C}_t(\mathbf{x}_t)). \tag{18}$$

DDPM defines a forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ by gradually injecting noise to the data. Please refer to Appendix A.4 for the detailed definition of $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$. We fit $p(\mathbf{x}_0)$ to the original data distribution $q(\mathbf{x}_0)$ by minimizing the KL divergence between the forward and reverse processes:

$$\min_p D_{KL}(q(\mathbf{x}_{0:T})||p(\mathbf{x}_{0:T})), \tag{19}$$

where $q(\mathbf{x}_{0:T}) = q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)$. In Theorem 4.2, we present the fixed-point solution of DDPM under diagonal covariance $\mathbf{C}_t(\mathbf{x}_t) = \mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]$ as used in (Nichol & Dhariwal, 2021).

Theorem 4.2 (Fixed-point solutions of DDPM)**.** *Let $\mathbf{C}_t(\mathbf{x}_t) = \mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]$ be a signal-dependent diagonal covariance for the reverse covariance. When $\tilde{\mu}_t, \beta_t, \bar{\alpha}_t, \bar{\beta}_t$ are determined by the forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, the optimal solutions $\mathbf{m}_t^*(\mathbf{x}_t)$ and $\mathbf{v}_t^{*2}(\mathbf{x}_t)$ to Eq. (19) are*

$$\mathbf{m}_t^*(\mathbf{x}_t) = \tilde{\mu}_t(\mathbf{x}_t, \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]), \tag{20}$$

$$\mathbf{v}_t^{*2}(\mathbf{x}_t) = \tilde{\beta}_t + (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t})^2 \cdot \mathbf{r}_t^{*2}(\mathbf{x}_t) \tag{21}$$

where $\mathbf{r}_t^{*2}(\mathbf{x}_t)$ is the optimal posterior variances determined by Eq. (16) under $s_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{\bar{\beta}_t/\bar{\alpha}_t}$, and $\tilde{\beta}_t = (\bar{\beta}_{t-1}/\bar{\beta}_t)\beta_t$.

Proof. Please refer to Appendix A.4. □

Theorem 4.2 implies that, given the reverse variances $\hat{\mathbf{v}}_t^2(\mathbf{x}_t)$ predicted by a pre-trained DDPM model at time t, the posterior variances are obtained according to Eq. (21) by

$$\hat{\mathbf{r}}_t^2(\mathbf{x}_t) = (\hat{\mathbf{v}}_t^2(\mathbf{x}_t) - \tilde{\beta}_t) \cdot \left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\right)^{-2}. \tag{22}$$

In fact, similar fixed-point solutions of DDPM have been proposed in (Bao et al., 2022b) and (Bao et al., 2022a) for determining optimal reverse variances. Our approach can be viewed as opposite direction to (Bao et al., 2022b) and (Bao et al., 2022a) for applying the fixed-point solutions.

4.2. Monte Carlo Estimation of Posterior Variances

We further develop a method using a model without providing the reverse variances prediction $\hat{\mathbf{v}}_t^2(\mathbf{x}_t)$ (*e.g.*, (Ho et al., 2020)). In this case, we suppose signal-independent $\Sigma_t(\mathbf{x}_t) = r_t^2\mathbf{I}$ is an isotropic covariance with a time-dependent variance r_t^2. We directly optimize r_t^2 by forcing the derivative of Eq. (12) w.r.t. r_t^2 to zero. Thus, we obtain that

$$r_t^{*2} = \frac{1}{d}\mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)}\left[\|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_2^2\right]. \tag{23}$$

Note that r_t^{*2} is the the expected reconstruction error of $D_t(\mathbf{x}_t)$, which can be estimated using Monte Carlo samples to calculate the empirical mean of $\|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|_2^2$ offline. In the experiments, we pre-compute the empirical mean for 1000 discrete time steps using 5% of the dataset, and for any t used in sampling, we use the pre-computed result with the nearest time step to t.

[4] Variances are the diagonal elements of the covariance matrix.

4.3. Modeling Pixel-Correlations With Latent Variances

We have achieved posterior covariance optimization in a plug-and-play fashion in Sections 4.1 and 4.2. In this section, we present a scalable way to model pixel correlations using orthonormal transforms, motivated by transform coding (Goyal, 2001) for image compression. We assume $\mathbf{x}_0$ can be represented by some orthonormal basis $\mathbf{\Psi}$, such that $\mathbf{x}_0 = \mathbf{\Psi}\theta_0$. According to the property of covariance,

$$\mathrm{Cov}[\mathbf{x}_0|\mathbf{x}_t] = \mathbf{\Psi}\mathrm{Cov}[\theta_0|\mathbf{x}_t]\mathbf{\Psi}^T. \tag{24}$$

Here, the elements of θ_0 are supposed to be mutually independent. Since conditioning on $\mathbf{x}_t$ is equivalent to conditioning on a perturbed version of θ_0 with isotropic Gaussian noise, the elements of $\theta_0|\mathbf{x}_t$ are also mutually independent. This motivates the idea that, $\mathrm{Cov}[\theta_0|\mathbf{x}_t]$ could be better approximated by a diagonal matrix than $\mathrm{Cov}[\mathbf{x}_0|\mathbf{x}_t]$, if the elements of θ_0 are more "closed to" mutually independent than $\mathbf{x}_0$ using a proper basis. Thereby, we parameterize the covariance of the variational Gaussian posterior $q_t(\mathbf{x}_0|\mathbf{x}_t)$ as

$$\Sigma_t(\mathbf{x}_t) = \mathbf{\Psi}\mathrm{diag}[\mathbf{r}_t^2(\mathbf{x}_t)]\mathbf{\Psi}^T. \tag{25}$$

Note that the number of parameters to predict is significantly reduced from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$, especially for image data with $d > 10^5$. The training procedure can be simply implemented as learning a diagonal Gaussian in the transform domain (see Appendix C.2 for details). We adopt the discrete wavelet transform (DWT) for $\mathbf{\Psi}$ in our experiments.

5. Related Work

Posterior covariance optimization. Concurrent to our work, Boys et al. (2023) leverage second-order Tweedie's formula to optimize posterior covariance for solving inverse problems. The optimal $\Sigma_t(\mathbf{x}_t)$ to (12) is solved using the Jacobian of $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$, *i.e.*, $\Sigma_t^*(\mathbf{x}_t) = \sigma_t^2\nabla_{\mathbf{x}_t}\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$[5]. However, this approach requires strong approximation to scale to high-dimensional data like images. In particular, row sum approximation for the Jacobian, *i.e.*, $\nabla_{\mathbf{x}_t}\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] \approx \mathrm{diag}[\nabla_{\mathbf{x}_t}\mathbf{1}^T D_t(\mathbf{x}_t)]$ is adopted for image experiments. Rout et al. (2023) use only the trace of $\nabla_{\mathbf{x}_t}D_t(\mathbf{x}_t)$. This is related to Section 4.2, since the trace of $\Sigma_t^*(\mathbf{x}_t)$ equals to $\mathbb{E}_{p_t(\mathbf{x}_0|\mathbf{x}_t)}[\|\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]\|_2^2]$. In contrast, our approach does not introduce additional computational cost for covariance prediction at the inference time.

High-order denoising score matching. The approach presented in Section 4.3 can be viewed as a variant of high-order denoising score matching (Meng et al., 2021; Lu et al., 2022) whose goal is to directly predict the high-order moments of the denoising posterior without exploiting high-order Tweedie's formulas. This is more practically appealing since small error in the first-order loss does not guarantee small estimation error in high-order scores and the Tweedie's approaches could suffer from large estimation error in practice (Meng et al., 2021). However, these approaches (Meng et al., 2021; Lu et al., 2022) can only be employed on small-scale datasets. To our best knowledge, our approach is the first scalable method applicable for large-scale datasets of high-resolution natural images.

[5]According to moment matching (Bishop, 2006), $\Sigma_t^*(\mathbf{x}_t)$ equals to the covariance of the denoising posterior $\mathrm{Cov}[\mathbf{x}_0|\mathbf{x}_t]$.

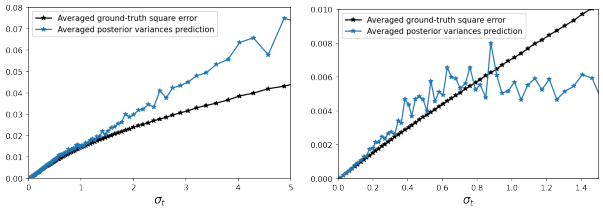

Figure 1: **Averaged values of $\mathbf{e}$ (black line) and $\mathbf{r}_t^2(\mathbf{x}_t)$ (blue line)** on FFHQ (Left) and ImageNet (Right).

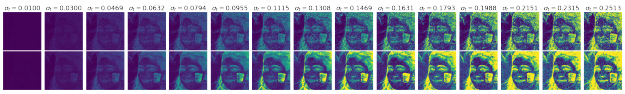

Figure 2: **Visualization of $\mathbf{e}$ and $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ of an example image $\mathbf{x}_0$ at different t.** Top: $\mathbf{e}$; Bottom: $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$. The results are averaged over RGB channels for better visualization.

6. Experiments

In this section, we evaluate the proposed methods on a range of inverse problems, including inpainting, debluring, and super resolution. To minimize the effect of different implementations for fair comparison, our techniques and prior works are all implemented on a unified codebase based on an open-source diffusion codebase k-diffusion[6]. Additional details and results are included in Appendix C. The source code is available at `https://github.com/xypeng9903/k-diffusion-inverse-problems`.

6.1. Sanity Check for Converting Reverse Variance

The derivation of Eq. (22) is based on the perfect model. However, estimation error of reverse variance in practice may lead to failure when applying Eq. (22). To validate the effectiveness of Eq. (22) in practice, we compare the ground-truth square errors made by the denoiser $\mathbf{e} = (\mathbf{x}_0 - D_t(\mathbf{x}_t))^2$ with the posterior variance prediction $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$. We compare $\mathbf{e}$ with $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ since the posterior variance $\mathbf{r}_t^{*2}(\mathbf{x}_t)$ is the MMSE estimator of $\mathbf{e}$ given $\mathbf{x}_t$ (assuming $D_t(\mathbf{x}_t) = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$), as suggested by Eq. (16). This implies that a good $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ should be a reliable predictor of $\mathbf{e}$. Thus, we compare $\mathbf{e}$ with $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ as a sanity check to gauge

[6]`https://github.com/crowsonkb/k-diffusion`

Dataset	Method	Inpaint (Random)			Deblur (Gaussian)			Deblur (Motion)			Super resolution (4×)		
		SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓
FFHQ	Convert (*Ours*)	**0.9279**	**0.0794**	**25.90**	0.7905	**0.1836**	**52.42**	0.7584	0.2156	**62.88**	0.7878	0.1962	58.37
	Analytic (*Ours*)	0.9272	0.0845	28.83	0.7926	0.1850	53.09	0.7579	0.2183	64.77	0.7878	0.1968	59.83
	DWT-Var (*Ours*)	0.9209	0.0863	28.54	**0.7968**	0.1837	57.52	**0.7677**	**0.2103**	65.34	**0.8025**	**0.1856**	**57.26**
	TMPD	0.8224	0.1924	70.92	0.7289	0.2523	76.52	0.7014	0.2718	83.49	0.7085	0.2701	79.58
	DPS	0.8891	0.1323	49.46	0.6284	0.3652	136.12	0.4904	0.4924	212.48	0.7719	0.2054	61.36
	ΠGDM	0.8784	0.1422	49.89	0.7890	0.1910	59.93	0.7543	0.2209	66.14	0.7850	0.2005	61.46
ImageNet	Convert (*Ours*)	0.8559	**0.1329**	**29.14**	0.6007	**0.3327**	95.23	**0.5634**	**0.3656**	**109.61**	0.5869	0.3477	96.76
	Analytic (*Ours*)	0.8481	0.1446	35.51	**0.6009**	0.3334	**93.21**	0.5611	0.3668	113.39	**0.5958**	0.3495	95.33
	TMPD	0.7011	0.2892	293.80	0.5430	0.4114	291.29	0.4773	0.4567	302.40	0.5186	0.4298	296.73
	DPS	**0.8623**	0.1490	36.58	0.4603	0.4630	173.77	0.3582	0.5554	282.21	0.5860	**0.3231**	**92.89**
	ΠGDM	0.7658	0.2328	64.96	0.5946	0.3429	102.89	0.5534	0.3781	113.89	0.5925	0.3552	100.36

Table 2: **Quantitative results on FFHQ and ImageNet dataset for Type I guidance.** We use **bold** and underline for the best and second best, respectively. Note that DPS here uses the step size of $1/(2\sigma^2)$ for pure covariance comparisons.

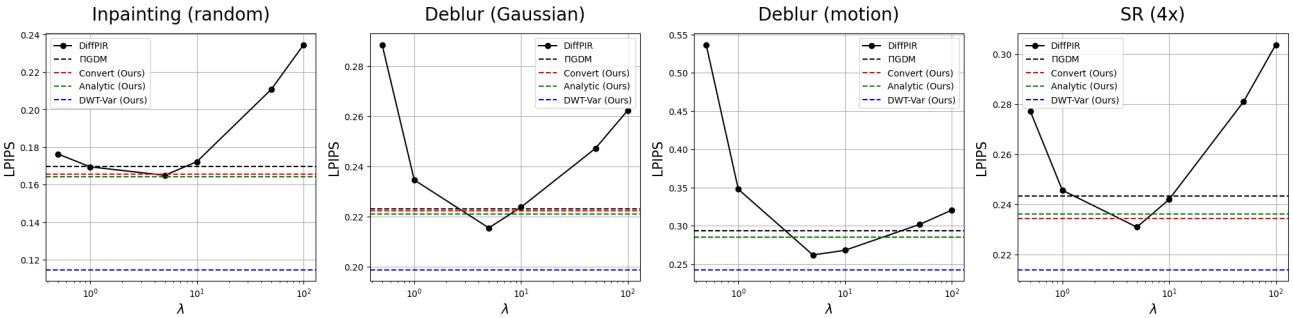

Figure 3: **LPIPS comparisons on FFHQ for Type II guidance.** For DiffPIR, we report LPIPS under different λ.

the effectiveness of Eq. (22) in practice. In Figure 1, we plot their averaged values over all the pixels and test images. The posterior variances prediction obtained via Eq. (22) is accurate only in the regions with low noise level. This is reasonable since the optimal reverse variance $\mathbf{v}^{*2}(\mathbf{x}_t)$ is bounded by the upper bound β_t and the lower bound $\tilde{\beta}_t$. The two bounds are almost equal in the regions with high noise level (Nichol & Dhariwal, 2021). When t is large, $\mathbf{v}^{*2}(\mathbf{x}_t) - \tilde{\beta}_t \approx \mathbf{0}$ and Eq. (22) becomes a $0/0$ limit that possesses high numerical instability. In Figure 2, we visualize $\mathbf{e}$ and $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ of an image $\mathbf{x}_0$ at different t and show that $\mathbf{e}$ can be well predicted by $\hat{\mathbf{r}}_t^2(\mathbf{x}_t)$ when t is small.

6.2. Quantitative Results

Experimental setup. Following (Chung et al., 2023a; Wang et al., 2023), we perform experiments on the FFHQ 256×256 and ImageNet 256×256 datasets to compare different methods with unconditional diffusion models from (Chung et al., 2023a) and (Dhariwal & Nichol, 2021), respectively. We learn optimal variance in DWT domain on FFHQ dataset by modifying the FFHQ model (see Appendix C.2 for details). The degradation models are specified mostly following (Zhu et al., 2023): (i) For inpainting, we randomly mask 50 percent of the total pixels. (ii) For Gaussian debluring and motion debluring, we use the same setup of the bluring kernels to (Chung et al., 2023a). (iii) For super resolution (SR), we consider bicubic downsampling. All measurements are corrupted by Gaussian noise with $\sigma = 0.05$. To evaluate different methods, we follow Chung et al. (2023a) to use three metrics: Structure Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS, Zhang et al. (2018)) and Frechet Inception Distance (FID, Heusel et al. (2017)). For the sampler setup, all Type I methods use the same Heun's 2nd deterministic sampler suggested in Karras et al. (2022) with 50 sampling steps, and all Type II methods use the same Heun's 2nd stochastic sampler ($S_{\text{churn}} = 80, S_{\text{tmin}} = 0.05, S_{\text{tmax}} = 50, S_{\text{noise}} = 1.003$, definition see Karras et al. (2022)) with 50 sampling steps since we found that Type II methods does not perform well using deterministic samplers[7].

[7]To understand how to leverage pre-trained DDPM model to perform sampling under perturbation kernels given in Section 2.2, please refer to Appendix C.4.

Dataset	Method	Inpaint (Random)			Deblur (Gaussian)			Deblur (Motion)			Super resolution (4×)		
		SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓
FFHQ	Convert (*Ours*)	**0.9241**	**0.0822**	**27.50**	0.7783	**0.1969**	**59.31**	0.7329	**0.2324**	**66.18**	**0.7632**	**0.2183**	**67.17**
	Analytic (*Ours*)	0.9232	0.0852	28.63	**0.7790**	0.1971	59.80	**0.7331**	0.2336	69.82	0.7622	0.2195	68.84
	ΠGDM	0.7078	0.2605	77.46	0.7221	0.2421	71.19	0.6977	0.2607	75.15	0.7205	0.2442	72.41
ImageNet	Convert (*Ours*)	**0.8492**	**0.1394**	**33.46**	**0.5770**	**0.3568**	**100.50**	**0.5341**	**0.3944**	131.48	**0.5613**	**0.3846**	**116.27**
	Analytic (*Ours*)	0.8417	0.1484	37.98	0.5722	0.3583	103.67	**0.5341**	0.3947	**126.39**	0.5537	0.3867	116.87
	ΠGDM	0.5102	0.4293	141.80	0.5071	0.4095	127.41	0.4887	0.4267	132.38	0.5150	0.4098	127.40

Table 3: **Results for ΠGDM with adaptive weight on FFHQ and ImageNet dataset.** We use **bold** and underline for the best and second best, respectively.

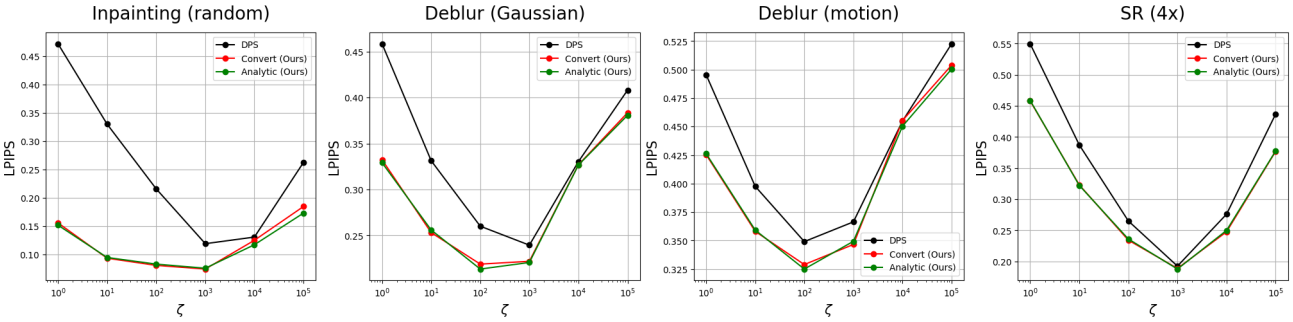

Figure 4: **LPIPS comparisons on FFHQ for DPS with heuristic step size $\zeta_t = \zeta/\|\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)\|_2$.** For comprehensive comparisons, we report LPIPS under different ζ.

In initial experiments, we observed that using the spatial variance predicted by the proposed plug-and-play methods (Section 4.1 and Section 4.2) for all sampling steps results in poor performance. This can be attributed to diagonal Gaussian variational $q_t(\mathbf{x}_0|\mathbf{x}_t)$ being an effective approximation to posterior $p_t(\mathbf{x}_0|\mathbf{x}_t)$ only for small noise level (Sohl-Dickstein et al., 2015; Xiao et al., 2022). To address this, we heuristically use the spatial variance only at the last several sampling steps with low noise level but use ΠGDM covariance for high noise level. We empirically find that sampling with spatial variance yields highly stable results when $\sigma_t < 0.2$ (*i.e.*, 12 out of 50 steps). Nevertheless, we observed that when using DWT variance (Section 4.3), the posterior approximation is accurate enough such that the variance can be used for all sampling steps. In the experiments, DWT variance is used when $\sigma_t < 1$ for efficiency, otherwise many CG steps are required to compute.

Results for pure covariance comparisons. Table 2 and Figure 3 summarize the results for Type I and Type II guidance. `DPS`, `ΠGDM`, `DiffPIR` refer to the posterior covariance in Table 1; `Convert`, `Analytic`, `DWT-Var`, and `TMPD` refer to the posterior covariance obtained using approaches presented in Sections 4.1, 4.2, 4.3, and $\sigma_t^2 \mathrm{diag}[\nabla_{\mathbf{x}_t} \mathbf{1}^T D_t(\mathbf{x}_t)]$ used in (Boys et al., 2023), respectively. Note that for Type I guidance, our methods achieve the best results in almost all tasks. Although DPS outperforms us in several cases, we observe that its performance is very unstable (see DPS performance in debluring tasks). In Figure 5, we visualiza reconstruction results of different methods. We observed that our methods reconstruct fine details of the image more faithfully compared to baselines. For Type II guidance, the plug-and-play methods obtain comparable performance to optimally-tuned DiffPIR and `DWT-Var` outperforms optimally-tuned DiffPIR by a significant margin without any hyper-parameter tuning.

Results with heuristics implementations. In addition to using Gaussian for denoising posterior approximation, prior works also propose several heuristics. Here we show that the methods with heuristics can also be improved by replacing the last sampling steps with our methods. In particular, ΠGDM introduces an *adaptive weight* to the likelihood score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ to adjust the guidance strength according to the timestep, which rescales the likelihood score by r_t^2 with $r_t^2 = \sigma_t^2/(1+\sigma_t^2)$. Table 3 summarizes the results for ΠGDM with adaptive weight. Obviously, ΠGDM achieves significant performance gain using Type I guidance with optimal variances at the last sampling steps. As for DPS, it introduces a trick to heuristically determine the step size of guidance, as mentioned in Section 3.1. Since this trick involves a hyper-parameter ζ, we report the performance under different ζ. Figure 4 summarizes the quantitative results for DPS with heuristic guidance strength. DPS

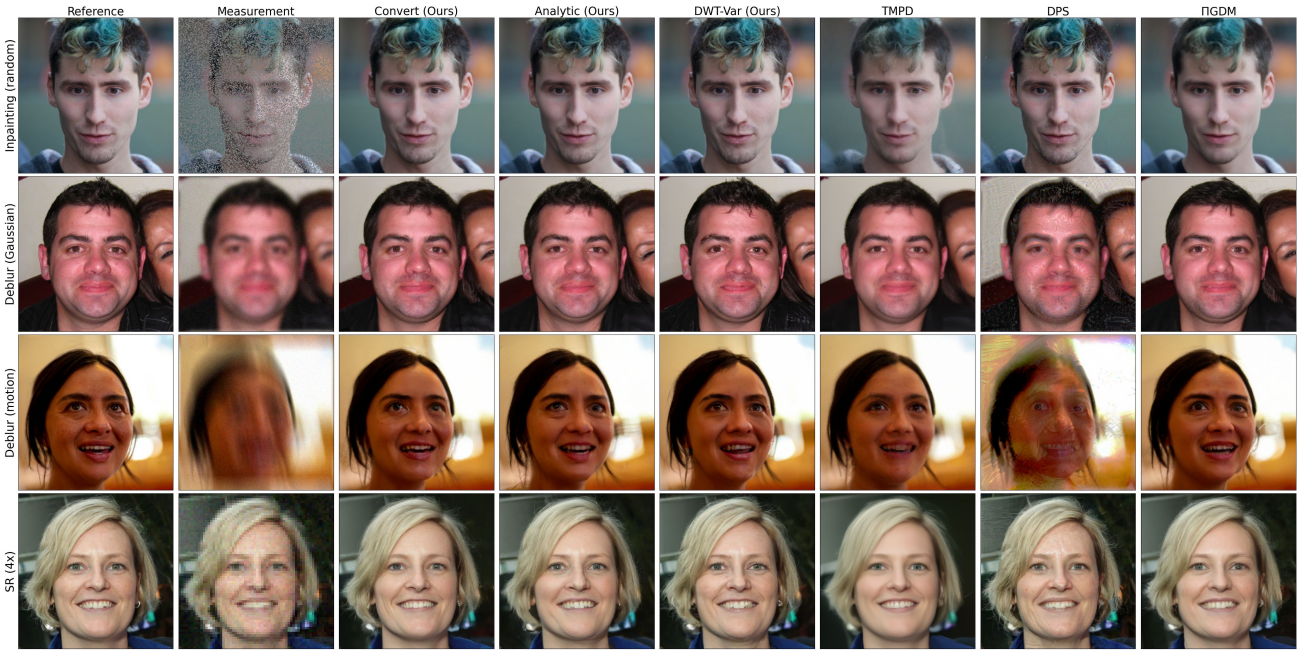

Figure 5: **Quantitative results for Table 2.** We observed that our methods reconstruct fine details of the image more faithfully compared to prior works.

with the last sampling steps replaced by our methods can achieve the best results across all hyper-parameters.

7. Conclusions

We reveal that recent methods can be uniformly interpreted as approximating the conditional posterior mean by employing Gaussian distribution with hand-crafted isotropic covariance to approximate the intractable denoising posterior. Inspired by this, we propose to enhance existing methods by using more principled covariance determined by maximum likelihood estimation. Experimental results show that the proposed methods significantly outperform existing methods and also eliminate the need for hyperparameter tuning.

Limitations. Despite the reasonable improvements our approaches offer over prior works, it is worth noting that, due to the diagonal constraint, the proposed covariance cannot match the optimum exactly even when the model is perfectly trained. Future research could explore better principles for designing the covariance (Dorta et al., 2018; Nehme et al., 2023), as well as leveraging nonlinear transformation (Ballé et al., 2020; Zhu et al., 2022; Li et al., 2024) for better correlation reduction. Additionally, methods like efficient approximation for Tweedie's approach could be investigated. These represent intriguing directions for future work.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62125109, Grant 61931023, Grant 61932022, Grant 62371288, Grant 62320106003, Grant 62301299, Grant T2122024, Grant 62120106007, Grant 62250055.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2020.

Bao, F., Li, C., Sun, J., Zhu, J., and Zhang, B. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 1555–1584, 2022a.

Bao, F., Li, C., Zhu, J., and Zhang, B. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion

probabilistic models. In *The 10th International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=0xiJLKH-ufZ.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Boys, B., Girolami, M., Pidstrigach, J., Reich, S., Mosca, A., and Akyildiz, O. D. Tweedie moment projected diffusions for inverse problems. *arXiv preprint arXiv:2310.06721*, 2023.

Chan, M. A., Young, S. I., and Metzler, C. A. SUD^2: Supervision by denoising diffusion models for image reconstruction. In *NeurIPS 2023 Deep Inverse Workshop*, 2023.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14347–14356, 2021.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In *The 11th International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Chung, H., Lee, S., and Ye, J. C. Fast diffusion sampler for inverse problems by geometric decomposition. *arXiv preprint arXiv:2303.05754*, 2023b.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34*, pp. 8780–8794, 2021.

Dorta, G., Vicente, S., Agapito, L., Campbell, N. D., and Simpson, I. Structured uncertainty prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5477–5485, 2018.

Feng, B. T., Smith, J., Rubinstein, M., Chang, H., Bouman, K. L., and Freeman, W. T. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10520–10531, 2023.

Goyal, V. K. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5):9–21, 2001.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pp. 6626–6637, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, pp. 6840–6851, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35*, pp. 26565–26577, 2022.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. In *Advances in Neural Information Processing Systems 34*, pp. 21696–21707, 2021.

Li, H., Li, S., Dai, W., Li, C., Zou, J., and Xiong, H. Frequency-aware transformer for learned image compression. In *The 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HKGQDDTuvZ.

Lu, C., Zheng, K., Bao, F., Chen, J., Li, C., and Zhu, J. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 14429–14460, 2022.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Luo, Z., Gustafsson, F. K., Zhao, Z., Sjölund, J., and Schön, T. B. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1680–1691, 2023.

Mardani, M., Song, J., Kautz, J., and Vahdat, A. A variational perspective on solving inverse problems with diffusion models. In *The 12th International Conference on Learning Representations*, 2024.

Meng, C., Song, Y., Li, W., and Ermon, S. Estimating high order gradients of the data distribution by denoising. In *Advances in Neural Information Processing Systems 34*, pp. 25359–25369, 2021.

Nehme, E., Yair, O., and Michaeli, T. Uncertainty quantification via neural posterior principal components. In *Advances in Neural Information Processing Systems 36*, pp. 37128–37141, 2023.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162–8171, 2021.

Pokle, A., Muckley, M. J., Chen, R. T., and Karrer, B. Training-free linear image inversion via flows. *arXiv preprint arXiv:2310.04432*, 2023.

Ravula, S., Levac, B., Jalal, A., Tamir, J. I., and Dimakis, A. G. Optimizing sampling patterns for compressed sensing MRI with diffusion generative models. *arXiv preprint arXiv:2306.03284*, 2023.

Rezende, D. J. and Viola, F. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

Rout, L., Chen, Y., Kumar, A., Caramanis, C., Shakkottai, S., and Chu, W.-S. Beyond first-order Tweedie: Solving inverse problems using latent diffusion. *arXiv preprint arXiv:2312.00852*, 2023.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceeding of the 35th International Conference on Machine Learning*, pp. 2256–2265, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *The 9th International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. Pseudoinverse-guided diffusion models for inverse problems. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32*, pp. 11918–11930, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *The 9th International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. In *The 10th International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vaRCHVj0uGI.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. In *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16293–16303, 2022.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion GANs. In *The 10th International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JprM0p-q0Co.

Zhang, K., Gool, L. V., and Timofte, R. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3217–3226, 2020.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zhu, Y., Yang, Y., and Cohen, T. Transformer-based transform coding. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=IDwN6xjHnK8.

Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., and Van Gool, L. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1219–1229, 2023.

A. Proofs

Lemma A.1 (Tweedie's formula)**.** *If the joint distribution between* $\mathbf{x}_0, \mathbf{x}_t$ *is given by* $p_t(\mathbf{x}_0, \mathbf{x}_t) = p(\mathbf{x}_0)p_t(\mathbf{x}_t|\mathbf{x}_0)$ *with* $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$ *, then* $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = \frac{1}{s_t^2\sigma_t^2}(s_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] - \mathbf{x}_t)$.

Proof.

$$\begin{aligned}
\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) &= \frac{\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t)}{p_t(\mathbf{x}_t)} && (26)\\
&= \frac{1}{p_t(\mathbf{x}_t)} \nabla_{\mathbf{x}_t} \int p(\mathbf{x}_0)p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (27)\\
&= \frac{1}{p_t(\mathbf{x}_t)} \int p(\mathbf{x}_0)\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (28)\\
&= \frac{1}{p_t(\mathbf{x}_t)} \int p(\mathbf{x}_0)p_t(\mathbf{x}_t|\mathbf{x}_0)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (29)\\
&= \int p_t(\mathbf{x}_0|\mathbf{x}_t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (30)\\
&= \mathbb{E}_{p_t(\mathbf{x}_0|\mathbf{x}_t)}[\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)] && (31)
\end{aligned}$$

For Gaussian perturbation kernel $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$, we have $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0) = \frac{1}{s_t^2\sigma_t^2}(s_t\mathbf{x}_0 - \mathbf{x}_t)$. Plug it into Eq. (31), we conclude the proof. □

Lemma A.2 (Conditional Tweedie's formula)**.** *If the joint distribution between* $\mathbf{x}_0, \mathbf{y}, \mathbf{x}_t$ *is given by* $p_t(\mathbf{x}_0, \mathbf{y}, \mathbf{x}_t) = p(\mathbf{x}_0)p(\mathbf{y}|\mathbf{x}_0)p_t(\mathbf{x}_t|\mathbf{x}_0)$ *with* $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$ *, then* $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \frac{1}{s_t^2\sigma_t^2}(s_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] - \mathbf{x}_t)$.

Proof.

$$\begin{aligned}
\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) &= \frac{\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t|\mathbf{y})}{p_t(\mathbf{x}_t|\mathbf{y})} && (32)\\
&= \frac{1}{p_t(\mathbf{x}_t|\mathbf{y})} \nabla_{\mathbf{x}_t} \int p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})p(\mathbf{x}_0|\mathbf{y})\mathrm{d}\mathbf{x}_0 && (33)\\
&= \frac{1}{p_t(\mathbf{x}_t|\mathbf{y})} \nabla_{\mathbf{x}_t} \int p_t(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0|\mathbf{y})\mathrm{d}\mathbf{x}_0 && (34)\\
&= \frac{1}{p_t(\mathbf{x}_t|\mathbf{y})} \int p(\mathbf{x}_0|\mathbf{y})\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (35)\\
&= \frac{1}{p_t(\mathbf{x}_t|\mathbf{y})} \int p(\mathbf{x}_0|\mathbf{y})p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (36)\\
&= \int p_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y})\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\mathrm{d}\mathbf{x}_0 && (37)\\
&= \mathbb{E}_{p_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y})}[\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)] && (38)
\end{aligned}$$

where Eq. (34) and Eq. (36) are due to the conditional independence between $\mathbf{x}_t$ and $\mathbf{y}$ given $\mathbf{x}_0$, such that $p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = p_t(\mathbf{x}_t|\mathbf{x}_0)$. For Gaussian perturbation kernel $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$, we have $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0) = \frac{1}{s_t^2\sigma_t^2}(s_t\mathbf{x}_0 - \mathbf{x}_t)$. Plug it into Eq. (38), we conclude the proof. □

A.1. Derivation of the Marginal Preserving Property of Diffusion ODEs

For the sake of completeness, here we prove that the ODEs given in Eq. (2) and Eq. (4) respectively maintain the exact same marginals to $p_t(\mathbf{x}_t)$ and $p_t(\mathbf{x}_t|\mathbf{y})$.

Proof. By borrowing the results from (Equation 4, (Karras et al., 2022)) and setting $s_t = 1, \sigma_t = t$, $\mathbf{x}_t$ determined by the following ODE preserves the marginal $p_t(\mathbf{x}_t)$ for all $t \in [0, T]$:

$$\mathrm{d}\mathbf{x}_t = -t\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T) \tag{39}$$

Using the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ to represent the score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ using Lemma A.1, we recover Eq. (2):

$$\mathrm{d}\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]}{t}\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T) \tag{40}$$

Likewise, the following ODE preserves the marginal $p_t(\mathbf{x}_t|\mathbf{y})$ for all $t \in [0, T]$:

$$\mathrm{d}\mathbf{x}_t = -t\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T|\mathbf{y}) \tag{41}$$

By Lemma A.2, we recover Eq. (4):

$$\mathrm{d}\mathbf{x}_t = \frac{\mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]}{t}\mathrm{d}t, \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T|\mathbf{y}) \tag{42}$$

□

A.2. Proof of Proposition 3.1

To relate $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ to $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$, we note that

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) \tag{43}$$

Using Lemma A.1 and Lemma A.2, we have

$$\frac{1}{s_t^2\sigma_t^2}(s_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] - \mathbf{x}_t) = \frac{1}{s_t^2\sigma_t^2}(s_t\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] - \mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) \tag{44}$$

and consequently,

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] + s_t\sigma_t^2\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) \tag{45}$$

A.3. Proof of Proposition 3.2

When $\sigma \to 0$, finding the minimizer of Eq. (10) is equivalent to solving the following hard-constraint optimization problem:

$$\min_{\hat{\mathbf{x}}_0} \|\hat{\mathbf{x}}_0 - D_t(\mathbf{x}_t)\|_2^2 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\hat{\mathbf{x}}_0 \tag{46}$$

We define the Lagrangian $\mathcal{L}(\hat{\mathbf{x}}_0, \lambda) = \frac{1}{2}\|\hat{\mathbf{x}}_0 - D_t(\mathbf{x}_t)\|_2^2 + \lambda^T(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_0)$, where λ is the Lagrangian multiplier. By the optimality condition, we have

$$\nabla_{\hat{\mathbf{x}}_0}\mathcal{L} = \hat{\mathbf{x}}_0 - D_t(\mathbf{x}_t) - \mathbf{A}^T\lambda = \mathbf{0} \tag{47}$$

$$\nabla_{\lambda}\mathcal{L} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_0 = \mathbf{0} \tag{48}$$

Multiplying $\mathbf{A}$ to Eq. (47) and combining the condition of Eq. (48) gives:

$$\mathbf{A}(\hat{\mathbf{x}}_0 - D_t(\mathbf{x}_t) - \mathbf{A}^T\lambda) = \mathbf{0} \tag{49}$$

$$\Rightarrow \mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t) - \mathbf{A}\mathbf{A}^T\lambda = \mathbf{0} \tag{50}$$

$$\Rightarrow \mathbf{A}\mathbf{A}^T\lambda = \mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t) \tag{51}$$

Multiplying $\mathbf{A}^\dagger$ to Eq. (51) and leveraging the property $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^T = \mathbf{A}^T$, we have

$$\mathbf{A}^T\lambda = \mathbf{A}^\dagger\mathbf{y} - \mathbf{A}^\dagger\mathbf{A}D_t(\mathbf{x}_t) \tag{52}$$

and consequently,

$$\hat{\mathbf{x}}_0 = D_t(\mathbf{x}_t) + \mathbf{A}^\dagger\mathbf{y} - \mathbf{A}^\dagger\mathbf{A}D_t(\mathbf{x}_t) \tag{53}$$

$$= \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})D_t(\mathbf{x}_t) \tag{54}$$

A.4. Proof of Theorem 4.2

We follow (Bao et al., 2022b) to derive the relationship between the optimal reverse variances $\mathbf{v}_t^2(\mathbf{x}_t)$ and optimal posterior variances $\mathbf{r}_t^2(\mathbf{x}_t)$ based on more general non-Markov forward process introduced by (Song et al., 2021a). To find the optimal solution to Eq. (19), we present a much simpler proof than (Bao et al., 2022b) using functional derivatives motivated by (Rezende & Viola, 2018). Given a noise schedule $\{\beta_t\}_{t=1}^T$ and $\alpha_t = 1 - \beta_t$, the forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is defined as

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = q(\mathbf{x}_T|\mathbf{x}_0)\prod_{t=2}^{T} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \tag{55}$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \lambda_t^2\mathbf{I}) \tag{56}$$

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 + \sqrt{\bar{\beta}_{t-1} - \lambda_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{\bar{\beta}_t}} \tag{57}$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ and $\bar{\beta}_t = 1 - \bar{\alpha}_t$. (Song et al., 2021a) show that for arbitrary choice of λ_t, the marginal distributions $q(\mathbf{x}_t|\mathbf{x}_0)$ maintain $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, \bar{\beta}_t\mathbf{I})$. DDPM forward process is a special case when $\lambda_t^2 = \tilde{\beta}_t$ with $\tilde{\beta}_t = \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\beta_t$, which is used in (Nichol & Dhariwal, 2021) for pre-training DDPM model.

To fit the data distribution $q(\mathbf{x}_0)$, we define the reverse process, given by

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p(\mathbf{x}_{t-1}|\mathbf{x}_t), \;\; p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}|\mathbf{m}_t(\mathbf{x}_t), \mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]) \tag{58}$$

To train $p(\mathbf{x}_0)$ we minimize the KL divergence between the forward and the reverse process:

$$\min_p D_{KL}(q(\mathbf{x}_{0:T})||p(\mathbf{x}_{0:T})), \;\; q(\mathbf{x}_{0:T}) = q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0) \tag{59}$$

which is equivalent to minimizing the variational bound $\mathbb{E}_q[L_{\text{vb}}]$ on negative log-likelihood of data distribution $q(\mathbf{x}_0)$ with L_{vb} given as follows:

$$L_{\text{vb}} = L_0 + L_1 + ... + L_T \tag{60}$$

$$L_0 = -\log p(\mathbf{x}_0|\mathbf{x}_1) \tag{61}$$

$$L_{t-1} = D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{x}_t)||p(\mathbf{x}_{t-1}|\mathbf{x}_t)) \tag{62}$$

$$L_T = D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) \tag{63}$$

For $t \in [2, T]$, L_{t-1} are KL divergences between two Gaussians, which possess analytical forms:

$$L_{t-1} \equiv \log \frac{|\mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]|}{|\lambda_t^2\mathbf{I}|} + \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mathbf{m}(\mathbf{x}_t)\|^2_{\mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]^{-1}} + \mathrm{tr}[\lambda_t^2 \mathrm{diag}[\mathbf{v}_t^2(\mathbf{x}_t)]^{-1}] \tag{64}$$

$$\equiv \sum_{i=1}^{d} \log \mathbf{v}_t^2(\mathbf{x}_t)_i + \frac{(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)_i - \mathbf{m}_t(\mathbf{x}_t)_i)^2}{\mathbf{v}_t^2(\mathbf{x}_t)_i} + \frac{\lambda_t^2}{\mathbf{v}_t^2(\mathbf{x}_t)_i} \tag{65}$$

where "$\equiv$" denotes "equals up to a constant and a scaling factor" and i indexes the elements of an vector.

Note that minimizing $\mathbb{E}_q[L_{\text{vb}}]$ can be decomposed into T independent optimization sub-problems:

$$\min_{\mathbf{m}_t, \mathbf{v}_t} \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)}[L_{t-1}], \;\; t \in [1, T] \tag{66}$$

The optimal $\mathbf{m}_t$ and $\mathbf{v}_t$ can be found by taking the functional derivatives of $\mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)}[L_{t-1}]$ w.r.t $\mathbf{m}_t$ and $\mathbf{v}_t^2$ then set to zero:

$$\frac{\delta \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)}[L_{t-1}]}{\delta \mathbf{m}_t(\mathbf{x}_t)_i} \equiv \mathbb{E}_{q(\mathbf{x}_0)}[q(\mathbf{x}_t|\mathbf{x}_0)\frac{\mathbf{m}_t(\mathbf{x}_t)_i - \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)_i}{\mathbf{v}_t^2(\mathbf{x}_t)_i}] = 0 \tag{67}$$

$$\frac{\delta \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t)}[L_{t-1}]}{\delta \mathbf{v}_t^2(\mathbf{x}_t)_i} \equiv \mathbb{E}_{q(\mathbf{x}_0)}[q(\mathbf{x}_t|\mathbf{x}_0)(\frac{1}{\mathbf{v}_t^2(\mathbf{x}_t)_i} - \frac{(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)_i - \mathbf{m}(\mathbf{x}_t)_i)^2}{(\mathbf{v}_t^2(\mathbf{x}_t)_i)^2} - \frac{\lambda_t^2}{(\mathbf{v}_t^2(\mathbf{x}_t)_i)^2})] = 0 \tag{68}$$

We can solve for optimal $\mathbf{m}_t(\mathbf{x}_t)_i$ by rearranging Eq. (67):

$$\mathbb{E}_{q(\mathbf{x}_0)}[q(\mathbf{x}_t|\mathbf{x}_0)]\mathbf{m}_t(\mathbf{x}_t)_i = \mathbb{E}_{q(\mathbf{x}_0)}[q(\mathbf{x}_t|\mathbf{x}_0)\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)_i] \tag{69}$$

$$\Rightarrow q(\mathbf{x}_t)\mathbf{m}_t(\mathbf{x}_t)_i = \int q(\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)_i \mathrm{d}\mathbf{x}_0 \tag{70}$$

$$\Rightarrow \mathbf{m}_t(\mathbf{x}_t)_i = \int q(\mathbf{x}_0|\mathbf{x}_t)\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)_i \mathrm{d}\mathbf{x}_0 \tag{71}$$

$$\Rightarrow \mathbf{m}_t(\mathbf{x}_t)_i = \tilde{\mu}_t(\mathbf{x}_t,\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])_i \tag{72}$$

where Eq. (72) is due to the linearity of $\tilde{\mu}_t$ w.r.t $\mathbf{x}_0$.

Likewise, rearranging Eq. (68) gives

$$\mathbf{v}_t^2(\mathbf{x}_t)_i = \lambda_t^2 + \mathbb{E}_{q(\mathbf{x}_0|\mathbf{x}_t)}[(\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0)_i - \mathbf{m}(\mathbf{x}_t)_i)^2] \tag{73}$$

By plugging the optimal $\mathbf{m}_t(\mathbf{x}_t)_i$ determined by Eq. (72) into Eq. (73) and dropping the element index i, we conclude the proof:

$$\begin{aligned}
\mathbf{v}_t^2(\mathbf{x}_t) &= \lambda_t^2 + \mathbb{E}_{q(\mathbf{x}_0|\mathbf{x}_t)}[(\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) - \tilde{\mu}_t(\mathbf{x}_t,\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]))^2] && (74)\\
&= \lambda_t^2 + \mathbb{E}_{q(\mathbf{x}_0|\mathbf{x}_t)}[(\tilde{\mu}_t(\mathbf{0},\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]))^2] && (75)\\
&= \lambda_t^2 + (\sqrt{\bar{\alpha}_{t-1}} - \sqrt{\bar{\beta}_{t-1} - \lambda_t^2}\sqrt{\frac{\bar{\alpha}_t}{\bar{\beta}_t}})^2 \cdot \mathbb{E}_{q(\mathbf{x}_0|\mathbf{x}_t)}[(\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])^2] && (76)\\
&= \lambda_t^2 + (\sqrt{\bar{\alpha}_{t-1}} - \sqrt{\bar{\beta}_{t-1} - \lambda_t^2}\sqrt{\frac{\bar{\alpha}_t}{\bar{\beta}_t}})^2 \cdot \mathbf{r}_t^2(\mathbf{x}_t) && (77)
\end{aligned}$$

We are often given a pre-trained DDPM model with learned reverse variances. The following Corollary of Theorem 4.2 gives a simplified relationship between optimal posterior variances and optimal reverse variances under DDPM case:

Corollary A.3. *For the DDPM forward process $\lambda_t^2 = \tilde{\beta}_t$, the optimal posterior variances $\mathbf{r}_t^{*2}(\mathbf{x}_t)$ and optimal reverse variances $\mathbf{v}_t^{*2}(\mathbf{x}_t)$ are related by*

$$\mathbf{v}_t^{*2}(\mathbf{x}_t) = \tilde{\beta}_t + (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t})^2 \cdot \mathbf{r}_t^{*2}(\mathbf{x}_t) \tag{78}$$

Remark A.4. According Theorem 4.2, to perform optimal ancestral sampling, we only need to provide the MMSE estimator $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ to compute the reverse mean $\mathbf{m}_t^*(\mathbf{x}_t)$ and the posterior variances $\mathbf{r}_t^{*2}(\mathbf{x}_t)$ to compute the reverse variances $\mathbf{v}_t^{*2}(\mathbf{x}_t)$, which can be both obtained from the proposed pre-training (Eq. (12)). This may provide an alternative way for pre-trainning DDPM model that differs from (Nichol & Dhariwal, 2021).

A.5. Proof of Proposition 4.1

Deriving the optimal solution to Eq. (12) under the diagonal posterior covariance case, i.e., $\Sigma_t(\mathbf{x}_t) = \mathrm{diag}[\mathbf{r}_t^2(\mathbf{x}_t)]$, is similar to Appendix A.4. Note that we seek for point-wise minimizer of Eq. (12), i.e., find the optimum of

$$\mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)}[\log q_t(\mathbf{x}_0|\mathbf{x}_t)] \equiv \mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)}\left[\sum_{i=1}^{d}\frac{1}{\mathbf{r}_t^2(\mathbf{x}_t)_i}(\mathbf{x}_0 - D_t(\mathbf{x}_t))_i^2 + \log \mathbf{r}_t^2(\mathbf{x}_t)_i\right] \tag{79}$$

For any t, taking the functional derivatives of $\mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)}[\log q_t(\mathbf{x}_0|\mathbf{x}_t)]$ w.r.t $D_t(\mathbf{x}_t)_i$ and $\mathbf{r}_t^2(\mathbf{x}_t)_i$ and then set to zero, we obtain the optimality conditions:

$$\mathbb{E}_{p(\mathbf{x}_0)}[p_t(\mathbf{x}_t|\mathbf{x}_0)(D_t(\mathbf{x}_t)_i - \mathbf{x}_{0i})] = 0 \tag{80}$$

$$\mathbb{E}_{p(\mathbf{x}_0)}[p_t(\mathbf{x}_t|\mathbf{x}_0)(-\frac{1}{(\mathbf{r}_t^2(\mathbf{x}_t)_i)^2}(\mathbf{x}_0 - D_t(\mathbf{x}_t))_i^2 + \frac{1}{\mathbf{r}_t^2(\mathbf{x}_t)_i}] = 0 \tag{81}$$

Combining the optimality conditions given by Eq. (80) and Eq. (81), we conclude the proof.

B. Practical Numerical Algorithms

In this section, we discuss practical algorithms for implementing efficient Type I and Type II guidance.

Type I guidance. We are required to obtain the likelihood score, which can be approximated via Jacobian-vector product (JVP) similar to (Song et al., 2023):

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) \approx \frac{\partial D_t(\mathbf{x}_t)}{\partial \mathbf{x}_t} \underbrace{\mathbf{A}^T(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}_{\mathbf{v}}. \tag{82}$$

Type II guidance. We are required to solve the following *auto-weighted* proximal problem:

$$\mathbf{x}_0^{(t)} = \arg\min_{\mathbf{x}_0} \|\mathbf{y} - \mathbf{A}\mathbf{x}_0\|^2 + \sigma^2 \|\mathbf{x}_0 - D_t(\mathbf{x}_t)\|^2_{\Sigma_t^{-1}(\mathbf{x}_t)} \tag{83}$$

which has general closed-form solution given by

$$\mathbf{x}_0^{(t)} = (\Sigma_t(\mathbf{x}_t)^{-1} + \frac{1}{\sigma^2}\mathbf{A}^T\mathbf{A})^{-1}(\Sigma_t(\mathbf{x}_t)^{-1}D_t(\mathbf{x}_t) + \frac{1}{\sigma^2}\mathbf{A}^T\mathbf{y}). \tag{84}$$

The derivation of Eq. (84) can be directly obtained by computing the mean of $q_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y})$ using the Bayes' theorem for Gaussian variables (Bishop, 2006). Leveraging the Woodbury matrix identity, we have an equivalent form[8] for Eq. (84), as follows:

$$\mathbf{x}_0^{(t)} = D_t(\mathbf{x}_t) + \Sigma_t(\mathbf{x}_t) \underbrace{\mathbf{A}^T(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}_{\mathbf{v}}. \tag{85}$$

To evaluate Eq. (82) and Eq. (85) for realizing Type I and Type II guidance, the computational challenge arises from evaluating the high-dimensional matrix inversion for $\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T$, which in general is $\mathcal{O}(d^3)$. To address this, note that Eq. (82) and Eq. (85) can be efficiently evaluated if we have fast matrix-vector multiplies available for $\mathbf{v} = \mathbf{A}^T(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))$. In the subsequent sections, we will discuss two methods based on closed-form solutions and CG for fast computation for $\mathbf{v}$, respectively designed for isotropic covariances and more general covariances.

Remark B.1. The discussions in this section indeed reveal a deeper connection in existing approaches. Suppose $D_t(\mathbf{x}_t) = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$. Then, by Proposition 3.1, Eq. (82), and the second-order Tweedie's formula $\sigma_t^2 \frac{\partial \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]}{\partial \mathbf{x}_t} = \Sigma_t^*(\mathbf{x}_t)$, the conditional posterior mean approximation in Type I guidance can be rewritten as:

$$\mathbf{x}_0^{(t)} = D_t(\mathbf{x}_t) + \Sigma_t^*(\mathbf{x}_t)\mathbf{A}^T(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)). \tag{86}$$

As can be seen, Eq. (86) of Type I and Eq. (85) of Type II can be uniformly considered as an approximation for the mean of the optimal approximated conditional denoising posterior $q_t(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}_0)\mathcal{N}(\mathbf{x}_0|\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t], \Sigma_t^*(\mathbf{x}_t))$, i.e.,

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] + \Sigma_t^*(\mathbf{x}_t)\mathbf{A}^T(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t^*(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]). \tag{87}$$

B.1. Using Closed-form Solutions for Isotropic Posterior Covariance

In this section, we provide efficient closed-form results for computing $\mathbf{v}$ under isotropic posterior covariance: $\Sigma_t(\mathbf{x}_t) = r_t^2(\mathbf{x}_t)\mathbf{I}$. Before we delve into the closed-form results, we first give some important notations. We define the downsampling operator given sampling position $\mathbf{m} \in \{0,1\}^{d\times 1}$ as $\mathbf{D}_{\mathbf{m}} \in \{0,1\}^{\|m\|_0 \times d}$, which selects rows of a given matrix that correspond to one in $\mathbf{m}$ and when performing left multiplication. We use $\mathbf{D}_{\downarrow s}$ to denote the standard s-folds downsampling operator, which is equivalent to $\mathbf{D}_{\mathbf{m}}$ when ones in $\mathbf{m}$ are spaced evenly. For image signal, it selects the upper-left pixel for each distinct $s \times s$ patch (Zhang et al., 2020) when performing left multiplication to the *vectorized* image. We use $\mathbf{D}_{\Downarrow s}$ to denote the distinct block downsampler, i.e., averaging s length d/s distinct blocks of a vector. For image signal, it averaging distinct $d/s \times d/s$ blocks (Zhang et al., 2020) when performing left multiplication to the *vectorized* image. We denote the Fourier transform matrix for d-dimensional signal as $\mathbf{F}$, Fourier transform matrix for d/s-dimensional signal as $\mathbf{F}_{\downarrow s}$, the Fourier transform of a vector $\mathbf{v}$ as $\hat{\mathbf{v}}$, and the complex conjugate of a complex vector $\mathbf{v}$ as $\bar{\mathbf{v}}$. We use $\odot$ to denote element-wise multiplication, and the divisions used below are also element-wise.

[8] Note that this form also share similar mathematical form to noisy version of DDNM, i.e., DDNM$^+$ (Eq. (17), Wang et al. (2023)).

Lemma B.2. *Performing s-fold standard downsampling in spacial domain is equivalent to performing s-fold block downsampling in frequency domain:* $\mathbf{D}_{\Downarrow s} = \mathbf{F}_{\downarrow s}\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}$.

Proof. Considering an arbitrary d-dimensional signal in frequency domain $\hat{\mathbf{x}}[k],\ k = 0, 2, .., d-1$. Multiplying $\mathbf{F}_{\downarrow s}\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}$ to $\hat{\mathbf{x}}$ is equivalent to letting $\hat{\mathbf{x}}$ go through the following linear system and obtain the output $\hat{\mathbf{y}}$:

$$\mathbf{x}[n] = \frac{1}{d}\sum_{k=0}^{d-1}\hat{\mathbf{x}}[k]e^{j\frac{2\pi}{d}kn} \tag{88}$$

$$\mathbf{x}_{\downarrow s}[n] = \mathbf{x}[ns] \tag{89}$$

$$\hat{\mathbf{y}}[k] = \sum_{n=0}^{d/s-1}\mathbf{x}_{\downarrow s}[n]e^{-j\frac{2\pi}{d/s}kn} \tag{90}$$

We now use $\hat{\mathbf{x}}$ to represent $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}}[k] = \sum_{n=0}^{d/s-1}\frac{1}{d}\sum_{k'=0}^{d-1}\hat{\mathbf{x}}[k']e^{j\frac{2\pi}{d}k'ns}e^{-j\frac{2\pi}{d/s}kn} \tag{91}$$

$$= \sum_{n=0}^{d/s-1}\frac{1}{d}e^{-j\frac{2\pi}{d/s}kn}\sum_{k'=0}^{d-1}\hat{\mathbf{x}}[k']e^{j\frac{2\pi}{d/s}k'n} \tag{92}$$

$$= \sum_{n=0}^{d/s-1}\frac{1}{d}e^{-j\frac{2\pi}{d/s}kn}(\sum_{k'=0}^{d/s-1}+\sum_{k'=d/s}^{2d/s-1}+\sum_{k'=2d/s}^{3d/s-1}+...+\sum_{k'=(s-1)d/s}^{sd/s-1})\hat{\mathbf{x}}[k']e^{j\frac{2\pi}{d/s}k'n} \tag{93}$$

$$= \sum_{n=0}^{d/s-1}\frac{1}{d}e^{-j\frac{2\pi}{d/s}kn}\sum_{k'=0}^{d/s-1}(\hat{\mathbf{x}}[k'] + \hat{\mathbf{x}}[k'+d/s] + ... + \hat{\mathbf{x}}[k'+(s-1)d/s])e^{j\frac{2\pi}{d/s}k'n} \tag{94}$$

$$= \sum_{n=0}^{d/s-1}e^{-j\frac{2\pi}{d/s}kn}\frac{1}{d/s}\sum_{k'=0}^{d/s-1}\frac{\hat{\mathbf{x}}[k'] + \hat{\mathbf{x}}[k'+d/s] + ... + \hat{\mathbf{x}}[k'+(s-1)d/s]}{s}e^{j\frac{2\pi}{d/s}k'n} \tag{95}$$

$$= \frac{\hat{\mathbf{x}}[k] + \hat{\mathbf{x}}[k+d/s] + ... + \hat{\mathbf{x}}[k+(s-1)d/s]}{s} \tag{96}$$

where Eq. (94) is because $e^{j\frac{2\pi}{d/s}k'n}$ has period of d/s in k' and Eq. (96) is because d/s-dimensional inverse Fourier transform and Fourier transform are canceled out. From Eq. (96) we have $\hat{\mathbf{y}} = \mathbf{D}_{\Downarrow s}\hat{\mathbf{x}}$. So $\mathbf{D}_{\Downarrow s}\hat{\mathbf{x}} = \mathbf{F}_{\downarrow s}\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}\hat{\mathbf{x}}$ for any $\hat{\mathbf{x}}$, and consequently, $\mathbf{D}_{\Downarrow s} = \mathbf{F}_{\downarrow s}\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}$. □

Inpainting. The observation model for image inpainting can be expressed as:

$$\mathbf{y} = \underbrace{\mathbf{D}_{\mathbf{m}}}_{\mathbf{A}}\mathbf{x}_0 + \mathbf{n}, \tag{97}$$

and the closed-form solution to $\mathbf{v}$ in image inpainting is given by the following:

$$\mathbf{v} = \frac{\tilde{\mathbf{y}} - \mathbf{m}\odot D_t(\mathbf{x}_t)}{\sigma^2 + r_t^2(\mathbf{x}_t)} \tag{98}$$

where $\tilde{\mathbf{y}} = \mathbf{D}_{\mathbf{m}}^T\mathbf{y} = \mathbf{m}\odot(\mathbf{x}_0 + \tilde{\mathbf{n}})$, $\tilde{\mathbf{n}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the zero-filling measurements that fills the masked region with zeros and processes the exact same size to $\mathbf{x}_0$. In practice, the measurements in inpainting are usually stored in the form of $\tilde{\mathbf{y}}$, while $\mathbf{y}$ used here is for mathematical convenience.

Proof.

$$\mathbf{v} = \mathbf{D}_{\mathbf{m}}^T(\sigma^2\mathbf{I} + \mathbf{D}_{\mathbf{m}} r_t^2(\mathbf{x}_t)\mathbf{I}\mathbf{D}_{\mathbf{m}}^T)^{-1}(\mathbf{y} - \mathbf{D}_{\mathbf{m}} D_t(\mathbf{x}_t)) \tag{99}$$

$$= \mathbf{D}_{\mathbf{m}}^T(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\mathbf{D}_{\mathbf{m}}\mathbf{D}_{\mathbf{m}}^T)^{-1}(\mathbf{y} - \mathbf{D}_{\mathbf{m}} D_t(\mathbf{x}_t)) \tag{100}$$

$$= \mathbf{D}_{\mathbf{m}}^T((\sigma^2 + r_t^2(\mathbf{x}_t))\mathbf{I})^{-1}(\mathbf{y} - \mathbf{D}_{\mathbf{m}} D_t(\mathbf{x}_t)) \tag{101}$$

$$= \frac{\mathbf{D}_{\mathbf{m}}^T(\mathbf{y} - \mathbf{D}_{\mathbf{m}} D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)} \tag{102}$$

$$= \frac{\tilde{\mathbf{y}} - \mathbf{m} \odot D_t(\mathbf{x}_t)}{\sigma^2 + r_t^2(\mathbf{x}_t)} \tag{103}$$

where Eq. (101) and Eq. (103) are because $\mathbf{D}_{\mathbf{m}}\mathbf{D}_{\mathbf{m}}^T = \mathbf{I}$ and $\mathbf{D}_{\mathbf{m}}^T\mathbf{D}_{\mathbf{m}} = \mathrm{diag}[\mathbf{m}]$. □

Debluring. The observation model for image debluring can be expressed as:

$$\mathbf{y} = \mathbf{x}_0 * \mathbf{k} + \mathbf{n}, \tag{104}$$

where $\mathbf{k}$ is the blurring kernel and $*$ is convolution operator. By assuming $*$ is a circular convolution operator, we can convert Eq. (104) to the canonical form $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}$ by leveraging the convolution property of Fourier transform:

$$\mathbf{y} = \underbrace{\mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]\mathbf{F}}_{\mathbf{A}}\mathbf{x}_0 + \mathbf{n}, \tag{105}$$

and the closed-form solution to $\mathbf{v}$ in image debluring is given by the following:

$$\mathbf{v} = \mathbf{F}^{-1}(\bar{\hat{\mathbf{k}}} \odot \frac{\mathbf{F}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)\bar{\hat{\mathbf{k}}} \odot \hat{\mathbf{k}}}). \tag{106}$$

Proof. Since $\mathbf{A}$ is a real matrix, we have $\mathbf{A}^T = \mathbf{A}^H = \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}$.

$$\mathbf{v} = \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}(\sigma^2\mathbf{I} + \mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]\mathbf{F}r_t^2(\mathbf{x}_t)\mathbf{I}\mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{107}$$

$$= \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{108}$$

$$= \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]\mathbf{F})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{109}$$

$$= \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}(\mathbf{F}^{-1}(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\mathrm{diag}[\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}])\mathbf{F})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{110}$$

$$= \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}(\mathbf{F}^{-1}\mathrm{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]\mathbf{F})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{111}$$

$$= \mathbf{F}^{-1}\mathrm{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{F}^{-1}\mathrm{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]^{-1}\mathbf{F}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{112}$$

$$= \mathbf{F}^{-1}(\bar{\hat{\mathbf{k}}} \odot \frac{\mathbf{F}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}}) \tag{113}$$

□

Super resolution. According to (Zhang et al., 2020), the observation model for image super resolution can be *approximately* expressed as:

$$\mathbf{y} = (\mathbf{x}_0 * \mathbf{k})_{\downarrow s} + \mathbf{n} \tag{114}$$

By leveraging the convolution property of Fourier transform, we can convert Eq. (114) to the canonical form $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}$:

$$\mathbf{y} = \underbrace{\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]\mathbf{F}}_{\mathbf{A}}\mathbf{x}_0 + \mathbf{n} \tag{115}$$

and the closed-form solution to $\mathbf{v}$ in image super resolution is given by the following:

$$\mathbf{v} = \mathbf{F}^{-1}(\bar{\hat{\mathbf{k}}} \odot_s \frac{\mathbf{F}_{\downarrow s}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)(\bar{\hat{\mathbf{k}}} \odot \hat{\mathbf{k}})_{\Downarrow s}}) \tag{116}$$

where $\odot_s$ denotes block processing operator with element-wise multiplication (Zhang et al., 2020).

Proof. Since $\mathbf{A}$ is a real matrix, we have $\mathbf{A}^T = \mathbf{A}^H = \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T$, then

$$\mathbf{v} = \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T(\sigma^2\mathbf{I} + \mathbf{D}_{\downarrow s}\mathbf{F}^{-1}\text{diag}[\hat{\mathbf{k}}]\mathbf{F}r_t^2(\mathbf{x}_t)\mathbf{I}\mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{117}$$

$$= \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}\text{diag}[\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{118}$$

$$= \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T(\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}(\sigma^2\mathbf{I} + r_t^2(\mathbf{x}_t)\text{diag}[\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}])\mathbf{F}\mathbf{D}_{\downarrow s}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{119}$$

$$= \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T(\mathbf{D}_{\downarrow s}\mathbf{F}^{-1}\text{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]\mathbf{F}\mathbf{D}_{\downarrow s}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{120}$$

By Lemma B.2, we have $\mathbf{D}_{\downarrow s}\mathbf{F}^{-1} = \mathbf{F}_{\downarrow s}^{-1}\mathbf{D}_{\Downarrow s}$. Taking the hermitian transpose to both side and leveraging $\mathbf{F}^{-1} = \frac{1}{d}\mathbf{F}^H$ and $\mathbf{F}_{\downarrow s}^{-1} = \frac{1}{d/s}\mathbf{F}_{\downarrow s}^H$, we also have $\mathbf{F}\mathbf{D}_{\downarrow s}^T = s\mathbf{D}_{\Downarrow s}^T\mathbf{F}_{\downarrow s}$. So

$$\mathbf{v} = \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]s\mathbf{D}_{\Downarrow s}^T\mathbf{F}_{\downarrow s}(\mathbf{F}_{\downarrow s}^{-1}\mathbf{D}_{\Downarrow s}\text{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}}]s\mathbf{D}_{\Downarrow s}^T\mathbf{F}_{\downarrow s})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{121}$$

$$= \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]s\mathbf{D}_{\Downarrow s}^T\mathbf{F}_{\downarrow s}(\mathbf{F}_{\downarrow s}^{-1}\text{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)(\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}})_{\Downarrow s}]\mathbf{F}_{\downarrow s})^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{122}$$

$$= \mathbf{F}^{-1}\text{diag}[\bar{\hat{\mathbf{k}}}]s\mathbf{D}_{\Downarrow s}^T\mathbf{F}_{\downarrow s}\mathbf{F}_{\downarrow s}^{-1}\text{diag}[\sigma^2 + r_t^2(\mathbf{x}_t)(\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}})_{\Downarrow s}]^{-1}\mathbf{F}_{\downarrow s}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)) \tag{123}$$

$$= \mathbf{F}^{-1}(\bar{\hat{\mathbf{k}}} \odot s\mathbf{D}_{\Downarrow s}^T\frac{\mathbf{F}_{\downarrow s}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)(\hat{\mathbf{k}} \odot \bar{\hat{\mathbf{k}}})_{\Downarrow s}}) \tag{124}$$

$$= \mathbf{F}^{-1}(\bar{\hat{\mathbf{k}}} \odot_s \frac{\mathbf{F}_{\downarrow_s}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t))}{\sigma^2 + r_t^2(\mathbf{x}_t)(\bar{\hat{\mathbf{k}}} \odot \hat{\mathbf{k}})_{\Downarrow_s}}) \tag{125}$$

□

B.2. Using Conjugate Gradient Method for General Posterior Covariance

For general posterior covariance, the closed-form solution for $\mathbf{v}$ is usually unavailable. However, we can still compute $\mathbf{v}$ efficiently by using numerical solutions of linear equations. Specifically, we first represent $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{A}^T\mathbf{u}, \;\; \mathbf{u} = (\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)^{-1}(\mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t)). \tag{126}$$

Note that $\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T$ is symmetric and positive-definite, and $\mathbf{u}$ can be represented as the solution of the following linear equations:

$$(\sigma^2\mathbf{I} + \mathbf{A}\Sigma_t(\mathbf{x}_t)\mathbf{A}^T)\mathbf{u} = \mathbf{y} - \mathbf{A}D_t(\mathbf{x}_t). \tag{127}$$

Therefore, $\mathbf{u}$ can be computed with acceptable precision using a sufficient number of CG iterates. In the experiments, we use the black box CG method implemented in `scipy.sparse.linalg.cg` with `tol=1e-4` for Convert (Section 4.1), DWT-Var (Section 4.3) and TMPD covariance. The other corvariances are isotropic, and therefore we use the closed-form solutions derived in Section B.1.

C. Additional Experimental Details and Results

C.1. Additional Quantitative Results

In this section, we report the SSIM (Figure 6) and FID (Figure 7) performance for supplementary results of Type II guidance

C.2. Training Objective for Latent Variance

In general, the sub-objective of Eq. (12) at time step t is equivalent to the following MLE objective:

$$\min_{D_t, \Sigma_t} \mathbb{E}_{p_t(\mathbf{x}_0, \mathbf{x}_t)} \left[(\mathbf{x}_0 - D_t(\mathbf{x}_t))^T\Sigma_t(\mathbf{x}_t)^{-1}(\mathbf{x}_0 - D_t(\mathbf{x}_t)) + \log\det\Sigma_t(\mathbf{x}_t)\right] \tag{128}$$

For general covariance Σ_t, evaluating Eq. (128) is hard due to the present of the inverse and log-determinant. Nevertheless, when parameterize Σ_t using fix orthonormal basis $\mathbf{\Psi}$ as $\Sigma_t(\mathbf{x}_t) = \mathbf{\Psi}\text{diag}[\mathbf{r}_t^2(\mathbf{x}_t)]\mathbf{\Psi}^T$, the inverse and log-determinant

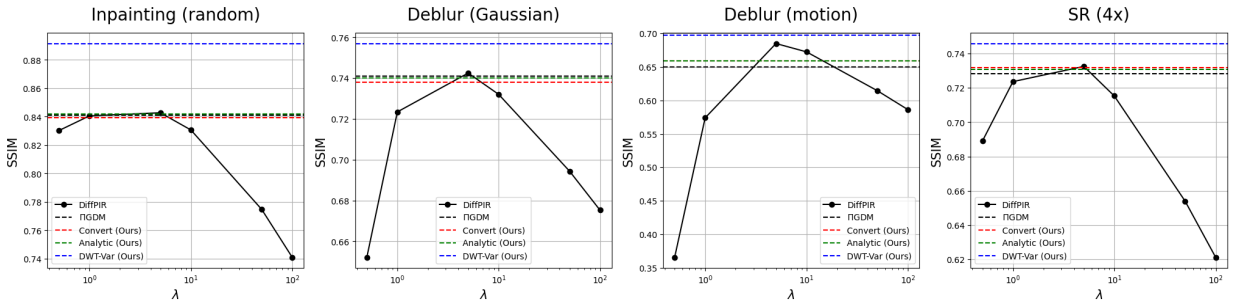

Figure 6: **SSIM comparisons on FFHQ for Type II guidance.** For DiffPIR, we report SSIM under different λ.

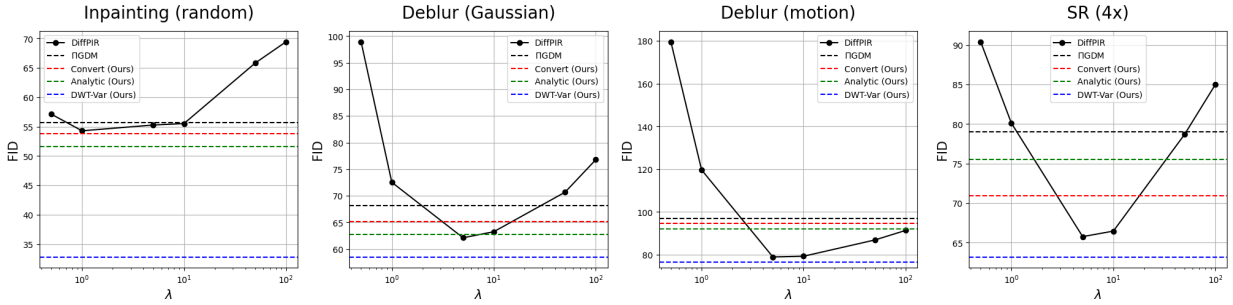

Figure 7: **FID comparisons on FFHQ for Type II guidance.** For DiffPIR, we report FID under different λ.

possess efficient forms as

$$\Sigma_t(\mathbf{x}_t)^{-1} = \mathbf{\Psi}\mathrm{diag}[\mathbf{r}_t^{-2}(\mathbf{x}_t)]\mathbf{\Psi}^T, \tag{129}$$

$$\log\det\Sigma_t(\mathbf{x}_t) = \sum_{i=1}^{d}\log\mathbf{r}_t^2(\mathbf{x}_t)_i. \tag{130}$$

And the training objective of Eq. (128) becomes mirrored to learning diagonal Gaussian posterior (Eq. (79)), but in the transform domain:

$$\mathbb{E}_{p_t(\mathbf{x}_0,\mathbf{x}_t)}\left[\sum_{i=1}^{d}\frac{1}{\mathbf{r}_t^2(\mathbf{x}_t)_i}(\mathbf{\Psi}^T\mathbf{x}_0 - \mathbf{\Psi}^T D_t(\mathbf{x}_t))_i^2 + \log\mathbf{r}_t^2(\mathbf{x}_t)_i\right] \tag{131}$$

We modify unconditional diffusion models of FFHQ dataset from (Chung et al., 2023a) with nine output channels for predicting three values: 1) posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$, 2) $\mathbf{r}_t^2(\mathbf{x}_t)$ when $\mathbf{\Psi}$ is set to identity matrix, referred to as spatial variance, and 3) $\mathbf{r}_t^2(\mathbf{x}_t)$ when $\mathbf{\Psi}$ is set to DWT basis, referred to as DWT variance. Example model predictions see Figure 8. As can be seen, similar to spatial variance, the DWT variance can act as a reliable predictor for the square errors of the DWT coefficients. In fact, it can be shown that the DWT variance is the MMSE estimator of the square errors of DWT coefficients.

C.3. DDNM as Noiseless DiffPIR

To validate Proposition 3.2, we re-implement DDNM under DiffPIR codebase. DiffPIR deals with noiseless inverse problems by setting σ to a relatively low value for ρ_t in Eq. (9) (*e.g.*, 0.001 in DiffPIR codebase). We demonstrate that directly using the DDNM solution in Eq. (11) produces similar results in the noiseless case. Table 4 reports the Mean Absolute Difference (MAD)[9] between the conditional posterior means in equations 9 and 11 by averaging over all the sampling steps and test images. Overall, the MAD is negligible in comparison to the data range of $[-1, 1]$. Debluring and super resolution yield larger MAD than inpainting, since several approximations are made for computing $\mathbf{A}^\dagger$, while for inpainting $\mathbf{A}^\dagger$ is exact. Details are given as follows:.

[9]MAD between $\mathbf{x}$ and $\mathbf{y}$ is defined by $\|\mathbf{x}-\mathbf{y}\|_1/d$.

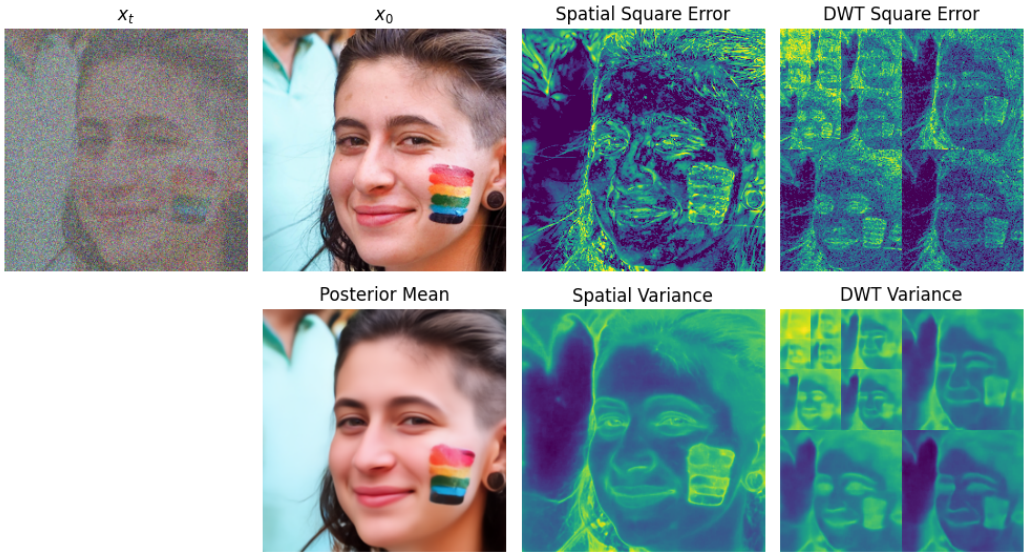

Figure 8: **Ground truths and model predictions.** σ_t is set to 1; Spatial Square Error refers to $(\mathbf{x}_0 - D_t(\mathbf{x}_t))^2$, and DWT Square Error refers to $(\mathbf{\Psi}^T\mathbf{x}_0 - \mathbf{\Psi}^T D_t(\mathbf{x}_t))^2$, where $\mathbf{\Psi}$ is the DWT basis; As can be seen, Spatial Square Error and DWT Square Error can be predicted by Spatial Variance and DWT Variance, respectively.

Task	Avg. MAD
Inpaint (Random)	1.6975×10^{-6}
Deblur (Gaussian)	6.4629×10^{-4}
Deblur (Motion)	1.5461×10^{-3}
Super resolution (4×)	1.1937×10^{-2}

Table 4: **DDNM v.s. DiffPIR in noiseless inverse problems.** We report the averaged MAD between their $\mathbf{x}_0^{(t)}$ averaged over all sampling steps and test images.

Inpainting. We construct the DDNM solution $\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}]$ for the inpainting case $\mathbf{A} = \mathbf{D_m}$ as follows:

$$\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] = \tilde{\mathbf{y}} + (1 - \mathbf{m}) \odot D_t(\mathbf{x}_t) \tag{132}$$

Proof. Since $\mathbf{D_m}\mathbf{D_m}^T = \mathbf{I}$ is non-singular, we can directly obtain $\mathbf{A}^\dagger$ by $\mathbf{A}^\dagger = \mathbf{D_m}^T(\mathbf{D_m}\mathbf{D_m}^T)^{-1} = \mathbf{D_m}^T$. Plug in Eq. (11), we have

$$\begin{aligned}
\mathbb{E}_q[\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}] &= \mathbf{D_m}^T\mathbf{y} + (\mathbf{I} - \mathbf{D_m}^T\mathbf{D_m})D_t(\mathbf{x}_t) && (133)\\
&= \tilde{\mathbf{y}} + (\mathbf{I} - \mathrm{diag}[\mathbf{m}])D_t(\mathbf{x}_t) && (134)\\
&= \tilde{\mathbf{y}} + \mathrm{diag}[1 - \mathbf{m}]D_t(\mathbf{x}_t) && (135)\\
&= \tilde{\mathbf{y}} + (1 - \mathbf{m}) \odot D_t(\mathbf{x}_t) && (136)
\end{aligned}$$

□

Debluring. We construct the pseudo inverse $\mathbf{A}^\dagger$ for the linear operator in the debluring case $\mathbf{A} = \mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]\mathbf{F}$ as follows

$$\mathbf{A}^\dagger = \mathbf{F}^{-1}\mathrm{diag}[\hat{\mathbf{k}}]^\dagger\mathbf{F} \tag{137}$$

where $\mathrm{diag}[\hat{\mathbf{k}}]^\dagger$ is defined as $\mathrm{diag}[\hat{\mathbf{k}}]^\dagger = \mathrm{diag}[[l_1, l_2, ...]^T]$ with $l_i = 0$ if $\hat{\mathbf{k}}_i = 0$[10] otherwise $l_i = 1/\hat{\mathbf{k}}_i$.

[10]Numerically, $|\hat{\mathbf{k}}_i|$ is always larger than zero. We threshold $\hat{\mathbf{k}}_i$ to zero when $|\hat{\mathbf{k}}_i| < 3 \times 10^{-2}$ similar to (Wang et al., 2023)

Proof. It is easy to see that with the above construction, $\mathbf{A}^\dagger$ satisfies $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$. □

Super resolution. We directly use `torch.nn.functional.interpolate` in place of $\mathbf{A}^\dagger$ for the super resolution case.

C.4. Converting Optimal Solutions Between Different Perturbation Kernels

Suppose we are given a family of optimal solutions q_t defined by the following MLE objectives for all $t \in [0, T]$:

$$\max_{q_t} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log q_t(\mathbf{x}_0 | \mathbf{x}_t = s_t(\mathbf{x}_0 + \sigma_t \epsilon)) \tag{138}$$

As can be seen, q_t equals the optimal solutions to Eq. (12) when the perturbation kernels $p_t(\mathbf{x}_t|\mathbf{x}_0)$ are set to $\mathcal{N}(\mathbf{x}_t | s_t\mathbf{x}_0, s_t^2\sigma_t^2\mathbf{I})$. Now, suppose we want to perform sampling based on the diffusion ODE (SDE) under the perturbation kernels $\mathcal{N}(\mathbf{x}_t | \tilde{s}_t\mathbf{x}_0, \tilde{s}_t^2\tilde{\sigma}_t^2\mathbf{I})$. This means that we are required to provide optimal solutions $\tilde{q}_t$ defined by the following objectives for all $t \in [0, T]$:

$$\max_{\tilde{q}_t} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log \tilde{q}_t(\mathbf{x}_0 | \mathbf{x}_t = \tilde{s}_t(\mathbf{x}_0 + \tilde{\sigma}_t \epsilon)) \tag{139}$$

The idea is that, $\tilde{q}_t$ can be directly represented by q_t, so we do not require to perform re-training:

$$\tilde{q}_t(\mathbf{x}_0 | \mathbf{x}_t = \mathbf{x}) = q_{t'}(\mathbf{x}_0 | \mathbf{x}_{t'} = \frac{s_{t'}}{\tilde{s}_t}\mathbf{x}), \quad \sigma_{t'} = \tilde{\sigma}_t \tag{140}$$

Proof. We only need to show that $\tilde{q}_t$ defined by Eq. (138) and Eq. (140) are equivalent to $\tilde{q}_t$ defined by Eq. (139). From Eq. (140), we actually know that $q_{t'}(\mathbf{x}_0|\mathbf{x}_{t'} = \mathbf{x}) = \tilde{q}_t(\mathbf{x}_0|\mathbf{x}_t = \frac{\tilde{s}_t}{s_{t'}}\mathbf{x})$, plug it in Eq. (138) at t' we have

$$\max_{\tilde{q}_t} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log \tilde{q}_t(\mathbf{x}_0 | \mathbf{x}_t = \tilde{s}_t(\mathbf{x}_0 + \sigma_{t'} \epsilon)), \quad \sigma_{t'} = \tilde{\sigma}_t \tag{141}$$

which is equivalent to Eq. (139). □

For example, suppose we are given optimal solutions q_t under DDPM perturbation kernels, i.e., $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \bar{\beta}_t\mathbf{I})$. We aim to convert these solutions to optimal solutions $\tilde{q}_t$ under perturbation kernels used in Section 2.2, i.e., $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_0, t^2\mathbf{I})$. We can realize it using following two steps: (1) finding t' such that $\sqrt{\frac{\bar{\beta}_{t'}}{\bar{\alpha}_{t'}}} = t$, and (2) scaling the input of $q_{t'}$ by the factor of $\sqrt{\bar{\alpha}_{t'}}$. Formally,

$$\tilde{q}_t(\mathbf{x}_0|\mathbf{x}_t = \mathbf{x}) = q_{t'}(\mathbf{x}_0|\mathbf{x}_{t'} = \sqrt{\bar{\alpha}_{t'}}\mathbf{x}), \quad \sqrt{\frac{\bar{\beta}_{t'}}{\bar{\alpha}_{t'}}} = t \tag{142}$$

C.5. Qualitative Results

In this section, we present additional visual examples for qualitative comparisons.

Figure 9: **Qualitative results for Table 3 on random inpainting.**

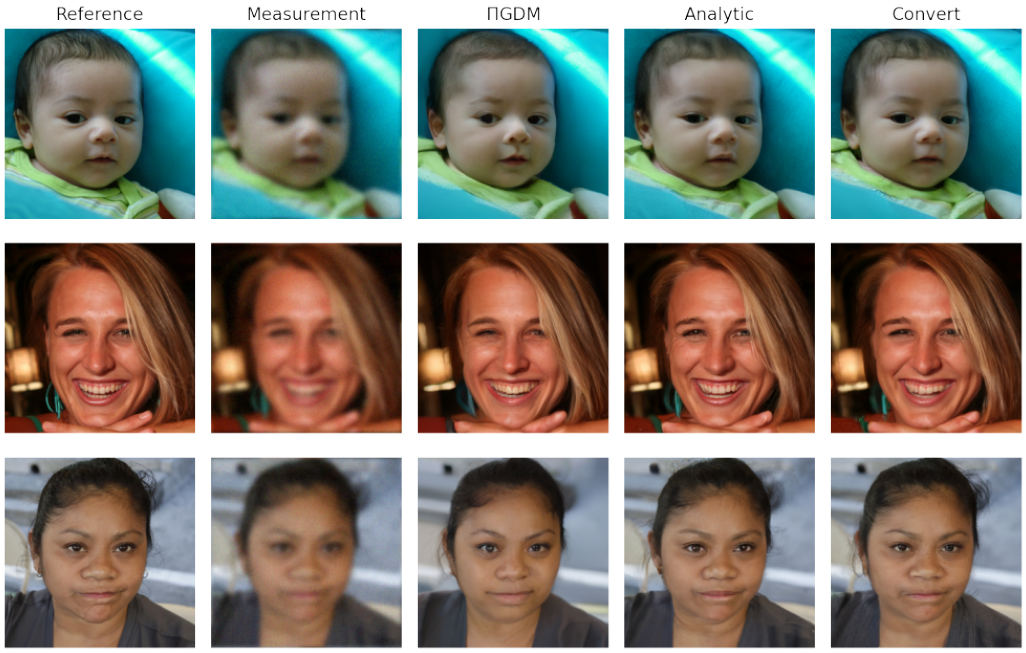

Figure 10: **Qualitative results for Table 3 on Gaussian debluring.**

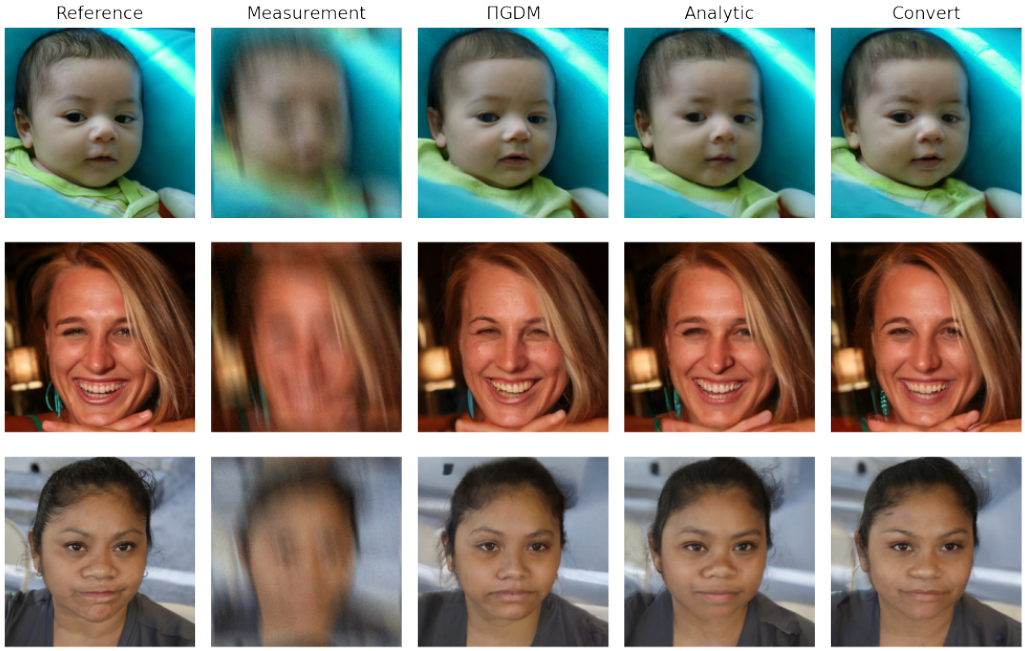

Figure 11: **Qualitative results for Table 3 on motion debluring.**

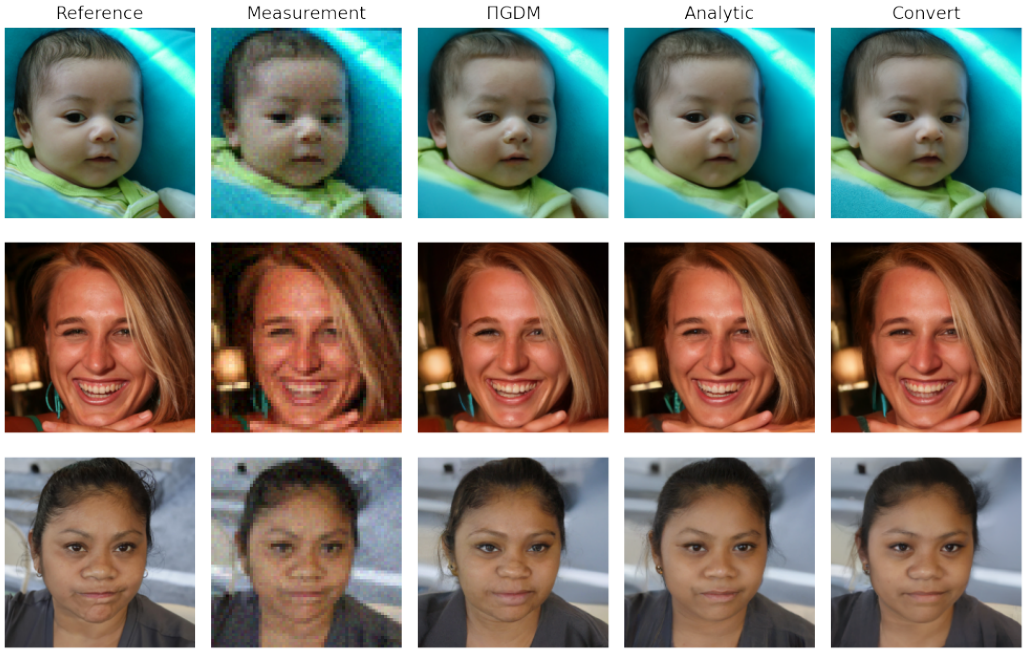

Figure 12: **Qualitative results for Table 3 on super resolution.**