Frequency-Guided Posterior Sampling for Diffusion-Based Image Restoration

Supplementary Material

A. Further Background

A.1. Diffusion Models

Recall that unconditional diffusion models learn the score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x})$ through denoising score matching. Assuming we have learned a neural network that approximates this score well, the backbone of most state of the art diffusion-based inverse problem solvers is the estimation of $\boldsymbol{\mu}_{0|t} = \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t]$ using the learned diffusion model. This is known as Tweedie's formula, which we state below.

Lemma 4. *(Tweedie's formula [17]) Suppose* $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$. *Then the posterior mean is*

$$\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t + (1-\alpha_t)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)) \tag{16}$$

Proof. We expand the score function as

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) = \frac{\nabla_{\mathbf{x}_t} p(\mathbf{x}_t)}{p(\mathbf{x}_t)} \tag{17}$$

$$= \frac{1}{p(\mathbf{x}_t)} \int \nabla_{\mathbf{x}_t} p(\mathbf{x}_t \mid \mathbf{x}_0) p(\mathbf{x}_0)\, d\mathbf{x}_0. \tag{18}$$

We can rewrite $\nabla_{\mathbf{x}_t} p(\mathbf{x}_t \mid \mathbf{x}_0)$ as $p(\mathbf{x}_t \mid \mathbf{x}_0)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{x}_0)$ and group $\frac{p(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}{p(\mathbf{x}_t)} = p(\mathbf{x}_0 \mid \mathbf{x}_t)$ to give

$$= \int p(\mathbf{x}_0 \mid \mathbf{x}_t) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{x}_0)\, d\mathbf{x}_0 \tag{19}$$

$$= \int p(\mathbf{x}_0 \mid \mathbf{x}_t) \frac{\sqrt{\alpha_t}\mathbf{x}_0 - \mathbf{x}_t}{1-\alpha_t}\, d\mathbf{x}_0 \tag{20}$$

$$= \frac{\sqrt{\alpha_t}\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] - \mathbf{x}_t}{1-\alpha_t}. \tag{21}$$

Rearranging, we have our final result that gives the posterior mean as a function of the unconditional score function. □

A.2. Signal Processing

The Fourier transform is a fundamental tool in signal processing that decomposes a signal into its constituent frequencies. For a continuous signal $\mathbf{x}(t)$, the Fourier transform is given by

$$\mathcal{F}_{\text{cont}}(\mathbf{x})[f] = \int_{-\infty}^{\infty} \mathbf{x}(t) e^{-2\pi i f t} dt, \tag{22}$$

where i denotes the complex root of unity. In discrete domains, such as digital images, we work with the Discrete Fourier Transform (DFT). For a signal $\mathbf{x} \in \mathbb{R}^n$, we denote its DFT as $\mathcal{F}(\mathbf{x})$, where $\mathcal{F}(\mathbf{x})[f_k]$ represents the frequency component at the k-th frequency

$$\mathcal{F}(\mathbf{x})[f_k] = \sum_{n=0}^{N-1} \mathbf{x}[n] e^{-2\pi i k n/N}. \tag{23}$$

The DFT can be expressed as a matrix operation $\mathbf{F} \in \mathbb{C}^{n \times n}$ where

$$\mathbf{F}_{jk} = \frac{1}{\sqrt{n}} e^{-2\pi i j k/n}. \tag{24}$$

This matrix $\mathbf{F}$ is unitary. A discrete convolution operation $\mathbf{x} \circledast \mathbf{h}$ can be represented as a matrix multiplication $\mathbf{C}_h\mathbf{x}$, where $\mathbf{C}_h$ is a circulant matrix constructed from the filter $\mathbf{h}$. A circulant matrix has the special property that each row is a cyclic shift of the previous row:

$$\mathbf{C}_h = \begin{bmatrix} h_0 & h_{n-1} & \cdots & h_1 \\ h_1 & h_0 & \cdots & h_2 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n-1} & h_{n-2} & \cdots & h_0 \end{bmatrix}. \tag{25}$$

A fundamental property of circulant matrices is that they can be diagonalized by the DFT matrix

$$\mathbf{C}_h = \mathbf{F}^* \text{diag}(\mathcal{F}(\mathbf{h}))\mathbf{F}. \tag{26}$$

This relationship explains why convolution in the spatial domain equals pointwise multiplication in the frequency domain

$$\mathcal{F}(\mathbf{x} \circledast \mathbf{h}) = \mathcal{F}(\mathbf{x}) \odot \mathcal{F}(\mathbf{h}). \tag{27}$$

Natural images typically exhibit a power law relationship in their frequency spectrum such that

$$|\mathcal{F}(\mathbf{x})[f_k]|^2 \propto \frac{1}{|f_k|^\alpha}, \tag{28}$$

where α is typically around 2. This relationship, often called the $1/f^2$ law, arises from the fundamental structure of natural scenes:

- Natural images tend to be locally smooth with occasional sharp transitions (edges)
- Objects in natural scenes exhibit self-similarity across scales
- Natural scenes contain hierarchical structures from fine to coarse details

This power law relationship provides a strong prior for image processing tasks, as it captures the statistical regularities present in natural images. The decay of frequency components according to this law explains why natural images are compressible and why high-frequency noise is particularly noticeable in image data.

B. Proofs for Section 5

We first prove Lemma 1, restated below.

Lemma 5. *Let $f_k = \frac{k}{n}$ for $k = 0, \dots, n-1$ denote the DFT sample frequencies for a signal of length n. There exists a covariance matrix $\mathbf{\Sigma}$ such that for $x \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, the following two properties hold. First, the signal $\mathbf{x}$ follows a power law in the frequency domain with parameters $c, \beta > 0$ i.e. it has a power spectral density $S(f_k) = c|f_k|^{-\beta}$ for the non-zero DFT sample frequencies f_k. Second, the eigenvalues of $\mathbf{\Sigma}$ are precisely $c|f_k|^{-\beta}$.*

Proof. We will first provide a construction for $\mathbf{\Sigma}$. Let $\mathbf{R}$ be a n-length signal that is the inverse Discrete Fourier Transform of $S(f_k) = c|f_k|^{-\beta}$ for the non-zero DFT sample frequencies f_k. We construct $\mathbf{\Sigma}$ as a circulant matrix whose first row (and column) is $\mathbf{R}$.

Next, we show the two properties of $\mathbf{\Sigma}$ needed to prove the lemma. First, let $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$. Then, $\mathbf{x}$ can be viewed as a finite stochastic process that is zero-mean, wide-sense, and stationary. It is zero-mean trivially because $\mathbf{x}$ is sampled from a zero-mean distribution. It is wide-sense stationary because $\mathbf{\Sigma}$ is circulant. Thus, the autocovariance function, which is exactly $\mathbf{R}$, depends only on the gap between two elements in the signal. From the discrete-time Wiener-Khinchin theorem, we have that

$$\mathbb{E}\left[\frac{|\mathcal{F}(\mathbf{x})[f_k]|^2}{n}\right] = \mathcal{F}(\mathbf{R})[f_k] \tag{29}$$

By construction, we have that $\mathcal{F}(\mathbf{R})[f_k] = S(f_k)$. This shows that in expectation, $\mathbf{x}$ follows a power spectral density of $S(f_k)$. Lastly, because $\mathbf{\Sigma}$ is circulant, we have that the eigenvalues of $\mathbf{\Sigma}$ are the Discrete Fourier Transform of the first row, which is $\mathbf{R}$. As before, by construction, we have that $\mathcal{F}(\mathbf{R})[f_k] = S(f_k)$. From Equation (29), we have that the eigenvalues of $\mathbf{\Sigma}$ are precisely $c|f_k|^{-\beta}$. □

Before we prove our main result, Theorem 2, we prove a useful lemma that calculates the posterior denoising distribution under a multivariate Gaussian assumption on the data.

Lemma 6. *(Posterior Denoising Distribution) Suppose $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Suppose $\mathbf{x}_t \mid \mathbf{x}_0 \sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$. Then,*

$$p(\mathbf{x}_0 \mid \mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_{0|t}, \mathbf{\Sigma}_{0|t}) \tag{30}$$

where $\boldsymbol{\mu}_{0|t} = \mathbf{\Gamma}_t \mathbf{x}_t$, $\mathbf{\Gamma}_t = \sqrt{\alpha_t}\mathbf{\Sigma}(\alpha_t\mathbf{\Sigma} + (1-\alpha_t)\mathbf{I})^{-1}$, and $\mathbf{\Sigma}_{0|t} = \mathbf{\Sigma} - \sqrt{\alpha_t}\mathbf{\Gamma}_t\mathbf{\Sigma}$.

Proof. Suppose $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{x}_t \mid \mathbf{x}_0 \sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$. This implies that

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + (1-\alpha_t)\boldsymbol{\epsilon} \tag{31}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, we can write the joint distribution of $\mathbf{x}_0$ and $\mathbf{x}_t$ as a multivariate Gaussian [2]:

$$\begin{bmatrix}\mathbf{x}_0 \\ \mathbf{x}_t\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{\Sigma} & \sqrt{\alpha_t}\mathbf{\Sigma} \\ \sqrt{\alpha_t}\mathbf{\Sigma} & \alpha_t\mathbf{\Sigma} + (1-\alpha_t)\mathbf{I}\end{bmatrix}\right). \tag{32}$$

Using properties of conditional Gaussian distributions, we have that $p(\mathbf{x}_0 \mid \mathbf{x}_t)$ is also a Gaussian distribution with conditional mean and covariance

$$\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \sqrt{\alpha_t}\mathbf{\Sigma}\left(\alpha_t\mathbf{\Sigma} + (1-\alpha_t)\mathbf{I}\right)^{-1}\mathbf{x}_t \tag{33}$$

and

$$\mathrm{Cov}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{\Sigma} - \sqrt{\alpha_t}\mathbf{\Sigma}\left(\alpha_t\mathbf{\Sigma} + (1-\alpha_t)\mathbf{I}\right)^{-1}\sqrt{\alpha_t}\mathbf{\Sigma}. \tag{34}$$

Letting $\mathbf{\Gamma}_t = \sqrt{\alpha_t}\mathbf{\Sigma}(\alpha_t\mathbf{\Sigma} + (1-\alpha_t)\mathbf{I})^{-1}$, $\boldsymbol{\mu}_{0|t} = \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t]$ and $\mathbf{\Sigma}_{0|t} = \mathrm{Cov}[\mathbf{x}_0 \mid \mathbf{x}_t]$, we have shown the lemma. □

Next, we prove our main theoretical result, Theorem 2, restated below.

Theorem 7. *Suppose $\mathbf{x}_0$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and we are given linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \mathbf{z}$, where $\mathcal{A}(\mathbf{x}_0) = \mathbf{A}\mathbf{x}_0$ and $\mathbf{z} \sim \mathcal{N}(0, \sigma_y^2 I)$. Suppose that the intermediate value $\mathbf{x}_t$ of the continuous-time reverse diffusion from Equation (5) can be written as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, we have that the true noisy likelihood score $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t)$ is*

$$(\mathbf{A}\mathbf{\Gamma}_t)^T(\mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T + \sigma_y^2\boldsymbol{I})^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{0|t}), \tag{35}$$

where $\mathbf{\Gamma}_t = \sqrt{\bar{\alpha}_t}\mathbf{\Sigma}(\bar{\alpha}_t\mathbf{\Sigma} + (1-\bar{\alpha}_t)\boldsymbol{I})^{-1}$, $\boldsymbol{\mu}_{0|t} = \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{\Gamma}_t\mathbf{x}_t$ and $\mathbf{\Sigma}_{0|t} = \mathrm{Cov}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{\Sigma} - \sqrt{\bar{\alpha}_t}\mathbf{\Gamma}_t\mathbf{\Sigma}$. Moreover, the FGPS approximation $\nabla_{\mathbf{x}_t} \log p(\mathbf{C}_t\mathbf{y} \mid \boldsymbol{\mu}_{0|t})$ can be analytically calculated as

$$(\mathbf{C}_t\mathbf{A}\mathbf{\Gamma}_t)^T\left(\sigma_y^2\mathbf{C}_t\mathbf{C}_t^T\right)^{-1}(\mathbf{C}_t\mathbf{y} - \mathbf{C}_t\mathbf{A}\boldsymbol{\mu}_{0|t}). \tag{36}$$

Proof. We denote $\overleftarrow{\mathbf{x}_t}$ as the iterate from the continuous-time reverse diffusion process such that $\overleftarrow{\mathbf{x}_t} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. From Lemma 6, we have that $p(\mathbf{x}_0 \mid \overleftarrow{\mathbf{x}_t}) = \mathcal{N}(\boldsymbol{\mu}_{0|t}, \mathbf{\Sigma}_{0|t})$ where $\boldsymbol{\mu}_{0|t} = \mathbf{\Gamma}_t\overleftarrow{\mathbf{x}_t}$ and $\mathbf{\Sigma}_{0|t} = \mathbf{\Sigma} - \sqrt{\alpha_t}\mathbf{\Gamma}_t\mathbf{\Sigma}$. Now, similar to the proof of Lemma 6, since $p(\mathbf{y} \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{A}\mathbf{x}_0, \sigma_y^2\mathbf{I})$, we can also calculate $p(\mathbf{y} \mid \overleftarrow{\mathbf{x}_t})$ in closed form as another Gaussian distribution. Specifically, first we can write the joint distribution of $\mathbf{x}_0$ and $\mathbf{y}$ conditioned on $\overleftarrow{\mathbf{x}_t}$ as a multivariate Gaussian [2]:

$$\begin{bmatrix}\mathbf{x}_0 \\ \mathbf{y}\end{bmatrix}\Big|\overleftarrow{\mathbf{x}_t} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{0|t} \\ \mathbf{A}\boldsymbol{\mu}_{0|t}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{0|t} & \mathbf{\Sigma}_{0|t}\mathbf{A}^T \\ \mathbf{A}\mathbf{\Sigma}_{0|t} & \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T + \sigma_y^2\mathbf{I}\end{bmatrix}\right). \tag{37}$$

Further, using properties of conditional Gaussian distributions, we have that $p(\mathbf{y} \mid \overleftarrow{\mathbf{x}_t}) = \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{0|t}, \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T +$

$\sigma_y^2\mathbf{I})$. Let $\mathbf{\Delta}_t = \mathbf{y} - \mathbf{A}\boldsymbol{\mu}_{0|t}$. Then, computing the gradient of $p(\mathbf{y} \mid \overleftarrow{\mathbf{x}_t})$ with respect to $\overleftarrow{\mathbf{x}_t}$, we have

$$\nabla_{\overleftarrow{\mathbf{x}_t}} \log p(\mathbf{y} \mid \overleftarrow{\mathbf{x}_t}) = \nabla_{\overleftarrow{\mathbf{x}_t}} \log \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{0|t}, \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T + \sigma_y^2\mathbf{I}) \tag{38}$$

$$= \nabla_{\overleftarrow{x_t}} - 0.5\mathbf{\Delta}_t^T(\mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T + \sigma_y^2\mathbf{I})^{-1}\mathbf{\Delta}_t \tag{39}$$

$$= \left(\mathbf{A}\frac{\partial \boldsymbol{\mu}_{0|t}}{\partial \overleftarrow{x_t}}\right)^T (\mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^T + \sigma_y^2\mathbf{I})^{-1}\mathbf{\Delta}_t \tag{40}$$

As $\boldsymbol{\mu}_{0|t} = \mathbf{\Gamma}_t\overleftarrow{\mathbf{x}_t}$, we have that $\frac{\partial \boldsymbol{\mu}_{0|t}}{\partial \overleftarrow{\mathbf{x}_t}} = \mathbf{\Gamma}_t$, which gives us Equation (35). The FGPS approximation to the true conditional score is $\nabla_{\overleftarrow{\mathbf{x}_t}} \log p(\mathbf{y} \mid \overleftarrow{\mathbf{x}_t}) \approx \nabla_{\overleftarrow{\mathbf{x}_t}} \log p(\mathbf{C}_t\mathbf{y} \mid \boldsymbol{\mu}_{0|t}) = \nabla_{\overleftarrow{\mathbf{x}_t}} \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{0|t}, \sigma_y^2\mathbf{I})$. Similar to above, we can also calculate its gradient $\nabla_{\overleftarrow{\mathbf{x}_t}} \log p(\mathbf{C}_t\mathbf{y} \mid \boldsymbol{\mu}_{0|t})$ with respect to $\overleftarrow{\mathbf{x}_t}$ as

$$(\mathbf{C}_t\mathbf{A}\mathbf{\Gamma}_t)^T (\sigma_y^2\mathbf{C}_t\mathbf{C}_t^T)^{-1}(\mathbf{C}_t\mathbf{y} - \mathbf{C}_t\mathbf{A}\boldsymbol{\mu}_{0|t}). \tag{41}$$

□

Corollary 3 can easily be proven by taking $\mathbf{C}_t$ as the identity matrix in the above proof.

C. Theoretical Investigation of Approximation Gap of FGPS and DPS

Our main theorem shows in the multivariate Gaussian setting, the true conditional score differs from approximations in the term $(\mathbf{I} + \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^\top)^{-1}$, which requires approximations in general. DPS approximates it at as the identity matrix. In this section, we show that in certain cases, FGPS is a significantly better approximation than the identity matrix. To see this, consider when $\mathbf{A}$ is a high-pass filter and the data covariance is $\mathbf{\Sigma}_f$ such that it follows a power law in the frequency domain. Then, the matrix $\mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^\top$ has significant energy in high-frequency directions. Specifically, in the Fourier basis, denoting the eigenvalues of $\mathbf{A}$, $\mathbf{\Sigma}_f$, $\mathbf{\Sigma}_{0|t}$ as a_i, λ_i and λ_i^t respectively, the eigenvalues of $\mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^\top$ take the form $a_i^2\lambda_i^t$, where a_i^2 grows with frequency (due to the high-pass nature of $\mathbf{A}$), and $\lambda_i^t = \frac{(1-\alpha_t)\lambda_i}{\alpha_t\lambda_i + (1-\alpha_t)}$. Although λ_i^t is small in high-frequency directions, the product $a_i^2\lambda_i^{\text{post}}$ can still be $\mathcal{O}(1)$ or larger due to the amplification by $\mathbf{A}$. Consequently, the eigenvalues of $(\mathbf{I} + \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^\top)^{-1}$ in those directions, given by $\frac{1}{1+a_i^2\lambda_i^{\text{post}}}$, are significantly less than 1, indicating strong suppression of high-frequency components especially when α_t is very small as in the beginning of the reverse process. In contrast, low-frequency directions (where $a_i^2 \approx 0$) are preserved. This shows that $(\mathbf{I} + \mathbf{A}\mathbf{\Sigma}_{0|t}\mathbf{A}^\top)^{-1}$ acts as a low-pass filter. The FGPS approximation, $\mathbf{C}_t^\top(\mathbf{C}_t\mathbf{C}_t^\top)^{-1}\mathbf{C}_t$, which is a projection matrix onto the low-frequency components, thus approximates this behavior more faithfully than the identity matrix, which uniformly preserves all directions.

D. Further Motivations for our Method

In Section 5, we argued that when the forward operator is convolution with a high-pass filter, existing methods have a large approximation between the conditional score and its approximation, which can lead to compromised sample quality in practice. High-pass filtering may seem like a contrived example, because after all, many inverse problem tasks considered in the literature have forward operators that are low-pass filters, such as Gaussian deblurring and superresolution tasks. Even though some natural imaging systems such as phase contrast microscopy, we argue that the high-pass filter effect can also show up in more complex image restoration tasks. For example, for motion deblurring, the forward operator can act as a high-pass filter in certain spatial directions of the image. In Figure 6, we examine this effect by looking at the log magnitude of the frequency domain of an image convolved with a simple directional blur kernel. We can clearly see that in the red circled direction of the Fourier domain, the filter retains high frequency components of the original image. This is also mathematically evident using the Fourier convolution theorem. The Fourier convolution theorem states that

$$\mathcal{F}(k \circledast x) = \mathcal{F}(k) \cdot \mathcal{F}(x), \tag{42}$$

where $\circledast$ denotes convolution and $\cdot$ denotes an elementwise product. Further, for a directional blur kernel, the Fourier transform in spatial directions has high frequency values in directions orthogonal to the direction of the blur. Thus, it is clear that in those directions, the motion blur will retain high frequency components. While the high-pass filter considered in Figure 1 was an extreme case of this, our analysis highlights a crucial deficiency of existing methods since high frequency components of the measurement can amplify approximation errors.

Lastly, we emphasize that our experimental results demonstrate a fascinating phenomenon where the performance gap between FGPS and baseline methods is even larger on more complex datasets, hinting at the fact that besides the frequency characteristics of the forward operator, the frequency characteristics and quality of the Tweedie estimate also greatly affects reconstruction. It would be an interesting theoretical direction to understand why the frequency schedule has larger benefits in this case.

E. Experimental Details

Below, we list the detailed setup for all experiments reported in the main paper. All the images for both FFHQ

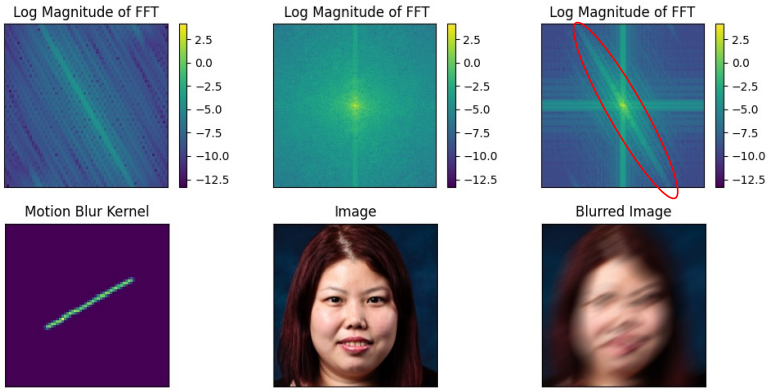

Figure 6. Directional motion blur can retain high frequency components in certain spatial directions orthogonal to the direction of the blur. The red circled direction shows an example of such a direction.

and ImageNet dataset are resized to 256×256, and we report results for 1000 images from the validation datasets for the FFHQ and Imagenet datasets. We will release our code upon publication.

E.1. Inverse Problem Task Descriptions

E.1.1. Linear Inverse Problems

For the first three image restoration problems we consider, the forward operators are linear operators defined as convolution with a given kernel. These kernels are all of size 61×61.

Gaussian Deblurring. The forward operator is convolution with a Gaussian blur kernel with standard deviation 3.

Motion Deblurring. The forward operator is convolution with a motion blur kernel generated from code[2] with intensity 0.5.

Deconvolution with High Pass Filter. the forward operator is a high-pass filter implemented as a Dirac kernel minus a Gaussian blur kernel of standard deviation 5.0.

E.1.2. Nonlinear Inverse Problems

Image Dehazing. The fourth image restoration is a nonlinear inverse problem, image dehazing, where the forward operator is a hazing operator with strength 1. The hazing operator models how light is scattered and attenuated in the atmosphere, based on the atmospheric scattering model

$$\mathcal{A}(\mathbf{x}) = \mathbf{x} \cdot t(\mathbf{x}) + L(1 - t(\mathbf{x})), \tag{43}$$

where $\mathbf{x}$ is the original clear image (scene radiance), $t(\mathbf{x})$ is the transmission map representing the portion of light that reaches the camera, and L is the atmospheric light value. The transmission map $t(\mathbf{x})$ can be modeled using the Beer-Lambert law, which states that

$$t(\mathbf{x}) = e^{-\beta d(\mathbf{x})}. \tag{44}$$

Above, β is the atmospheric scattering coefficient, and $d(\mathbf{x})$ is the scene depth map. In our experiments, we set $L = 1, \beta = 1$, and set $d(\mathbf{x})$ be Euclidean distance of each pixel to the center of the image.

E.2. Baselines

DPS [9] We use the code from https://github.com/DPS2022/diffusion-posterior-sampling using the default hyperparameter settings for Gaussian deblurring and motion deblurring for both FFHQ and ImageNet. For high-pass filter operator, we used the same hyperparameters used for motion deblurring.

MCG [10] We use the code from https://github.com/HJ-harry/MCG_diffusion using the default hyperparameter settings for Gaussian deblurring and motion deblurring for both FFHQ and ImageNet. For high-pass filter operator, we used the same hyperparameters used for motion deblurring.

DSG [42] We use the code from https://github.com/LingxiaoYang2023/DSG2024 using the default hyperparameter settings for Gaussian deblurring for both FFHQ and ImageNet. The only hyperparameter change was that we set $interval = 10$ as we observed this worked better in practice. For high-pass filter deconvolution and motion deblurring operator, we used the same hyperparameters used for Gaussian deblurring.

Score-SDE/ILVR [8, 37] Generally, we group Score-SDE and ILVR as methods that use a sequence of noisy measurements to approximate the conditional score, as mentioned in [12] and [9]. This is a generalization of the

[2] https://github.com/LeviBorodenko/motionblur

methods in [37] and [8] as their methods only were presented for the inpainting and superresolution tasks. To consider general tasks, we sample a sequence of measurements $\mathbf{y}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{y}, (1-\bar{\alpha}_t\mathbf{I}))$. Then, we approximate the conditional score as $-\eta\nabla_{\mathbf{x}_t}\|\mathbf{y}_t - \mathcal{A}(\mathbf{x}_t)\|_2^2$. We use the step size η from the DPS method [9].

AOD-Net. We use the code from https://github.com/MayankSingal/PyTorch-Image-Dehazing and train the network using the default parameters on 10000 images from the FFHQ training dataset.

DoubleDIP. We use the code from https://github.com/yossigandelsman/DoubleDIP using all default parameters. This method is unsupervised and does not require any retraining.

E.3. Details for Figure 1

In Figure 1, we demonstrated the approximation gap of DPS and our method on synthetic data. Below, we describe the precise experimental details of our experiments.

Forward Operators. We consider a high-pass filter operators in our experiments. a low-pass filter and a high-pass filter. First, we create a Gaussian blur kernel of width σ i.e. this kernel is $e^{-\mathbf{x}^2/(2\sigma^2)}$ and normalized to have sum 1. For the high-pass filter, we simply subtract the Gaussian blur kernel of width σ from a Dirac kernel, which has 1 at the center and zeroes elsewhere. The filter is a circulant matrix constructed from this kernel. Importantly, the first row of the circulant matrix corresponding to this kernel is normalized to have sum zero, such that the kernel is indeed a high-pass filter.

Data Generation. Next, we describe the data generation process. Recall from Lemma 1 the construction of a covariance matrix $\mathbf{\Sigma}_f$ such that data drawn from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_f)$ follows a power law in the frequency domain with parameters c and β. We let $c = 1$ and $\beta = 2.5$, which is the typical range for natural image data [34]. We draw 10000 signals $\mathbf{x}^{(i)}$ of length 2000 from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_f)$. For each signal, we generate a corresponding measurement $\mathbf{y}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{A}$ is the circulant matrix corresponding to the low-pass filter or the high-pass filter.

Approximation Gap. For each $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ pair, we calculate $\mathbf{x}_t^{(i)} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}^{(i)}, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t$ is the variance schedule from the DDPM paper [20]. We report the average approximation gap over the 10000 signals. Below, we give the exact approximation that our method, FGPS, makes for the noisy likelihood score. From Theorem 2, we know that under a multivariate normal assumption on the data, the posterior mean $\boldsymbol{\mu}_{0|t}$ is $\mathbf{\Gamma}_t\mathbf{x}_t$. Under the same assumptions as Theorem 2, for linear inverse problems where $\mathcal{A}(\mathbf{x}_0) = \mathbf{A}\mathbf{x}_0$ we analytically compute our approximation to the noisy likelihood score given in Equation (12) as

$$(\mathbf{C}_t\mathbf{A})^T(\sigma_y^2\mathbf{C}_t\mathbf{C}_t^T)^{-1}(\mathbf{C}_t\mathbf{y} - \mathbf{C}_t\mathbf{A}\boldsymbol{\mu}_{0|t}). \tag{45}$$

Frequency Curriculum. For our method, we have the choices of the frequency curriculum described through τ_T and τ_1, the frequency cutoffs for the time-dependent low pass filter. For simplicity, we use a frequency curriculum from the variance schedule of the diffusion model. Specifically, let $\sigma_t = \frac{1}{\bar{\alpha}_t} - 1$, which corresponds to the corresponding variance-exploding parameterization of the SDE (see [24], Appendix B). Then we set τ_t be the max frequency value f_k such that $S(f_k) \geq \max\{\sigma_y^2, \sigma_t^2\}$, where $S(f_k)$ denotes the power spectral density from Equation (8). This utilizes the intuition that for $\mathbf{x}_t^{(i)}$, the frequency components of the original signal $\mathbf{x}^{(i)}$ still present in the image are the frequencies below τ_t.

E.4. Implementation Details for Our Method

In practice, we let τ_1 and τ_T be hyperparameters and consider various schedules to interpolate between them. However, our theoretical results can guide us on setting them in a data-dependent way. Recall that natural images tend to follow a radially averaged power law in the frequency domain such that each non-zero frequency f_k has power $S(f_k) \approx c|f_k|^{-\beta}$ for some constants $c, \beta > 0$. We employ three heuristics for setting our frequency schedule.

1. (Setting τ_1 to improve robustness to noise) When the measurement noise has variance σ_y^2, higher frequency components of $\mathbf{y}$ are affected, specifically those frequencies f such that $S(f) \leq \sigma_y^2$. Denoting f_{noise} as the minimum frequency value that satisfies this constraint, we can set τ_1 less than f_{noise} and make our method robust to additive Gaussian noise of a known variance.
2. (Frequency Schedule as Function of Power Law Decay) For data that follows a faster decaying power law (when β is larger), we can set a schedule for τ_t that also increases faster as a function of t.
3. (Setting τ_T) For natural images, it is sufficient to set τ_T to be a small percentage of the overall frequency range, as this contains most of the information present in the measurement $\mathbf{y}$. This is the same intuition behind JPEG compression. In our experiments on natural images, we take τ_T to be roughly 30% of the overall frequency range.

To implement a frequency schedule, we use a binary mask applied in the frequency domain. Specifically, we utilize the Fast Fourier Transform (FFT) to transform the image data into the frequency domain. A low-pass filter mask is then created based on a specified cutoff frequency τ_t. We utilize the Euclidean distance from the center of the frequency domain to create the low-pass filter mask, which

selectively retains frequencies below the cutoff value. As the reverse process progresses, the cutoff is adjusted to allow more high-frequency components, thereby refining the image details. This method efficiently integrates frequency control into the reverse diffusion process, contributing to improved image restoration. In our work, we consider two frequency schedules, an exponential schedule and a linear schedule, which interpolate between τ_T and τ_1 in different ways. Precisely, the exponential schedule follows

$$\tau_t = \tau_1 - (\tau_1 - \tau_T)\exp\left(-\frac{5t}{T}\right), \tag{46}$$

and the linear schedule follows

$$\tau_t = \tau_T + \frac{t}{T}(\tau_1 - \tau_T). \tag{47}$$

Further, we set the step size $\mathbf{S}_t$ to be $\mathbf{S}_t = \frac{\kappa_t}{\left\|\phi_t(\mathbf{y}) - \phi(\mathcal{A}(\boldsymbol{\mu}_{0|t}))\right\|_2}\mathbf{I}$ for a scalar time-dependent hyperparameter κ_t. We set this schedule to smoothly transition from κ_T at the beginning of the reverse process to κ_1 at the end. Precisely, this schedule follows

$$\kappa_t = \frac{1}{2}(\kappa_T + \kappa_1) + \frac{1}{2}(\kappa_T - \kappa_1)\cos\left(\frac{\pi t}{T}\right), \tag{48}$$

where t denotes the current time step, and T is the total number of time steps (e.g. 1000 in our case).

Next, we detail all the hyperparameter settings used for our experiments on the FFHQ and ImageNet datasets.

- FFHQ
  - Gaussian Deblurring:
    - Step Size: $\kappa_T = 3.0, \kappa_1 = 0.6$
    - Frequency Schedule: Exponential
  - Motion Deblurring:
    - Step Size: $\kappa_T = 5.0, \kappa_1 = 1.0$
    - Frequency Schedule: Exponential
  - High pass:
    - Step Size: $\kappa_T = 5.1, \kappa_1 = 1.1$
    - Frequency Schedule: Linear
  - Haze:
    - Step Size: $\kappa_T = 5.0, \kappa_1 = 1.0$
    - Frequency Schedule: Exponential
- ImageNet
  - Gaussian Deblurring:
    - Step Size: $\kappa_T = 2.0, \kappa_1 = 0.01$
    - Frequency Schedule: Linear
  - Motion Deblurring:
    - Step Size: $\kappa_T = 3.0, \kappa_1 = 0.1$
    - Frequency Schedule: Linear
  - High pass:
    - Step Size: $\kappa_T = 3.5, \kappa_1 = 0.6$
    - Frequency Schedule: Linear

In Appendix G, we provide ablation studies to understand the effects of these hyperparameters on the generation quality.

E.5. Computational Overhead of Our Method

In this section, we evaluate the computational efficiency of our method by measuring the average runtime per image on the FFHQ dataset. We run each method from Table 1 on 100 images, and report the average time taken per image, given in Table 3. We see that our method introduces minimal computational overhead compared to the DPS method, demonstrating that our modifications can be implemented in an efficient manner and easily integrated into existing frameworks.

Method	Time (seconds)
DSG	45.13744
Score-SDE/ILVR	41.411510
MCG	93.20647
DPS	90.641704
FPGS (Ours)	92.383272

Table 3. Average Wall Clock Runtime Per Image (seconds) on FFHQ Dataset

F. Comparison to ILVR/Score-SDE

Our method is closely related to the Score-SDE and ILVR works that consider a sequence of noisy measurements y_t such that $y_t = \mathcal{N}(\sqrt{\alpha_t}y_0, (1-\alpha)\mathbf{I})$ [8, 37]. These methods consider approximations to the noisy likelihood score that are typically of the form $\mathbf{L}_{\mathcal{A}}(\mathbf{y}_t - \mathcal{A}(\mathbf{x}_t))$, where $\mathbf{L}_{\mathcal{A}}$ is a fixed matrix that depends on the measurement operator $\mathcal{A}$. There are three crucial differences between these methods and our method. First, our approximation to the noisy likelihood score considers a time-varying model likelihood at each diffusion timestep $p(\mathbf{y} \mid \boldsymbol{\mu}_{0|t})$ as given in Equation (10) as opposed to simply varying $\mathbf{y}_t$. Further, we leverage the powerful denoising capabilities of the pretrained diffusion model by using the Tweedie estimate $\boldsymbol{\mu}_{0|t}$ instead of the noisy $\mathbf{x}_t$. Lastly, we propose a frequency curriculum that can differ from the variance schedule of the diffusion model and be adapted to the frequency characteristics of the data. These differences lead to a drastically improved performance of our method compared to Score-SDE/ILVR as seen in Table 1.

G. Further Experiments and Ablation Studies

G.1. Visualizing the Transformed Measurements

In Figure 7, we show the transformed measurements at three timesteps in the reverse diffusion process when applying our frequency curriculum on the FFHQ dataset on all the

forward operators we considered. Our theoretical results indicate that the initial reverse process steps are very important in order to obtain coarse alignment with the given measurement, so it is reasonable to align the measurements with this coarse-to-fine strategy.

G.2. Effect of Frequency Curriculum

Impact of Time-Dependent Curriculum. As observed in Figure 5, we observe that the time-dependent frequency curriculum helps the stability of the method. Namely, when the operator is a high pass filter, the time-dependent frequency curriculum results in high-quality reconstructions. On the other hand, for a fixed curriculum, we observe that over different generations, most random samples of the diffusion model fall off the natural image manifold in the early steps of the reverse process, which results in unnatural looking images as in Figure 5.

Next, to assess the impact of different frequency curricula on image quality, we provide a visual comparison using images generated with linear and exponential time-dependent curricula on the FFHQ and ImageNet datasets.

FFHQ Motion Deblurring. The images produced with the exponential frequency schedule exhibit superior quality, as shown in Figure 9. The exponential increase in high-frequency components helps to better capture fine facial features and intricate details, resulting in more visually appealing images.

ImageNet Motion Deblurring. The ImageNet dataset benefits more from the linear frequency schedule, as illustrated in Figure 10. We hypothesize this is because the diversity and detail of ImageNet scenes require more careful consideration at many frequency ranges to obtain a high-quality reconstruction.

Motivation for Frequency Schedule Selection. Our observations indicate that the choice of frequency schedule should be informed by the characteristics of the underlying data. The exponential schedule is well-suited for datasets with structured and hierarchical features, such as faces, where it is crucial to refine high-frequency details over an extended number of reverse process time steps. In contrast, datasets like ImageNet, which contain diverse and complex textures, benefit from a more uniform and gradual frequency progression, ensuring consistent detail incorporation throughout the sampling process. We note though that while our method offers the flexibility to incorporate different frequency curricula, it is not a necessity to obtain high-quality reconstructions, and even a simple exponential curriculum can already obtain results that are significantly better than baseline methods. We simply highlight that this can be further improved by carefully considering the frequency distribution of the data.

G.3. Effect of Step Size Schedule

Recall that in Equation (12), we had that our approximation to the score should be scaled by $\mathbf{S}_t$, which can be thought of as a step size to the gradient terrm. In practice, instead of using this exact quantity, which would require using a potentially large circulant matrix and its inverse, we set $\mathbf{S}_t = \frac{\kappa}{\left\|\phi_t(\mathbf{y}) - \phi(\mathcal{A}(\boldsymbol{\mu}_{0|t}))\right\|_2}\mathbf{I}$ for a scalar hyperparameter κ, which works well in practice. This is similar to the step size chosen in DPS. The intuition behind this step size choice is to control the approximation error. For example, when the approximation error is high, the step size should be smaller as the conditional score is noisy. These schedules play a crucial role in the effectiveness and stability of the optimization process. In this section, we study the effect of different step size schedules to motivate our choice. Specifically, we study three step size schedules (where we vary κ) on the motion deblurring task on the FFHQ dataset. These three schedules are:

1. Fixed small step size (used by DPS)
2. Fixed large step size
3. Cosine Annealed step size according to Equation (48). This schedule starts with a relatively high step size, which facilitates rapid progress in the initial stages, and gradually decreases to smaller step sizes as the optimization nears convergence.

In Table 4 and Figure 8, we demonstrate the quantitative and qualitative differences between the three schedules. We clearly see that the cosine annealed step size strongly outperforms the other two schedules, which is why we adopt it for our experiments. We hypothesize that the cosine schedule allows the model to quickly capture the coarse structure of the image early on, while the progressively smaller steps enable fine-tuning of details, leading to a more refined final reconstruction. This is evident from Figure 8. For example, in the first row, the cosine schedule successfully captures the necklace detail around the subject's neck, while the fixed small step size schedule completely misses this feature, resulting in a blurred and oversmoothed output. The fixed large step size schedule manages to retain some necklace details but introduces noticeable artifacts, which degrade the overall image quality. The large step size likely destabilizes the optimization and pushes the image off the natural image manifold, which compromises image quality.

Step Size Schedule	FID↓	LPIPS↓	PSNR↑	SSIM↑
Cosine Annealing	**49.66**	**0.1254**	**25.53**	**0.724**
Fixed Lower Bound	70.73	0.1790	24.22	0.677
Fixed Upper Bound	59.94	0.1559	24.22	0.686

Table 4. Performance comparison of different scale schedules on the FFHQ dataset. Metrics include FID, LPIPS, PSNR, and SSIM.

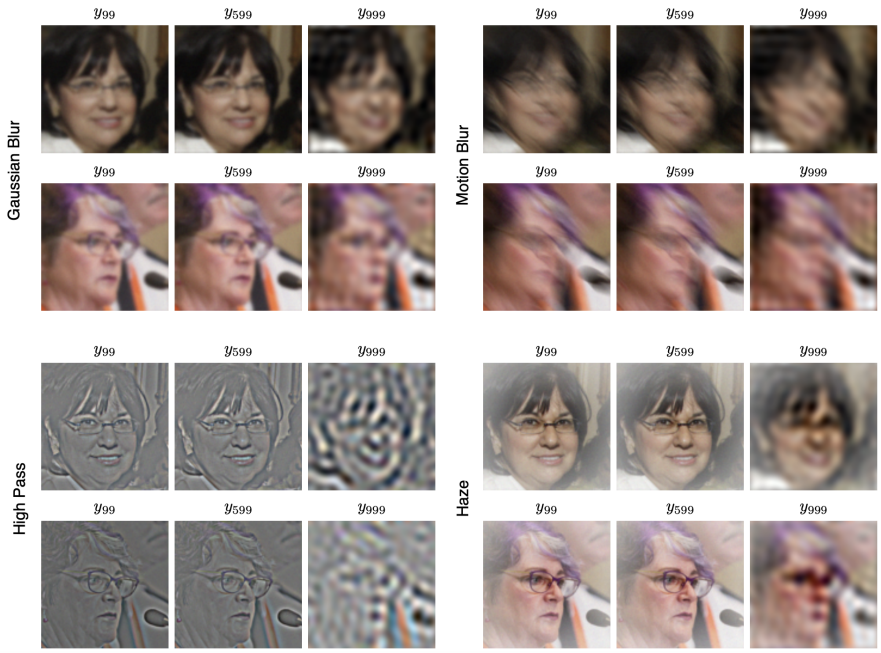

Figure 7. We demonstrate the visual effects of our frequency curriculum on the transformed measurements. In the early stages of the reverse process, only very coarse features of the measurement are retained.

Figure 8. Qualitative comparison of different scale schedules on the FFHQ dataset. The first row highlights how the cosine schedule accurately captures details, such as the necklace, while the fixed small step size ($\kappa = 1$) fails and the fixed large step size ($\kappa = 5$) introduces artifacts.

G.4. Further Qualitative Results

Comparison to Baselines. In Figure 11, we see the zoomed in version from Figure 3, highlighting the intricate details of the image captured by our reconstructions. In Figure 12, we see a comprehensive comparison to all the baselines reported in Table 1 on the motion blurring task on the FFHQ dataset. We observe that Score-SDE/ILVR often captures the correct structure of the image but fills in different facial features. MCG usually overfits to the Gaussian measurement noise as reported in [9]. DPS results in good quality reconstructions, but the images are usually smoothed out and lose small facial features. DSG performs very well on the motion deblurring task, as shown also in Table 1. How-

ever, DSG often gives grainy reconstructions and artifacts that are clearly visible. In contrast, our method is significantly more stable and is able to give images that have higher perceptual quality than all the baselines.

FFHQ Additional Results. Figures 13 and 14 show additional results of our method on the motion deblurring and image dehazing task on the FFHQ dataset.

ImageNet Additional Results. Figures 15, 16, and 17 show additional results of our method on the Gaussian deblurring, motion deblurring, and high-pass filter deconvolution tasks on the ImageNet dataset. On the high-pass filter task, we note that there is sometimes a color shift which results from the loss of color information in the measurement. We observe that compared to DPS and a fixed-time frequency curriculum (as in Figure 5), our method is much more stable and usually gives visually plausible reconstructions instead of strong color artifacts that dominate the reconstruction.

H. Limitations and Future Work

While FGPS requires little frequency schedule tuning, the step size still plays a large role in dictating image quality and needs to be carefully tuned. Further, the frequency curriculum is applicable only for image restoration tasks where the measurement is still an image. Lastly, similar to DPS, FGPS requires knowledge of the forward operator during the reverse process, which restricts it to non-blind inverse problems.

It is important to note that our theoretical results do not paint the full picture of the success of FGPS in practice. Empirically, the Tweedie estimate $\boldsymbol{\mu}_{0|t}$ used for the conditional score approximation behaves in complex ways and its frequency structure is not as simple as the form in our theoretical results, $\boldsymbol{\Gamma}_t \mathbf{x}_0$. We conjecture it is still important to explicitly align the frequency structure of the unconditional score and noisy likelihood score, which is why FGPS outperforms DPS for low-frequency measurements on FFHQ and Imagenet. In addition to the role of the Tweedie estimate, we find an intriguing role of the dataset where FGPS performs even better on harder datasets like ImageNet due to more complex frequency structure. Explaining both these phenomena theoretically is an interesting direction for future work. That being said, our empirical findings demonstrate that the core idea behind FGPS, aligning the spectral structure of the measurement with the score function, remains effective for complex data. Our curriculum strategy reflects a coarse-to-fine alignment of frequencies, motivated by both the empirical behavior of diffusion models and spectral properties of natural images. We believe this insight opens avenues for more principled guidance mechanisms utilizing the structure of the score function.

Future work would include a rigorous analysis of the step size and how it affects the approximation error. It would also be useful to consider several competing works and their introduced approximation errors using our theoretical analysis as a backbone. Lastly, we hope to extend FGPS to other inverse problems, both blind and non-blind, where the measurement is not an image such as medical imaging tasks.

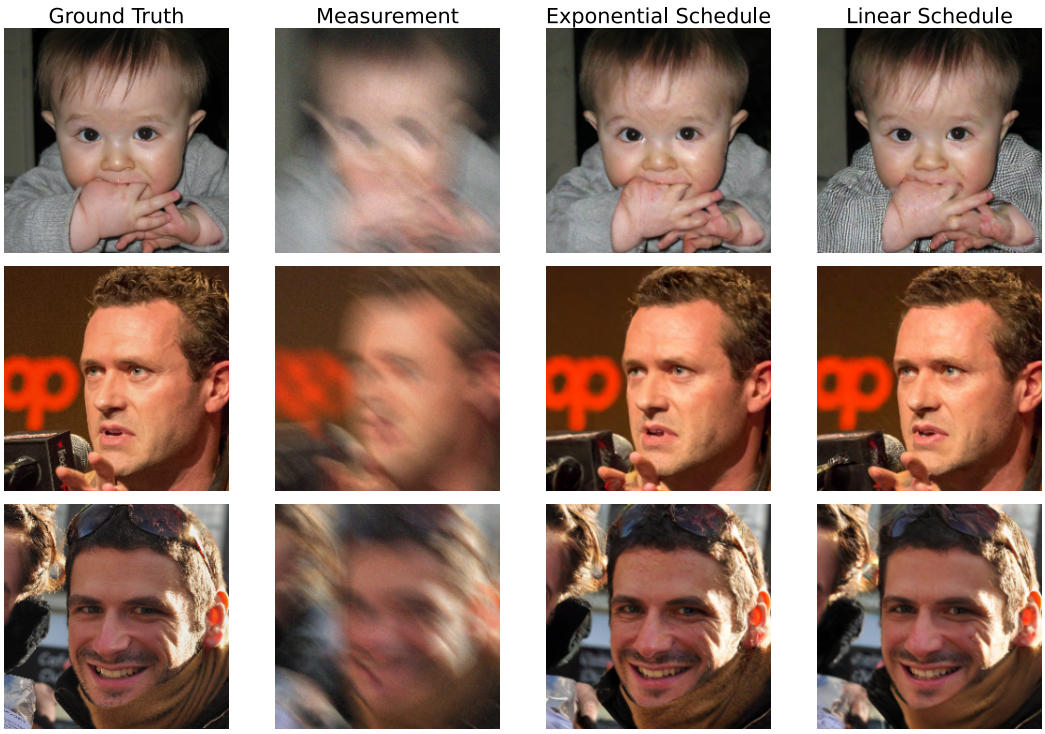

Figure 9. Comparison of images on the FFHQ dataset using exponential and linear frequency schedules. The exponential schedule produces higher-quality images with refined details and realistic textures compared to the linear schedule.

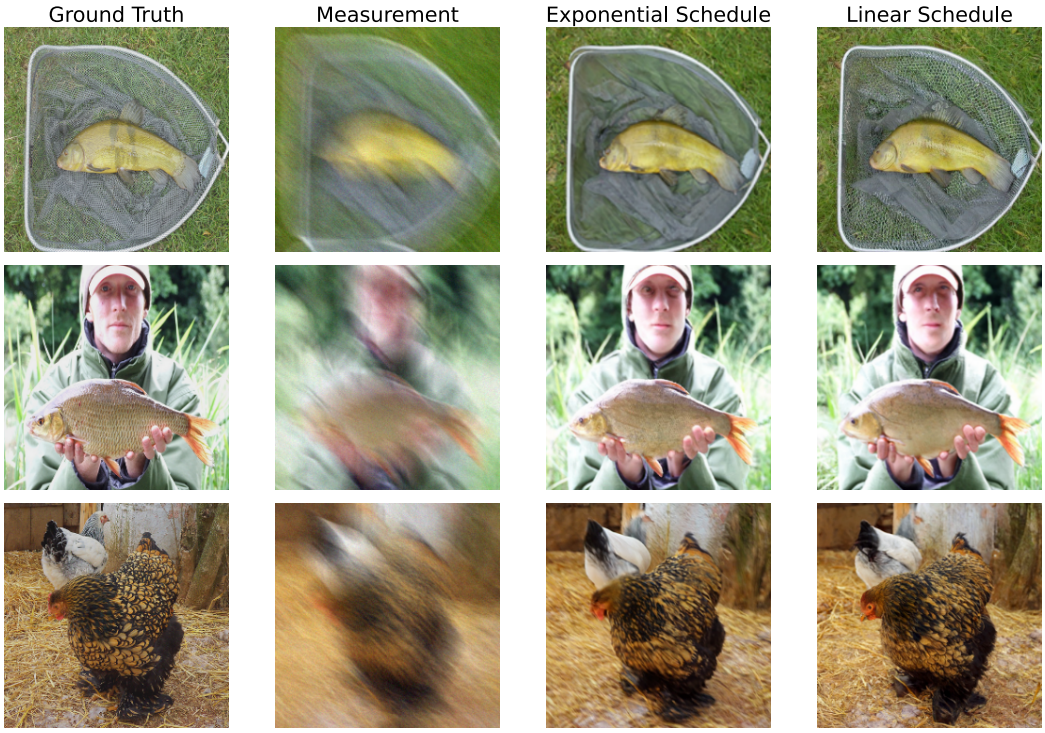

Figure 10. Comparison of images on the ImageNet dataset using exponential and linear frequency schedules. The linear schedule preserves object shapes and structural details better, resulting in clearer images.

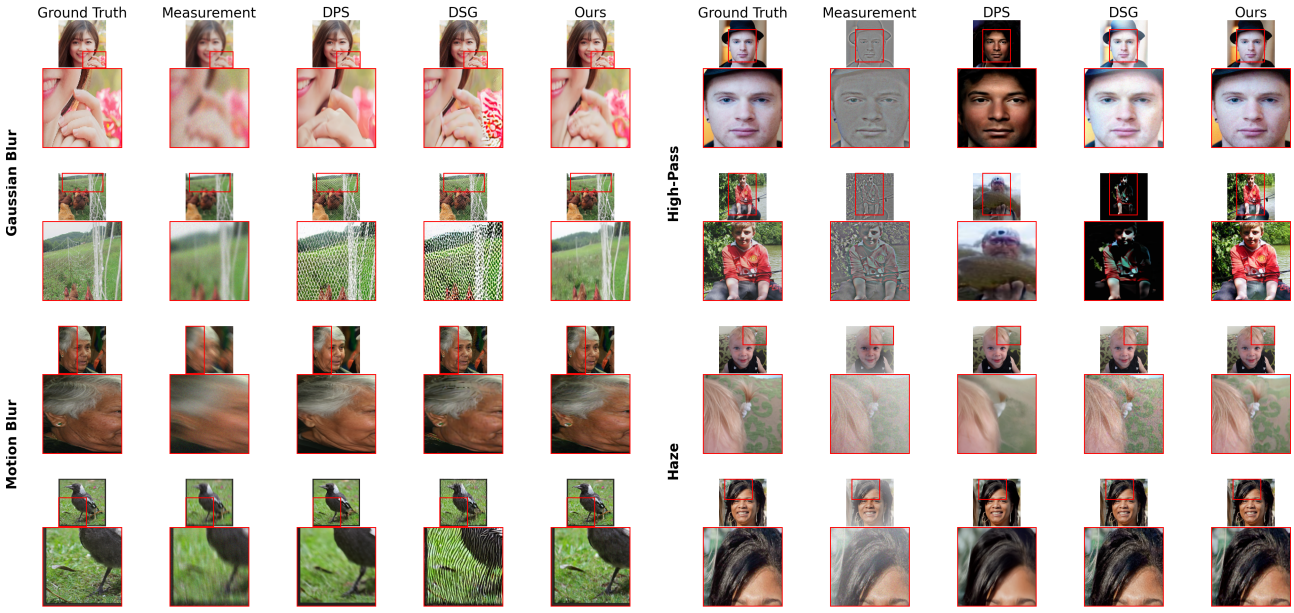

Figure 11. Qualitative results of our method with zoomed in portions of images from Figure 3. Our method successfully preserves finer details like background pattern.

Figure 12. Qualitative motion deblurring results on FFHQ dataset for all baselines we report in Table 1. The same blur kernel is applied to each image.

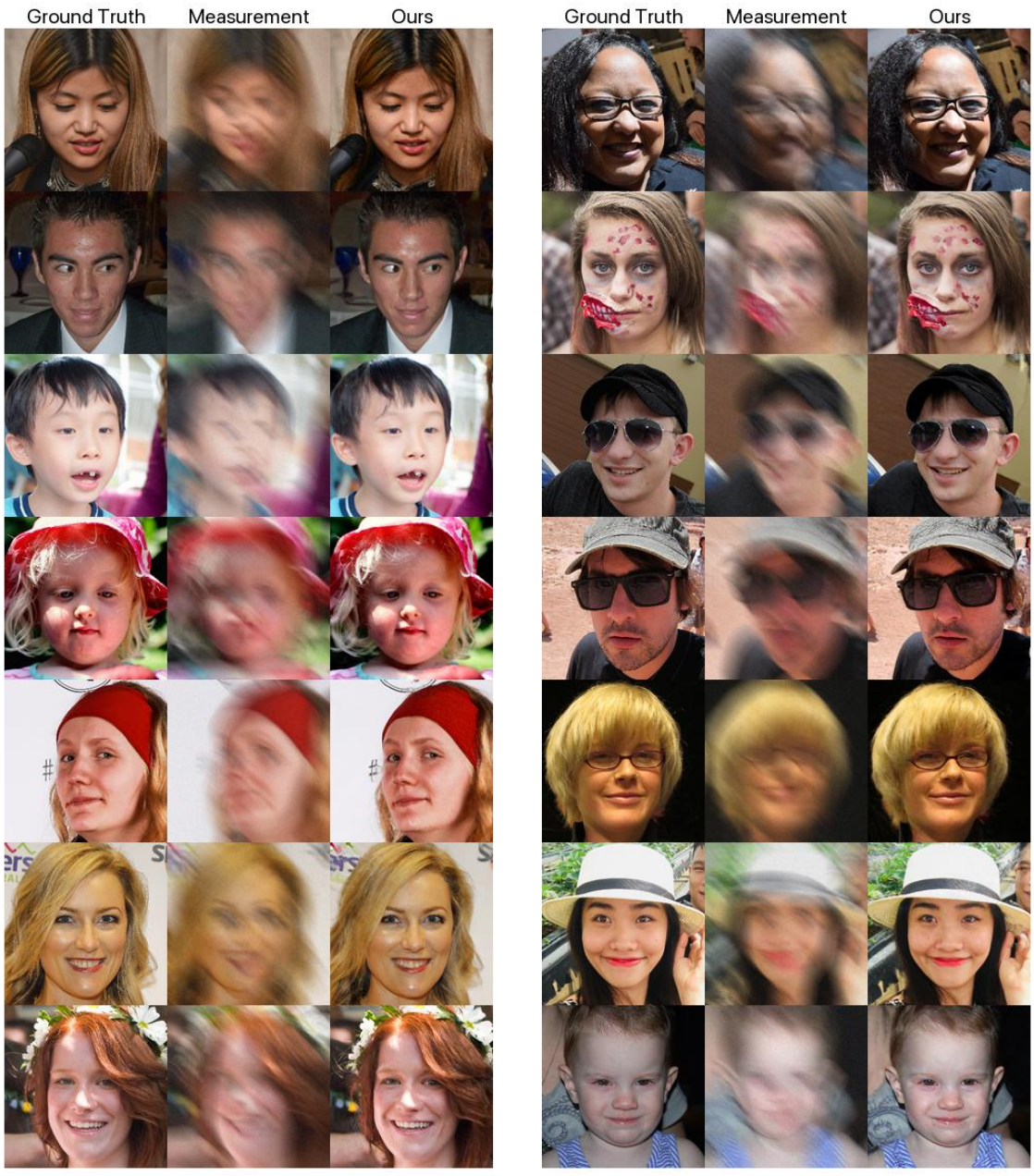

Figure 13. Qualitative motion deblurring results on FFHQ dataset. The same blur kernel is applied to each image.

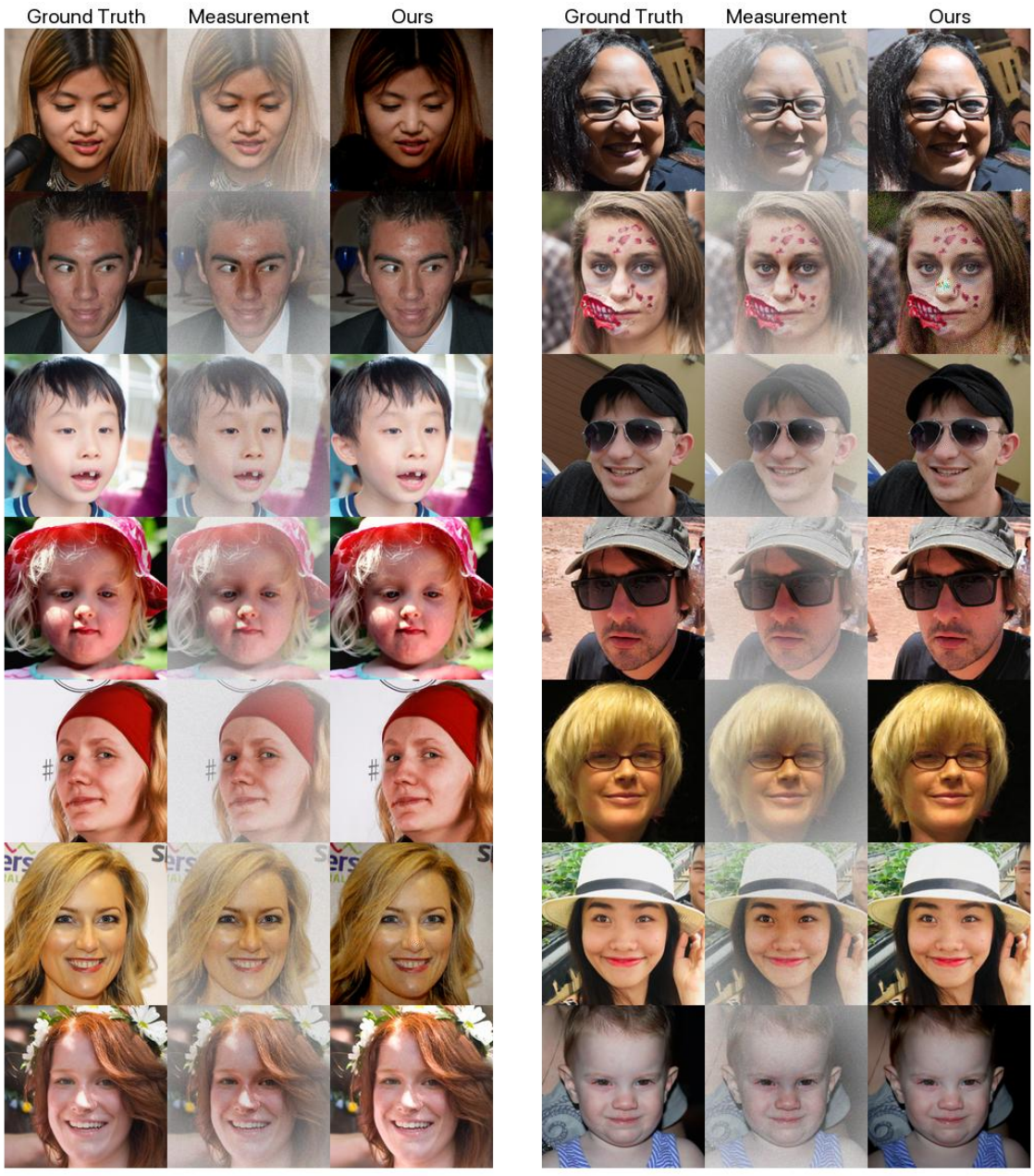

Figure 14. Qualitative image dehazing results on FFHQ dataset.

Figure 15. Qualitative Gaussian deblurring results on Imagenet dataset.

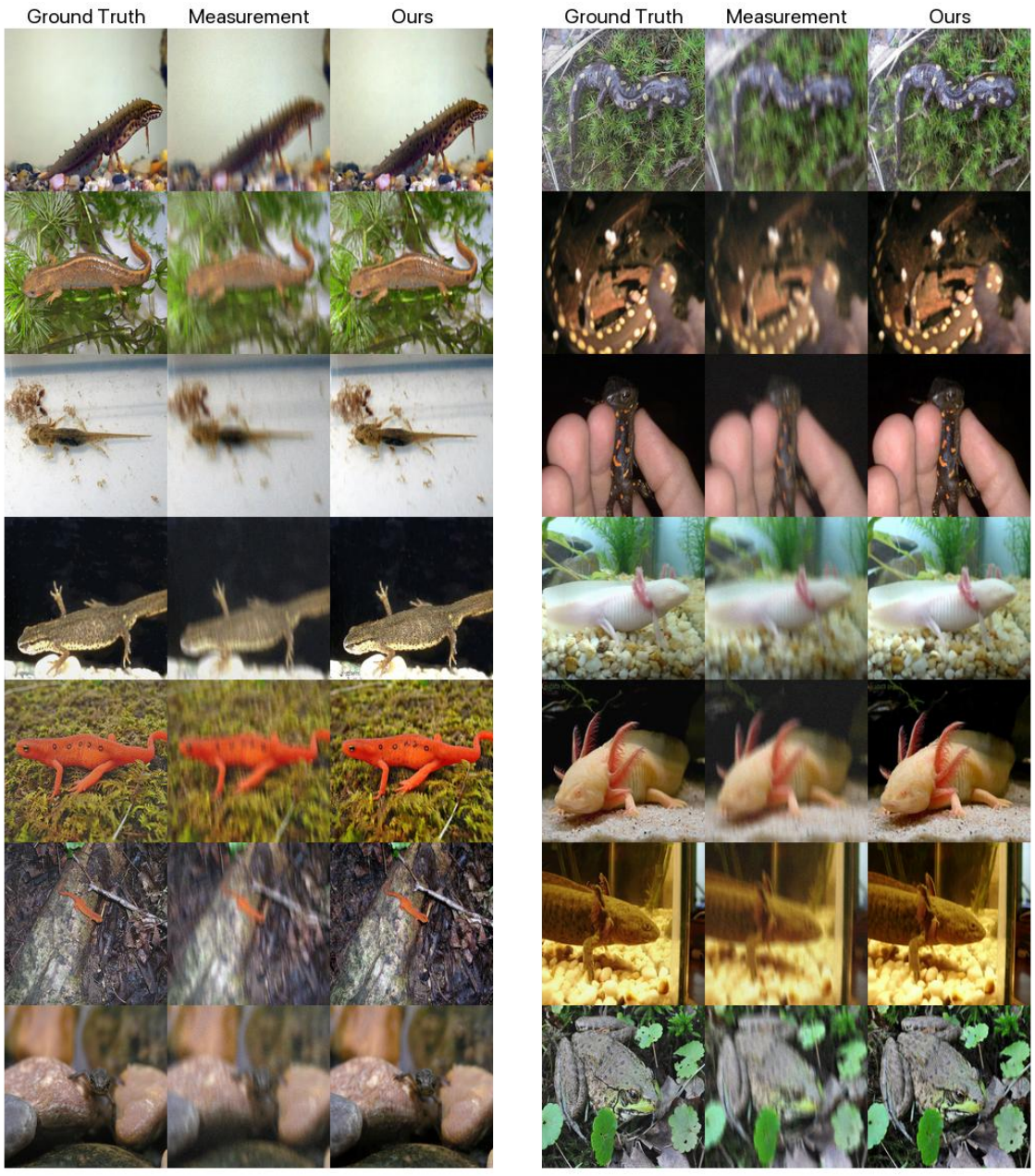

Figure 16. Qualitative motion deblurring results on Imagenet dataset. The same blur kernel is applied to each image.

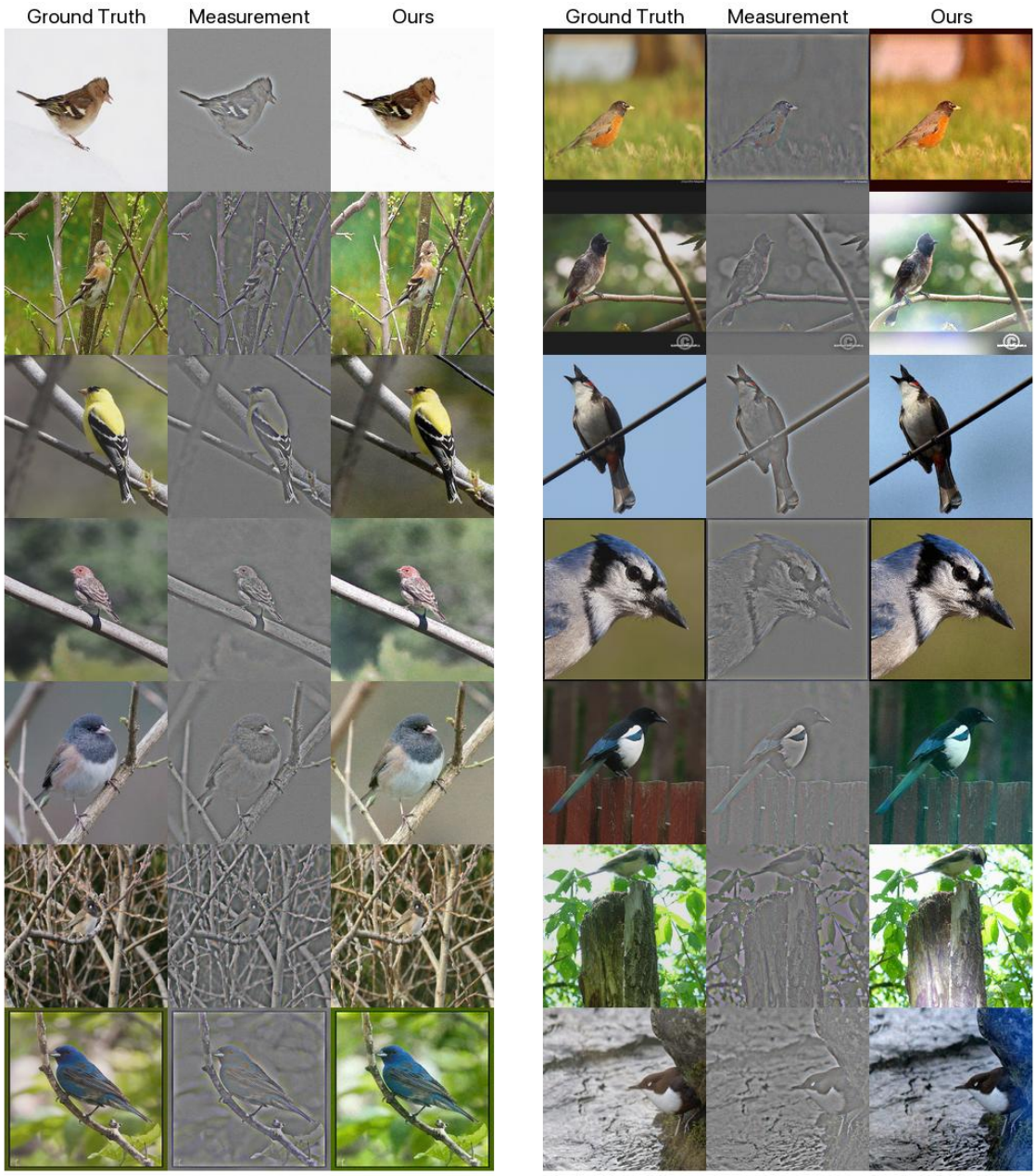

Figure 17. Qualitative results on Imagenet dataset when the measurement is a high-pass filter applied to the ground truth image.