Traversing Distortion-Perception Tradeoff using a Single Score-Based Generative Model

Supplementary Material

In this supplementary material, we first provide the approximation for the reverse posterior $p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y})$ [29] and its connection to conditional score in Appendix A. Then we prove Theorem 1 and 2 in Appendix B and C respectively. Appendix D provides the derivation of posterior distribution and MMSE for the mixture Gaussian case. Finally, the experimental details and more results on two-dimensional datasets and the FFHQ dataset are included in Appendix E.

A. Approximation of Reverse Posterior Distribution

In this section, we will first include the deviation of posterior mean and variance in [29] for self-contained. We will use a modified proof to reveal the relationship between the posterior mean and conditional score.

With the Bayes' rule, we have for VP-diffusion

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}) &= \frac{p(\mathbf{x}_{k+1}|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y})}{p(\mathbf{x}_{k+1}|\mathbf{y})} \\ &\propto \exp\Big(-\frac{1}{2\beta_{k+1}}||\mathbf{x}_{k+1}-(1-\frac{1}{2}\beta_{k+1})\mathbf{x}_k||^2+\log p(\mathbf{x}_k|\mathbf{y})-\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big). \end{aligned} \tag{14}$$

We can approximate $\log p(\mathbf{x}_k|\mathbf{y})$ by Taylor's expansion on point $\mathbf{x}_{k+1}$. When $T\to\infty$,

$$\log p(\mathbf{x}_k|\mathbf{y}) \approx \log p(\mathbf{x}_{k+1}|\mathbf{y}) + (\mathbf{x}_k-\mathbf{x}_{k+1})^\top \nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y}) + \mathcal{O}(\|\mathbf{x}_k-\mathbf{x}_{k+1}\|^2).$$

Then, (14) can be written as

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}) &\propto \exp\Big(-\frac{1}{2\beta_{k+1}}||\mathbf{x}_{k+1}-(1-\frac{1}{2}\beta_{k+1})\mathbf{x}_k||^2+(\mathbf{x}_k-\mathbf{x}_{k+1})^T\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big) \\ &= \exp\Big(-\frac{1}{2\beta_{k+1}}\Big(\mathbf{x}_{k+1}^\top\mathbf{x}_{k+1}-2(1-\frac{1}{2}\beta_{k+1})\mathbf{x}_{k+1}^\top\mathbf{x}_k+(1-\frac{1}{2}\beta_{k+1})^2\mathbf{x}_k^\top\mathbf{x}_k\Big) \\ &\qquad\qquad + \mathbf{x}_k^\top\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}\mid\mathbf{y}) - \mathbf{x}_{k+1}^\top\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big) \\ &\propto \exp\Big(-\frac{1-\beta_{k+1}}{2\beta_{k+1}}\mathbf{x}_k^\top\mathbf{x}_k+\Big(\frac{1-\frac{1}{2}\beta_{k+1}}{\beta_{k+1}}\mathbf{x}_{k+1}+\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big)^\top\mathbf{x}_k\Big) \\ &\propto \exp\Big(-\frac{1-\beta_{k+1}}{2\beta_{k+1}}\Big(\mathbf{x}_k^\top\mathbf{x}_k-2(\frac{1-\frac{1}{2}\beta_{k+1}}{1-\beta_{k+1}}\mathbf{x}_{k+1}-\frac{\beta_{k+1}}{1-\beta_{k+1}}\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y}))^\top\mathbf{x}_k\Big)\Big) \\ &\propto \exp\Big(-\frac{1-\beta_{k+1}}{2\beta_{k+1}}\Big\|\mathbf{x}_k-\frac{1}{\sqrt{\alpha_{k+1}}}\Big(\mathbf{x}_{k+1}+(1-\alpha_{k+1})\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big)\Big\|^2\Big), \end{aligned} \tag{15}$$

where the last step utilizes the equivalent infinitesimal $\sqrt{\alpha_k}=\sqrt{1-\beta_k}=1-\frac{1}{2}\beta_k$ when $T\to\infty$. From (15), we can see that $p(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y})$ has mean

$$\boldsymbol{\mu}_k(\mathbf{x}_{k+1},\mathbf{y}) \triangleq \frac{1}{\sqrt{\alpha_{k+1}}}\Big(\mathbf{x}_{k+1}+(1-\alpha_{k+1})\nabla_{\mathbf{x}_{k+1}}\log p(\mathbf{x}_{k+1}|\mathbf{y})\Big). \tag{16}$$

For VP diffusion, the expectation and covariance of $p(\mathbf{x}_k|\mathbf{y})$ can be computed as [29]

$$\begin{aligned}
\boldsymbol{\mu}_k = \mathbb{E}_{p(\mathbf{x}_k|\mathbf{y})}\left[X_k\right] &= \int \mathbf{x}_k p\left(\mathbf{x}_k \mid \mathbf{y}\right) d\mathbf{x}_k \\
&= \int \mathbf{x}_k \int p\left(\mathbf{x}_k \mid \mathbf{x}_0\right) p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 d\mathbf{x}_k \\
&= \iint \mathbf{x}_k p\left(\mathbf{x}_k \mid \mathbf{x}_0\right) d\mathbf{x}_k p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 \\
&= \int \sqrt{\bar{\alpha}_k}\mathbf{x}_0 p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 = \sqrt{\bar{\alpha}_k}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right] \qquad (17)\\
\boldsymbol{\Sigma}_k = \operatorname{Cov}_{p(\mathbf{x}_k|\mathbf{y})}\left[X_k\right] &= \int \mathbf{x}_k\mathbf{x}_k^\top p\left(\mathbf{x}_k \mid \mathbf{y}\right) d\mathbf{x}_k - \bar{\alpha}_k \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top \\
&= \int \mathbf{x}_k\mathbf{x}_k^\top \int p\left(\mathbf{x}_k \mid \mathbf{x}_0\right) p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 d\mathbf{x}_k - \bar{\alpha}_k \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top \\
&= \iint \mathbf{x}_k\mathbf{x}_k^\top p\left(\mathbf{x}_k \mid \mathbf{x}_0\right) d\mathbf{x}_k p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 - \bar{\alpha}_k \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top \\
&= \int \left(\bar{\alpha}_k\mathbf{x}_0\mathbf{x}_0^\top + \left(1-\bar{\alpha}_k\right)\mathbf{I}\right) p\left(\mathbf{x}_0 \mid \mathbf{y}\right) d\mathbf{x}_0 - \bar{\alpha}_k \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top \\
&= \left(1-\bar{\alpha}_k\right)\mathbf{I} + \bar{\alpha}_k\left(\operatorname{Cov}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right] + \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top\right) - \bar{\alpha}_k \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]^\top \\
&= \left(1-\bar{\alpha}_k\right)\mathbf{I} + \bar{\alpha}_k \operatorname{Cov}_{p(\mathbf{x}_0|\mathbf{y})}\left[X_0\right]. \qquad (18)
\end{aligned}$$

Suppose that $p(\mathbf{x}_k|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, i.e., $\nabla_{x_k}\log p(\mathbf{x}_k|\mathbf{y}) = -\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k)$. We can obtain an approximation of the posterior mean $\boldsymbol{\mu}_{k-1}(\mathbf{x}_k, \mathbf{y})$ from (16):

$$\begin{aligned}
\boldsymbol{\mu}_{k-1}(\mathbf{x}_k, \mathbf{y}) &= \frac{1}{\sqrt{\alpha_k}}(\mathbf{x}_k + (1-\alpha_k)\nabla_{\mathbf{x}_k}\log p(\mathbf{x}_k|\mathbf{y})) \\
&= \frac{1}{\sqrt{\alpha_k}}(\mathbf{x}_k - (1-\alpha_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k)) \\
&= \frac{1}{\sqrt{\alpha_k}}(\mathbf{I} - (1-\alpha_k)\boldsymbol{\Sigma}_k^{-1})\mathbf{x}_k + \frac{1}{\sqrt{\alpha_k}}(1-\alpha_k)\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\
&= \frac{1}{\sqrt{\alpha_k}}\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] - (1-\alpha_k)\mathbf{I}\big)\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbf{x}_k \\
&\quad + \frac{1}{\sqrt{\alpha_k}}(1-\alpha_k)\boldsymbol{\Sigma}_k^{-1}\sqrt{\bar{\alpha}_k}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= \frac{1}{\sqrt{\alpha_k}}\alpha_k\big((1-\bar{\alpha}_{k-1})\mathbf{I} + \bar{\alpha}_{k-1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbf{x}_k \\
&\quad + (1-\alpha_k)\boldsymbol{\Sigma}_k^{-1}\sqrt{\bar{\alpha}_{k-1}}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= \big((1-\bar{\alpha}_{k-1})\mathbf{I} + \bar{\alpha}_{k-1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\sqrt{\alpha_k}\mathbf{x}_k \\
&\quad + (1-\alpha_k)\boldsymbol{\Sigma}_k^{-1}\sqrt{\bar{\alpha}_{k-1}}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],
\end{aligned}$$

which is the mean derived in [29].

We can also utilize the Gaussian assumption to compute the posterior distribution $p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y})$ with the following lemma.

Lemma 3. *[2, Section 2.3.3] Given a marginal Gaussian distribution for X and a conditional Gaussian distribution for Y given X in the form*

$$\begin{gathered}
p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right), \\
p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right),
\end{gathered}$$

the marginal distribution of Y and the conditional distribution of X given Y are given by

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\right)$$
$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right),$$

where

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A}\right)^{-1}.$$

Suppose that $p(\mathbf{x}_k|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, i.e., $\nabla_{\mathbf{x}_k} \log p(\mathbf{x}_k|\mathbf{y}) = -\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k)$. Together with $p(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\sqrt{1-\beta_{k+1}}\mathbf{x}_k, \beta_{k+1}\mathbf{I})$, we can directly obtain that $p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y})$ is a Gaussian with mean $\boldsymbol{\mu}_k(\mathbf{x}_{k+1}, \mathbf{y}) = \mathbf{U}_k\mathbf{x}_{k+1} + \mathbf{V}_k\mathbf{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]$ and variance $\mathbf{C}_k$ where

$$\begin{aligned}
\mathbf{U}_k &:= \sqrt{\alpha_{k+1}}\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\big((1-\bar{\alpha}_{k+1})\mathbf{I} + \bar{\alpha}_{k+1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} \\
&= (1 - \frac{1}{2}\beta_{k+1})\boldsymbol{\Sigma}_k\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-1} && (19) \\
\mathbf{V}_k &:= \sqrt{\bar{\alpha}_k}(1-\alpha_{k+1})\big((1-\bar{\alpha}_{k+1})\mathbf{I} + \bar{\alpha}_{k+1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} \\
&= \beta_{k+1}\sqrt{\bar{\alpha}_k}\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-1} && (20) \\
\mathbf{C}_k &:= \frac{\beta_{k+1}}{1-\beta_{k+1}}\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\Big(\big(\frac{\beta_{k+1}}{1-\beta_{k+1}} + 1 - \bar{\alpha}_k\big)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\Big)^{-1} \\
&= \beta_{k+1}\boldsymbol{\Sigma}_k\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-1}. && (21)
\end{aligned}$$

For each parameter $\mathbf{U}_k$, $\mathbf{V}_k$, and $\mathbf{C}_k$, the first expression is used in [29]. The second expression is equivalent when considering the equivalent infinitesimal $\sqrt{1-\beta_k} = 1 - \frac{1}{2}\beta_k$ as $T \to \infty$, and will be used in the following proofs for convenience.

B. Proof of Theorem 1

Since $p_\lambda(\mathbf{x}_{T-1}|\mathbf{x}_T, \mathbf{y}) = \mathcal{N}\Big(\mathbf{U}_{T-1}\mathbf{x}_T + \mathbf{V}_{T-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0], \lambda\mathbf{C}_{T-1}\Big)$ and $p_\lambda(\mathbf{x}_T|\mathbf{y}) = \mathcal{N}(0, \mathbf{I})$, from Lemma 3 we have that

$$p_\lambda(\mathbf{x}_{T-1}|\mathbf{y}) = \mathcal{N}\Big(\mathbf{V}_{T-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],\ \lambda\mathbf{C}_{T-1} + \mathbf{U}_{T-1}\mathbf{U}_{T-1}^\top\Big).$$

By simplifying the mean and variance, we have that

$$\begin{aligned}
\boldsymbol{\mu}_{T-1}^\lambda &= \mathbf{V}_{T-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0], \\
\boldsymbol{\Sigma}_{T-1}^\lambda &= \lambda\mathbf{C}_{T-1} + \mathbf{U}_{T-1}^\top\mathbf{U}_{T-1} \\
&= \lambda\beta_T\boldsymbol{\Sigma}_{T-1}\big((1-\beta_T)\boldsymbol{\Sigma}_{T-1} + \beta_T\mathbf{I}\big)^{-1} + (1-\beta_T)\boldsymbol{\Sigma}_{T-1}\big((1-\beta_T)\boldsymbol{\Sigma}_{T-1} + \beta_T\mathbf{I}\big)^{-1}\big((1-\beta_T)\boldsymbol{\Sigma}_{T-1} + \beta_T\mathbf{I}\big)^{-\top}\boldsymbol{\Sigma}_{T-1}^\top \\
&= \lambda\beta_T\boldsymbol{\Sigma}_{T-1}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} + \boldsymbol{\Sigma}_{T-1}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} \\
&\qquad \cdot \big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-\top}\big((\alpha_T - \bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^\top && (22) \\
&= (\lambda\beta_T + \alpha_T)\boldsymbol{\Sigma}_{T-1}, \quad \text{as } \bar{\alpha}_T \to 0,
\end{aligned}$$

where (22) follows from $\boldsymbol{\Sigma}_{T-1} = (1-\bar{\alpha}_{T-1})\mathbf{I} + \bar{\alpha}_{T-1}\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[\mathbf{x}_0]$.

Then at time $T-2$, since $p_\lambda(\mathbf{x}_{T-2}|\mathbf{x}_{T-1}, \mathbf{y}) = \mathcal{N}\Big(\mathbf{U}_{T-2}\mathbf{x}_{T-1} + \mathbf{V}_{T-2}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0], \lambda\mathbf{C}_{T-2}\Big)$, we have that

$$\begin{aligned}
p_\lambda(\mathbf{x}_{T-2}|\mathbf{y}) &= \mathcal{N}\Big((\mathbf{U}_{T-2}\mathbf{V}_{T-1} + \mathbf{V}_{T-2})\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],\ \lambda\mathbf{C}_{T-2} + \mathbf{U}_{T-2}(\lambda\mathbf{C}_{T-1} + \mathbf{U}_{T-1}\mathbf{U}_{T-1}^\top)\mathbf{U}_{T-2}^\top\Big) \\
&= \mathcal{N}\Big((\mathbf{U}_{T-2}\mathbf{V}_{T-1} + \mathbf{V}_{T-2})\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],\ \lambda\mathbf{C}_{T-2} + \mathbf{U}_{T-2}(\lambda\beta_T + \alpha_T)\boldsymbol{\Sigma}_{T-1}\mathbf{U}_{T-2}^\top\Big),
\end{aligned}$$

and we can further simplify the mean and variance as

$$
\begin{aligned}
\boldsymbol{\mu}_{T-2}^{\lambda} &= (\mathbf{U}_{T-2}\mathbf{V}_{T-1} + \mathbf{V}_{T-2})\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= \Big(\sqrt{\alpha_{T-1}}\sqrt{\bar{\alpha}_{T-1}}(1-\alpha_T)\big((1-\bar{\alpha}_{T-2})\mathbf{I} + \bar{\alpha}_{T-2}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} \\
&\qquad\qquad + \sqrt{\bar{\alpha}_{T-2}}(1-\alpha_{T-1})\mathbf{I}\Big) \cdot \big((1-\bar{\alpha}_{T-1})\mathbf{I} + \bar{\alpha}_{T-1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{T-1}\alpha_T)\sqrt{\bar{\alpha}_{T-2}}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{T-1}\alpha_T)\sqrt{\bar{\alpha}_{T-2}}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \quad \text{as } \bar{\alpha}_T \to 0, \\
\boldsymbol{\Sigma}_{T-2}^{\lambda} &= \lambda\mathbf{C}_{T-2} + \mathbf{U}_{T-2}(\lambda\beta_T + \alpha_T\boldsymbol{\Sigma}_{T-1})\mathbf{U}_{T-2}^{\top} \\
&= \lambda\beta_{T-1}\boldsymbol{\Sigma}_{T-2}\big((1-\beta_{T-1})\boldsymbol{\Sigma}_{T-2} + \beta_{T-1}\mathbf{I}\big)^{-1} + (1-\beta_{T-1})\boldsymbol{\Sigma}_{T-2}\big((1-\beta_{T-1})\boldsymbol{\Sigma}_{T-2} + \beta_{T-1}\mathbf{I}\big)^{-1}(\lambda\beta_T + \alpha_T)\boldsymbol{\Sigma}_{T-1} \\
&\qquad \cdot \big((1-\beta_{T-1})\boldsymbol{\Sigma}_{T-2} + \beta_{T-1}\mathbf{I}\big)^{-\top}\boldsymbol{\Sigma}_{T-2}^{\top} \\
&= \boldsymbol{\Sigma}_{T-2}\big(\lambda\beta_{T-1}\boldsymbol{\Sigma}_{T-1}^{-1} + (1-\beta_{T-1})(\lambda + (1-\lambda)\alpha_T)\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{T-2}\big) \\
&= \boldsymbol{\Sigma}_{T-2}\big(\lambda\boldsymbol{\Sigma}_{T-1}^{-1}(\beta_{T-1}\mathbf{I} + (1-\beta_{T-1})\boldsymbol{\Sigma}_{T-2} - \alpha_{T-1}\alpha_T\boldsymbol{\Sigma}_{T-2}) + \alpha_{T-1}\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{T-2}\big) \\
&= \boldsymbol{\Sigma}_{T-2}\big(\lambda\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{T-1} + (1-\lambda)\alpha_{T-1}\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{T-2}\big) \\
&= \boldsymbol{\Sigma}_{T-2}\big(\lambda\mathbf{I} + (1-\lambda)\alpha_T\alpha_{T-1}\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{T-2}\big), \quad \text{as } \bar{\alpha}_T \to 0.
\end{aligned}
$$

Now, let's prove the general case by induction. For $0 \le k \le T-3$, suppose that the variance of $p_\lambda(\mathbf{x}_{k+1}|\mathbf{y})$ is

$$
\boldsymbol{\Sigma}_{k+1}^{\lambda} = \boldsymbol{\Sigma}_{k+1}\Big(\lambda\mathbf{I} + (1-\lambda)\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{k+1}\Big),
$$

and the expectation is

$$
\begin{aligned}
\boldsymbol{\mu}_{k+1}^{\lambda} &= \Big(\mathbf{U}_{k+1}\big(\mathbf{U}_{k+2}(\cdots(\mathbf{U}_{T-2}\mathbf{V}_{T-1} + \mathbf{V}_{T-2})\cdots) + \mathbf{V}_{k+2}\big) + \mathbf{V}_{k+1}\Big)\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T)\sqrt{\bar{\alpha}_{k+1}}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T)\sqrt{\bar{\alpha}_{k+1}}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0], \quad \text{as } \bar{\alpha}_T \to 0.
\end{aligned}
$$

By Lemma 3 and $p_\lambda(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}) = \mathcal{N}\Big(\mathbf{U}_k\mathbf{x}_{k+1} + \mathbf{V}_k\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],\ \lambda\mathbf{C}_k\Big)$, we have mean and variance of $p_\lambda(\mathbf{x}_k|\mathbf{y})$ as

$$
\begin{aligned}
\boldsymbol{\mu}_{k}^{\lambda} &= \Big(\mathbf{U}_k(1-\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T)\sqrt{\bar{\alpha}_{k+1}}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1} + \mathbf{V}_k\Big)E_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= \sqrt{\bar{\alpha}}_k\Big((1-\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T)\big((1-\bar{\alpha}_k)\mathbf{I} + \bar{\alpha}_k\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big) + (1-\alpha_{k+1})\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\Big) \\
&\qquad \cdot \big((1-\bar{\alpha}_{k+1})\mathbf{I} + \bar{\alpha}_{k+1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= \sqrt{\bar{\alpha}}_k\Big((1-\bar{\alpha}_{k+1})(1-\alpha_{k+1}\alpha_{k+2}\cdots\alpha_T)\mathbf{I} + (1-\alpha_{k+1}\alpha_{k+2}\cdots\alpha_T)\bar{\alpha}_{k+1}\,\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[\mathbf{x}_0]\Big) \\
&\qquad \cdot \big((1-\bar{\alpha}_{k+1})\mathbf{I} + \bar{\alpha}_{k+1}\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{k+1}\alpha_{k+2}\cdots\alpha_T)\sqrt{\bar{\alpha}_k}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)^{-1}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \\
&= (1-\alpha_{k+1}\alpha_{k+2}\cdots\alpha_T)\sqrt{\bar{\alpha}_k}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \quad \text{as } \bar{\alpha}_T \to 0,
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}_{k}^{\lambda} &= \lambda\mathbf{C}_k + \mathbf{U}_k\boldsymbol{\Sigma}_{k+1}^{\lambda}\mathbf{U}_k^{\top} \\
&= \lambda\beta_{k+1}\boldsymbol{\Sigma}_k\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-1} + (1-\beta_{k+1})\boldsymbol{\Sigma}_k\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-1} \\
&\qquad \cdot \boldsymbol{\Sigma}_{k+1}\big(\lambda\mathbf{I} + (1-\lambda)\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_{k+1}\big)\big((1-\beta_{k+1})\boldsymbol{\Sigma}_k + \beta_{k+1}\mathbf{I}\big)^{-\top}\boldsymbol{\Sigma}_k^{\top} \\
&= \boldsymbol{\Sigma}_k\Big(\lambda\beta_{k+1}\boldsymbol{\Sigma}_{k+1}^{-1} + (1-\beta_{k+1})\big(\lambda\boldsymbol{\Sigma}_{k+1}^{-1}\boldsymbol{\Sigma}_k^{\top} + (1-\lambda)\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k^{\top}\big)\Big) \\
&= \boldsymbol{\Sigma}_k\Big(\lambda\boldsymbol{\Sigma}_{k+1}^{-1}\big(\beta_{k+1}\mathbf{I} + (1-\beta_{k+1})\boldsymbol{\Sigma}_k - \alpha_{k+1}\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T\boldsymbol{\Sigma}_{k+1}\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k\big) + \alpha_{k+1}\alpha_{k+2}\alpha_{k+3}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k\Big) \\
&= \boldsymbol{\Sigma}_k\Big(\lambda\big(\mathbf{I} - \alpha_{k+1}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k\big) + \alpha_{k+1}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k\Big) \\
&= \boldsymbol{\Sigma}_k\Big(\lambda\mathbf{I} + (1-\lambda)\alpha_{k+1}\alpha_{k+2}\cdots\alpha_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_k\Big).
\end{aligned}
$$

In particular, when $\bar{\alpha}_T \to 0$, the variance of $p_\lambda(\mathbf{x}_0|\mathbf{y})$ is

$$\boldsymbol{\Sigma}_0^\lambda = \boldsymbol{\Sigma}_0\Big(\lambda\mathbf{I} + (1-\lambda)\bar{\alpha}_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_0\Big) \to \lambda\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0],$$

and the mean is

$$\boldsymbol{\mu}_0^\lambda = (1-\bar{\alpha}_T)\sqrt{\bar{\alpha}_0}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \to \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0].$$

C. Proof of Theorem 2

Optimality: First, we shall show that there is no loss of optimality in assuming that $\hat{X}$ is jointly Gaussian with X given $\mathbf{y}$. Let $\hat{X}_G$ be a random variable with the same first and second-order statistics as $\hat{X}$, and $p_{\hat{X}_G|Y}(\hat{\mathbf{x}}_G|\mathbf{y})$ be a Gaussian distribution, i.e., $p_{\hat{X}_G|Y}(\hat{\mathbf{x}}_G|\mathbf{y}) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_\mathbf{y}, \hat{\boldsymbol{\Sigma}}_\mathbf{y})$. Since the first and second-order statistics are the same, we have $\mathbb{E}[||X - \hat{X}||^2] = \mathbb{E}[||X - \hat{X}_G||^2]$. Meanwhile, by [15, Proposition 1.6.5], $W_2^2(p_{X|Y}(\mathbf{x}|\mathbf{y}), p_{\hat{X}|Y}(\hat{\mathbf{x}}|\mathbf{y})) \geq ||\boldsymbol{\mu}_\mathbf{y} - \hat{\boldsymbol{\mu}}_\mathbf{y}||_2^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y} + \hat{\boldsymbol{\Sigma}}_\mathbf{y} - 2\big(\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\big)^{\frac{1}{2}}) = W_2^2(p_{X|Y}(\mathbf{x}|\mathbf{y}), p_{\hat{X}_G|Y}(\hat{\mathbf{x}}_G|\mathbf{y}))$, where $W_2(p,q)$ denotes the Wasserstein-2 (W2) distance between two distributions p and q.

Thus, we can assume that the construction $\hat{X}$ is jointly Gaussian with X given $\mathbf{y}$. Together with the Markov chain $X - Y - \hat{X}$, i.e., $p_{X,\hat{X}|Y}(\mathbf{x},\hat{\mathbf{x}}|\mathbf{y}) = p_{X|Y}(\mathbf{x}|\mathbf{y})p_{\hat{X}|Y}(\hat{\mathbf{x}}|\mathbf{y})$, the optimization problem (11) in Theorem 2 becomes

$$\begin{aligned} D(P) = \min_{\hat{\boldsymbol{\mu}}_\mathbf{y}, \hat{\boldsymbol{\Sigma}}_\mathbf{y}} & \; ||\boldsymbol{\mu}_\mathbf{y} - \hat{\boldsymbol{\mu}}_\mathbf{y}||_2^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_\mathbf{y}) \\ & \text{s.t. } ||\boldsymbol{\mu}_\mathbf{y} - \hat{\boldsymbol{\mu}}_\mathbf{y}||_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_\mathbf{y} + \hat{\boldsymbol{\Sigma}}_\mathbf{y} - 2\big(\boldsymbol{\Sigma}_y^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) \leq P^2. \end{aligned}$$

Without loss of optimality, we set $\hat{\boldsymbol{\mu}}_\mathbf{y} = \boldsymbol{\mu}_\mathbf{y}$. Consider the KKT condition with dual variable ν:

$$\nabla_{\hat{\boldsymbol{\Sigma}}_\mathbf{y}}\Big(\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_\mathbf{y}) + \nu\big(\mathrm{Tr}\left(\boldsymbol{\Sigma}_\mathbf{y} + \hat{\boldsymbol{\Sigma}}_\mathbf{y} - 2\big(\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right)\big)\Big) = \mathbf{I} + \nu\mathbf{I} - \nu\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\left(\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}} = 0, \tag{23}$$

$$\nu\big(\mathrm{Tr}\left(\boldsymbol{\Sigma}_\mathbf{y} + \hat{\boldsymbol{\Sigma}}_\mathbf{y} - 2\big(\boldsymbol{\Sigma}_y^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) - P^2\big) = 0, \tag{24}$$

$$\nu \geq 0. \tag{25}$$

With (23), we have $\hat{\boldsymbol{\Sigma}}_\mathbf{y} = \big(\frac{\nu}{1+\nu}\big)^2\boldsymbol{\Sigma}_\mathbf{y}$. Plugging in (24), we have

$$\nu\Big(\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + (\frac{\nu}{1+\nu})^2\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) - 2(\frac{\nu}{1+\nu})\mathrm{Tr}\left(\big(\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\boldsymbol{\Sigma}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) - P^2\Big) = \nu\Big(\frac{1}{(1+\nu)^2}\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) - P^2\Big) = 0.$$

When $P > \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})}$, ν should be zero. When $P \leq \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})$, we have $\nu = \sqrt{\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})}{P^2}} - 1$, and the distortion level is

$$D(P) = \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + (\frac{\nu}{\nu+1})^2\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) = \Big(1 + (1 - \sqrt{\frac{P^2}{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})}})^2\Big)\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) = \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + \big(\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})} - P\big)^2.$$

In summary, the optimal conditional distortion-perception tradeoff with MSE and W2 constraint is

$$D(P) = \begin{cases} \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}) + \big(\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})} - P\big)^2, \text{ for } P \leq \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})} \\ \mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y}), \text{ for } P > \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{y})}. \end{cases} \tag{11}$$

Achievability: In Theorem 1, we have shown that when $\bar{\alpha}_T \to 0$, the output distribution $p_\lambda(\mathbf{x}_0|\mathbf{y})$ of the proposed reverse diffusion process (10) is multivariate Gaussian with variance

$$\boldsymbol{\Sigma}_0^\lambda = \boldsymbol{\Sigma}_0\Big(\lambda\mathbf{I} + (1-\lambda)\bar{\alpha}_T\boldsymbol{\Sigma}_{T-1}^{-1}\boldsymbol{\Sigma}_0\Big) \to \lambda\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] = \lambda\boldsymbol{\Sigma}_\mathbf{y},$$

and mean

$$\boldsymbol{\mu}_0^\lambda = (1-\bar{\alpha}_T)\sqrt{\bar{\alpha}_0}\big((1-\bar{\alpha}_T)\mathbf{I} + \bar{\alpha}_T\mathrm{Cov}_{p(\mathbf{x}_0|\mathbf{y})}[X_0]\big)\mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] \to \mathbb{E}_{p(\mathbf{x}_0|\mathbf{y})}[X_0] = \boldsymbol{\mu}_\mathbf{y}.$$

Denote the reconstruction associated with λ as X_0^λ for $0 \leq \lambda \leq 1$, and $p_{X_0^\lambda|Y}(\mathbf{x}_0^\lambda|\mathbf{y}) \triangleq p_\lambda(\mathbf{x}_0^\lambda|\mathbf{y})$. Since both $p_{X_0^\lambda|Y}(\mathbf{x}_0^\lambda|\mathbf{y})$ and $p_{X|Y}(\mathbf{x}|\mathbf{y})$ are Gaussian, the Wasserstein-2 distance for two conditional distributions can be computed as

$$\begin{aligned} W_2^2(p_{X|Y}(\mathbf{x}|\mathbf{y}), p_{X_0^\lambda|Y}(\mathbf{x}_0^\lambda|\mathbf{y})) &= \operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y}) + \lambda \operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y}) - 2\sqrt{\lambda} \operatorname{Tr}((\mathbf{\Sigma}_\mathbf{y}^{\frac{1}{2}} \mathbf{\Sigma}_\mathbf{y} \mathbf{\Sigma}_\mathbf{y}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= (1-\sqrt{\lambda})^2 \operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y}). \end{aligned}$$

For the distortion, we have

$$\begin{aligned} \mathbb{E}_{p_{X_0^\lambda, X|Y}(\mathbf{x}_0^\lambda, \mathbf{x}|\mathbf{y})}[||X_0^\lambda - X||] &= \mathbb{E}_{p_{X_0^\lambda, X|Y}(\mathbf{x}_0^\lambda, \mathbf{x}|\mathbf{y})}[||X||^2 + ||X_0^\lambda||^2 - 2XX_0^\lambda] \\ &= \mathbb{E}_{p_{X|Y}(\mathbf{x}|\mathbf{y})}[||X||^2] + \mathbb{E}_{p_{X_0^\lambda|Y}(\mathbf{x}_0^\lambda|y)}[||X_0^\lambda||^2] - 2\mathbb{E}_{p_{X_0^\lambda|Y}(\mathbf{x}_0^\lambda|y)p_{X|Y}(\mathbf{x}|\mathbf{y})}[XX_0^\lambda] \\ &= \boldsymbol{\mu}_\mathbf{y}^\top \boldsymbol{\mu}_\mathbf{y} + \operatorname{Tr}(\mathbf{\Sigma}_y) + \boldsymbol{\mu}_\mathbf{y}^{\lambda\top} \boldsymbol{\mu}_\mathbf{y}^\lambda + \operatorname{Tr}(\mathbf{\Sigma}_y^\lambda) - 2\boldsymbol{\mu}_\mathbf{y}^\top \boldsymbol{\mu}_\mathbf{y}^\lambda \\ &= (1+\lambda) \operatorname{Tr}(\mathbf{\Sigma}_y^\lambda). \end{aligned}$$

Thus, the conditional distortion-perception tradeoff given by the scaled reverse diffusion process (10) is

$$\begin{aligned} D(\lambda) &= (1+\lambda) \operatorname{Tr}(\mathbf{\Sigma}_y), \\ P^2(\lambda) &= (1-\sqrt{\lambda})^2 \operatorname{Tr}(\mathbf{\Sigma}_y), \end{aligned}$$

which by eliminating λ is equivalent to

$$D(P) = \operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y}) + \left(\sqrt{\operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y})} - P\right)^2, \text{ for } P \leq \sqrt{\operatorname{Tr}(\mathbf{\Sigma}_\mathbf{y})}.$$

Hence, the achieved tradeoff coincides with the optimal tradeoff (11).

D. Derivation of Mixture Gaussian Example

Consider the mixture Gaussian distribution $X_0 \sim p(x_0)$ with two components, where

$$p(x_0) = w_1 \underbrace{\mathcal{N}(\mu, \sigma_1^2)}_{p_1(x_0)} + w_2 \underbrace{\mathcal{N}(\mu, \sigma_2^2)}_{p_2(x_0)}.$$

The noisy observation is obtained by $Y = aX_0 + \sigma_0 \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$, i.e., $p(y|x_0) = \mathcal{N}(ax_0, \sigma_0^2)$. The joint distribution of (Y, X_0) is

$$p(y, x_0) = p(y|x_0)p(x_0) = w_1 \underbrace{\mathcal{N}\left(\begin{bmatrix} y, \\ x_0 \end{bmatrix}; \begin{bmatrix} a\mu_1, \\ \mu_1 \end{bmatrix}, \begin{bmatrix} a^2\sigma_1^2 + \sigma_0^2, & a\sigma_1^2 \\ a\sigma_1^2, & \sigma_1^2 \end{bmatrix} \right)}_{f_1(x_0, y)} + w_2 \underbrace{\mathcal{N}\left(\begin{bmatrix} y, \\ x_0 \end{bmatrix}; \begin{bmatrix} a\mu_2, \\ \mu_2 \end{bmatrix}, \begin{bmatrix} a\sigma_2^2 + \sigma_0^2, & a\sigma_2^2 \\ a\sigma_2^2, & \sigma_2^2 \end{bmatrix} \right)}_{f_2(x_0, y)}.$$

Then the marginal distribution of Y is

$$p(y) = w_1 \underbrace{\mathcal{N}(a\mu_1, a^2\sigma_1^2 + \sigma_0^2)}_{p_1(y)} + w_2 \underbrace{\mathcal{N}(a\mu_2, a^2\sigma_2^2 + \sigma_0^2)}_{p_2(y)}.$$

For component $f_1(x_0, y)$, it is a bivariate Gaussian distribution with marginals as $p_1(x_0) = \mathcal{N}(\mu_1, \sigma_1^2)$, and $p_1(y) = \mathcal{N}(a\mu_1, a^2\sigma_1^2 + \sigma_0^2)$, with correlation $\rho = \frac{a\sigma_1}{\sqrt{a^2\sigma_1^2 + \sigma_0^2}}$.

Then $f_1(x_0, y)$ can be written as

$$\begin{aligned} f_1(x_0, y) &= \frac{1}{2\pi\sigma_0\sigma_1} \exp\Big(- \frac{a^2\sigma_1^2 + \sigma_0^2}{2\sigma_0^2} \Big[(\frac{x-\mu_1}{\sigma_1})^2 - 2\frac{a\sigma_1}{\sqrt{a^2\sigma_1^2 + \sigma_0^2}} (\frac{x-\mu_1}{\sigma_1})(\frac{y-a\mu_1}{\sqrt{a^2\sigma_1^2 + \sigma_0^2}}) + (\frac{y-a\mu_1}{\sqrt{a^2\sigma_1^2 + \sigma_0^2}})^2 \Big] \Big) \\ &= p_1(x_0|y)p_1(y), \end{aligned}$$

where $p_1(x_0|y) = \mathcal{N}\big((\frac{\mu_1}{\sigma_1^2}+\frac{ay}{\sigma_0^2})/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}), 1/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2})\big)$ and $p_1(y) = \mathcal{N}(a\mu_1, a\sigma_1^2+\sigma_0^2)$. Similarly, we can write $f_2(x_0, y)$ as $p_2(x_0|y)p_2(y)$ where, $p_2(x_0|y) = \mathcal{N}\big((\frac{\mu_2}{\sigma_2^2}+\frac{ay}{\sigma_0^2})/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_2^2}), 1/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_2^2})\big)$, and $p_1(y) = \mathcal{N}(a\mu_2, a\sigma_2^2+\sigma_0^2)$.

Then, the posterior distribution of x_0 given y can be computed as

$$\begin{aligned} p(x_0|y) &= \frac{p(x_0,y)}{p(y)} = \frac{w_1 f_1(x_0,y) + w_2 f_2(x_0,y)}{w_1 p_1(y) + w_2 p_2(y)} \\ &= \frac{w_1 p_1(x_0|y)p_1(y) + w_2 p_2(x_0|y)p_2(y)}{w_1 p_1(y) + w_2 p_2(y)} \\ &= \underbrace{\frac{w_1 p_1(y)}{w_1 p_1(y) + w_2 p_2(y)}}_{a_1(y)} p_1(x_0|y) + \underbrace{\frac{w_2 p_2(y)}{w_1 p_1(y) + w_2 p_2(y)}}_{a_2(y)} p_2(x_0|y) \\ &= a_1(y)\mathcal{N}\Big(\frac{\frac{\mu_1}{\sigma_1^2}+\frac{ay}{\sigma_0^2}}{\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}}, \frac{1}{\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}}\Big) + a_2(y)\mathcal{N}\Big(\frac{\frac{\mu_2}{\sigma_2^2}+\frac{ay}{\sigma_0^2}}{\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}}, \frac{1}{\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}}\Big). \end{aligned}$$

Thus, the MMSE estimator is $\mathbb{E}_{p(x_0|y)}[X_0] = a_1(y)(\frac{\mu_1}{\sigma_1^2}+\frac{ay}{\sigma_0^2})/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2}) + a_2(y)(\frac{\mu_2}{\sigma_2^2}+\frac{ay}{\sigma_0^2})/(\frac{a^2}{\sigma_0^2}+\frac{1}{\sigma_1^2})$.

E. Experimental Details and More Experimental Results

1. Experimental Details

1.1 Two-dimensional datasets

Here, we list the architecture design and choices of hyperparameters for the two-dimensional datasets.

Network architecture: We use a simple architecture modified from [4]. For the score network, the input point $\mathbf{x}$ and the time index k are fed to an MLP Block, respectively, where each MLP Block is a multilayer perceptron network. Then, we concatenate the outputs of two MLP Blocks and then feed the concatenated output into a third MLP Blocks. For PSCGAN, the generator of CGAN is also built upon MLP Blocks. Specifically, the noisy observation $\mathbf{y}$ and initial noise $\mathbf{z}$ are fed to an MLP Block respectively, and the concatenated output is fed to another MLP Block. The discriminator of CGAN involves five linear layers, and leaky Relu is used for the activation function. Note that the number of parameters for the generator and discriminator are 25682 and 25025, respectively. The total number of parameters for the score network is 26498.

Choices of hyperparameters: We set $T = 1000$ and a linear schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. Meanwhile, $\tilde{\sigma}_k$ is set to be β_k. For pinwheel dataset, the $\zeta_{k,\lambda}$ is set to be $1.2 + 1.8\lambda$, and for S-curve and moon datasets, $\zeta_{k,\lambda}$ is set to be $1 + 1\lambda$ for all $k = 0, 1, \cdots, T$. For PSCGAN, we follow the setup shown in the original paper [14]

All experiments for two-dimensional datasets were conducted on a single NVIDIA RTX A6000 GPU.

1.2 FFHQ dataset

Here we list the choices of hyperparameters for the FFHQ dataset. Note that the score network for our sampling method was taken from [5], which was trained from scratch using 49k training data for 1M steps. The pre-trained model for PSCGAN is taken from the original paper [14]

Choices of hyperparameters: We set $T = 1000$ and a linear schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. Meanwhile, $\tilde{\sigma}_k$ is set to be $\beta_k \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}$. The choices of $\{\zeta_{k,\lambda}\}_{k=0}^T$ are heuristic and may be slightly different for different devices to get the best results. Recall that $\{\zeta_{k,\lambda}\}_{k=0}^T$ control the weight of the conditional score. Theoretically, if we directly follow Bayes' rule and set the weight of $\nabla_{\mathbf{x}_k} p(\mathbf{x}_k)$ and $\nabla_{\mathbf{x}_k} p(\mathbf{y}|\mathbf{x}_k)$ to be equal, we can obtain the theoretical value as $\zeta_k^{'} = \frac{1-\alpha_k}{2\sqrt{\alpha_k}\sigma_n^2}$. However, the choice of $\zeta_k^{'}$ is not practical. Since $s_\theta(\mathbf{x}_k, k)$ is usually much larger than $\hat{c}(\hat{\mathbf{x}}_0)$ in Algorithm 1, $\zeta_k^{'}$ is too small to reflect information on the conditional score properly. Thus, we still use the heuristic choices of $\zeta_{k,\lambda}$.

In general, for small λ's (e.g., ≤ 0.6), the $\{\zeta_{k,\lambda}\}_{k=0}^T$ need to be set large to get good reconstruction, while for λ close to 1, small $\{\zeta_{k,\lambda}\}_{k=0}^T$ leads to better images. Large $\{\zeta_{k,\lambda}\}_{k=0}^T$ for $\lambda > 0.7$ would result in degraded reconstructions. The possible reason is that for small λ (with less stochasticity), the conditional information becomes more important in constructing a good image, leading to a greater reliance on the conditional score. When λ is large, too much conditional information may conflict with the great stochasticity. In this paper, we mainly focus on tuning $\{\zeta_{k,\lambda}\}_{k=0}^T$ as a function of λ. Thus, $\{\zeta_{k,\lambda}\}_{k=0}^T$

Table 1. Choices of $\{\zeta_{k,\lambda}\}_{k=0}^{T}$ on Gaussian deblurring task with $\sigma_n = 0.3$ for discrete λ's.

$\lambda=0$	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$	$\lambda=1$
39	24	24	26	26	40	22	18	12	12	6

Table 2. Choices of $\{\zeta_{k,\lambda}\}_{k=0}^{T}$ on Gaussian deblurring task with $\sigma_n = 0.5$ for discrete λ's.

$\lambda=0$	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$	$\lambda=1$
33	33	33	37	40	40	33	23	15	10	6.5

Table 3. Choices of $\{\zeta_{k,\lambda}\}_{k=0}^{T}$ on super-resolution task with scale factor 8 for discrete λ's.

$\lambda=0$	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$	$\lambda=1$
26	24	24	24	30	24	20	15	12	12	10

is a constant for all k and $\mathbf{y}$. It is possible to further tune the parameters as a function of k or $\|\mathbf{y} - \mathcal{A}(\mathbf{x}_0)\|_2^2$ [5]. In practice, the choices in Table 1 , 2 and 3 could be considered for discrete $\lambda \in \{0, 0.1, \cdots, 1\}$.

For PSCGAN and DiffPIR, we use the hyperparameters according to the suggested values in the respective papers. All experiments are conducted on a single NVIDIA A100 GPU.

2. More Experimental Results

2.1 Two-dimensional datasets

We provide additional experiments on two-dimensional datasets, including more data distributions and validation of adjusting the variance scale.

More data distributions: Other than pinwheel data points shown in Section 4.1, we illustrate the results on S-curve and moon-type data distributions. Fig. 9 shows the original distributions, noisy distributions, as well as the reconstructions for each dataset. The numerical DP tradeoffs are depicted in Fig. 10. Similar to the pinwheel case, our score-based method achieves a much larger range of tradeoffs compared to the GAN-based approach, revealing great effectiveness and optimality.

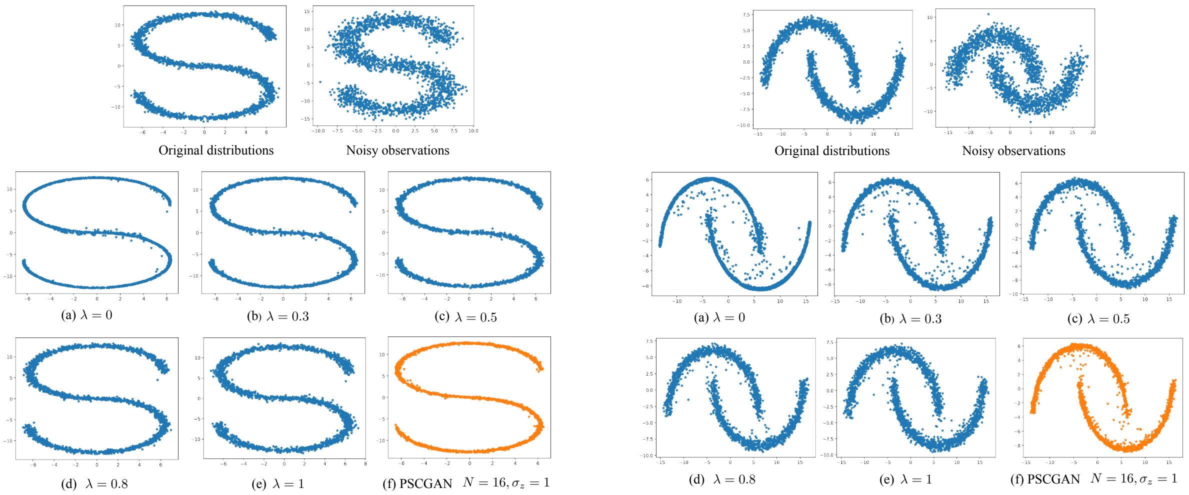
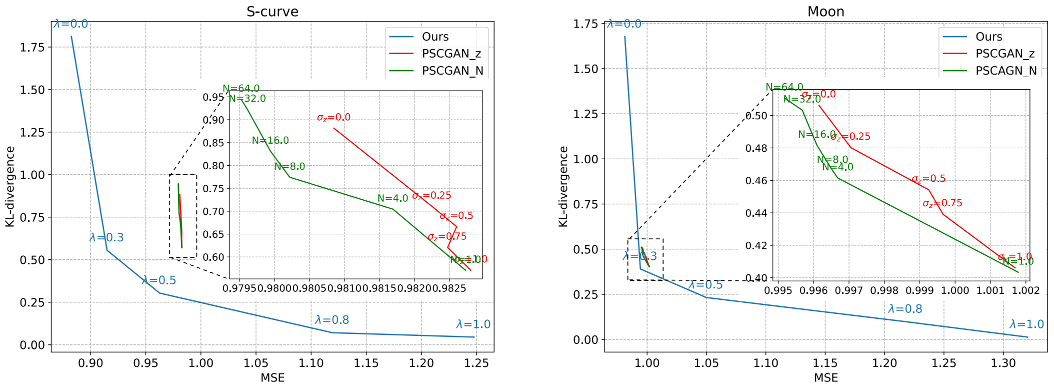

Original distributions
Noisy observations
(a) $\lambda = 0$
(b) $\lambda = 0.3$
(c) $\lambda = 0.5$
(d) $\lambda = 0.8$
(e) $\lambda = 1$
(f) PSCGAN $N = 16, \sigma_z = 1$
Original distributions
Noisy observations
(a) $\lambda = 0$
(b) $\lambda = 0.3$
(c) $\lambda = 0.5$
(d) $\lambda = 0.8$
(e) $\lambda = 1$
(f) PSCGAN $N = 16, \sigma_z = 1$

Figure 9. Experiments on the S-curve dataset (left) and Moon dataset (right). The first row illustrates the original distribution and the noisy observation Y, given by $Y = aX + N$ for $N \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$ (right). The second and third row shows the reconstructions on each dataset: (a)-(e) variance-scaled reverse diffusion process with different λ's; (f) PSCGAN with $N = 16, \sigma_z = 1$.



(a) S-curve

(b) Moon

Figure 10. DP tradeoff on S-curve (left) and moon-type (right) datasets traversed by our variance-scaled reverse diffusion process and PSCGAN.

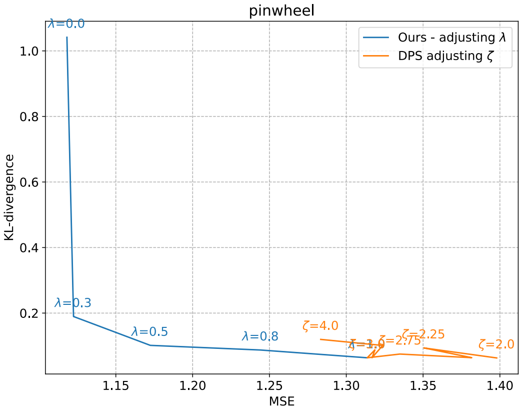


Figure 11. DP tradeoff on S-curve (left) and moon-type (right) datasets traversed by our variance-scaled reverse diffusion process and PSCGAN.

Validation of adjusting the variance scale: In the original DPS sampling procedure [7, Algorithm 1], there is a hyperparameter ζ_k controlling the weight that is given to the likelihood $\nabla_{\mathbf{x}_k} ||y - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_k))||_2^2$, which may also affect the distortion-perception performance. Theorem 1 and 1 show that the proposed variance-scaled diffusion process serves as the optimal solution to the DP tradeoff for conditional multivariate Gaussian. In contrast, there is no theoretical guarantee that adjusting the DPS weight ζ_l in Algorithm 1 can traverse the optimal DP tradeoff. We conduct a simple experiment on the pinwheel dataset, which compares the performance of the proposed variance-scaled reverse diffusion process and the DPS sampling procedure with adjusted ζ_k. Fig. 11 demonstrates that adjusting ζ_k for fixed $\lambda = 1$ is inferior to our variance-scaled method and unable to traverse the tradeoff.

2.2 FFHQ dataset

We provide more experimental results on the FFHQ dataset, including the effect of increasing stochasticity, more metrics, and more examples.

Increasing stochasticity: It is observed in the mixture Gaussian example (Section 3.4) that for $\lambda = 0$, the trajectories are deterministic and converge to the MMSE point given an initial x_T. When λ increases, the generated trajectories follow the form of the posterior distribution and show more stochasticity. This phenomenon can also be observed in real-world datasets. As shown in Fig. 12, the reconstructions show more stochasticity with λ increasing. Specifically, details such as hairs, eye expressions, and the shape of the mouth exhibit more variations. The images become sharper with the increase in MSE.

More metrics: We report more metrics of Gaussian deblur task with additive noise of $\sigma_n = 0.3$ on FFHQ dataset,

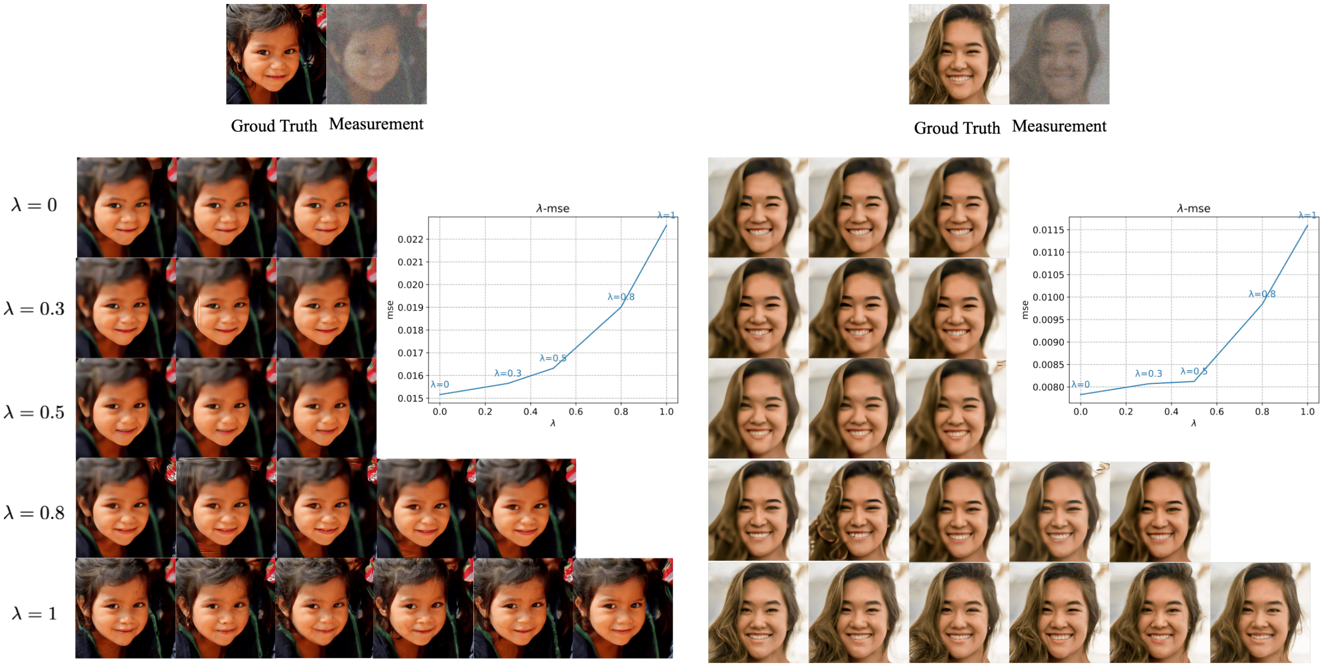


Figure 12. Multiple samples with different λ's. As λ increases, the reconstructions show more stochasticity, and the MSE increases.

Metrics	Ours					PSCGAN		DiffPIR
	$\lambda = 0$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.8$	$\lambda = 1$	$N = 1$	$N = 64$	
PSNR↑	25.27	24.93	24.80	24.47	24.40	22.10	24.39	22.73
LPIPS↓	0.368	0.337	0.329	0.312	0.263	0.304	0.350	0.262

Table 4. Quantitative evaluation (PSNR, LPIPS) of Gaussian deblur task with additive noise of $\sigma_n = 0.3$ on FFHQ dataset.

including PSNR for distortion and LPIPS [30] for perception measure. It can be shown in Table 4 that when λ increases, PSNR becomes worse while LPIPS becomes better. This phenomenon coincides with the results of MSE and FID, indicating that the proposed method can effectively traverse the tradeoff between distortion and perception.

More examples for different tasks: Fig. 13 shows more samples from the FFHQ dataset on the Gaussian deblurring task. We test the methods on different noise levels. Note that the PSCGAN is trained on $\sigma_n = 0.3$. We can see that with a single score network, our method can robustly traverse DP on different noise levels. The PSCGAN trained on $\sigma_n = 0.3$ fails to generate valid images when $\sigma_n = 0.5$. More examples of the super-resolution task are shown in Fig. 14.

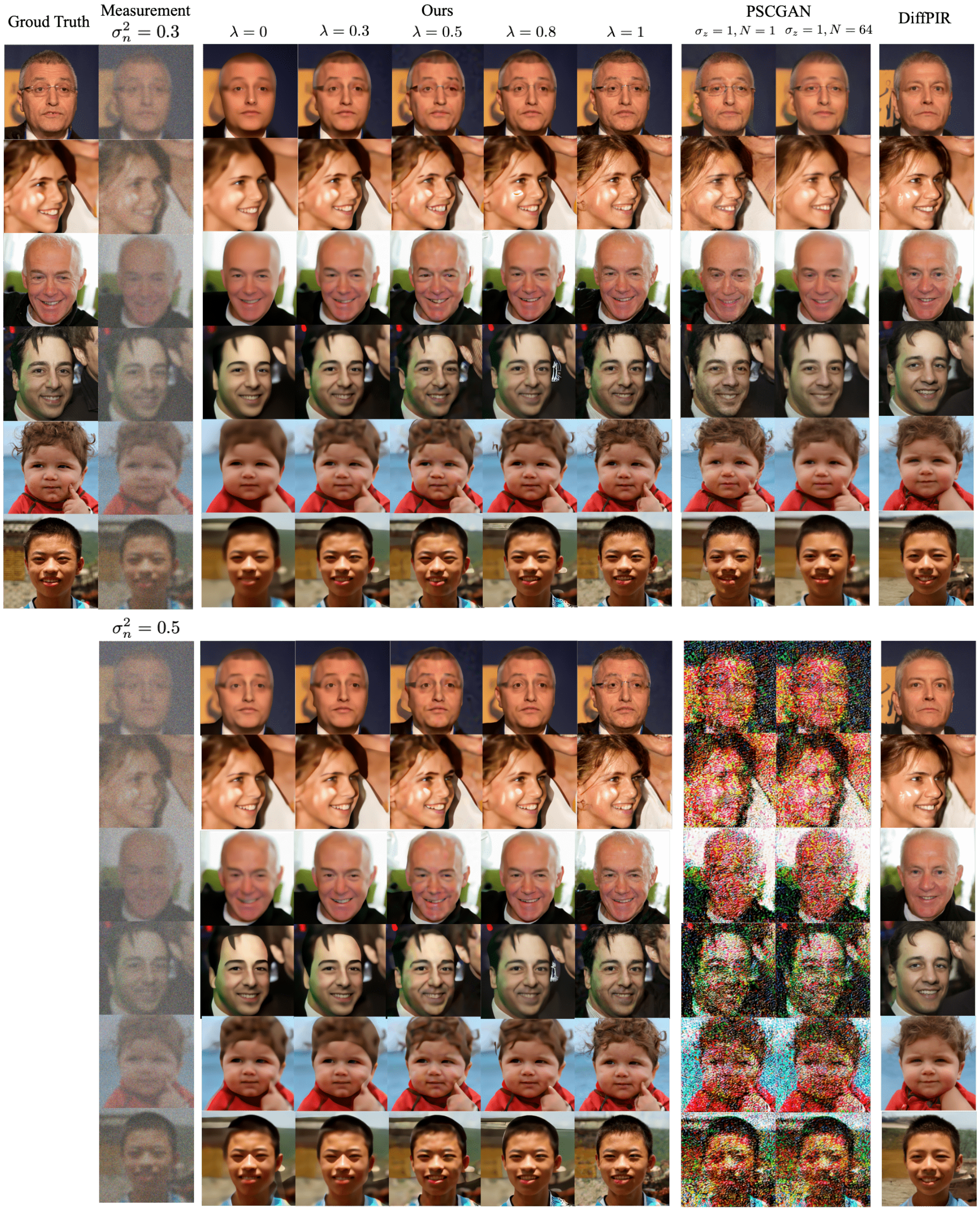


Figure 13. More examples on FFHQ dataset of Gaussian deblurring for both $\sigma_n = 0.3$ and $\sigma_n = 0.5$. Note that for each method, we use the same pre-trained model for both noise levels. With a single score network, our method can robustly traverse DP on different noise levels. The PSCGAN trained on $\sigma_n = 0.3$ fails to generate valid images when $\sigma_n = 0.5$.

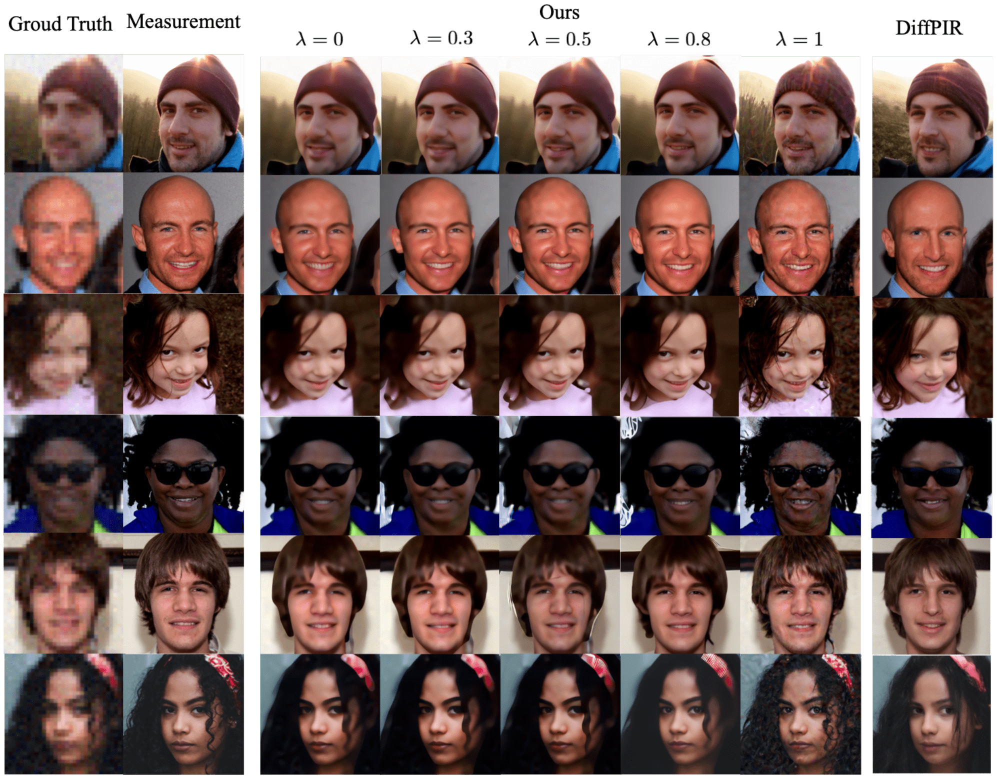


Figure 14. More examples on FFHQ dataset of super-resolution with downsampling scale 8.