Guidance with Spherical Gaussian Constraint for Conditional Diffusion

Lingxiao Yang [1 2] **Shutong Ding** [1 2] **Yifan Cai** [1 2] **Jingyi Yu** [1 2] **Jingya Wang** [1 2] **Ye Shi** [1 2 *]

Abstract

Recent advances in diffusion models attempt to handle conditional generative tasks by utilizing a differentiable loss function for guidance without the need for additional training. While these methods achieved certain success, they often compromise on sample quality and require small guidance step sizes, leading to longer sampling processes. This paper reveals that the fundamental issue lies in the manifold deviation during the sampling process when loss guidance is employed. We theoretically show the existence of manifold deviation by establishing a certain lower bound for the estimation error of the loss guidance. To mitigate this problem, we propose Diffusion with Spherical Gaussian constraint (DSG), drawing inspiration from the concentration phenomenon in high-dimensional Gaussian distributions. DSG effectively constrains the guidance step within the intermediate data manifold through optimization and enables the use of larger guidance steps. Furthermore, we present a closed-form solution for DSG denoising with the Spherical Gaussian constraint. Notably, DSG can seamlessly integrate as a plugin module within existing training-free conditional diffusion methods. Implementing DSG merely involves a few lines of additional code with almost no extra computational overhead, yet it leads to significant performance improvements. Comprehensive experimental results in various conditional generation tasks validate the superiority and adaptability of DSG in terms of both sample quality and time efficiency.

1. Introduction

The past few years have witnessed great successes of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020b), especially conditional diffusion models (CDMs) (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Rombach et al., 2022; Zhang et al., 2023), for their powerful expressive and re-editing abilities in generative tasks such as image generation (Rombach et al., 2022; Nichol & Dhariwal, 2021; Song & Ermon, 2020; Song et al., 2021), image inpainting (Chung et al., 2022; Lugmayr et al., 2022; Chung et al., 2023), super-resolution (Daniels et al., 2021; Saharia et al., 2022), image editing (Choi et al., 2021; Meng et al., 2021), and human motion generation (Tevet et al., 2022; Zhang et al., 2022).

Generally, there are two principal techniques in conditional diffusion models, known as classifier-guided (Dhariwal & Nichol, 2021) and classifier-free (Ho & Salimans, 2022) diffusion models respectively. However, there still exist challenges in learning cost and model generality since these two methods both need extra training and data to endow the diffusion model with conditional generation ability. Hence, recent advances (Chung et al., 2022; 2023; Zhu et al., 2023; Bansal et al., 2023; Yu et al., 2023; He et al., 2024) in conditional diffusion models develop training-free methods, which only use the off-the-shelf loss guidance during the denoising process. While these training-free methods avoid extra training, they may sacrifice the realism and quality of the generated samples. This weakness is probably due to the deviation between the intermediate data sample x_t from the intermediate data manifold during the denoising process. Due to this issue, existing methods (Chung et al., 2023; Yu et al., 2023; Bansal et al., 2023) usually had to adopt small loss-guided step sizes but this would further increase the number of sampling steps and make their algorithms slow.

In this paper, we reveal the occurrence of manifold deviation during the diffusion sampling process by establishing a certain lower bound for the estimation error of the loss guidance. To address the manifold deviation issue, we propose **D**iffusion with **S**pherical **G**aussian constraint (DSG[1]), a novel training-free method for the conditional diffusion model. The principle idea of DSG is to restrict the guidance step within the intermediate data manifold via the Spherical Gaussian constraint. Concretely, the Spherical Gaussian constraint is a spherical surface determined by the interme-

[1]ShanghaiTech University [2]MoE Key Laboratory of Intelligent Perception and Human Machine Collaboration. Correspondence to: Ye Shi <shiye@shanghaitech.edu.cn>.

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

[1]Code is available at https://github.com/LingxiaoYang2023/DSG2024.

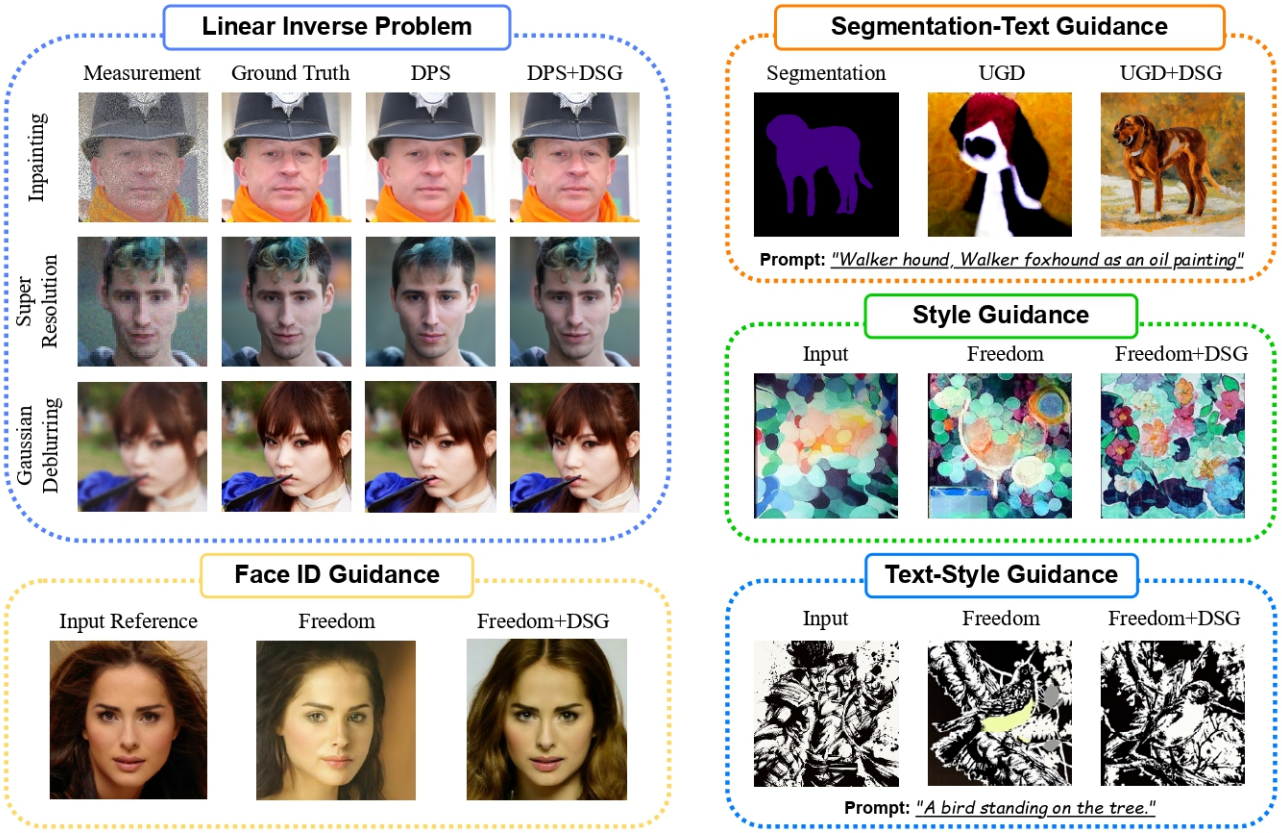

Figure 1. DSG mitigates manifold deviation problem by introducing Spherical Gaussian constraint without relying on the linear manifold assumption. Simultaneously, DSG enables the use of larger step sizes, significantly reducing inference time while enhancing sample quality. The integration of DSG into existing training-free CDMs only incurs a few additional lines of code. Here we briefly show the performance of integrating DSG into recent CDMs, such as DPS (Chung et al., 2023) for Inpainting, Super Resolution, Gaussian Deblurring, UGD (Bansal et al., 2023) for Segmentation-text Guidance, and Freedom (Yu et al., 2023) for Style Guidance tasks. DSG imposes nearly negligible computational overhead while delivering substantial performance enhancements.

diate data manifold, which is the high-confidence region of the unconditional diffusion step. Then, we formulate the calculation of guidance as an optimization problem with the Spherical Gaussian constraint and the guided-loss objective. In that case, the manifold deviation can be avoided and a sufficiently large guidance step size can also be adopted. Besides, we find that this optimization problem can be effectively solved with a closed-form solution. This means DSG imposes almost no extra computational costs in each diffusion step. Hence, our model achieves better performances in both time efficiency and sample quality compared with previous training-free diffusion models. To summarize, our contribution is summarized as follows:

1) Reveal the manifold deviation issue. We illustrate the intermediate data manifold deviation problem that exists widely in previous training-free conditional diffusion models by showing the certain lower bound for the estimation error of the loss guidance, as induced by the Jensen gap.

2) Spherical-Gaussian-constrained guidance. We introduce the Diffusion with Spherical Gaussian Constraint (DSG) method, a plug-and-play module for training-free conditional diffusion models. DSG is designed based on the high-dimensional Gaussian distribution concentration phenomenon, effectively mitigating the manifold deviation issue.

3) A few lines of code to boost performance. We provide a closed-form solution for the DSG denoising process, simplifying the integration of DSG into existing training-free conditional diffusion models to just a few additional lines of code. DSG incurs nearly negligible computational overhead while delivering substantial performance enhancements.

4) Applicable for various tasks. We conduct experiments on various conditional generation tasks, involving Inpainting, Super Resolution, Gaussian Deblurring, Text-Segmentation Guidance, Style Guidance, Text-Style Guidance, and FaceID Guidance. These results robustly confirm the superiority and adaptability of our method in terms of both sample quality and time efficiency.

2. Related Work

In this section, we briefly review the literature on training-required and training-free conditional diffusion models (CDMs).

Training-required CDMs. Generally, training-required CDMs are divided into two principal branches. One of them is the classifier-guided diffusion model (Dhariwal & Nichol, 2021), which trains a time-dependant classifier, based on an off-the-shelf diffusion model, to approximate an unbiased posterior probability $p(y|x_t)$. Another branch is the classifier-free diffusion model (Ho & Salimans, 2022) such as the well-known Stable Diffusion (Rombach et al., 2022) and variants of it (Zhang et al., 2023; Mou et al., 2023). Different from the classifier-guided model, The key idea of it is to utilize the data pairs and train a conditional denoiser directly. While these models can ensure the realism and quality of the generated samples with the given condition, it is obvious that the extra cost of data-pair collection and training time training is unavoidable. This makes it hard to deploy them in some real-world scenarios without extra data pairs or sufficient computation resources.

Training-free CDMs. To overcome the issue, recent research works attempt to handle conditional generation in a training-free manner. Generally, instead of training the time-dependent classifier, they utilize the off-the-shelf classifier guidance to approximate $\nabla_{x_t} \log(y|x_t)$. Specifically, MCG (Chung et al., 2022) approximated the guidance with the aid of Tweedie's formula to solve linear inverse problems, and then this idea is extended to general conditional generation tasks by DPS(Chung et al., 2023), Freedom (Yu et al., 2023), UGD (Bansal et al., 2023). Besides, LGD-MC (Song et al., 2023) attempted to use multiple samples from an inaccurate Gaussian distribution to reduce the estimation bias but it still suffers from the computation cost for Monte Carlo simulation. Other works like UGD, DiffPIR (Bansal et al., 2023; Zhu et al., 2023) performed the guidance step on the clean data sample (i.e., x_0) and then projected the updated clean data sample to the corresponding intermediate data manifold (i.e., x_t). Very recently, a concurrent work called MPGD (He et al., 2024) further studied this idea with an extra auto-encoder via enforcing the guidance step within the tangent space of the clean data manifold. While it makes some improvements, it depends on the strong linear assumption of the data manifold, which does not hold in most real-world applications. Additionally, MPGD heavily relies on pre-trained autoencoders for manifold projections. While it performs well in image generation, it cannot generalize to scenarios where pre-trained autoencoders are not available. In contrast, the proposed DSG method can handle the manifold deviation problem properly without such strong assumptions and improve the performance on both sample quality and time efficiency.

3. Preliminaries

3.1. Diffusion Models

Diffusion models (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020b) are a class of likelihood-based generative models that outperform other generative models(e.g. GAN (Goodfellow et al., 2020), VAE (Kingma & Welling, 2013)) in various tasks. It has predefined forward processes, where Gaussian noise is incrementally added to clean data x_0 until it evolves into pure noise $x_T \in \mathcal{N}(0, I)$. In the formulation of DDPM (Ho et al., 2020), the posterior distribution of any $x_t, t \in [0, T]$ given x_0 is defined as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)I). \tag{1}$$

Here $\{\alpha_0, \cdots, \alpha_T\}$ is the predefined noise schedule. The objective is to train a denoiser, denoted as $\epsilon_\theta(x_t, t)$, to approximate Gaussian noise ϵ_t at each time step t. This is achieved by minimizing the loss function L_{simple}, which is equivalent to matching the denoising scores $\nabla_{x_t} \log p_t(x_t)$ (Ho et al., 2020).

$$L_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon_t}[||\epsilon_t - \epsilon_\theta(x_t, t)||^2]. \tag{2}$$

During the sampling stage, the joint distribution $p(x_{0:T})$ in the reverse process can be defined as:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{4}$$

In this work, we employ the update rule of DDIM (Song et al., 2020a) for calculating $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t))$:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right)}_{\text{" predicted } x_0 \text{ "}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t)}_{\text{"direction pointing to } x_t \text{ "}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}, \tag{5}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1}/(1-\alpha_t))}\sqrt{1-\alpha_t/\alpha_{t-1}}$. When $\eta = 1$, the generative process becomes DDPM, while $\eta = 0$ corresponds to deterministic sampling.

3.2. Training-Free Conditional Diffusion Methods

Classifier guidance (Dhariwal & Nichol, 2021) is the first work that utilizes the pre-trained diffusion model in conditional generative tasks. Specifically, considering the Bayes rule $p(x|y) = p(y|x)p(x)/p(y)$, it introduces the given condition with an additional likelihood term $p(x_t|y)$:

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} p(y|x_t), \tag{6}$$

where y represents the condition or measurement. Recently, many works, known as training-free methods, attempt to employ the pre-trained diffusion models for the conditional generation without re-training or fine-tuning the diffusion prior. Without training the time-dependent classifier to estimate $p(y|x_t)$, these current training-free guided diffusion models just need a pre-trained diffusion prior D_θ and a differentiable loss function $L(x_0, y)$ which is defined on the support of x_0 using an off-the-shelf neural network. They use Tweedie's formula to calculate $\nabla_{x_t} \log p(y|x_t)$ by estimating $\hat{x}_0$ based on x_t:

$$\hat{x}_0(x_t) \approx \mathbb{E}[x_0|x_t] = (x_t + b_t^2 \nabla_{x_t} \log p(x_t))/a_t, \quad (7)$$

$$\begin{aligned} \nabla_{x_t} \log p(y|x_t) &\approx \nabla_{x_t} \log p(y|\hat{x}_0(x_t)), \\ &= \gamma \nabla_{x_t} L(\hat{x}_0(x_t), y) \end{aligned} \quad (8)$$

where $x_t \sim \mathcal{N}(a_t x_0, b_t^2 I)$ and γ is the step size. In the formulation of DDPM (Ho et al., 2020), $a_t = \sqrt{\alpha_t}$, $b_t^2 = \sqrt{1-\alpha_t}$. Then they use the estimated likelihood of $L(x_0, y)$ for additional correction step:

$$x_{t-1} = \underbrace{DDIM(x_t, \epsilon_\theta(x_t, t))}_{\text{sampling step}} - \underbrace{\gamma \nabla_{x_t} L(\hat{x}_0(x_t), y)}_{\text{correction step}}. \quad (9)$$

3.3. Low-dimensional Manifold Assumption

Suppose *clean data manifold* $\mathcal{M}_0$ is the set consisting of all data points. Typically, the data are assumed to lie in a low-dimensional space rather than a whole ambient space:

Assumption 3.1. (Low-dimensional Manifold Assumption). The data manifold $\mathcal{M}_0$ lies in the k-dimensional subspace $\mathbb{R}^k$ with $k \ll n$.

Given this assumption, recent work (Chung et al., 2022) has shown that the set of noisy data x_t, a.k.a. *intermediate data manifold* in this paper, is concentrated on a $(n-k)$-dimensional manifold $\mathcal{M}_t : \{x \in \mathbb{R}^n | d(x, \sqrt{\alpha_t}\mathcal{M}) = \sqrt{(1-\alpha_t)(n-k)}\}$.

4. Method

In this section, we will first present the limitations of previous training-free CDMs, including the strong manifold assumption and estimation error of the loss guidance. After that, our proposed algorithm, DSG, will be proposed to effectively address and resolve these issues by applying guidance with the spherical Gaussian constraint.

4.1. Manifold Deviation by Loss Guidance

Although previous training-free work has achieved great success in various image fields due to its plug-and-play characteristics, they will sacrifice the quality of generated samples. There are two main issues for this phenomenon: linear manifold assumption about $\mathcal{M}_0$ and the estimation error of loss guidance induced by the Jensen gap, which leads to the manifold deviation problem.

Linear manifold assumption. Except for the low-dimensional manifold assumption, some of the previous training-free conditional diffusion methods (Chung et al., 2022; He et al., 2024) further require a linear subspace assumption for $\mathcal{M}_0$. MCG (Chung et al., 2022) employed this linear subspace assumption to establish the estimated gradient $\nabla_{x_t} L(\hat{x}_0(x_t), y)$. A concurrent work, MPGD (He et al., 2024), projected the gradient of the loss on $\hat{x}_0$ to the tangent space of M_0 to keep the sample located on the intermediate manifold $\mathcal{M}_t$ for any $t \in [0, T]$. But this projection is also based on the linear manifold assumption. However, the linear manifold assumption is quite strong, so in practice, both MCG and MPGD will inevitably introduce errors in estimating $p(y|x_t)$, leading to deviations of generated samples from the clean data manifold $\mathcal{M}_0$.

Jensen gap. As mentioned in DPS (Chung et al., 2023), the estimation error of loss guidance results in previous works results from the approximation of $\mathbb{E}_{x_0|x_t}[L(x_0)] \approx L\left(\mathbb{E}_{x_0|x_t}[x_0]\right) = L(\hat{x}_0(x_t))$. Actually, the gap between $\mathbb{E}_{x_0|x_t}[L(x_0)]$ and $L(\hat{x}_0(x_t))$ is related to Jensen's inequality, which is known as the Jensen gap.

Definition 4.1. (Jensen gap). Let x be a random variable with distribution $p(x)$. For function f that is convex or non-convex, the Jensen gap is defined as:

$$\mathcal{J}(f, x \sim p(x)) = \mathbb{E}[f(x)] - f(\mathbb{E}[x]), \quad (10)$$

where the expectation is taken over $p(x)$.

Although DPS (Chung et al., 2023) provides an upper bound for the Jensen gap, it depends on the assumption on the finite derivatives of guided loss, which is obviously unavailable at least in linear inverse problems. In fact, the lower bound of this gap is very large in high-dimensional generative tasks, as shown in Proposition 4.2, and it further leads to the serious manifold deviation problem.

Proposition 4.2. *(Lower bound for Jensen gap). For the β-strongly convex function f and the random variable $x \in \mathbb{R}^n \sim \mathcal{N}(\mu, \Sigma)$, we can have the lower bound for the Jensen Gap:*

$$\mathcal{J} \geq \frac{1}{2}\beta \sum_{i=1}^{n} \lambda_i.$$

where $\Sigma = P^T \Lambda P$ via spectral decomposition and $\lambda_1, \cdots, \lambda_n$ are positive diagonal elements in Λ.

Please refer to the Appendix for our proof. In specific, the function f is the off-the-shelf guided loss L and the random variable x is $x_0 \,|x_t$. Notably, here we utilize the assumption on $x_0|x_t$ mentioned in LGD-MC (Song et al., 2023) and

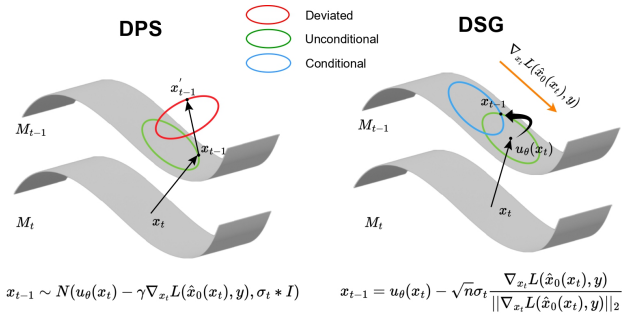

Figure 2. Illustration of the manifold deviation problem in DPS (left) and a schematic overview of how our DSG restricts the guidance within the manifold (right). The red circular ring represents the concentration region of samples under deviated conditional guidance, the blue circular ring indicates the concentration region of samples under accurate conditional guidance, and the green circular ring represents the concentration region of samples without condition.

extend it to general Gaussian distribution for more accurate estimation. Besides, the assumption on strong convexity of L is also rational. On one hand, L is usually a quadratic function in linear inverse problems, which is β-strongly convex. On the other hand, in practical scenarios, L tends to possess at least local β-strong convexity. Considering the probability of Gaussian distribution can be ignored when x is far from $\hat{x}_0$, the lower bound could still hold in general cases. Hence, the Jensen gap increases linearly with the number of sample dimensions, and it will be very large in high-dimensional generative tasks like image generation according to Proposition 4.2. In that case, such a large gap leads to the deviation of x_{t-1} from the intermediate data manifold $\mathcal{M}_{t-1}$, and thus causes the loss of image structural information.

4.2. Diffusion with Spherical Gaussian Constraint

According to our theoretical analysis of previous methods in Section 4.1, both the Jensen gap and the linear manifold assumption will introduce additional errors, leading to a substantial decline in the authenticity of generated samples in comparison to unconditional generation.

Taking an alternative perspective, considering the inevitable error in estimating the conditional bias in every denoising step, why not start from the pre-existing unconditional intermediate manifold $\mathcal{M}_t$ and subsequently find the point closest to the conditional sampling?

To this end, we propose diffusion with spherical Gaussian constraint (DSG), an optimization method that performs guidance within high confidence intervals of unconditional intermediate manifold $\mathcal{M}_t$:

$$\begin{aligned} &\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x' - x_t) \\ &\text{s.t. } x' \in CI_{1-\delta} \end{aligned} \tag{11}$$

Algorithm 1 Diffusion with Spherical Gaussian constraint

Input: pure noise $x_T \sim N(0, I)$, guidance interval i, guidance rate g_r
for $t = T$ **to** 1 **do**
  $\epsilon_t \sim N(0, I)$
  $\hat{x}_0(x_t) = (x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t))/\sqrt{\alpha_t}$
  $\mu_\theta(x_t) = \sqrt{\alpha_{t-1}}\hat{x}_0(x_t) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(\boldsymbol{x}_t, t)$
  // Guidance with Spherical Gaussian Constraint
  $d^* = -\sqrt{n}\sigma_t \cdot \frac{\nabla_{x_t} L(\hat{x}_0(x_t), y)}{||\nabla_{x_t} L(\hat{x}_0(x_t), y)||}$
  $d^{sample} = \sigma_t \epsilon_t$
  $d_m = d^{sample} + g_r(d^* - d^{sample})$
  $x_{t-1} = \mu_\theta(x_t) + r\frac{d_m}{||d_m||}$
end for
Return x_0

where $CI_{1-\delta}$ represents the $(1-\delta)$ confidence intervals for Gaussian distribution in Equation (4).

In this optimization problem, the objective encourages sampling in the direction of gradient descent, while the constraint enforces sampling within high-confidence intervals for a Gaussian distribution.

Nevertheless, when high-confidence intervals encompass substantial regions in n-dimensional space, the optimization problem may appear to be challenging and infeasible. Fortunately, in the case of high-dimensional isotropic Gaussian distribution, where high-confidence intervals are concentrated on a hypersphere, we can simplify the constraint by approximating it with this hypersphere, referred to as the spherical Gaussian constraint.

Proposition 4.3. *(Concentration of high-dimensional isotropy Gaussian distribution). For any n-dimensional isotropy Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2 I)$:*

$$P(||x-\mu||_2^2 \geq x_{lower} = n\sigma^2 + 2n\sigma^2(\sqrt{\epsilon} + \epsilon)) \leq e^{-n\epsilon},$$
$$P(||x-\mu||_2^2 \leq x_{upper} = n\sigma^2 - 2n\sigma^2\sqrt{\epsilon}) \leq e^{-n\epsilon},$$

When n is sufficiently large, it is close to the uniform distribution on the hypersphere of radius $\sqrt{n}\sigma$.

Please refer to Appendix A.2 for further proof. According to Proposition 4.3, since the probability of both $P(||x-\mu||_2^2 \geq x_{lower})$ and $P(||x-\mu||_2^2 \leq x_{upper})$ decrease sharply with the increase of ϵ, the posterior Gaussian distribution in Equation 5 can be approximated to hypersphere $S^n_{\mu,r} = S^n_{\mu_\theta(x_t),\sqrt{n}\sigma} = \{x \in \mathbb{R}^n : ||x - \mu_\theta(x_t)||_2^2 = r^2 = n\sigma_t^2\}$. Hence, the optimization problem can be approximated as:

$$\begin{aligned} &\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x' - x_t) \\ &\text{s.t. } x' \in S^n_{\mu_\theta(x_t),\sqrt{n}\sigma_t} \end{aligned} \tag{12}$$

While this optimization problem is inherently non-convex owing to the spherical Gaussian constraint, we can obtain a

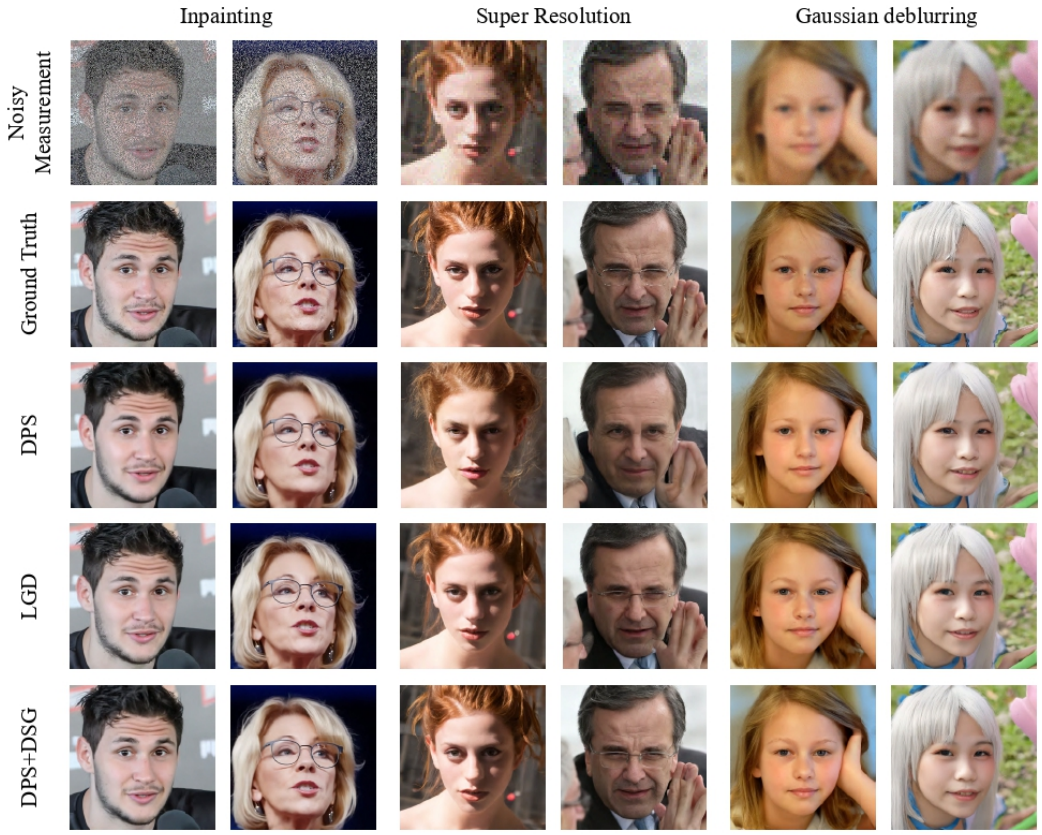

Figure 3. Result in solving three linear inverse problems (Inpainting, Super-resolution, Gaussian deblurring) using DSG using the pre-trained FFHQ diffusion model.

closed-form solution:

$$x^*_{t-1} = \mu_\theta(x_t) - \sqrt{n}\sigma_t \frac{\nabla_{x_t} L(\hat{x}_0(x_t), y)}{||\nabla_{x_t} L(\hat{x}_0(x_t), y)||_2}. \quad (13)$$

Detailed derivation is provided in the Appendix A.3.

Remark 1. It's worth noting that our DSG can be a plugin module to existing training-free CDMs, such as DPS (Chung et al., 2023), UGD (Bansal et al., 2023), Freedom (Yu et al., 2023), etc, while incurring almost no extra computational overhead. Notably, this method only requires modifying a few lines of code to generate more realistic images with accelerated sampling speed. Using DPS (Chung et al., 2023) as an example, we highlight the key difference of code between DPS and our DSG as follows:

```
diff=measure-self.forward(x_0_hat)
dist=norm(diff)
grad=autograd.grad(
         outputs=distance, inputs=x_t)[0]
# DPS: x_t=x_t_mean+sigma_t*eps_t-
    grad*self.gamma
# Ours:
norm=norm(grad, dim=[1, 2, 3])
x_t=x_t_mean-sqrt(n)*sigma_t*grad/norm
```

Remark 2. Another intuitive explanation for the optimal solution obtained is that we directly apply the gradient descent to $\mu_\theta(x_t)$ instead of sampling point x_t. While the standard deviation σ_t of DDIM generally decreases as time t gets smaller, our method can be seen as adaptive gradient descent which decreases fast in the early stage and then coverage in the final stage. In practice, we can use extremely large step sizes (exceed 400× when we use $\eta = 1$ and $t = T$) compared to current training-free methods, thus achieving better alignment. Therefore, our method is robust to smaller DDIM steps while the performance of DPS quickly degrades, which will be evaluated in Sec 5.3.

In practice, when employing DSG to enhance alignment and authenticity, it may sacrifice sample diversity. Therefore, we use the weighted direction of the optimized gradient and unconditional sampling like the Classifier-free (Ho & Salimans, 2022) manner to enhance diversity:

$$d_m = d^{sample} + g_r(d^* - d^{sample}). \quad (14)$$

$$x_{t-1} = \mu_\theta(x_t) + r\frac{d_m}{||d_m||}. \quad (15)$$

Here $d^{sample} = \sigma_t \epsilon_t$ represents the unconditional sampling direction, $d^* = -\sqrt{n}\sigma_t \cdot \frac{\nabla_{x_t} L(\hat{x}_0(x_t),y)}{||\nabla_{x_t} L(\hat{x}_0(x_t),y)||}$ represents the steepest gradient descent direction, and d_m represents the weighted direction which will be scaled to satisfy the Spherical Gaussian constraint. The detailed procedure is shown in Algorithm 1.

5. Experiments

In this section, we evaluate our methods in various tasks, including three linear inverse problems (Inpainting, Super-resolution, Gaussian deblurring), Style Guidance, Text-Segmentation Guidance, FaceID Guidance, and Text-Style Guidance. We demonstrate that our method can plug and play easily into various training-free methods including DPS (Chung et al., 2023), Freedom (Yu et al., 2023), and UGD (Bansal et al., 2023), and improve their performance significantly.

5.1. Implementation Details

Linear Inverse Problems For linear inverse problem $y = Ax + \epsilon$ including Inpainting, Super-resolution, and Gaussian deblurring, we evaluate our method in 1k images of FFHQ256×256 (Karras et al., 2019) and Imagenet256×256 validation dataset (Deng et al., 2009) using pre-trained diffusion models taken from (Chung et al., 2023; Dhariwal & Nichol, 2021). For comparison, we choose MCG (Chung et al., 2022), PnP-ADMM (Chan et al., 2016), Score-SDE (Song et al., 2020b), DDRM (Kawar et al., 2022), DPS(Chung et al., 2023) and LGD (Song et al., 2023) as the baseline. We generate noisy measurements by introducing Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.05)$. The noisy measurements are then acquired through various forward models $A(\cdot)$ customized for specific tasks: (i) In the context of image Inpainting, we randomly mask approximately 92% of the total pixels within the RGB channel. (ii) For Super-resolution, we employ bicubic downsampling to achieve a 4x reduction in resolution. (iii) For Gaussian deblurring, we apply a 61x61 kernel size Gaussian blur with a standard deviation of 3.0. The loss guidance can be expressed as:

$$\text{Loss}(x_0, y) = ||A\hat{x}_0(x_t) - y||_2^2, \tag{16}$$

where y represents the noisy measurement and x_0 donates the image we aim to reconstruct.

Table 1. Quantitative evaluation of the FFHQ dataset in Linear Inverse Problem

Methods	Inpainting		Super resolution		Gaussian deblurring	
	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓
DDRM (Kawar et al., 2022)	0.665	114.9	0.339	59.57	0.427	63.02
MCG (Chung et al., 2022)	0.414	39.19	0.637	144.5	0.550	95.04
PnP-ADMM (Chan et al., 2016)	0.677	114.7	0.433	97.27	0.519	100.6
Score-SDE (Song et al., 2020b) (ILVR (Choi et al., 2021))	0.659	127.1	0.701	170.7	0.667	120.3
DPS (Chung et al., 2023)	0.212	21.19	0.214	39.35	0.257	44.05
LGD (Song et al., 2023)	0.159	28.21	0.231	34.44	0.229	32.57
DPS+DSG(Ours)	**0.115**	**15.77**	**0.211**	**30.30**	**0.208**	**28.22**

Style Guidance For Style Guidance, our test set consists of 1,000 paintings by various artists collected from WikiArt (Saleh & Elgammal, 2015), and we employ LDM (Rombach et al., 2022) as the pre-trained diffusion prior. Freedom (Yu et al., 2023) and LGD (Song et al., 2023) are chosen as the baseline. For an input image x_{in}, the loss guidance can be represented as:

$$\text{Loss}(\hat{x}_0(x_t), x_{in}) = ||E(\hat{x}_0(x_t)) - E(x_{in})||_F^2, \tag{17}$$

where $E(\cdot)$ is the Gram matrix of the 3rd feature map extracted from the CLIP (Radford et al., 2021) image encoder and $||\cdot||_F^2$ denotes the Frobenius norm.

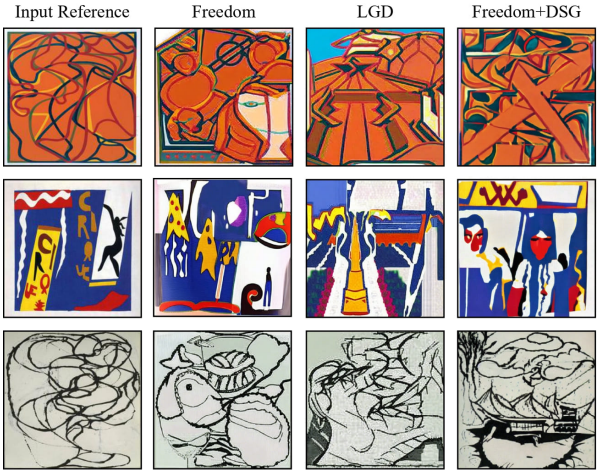

Figure 4. Qualitative results of Style Guidance using pre-trained Stable Diffusion.

Text-Segmentation Guidance In the context of Text-Segmentation Guidance, we apply our method to a more challenging task, incorporating multiple conditions to guide the conditional generation process. For a given text description t and a segmentation map M as input, the objective is to generate an image x_0 that aligns with both the input text t and the provided segmentation mask M within 500 denoising steps. To achieve this, we utilize LDM (Rombach et al., 2022) as the diffusion prior, allowing us to integrate the original text conditions through its conditional denoiser, and use the additional plug-and-play segmentation guidance:

$$\text{Loss}(\hat{x}_0(x_t)), M) = CE(E(\hat{x}_0(x_t)), M). \tag{18}$$

Here $CE(\cdot)$ represents the Cross-Entropy Loss and $E(\cdot)$ represents the MobileNetV3-Large segmentation network.

FaceID Guidance In FaceID Guidance, we utilized a pre-trained diffusion model from CelebA-HQ256*256 provided by (Yu et al., 2023) and employed the CelebA-HQ test set as the input reference image for generating 1k images in 100 denoising steps. For an input reference image I, we use the FaceID loss for guidance:

$$\text{Loss}(\hat{x}_0(x_t)), I) = CE(E(\hat{x}_0(x_t)), E(I)), \tag{19}$$

where $E(\cdot)$ represents ArcFace (Deng et al., 2019), a human face recognition network used to extract features from the facial image.

Text-Style Guidance In Text-Style Guidance, for a given text description t and a reference image I as input, we utilize

Table 2. Quantitative evaluation with different DDIM step in Linear Inverse Problem

Methods	Inpainting			Super resolution			Gaussian deblurring			DDIM steps
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	
DPS (Chung et al., 2023)	0.815	27.73	0.233	0.649	22.99	0.329	0.662	23.43	0.310	100
DPS+DSG(Ours)	**0.899**	**30.95**	**0.132**	**0.784**	**26.82**	**0.227**	**0.762**	**26.16**	**0.239**	100
DPS (Chung et al., 2023)	0.704	22.74	0.314	0.568	19.57	0.405	0.543	18.59	0.428	50
DPS+DSG(Ours)	**0.886**	**30.67**	**0.176**	**0.790**	**26.96**	**0.231**	**0.749**	**25.76**	**0.264**	50
DPS (Chung et al., 2023)	0.408	12.08	0.590	0.337	10.93	0.614	0.353	10.97	0.639	20
DPS+DSG(Ours)	**0.812**	**27.53**	**0.267**	**0.771**	**26.68**	**0.272**	**0.771**	**26.68**	**0.272**	20

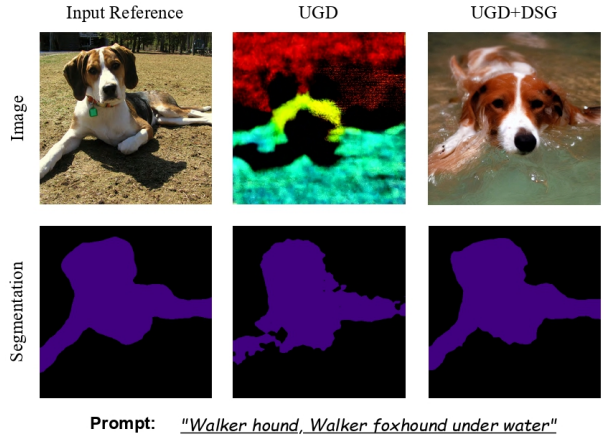

Figure 5. Qualitative result of Text-Segmentation Guidance using 500 denoising steps with Stable Diffusion.

LDM (Rombach et al., 2022) as the pre-trained conditional diffusion prior and utilize the style loss for guidance, which is the same as Style Guidance:

$$\text{Loss}(\hat{x}_0(x_t), x_{in}) = ||E(\hat{x}_0(x_t)) - E(x_{in})||_F^2, \quad (20)$$

Please refer to the Appendix C for further experimental details and additional results for FaceID Guidance and Text-Style Guidance.

5.2. Comparison with the State-of-the-art

The qualitative and quantitative results are shown in Table 1, 3 and Figure 3, 4, 5, 6. The result in Table 1, 3 demonstrates that our method outperforms the current state-of-the-art training-free method by a significant margin. Our method produces samples that exhibit both increased realism and better alignment with the input conditions.

In the context of linear inverse problems, our method excels in preserving intricate details, such as facial features and hair, whereas DPS tends to apply smoothing and blurring effects to these details. With FaceID Guidance, our model generates faces that more closely match the input reference compared to baseline methods. For Style Guidance, Text-Segmentation guidance, and Text-Style Guidance, the images generated by our method also exhibit improved alignment with the input conditions while reducing artifacts.

Table 3. The quantitative result in CelebA-HQ test set for FaceID guidance

Method	FaceID	FID	KID
Freedom	0.545	39.61	0.0165
LGD	0.533	39.13	0.0170
Freedom+DSG	**0.371**	**34.29**	**0.0092**

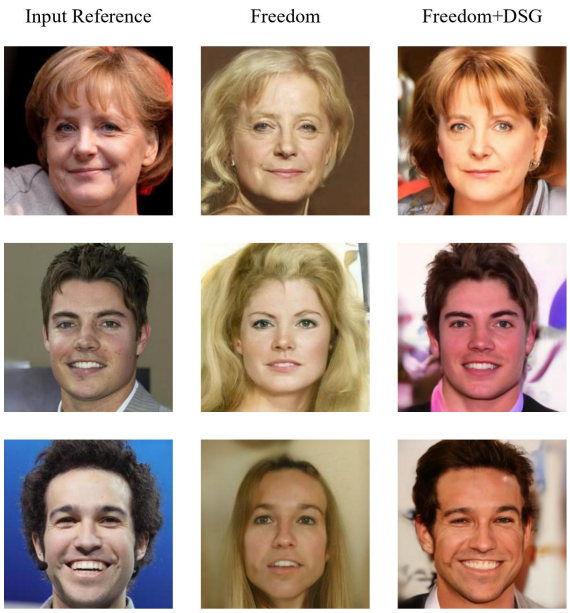

Figure 6. The qualitative results in FaceID Guidance using a diffusion model pre-trained from CelebA-HQ256*256.

5.3. Ablation Study

Robustness with fewer denoising steps When reducing the number of denoising steps (Table 2 and Figure 7), we observe a more pronounced performance gap. This phenomenon can be attributed to the limitations of DPS, which restricts the small step sizes for not deviating far from the intermediate manifold $\mathcal{M}_t$. Consequently, with fewer guidance steps, aligning with the measurements becomes increasingly challenging. In contrast, our model experiences only a slight decrease in performance. This can be attributed to our approach, which allows for larger step sizes while still preserving the underlying manifold structure. As a result, even with a reduced number of denoising steps, we can still achieve accurate alignment with the measurements while

generating realistic samples, as shown in Figure 7. Please refer to Appendix B.2 for further details.

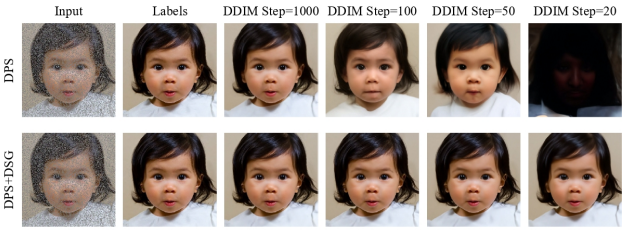

Figure 7. Comparison between DPS and DPS+DSG in different denoising steps in solving inpainting. With the decreasing of denoising steps, DPS can not align with the input due to the small step size. However, when plugging in DSG, it can allow larger step sizes, thus achieving better alignment.

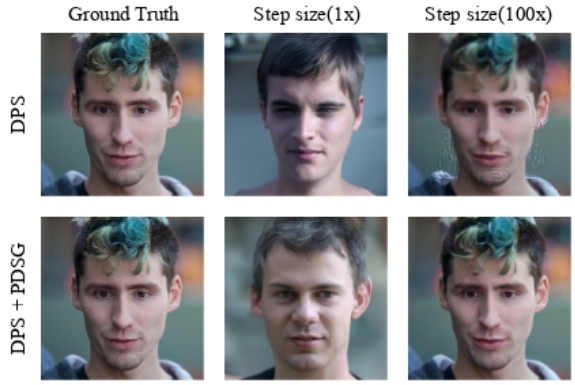

Figure 8. Comparison between DPS and DPS+PDSG (Another version of DSG by projecting DPS into Spherical Gaussian constraint) using 50 denoising steps.

Advantanges of guidance with Spherical Gaussian constraint. We performed additional comparative experiments to demonstrate the advantages of constraining the guidance with the spherical Gaussian constraint. To demonstrate the benefits of spherical Gaussian constraint, we implemented another version of DSG(PDSG) by projecting the x_{t-1} obtained by DPS onto the hyperphere $S^n_{\mu_\theta(x_t),\sqrt{n}\sigma}$ in Equation 12. With this projection, DPS can have a larger step size(100x) than the original setting with no artifacts. See Appendix B.3 for further details.

6. Limitations

Although DSG addresses the manifold deviation problem, allowing for larger guidance steps that enhance time efficiency, its sampling strategy may compromise sample diversity. This arises for two main reasons: Firstly, DSG restricts Gaussian sampling to the proximity of the gradient descent direction, while baseline methods (such as DPS) sample random noise across all directions. Secondly, approximating the high-dimensional Gaussian in DSG to a hypersphere could potentially diminish sample diversity, even though the possibility of sampling points outside the hypersphere is very small.

7. Conclusion

In this paper, we have revealed a crucial issue in the training-free conditional diffusion models: the occurrence of manifold deviation during the sampling process when employing loss guidance. This phenomenon is substantiated by establishing a certain lower bound for the estimation error of the loss guidance. To tackle this issue, we have proposed Diffusion with Spherical Gaussian constraint (DSG), inspired by the concentration phenomenon in high-dimensional Gaussian distributions. DSG effectively constrains the guidance step within the intermediate data manifold through optimization, thereby mitigating the manifold deviation problem and enabling the utilization of larger guidance steps. Furthermore, we presented a closed-form solution for DSG denoising with the Spherical Gaussian constraint. It's worth noting DSG serves as a plug-and-play module for training-free conditional diffusion models (CDMs). Integrating DSG into these CDMs merely involves modifying a few lines of code with almost no extra computational cost, but yields significant performance improvements. We have integrated DSG into several recent new CDMs for various conditional generative tasks. The experimental results validate the superiority and adaptability of DSG in terms of both sample quality and time efficiency.

Notably, the improvements of DSG on both sample quality and time efficiency come at the expense of sample diversity. This is because DSG replaces the random Gaussian noise term in the reverse diffusion process with the deterministic conditional guidance step. Although we use the guidance rate g_r to alleviate this problem in practice, it cannot be avoided inherently. Enhancing DSG diversity while preserving its generation quality is under consideration in our future work.

Acknowledgement

This work was supported by NSFC (No.62303319), Shanghai Sailing Program (22YF1428800, 21YF1429400), Shanghai Local College Capacity Building Program (23010503100), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging. Additionally, we would like to express our sincere gratitude to Professor Xuming He for his invaluable suggestions during the revision of this paper.

Impact Statement

There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Chan, S. H., Wang, X., and Elgendy, O. A. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. in 2021 ieee. In *CVF international conference on computer vision (ICCV)*, pp. 14347–14356, 2021.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *International Conference on Learning Representation*, 2023.

Daniels, M., Maunu, T., and Hand, P. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34: 12955–12965, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim, D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., et al. Manifold preserving guided diffusion. *International Conference on Learning Representation*, 2024.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Saleh, B. and Elgammal, A. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arxiv 2015. *arXiv preprint arXiv:1505.00855*, 2015.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.-Y., Kautz, J., Chen, Y., and Vahdat, A. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

Yu, J., Wang, Y., Zhao, C., Ghanem, B., and Zhang, J. Freedom: Training-free energy-guided conditional diffusion model. *arXiv preprint arXiv:2303.09833*, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

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

A. Proofs

A.1. Proof of Proposition 4.1

Proposition A.1. *(Lower bound of Jensen gap). For the β-strongly convex function f and the random variable $X \in \mathbb{R}^n \sim \mathcal{N}(\mu, \Sigma)$, we can have the lower bound of the Jensen Gap:*

$$\mathcal{J} \geq \frac{1}{2}\beta \sum_{i=1}^{n} \lambda_i.$$

Proof: Since $f(x)$ is a β-strongly convex function, we have

$$f(\frac{1}{2}(x_1 + x_2)) \leq \frac{1}{2}\left[f(x_1) + (1-\lambda)f(x_2)\right] - \frac{\beta}{8}\|x_1 - x_2\|^2,$$

Then, for convenience, let random variable $\delta = x - \mu$, and $\delta \sim \mathcal{N}(0, \Sigma)$. we obtain

$$\begin{aligned}
\mathcal{J}(f, x \sim \mathcal{N}(\mu, \Sigma)) =& \mathbb{E}\left[f(x)\right] - f(\mathbb{E}\left[x\right]) \\
=& \int p(x) f(x) dx - f(\mu) \\
=& \frac{1}{2}\int p(\mu+\delta) f(\mu+\delta) d\delta + \frac{1}{2}\int p(\mu-\delta) f(\mu-\delta) d\delta - f(\mu) \\
=& \int p(\mu+\delta)\left(\frac{1}{2}f(\mu+\delta) + \frac{1}{2}f(\mu-\delta) - f(\mu)\right) d\delta \quad \diamondsuit \text{ symmetry of Gaussian distribution} \\
\geq& \frac{\beta}{2}\int p(\mu+\delta)\|\delta\|^2 d\delta \quad \diamondsuit\ \beta\text{-strongly convexity} \\
=& \frac{\beta}{2}\mathbb{E}\left[\|\delta\|^2\right].
\end{aligned}$$

Since here the covariance matrix Σ is symmetric, we can use the spectral theorem here and write $\Sigma = P^T \Lambda P$ where P is an orthogonal matrix and Λ is diagonal with positive diagonal elements $\lambda_1, \cdots, \lambda_n$. Then, let $z = P\Sigma^{-\frac{1}{2}}\delta$ and $z \sim \mathcal{N}(0, I)$. Thus, we have

$$\|\delta\|^2 = \delta^T \delta = \left(\Sigma^{-\frac{1}{2}}\delta\right)^T \Sigma \left(\Sigma^{-\frac{1}{2}}\delta\right) = \left(\Sigma^{-\frac{1}{2}}\delta\right)^T P^T \Lambda P \left(\Sigma^{-\frac{1}{2}}\delta\right) = \|z\|_\Lambda^2$$

Consider each component in z is independent.

$$E\left[\|\delta\|^2\right] = \mathbb{E}\left[\|z\|_\Lambda^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\lambda_i z_i^2\right] = \sum_{i=1}^{n} \lambda_i.$$

Therefore, we have the lower bound of the Jensen gap that

$$\mathcal{J} \geq \frac{1}{2}\beta \sum_{i=1}^{n} \lambda_i.$$

A.2. Proof of Proposition 4.3

Proposition A.2. *(Concentration of high-dimensional isotropy Gaussian distribution). For any n-dimensional isotropy Gaussian distribution $x \sim N(\mu, \sigma^2 I)$:*

$$\begin{aligned}
P(||x-\mu||_2^2 \geq x_{lower} = n\sigma^2 + 2n\sigma^2(\sqrt{\epsilon} + \epsilon)) \leq e^{-n\epsilon}, \\
P(||x-\mu||_2^2 \leq x_{upper} = n\sigma^2 - 2n\sigma^2\sqrt{\epsilon}) \leq e^{-n\epsilon}.
\end{aligned}$$

When n is sufficiently large, it is close to the uniform distribution on the hypersphere of radius $\sqrt{n}\sigma$.

Proof:

First of all, we can find the expected value of the $||x-\mu||_2^2$ in the following way:

$$E(||x-\mu||_2^2) = \sum_{i=1}^{n} E((x_i-\mu)^2) = \sum_{i=1}^{n} E^2((x_i-\mu)) + D(x_i-\mu) = n\sigma^2, \tag{21}$$

where x_i is ith element of x.

Then, the concentration of $||x-\mu||_2^2$ can be proved using **standard Laurent-Massart bound** for a chi-square distribution. If Z is a chi-square distribution with n degrees of freedom,

$$\begin{aligned} P[Z-n \geq 2\sqrt{nt}+2t] \leq e^{-t}, \\ P[Z-n \leq -2\sqrt{nt}] \leq e^{-t}. \end{aligned} \tag{22}$$

By substituting $Z = \sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$ and $t = n\epsilon$,

$$\begin{aligned} P(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \geq n + 2n(\sqrt{\epsilon}+\epsilon)) \leq e^{-n\epsilon}, \\ P(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \leq n - 2n\sqrt{\epsilon}) \leq e^{-n\epsilon}. \end{aligned} \tag{23}$$

Therefore,

$$P(r\sqrt{1-2\sqrt{\epsilon}} \leq ||x-\mu||_2 \leq r\sqrt{1+2\sqrt{\epsilon}+2\epsilon}) \geq 1-2e^{-n\epsilon}. \tag{24}$$

By choosing $\epsilon_{min} = min\{1-\sqrt{1-2\sqrt{\epsilon}}, \sqrt{1+2\sqrt{\epsilon}+2\epsilon}-1\}$,

$$P((1-\epsilon_{min})r \leq ||x-\mu||_2 \leq (1+\epsilon_{min})r) \geq 1-2e^{-n\epsilon}, \tag{25}$$

which proves the concentration of $||x-\mu||_2$.

A.3. Closed-form Solution of Equation 13

Given the optimization problem

$$\begin{aligned} \arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x'-x_t) \\ \text{s.t. } x' \in S^d_{\mu_\theta(x_t),\sqrt{n}\sigma_t}, \end{aligned} \tag{26}$$

where $S^d_{\mu,r} = S^d_{\mu_\theta(x_t),\sqrt{n}\sigma} = \{x : ||x-\mu_\theta(x_t)||_2^2 = n\sigma_t^2\}$, the optimal solution can be derived as follows:

$$\begin{aligned} &\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x'-x_t) \\ =&\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T ((x'-\mu_\theta(x_t)) + (\mu_\theta(x_t)-x_t)) \\ =&\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x'-\mu_\theta(x_t)). \end{aligned} \tag{27}$$

Table 4. Hyperparameter settings of DSG for Linear Inverse Problems in FFHQ

Task	Interval	Guidance Rate	Denoising steps
Inpainting	5	0.2	1000
Super-Resolution	20	0.2	1000
Gaussian-deblurring	5	0.2	1000
Inpainting	1	0.2	100
Super-Resolution	2	0.1	100
Gaussian-deblurring	1	0.1	100
Inpainting	1	0.2	50
Super-Resolution	1	0.1	50
Gaussian-deblurring	1	0.1	50
Inpainting	1	0.2	20
Super-Resolution	1	0.2	20
Gaussian-deblurring	1	0.2	20

By reparameterizing $x' = \mu_\theta(x_t) + \sqrt{n}\sigma_t d$ where $||d||_2^2 = 1$ using constraint,

$$\begin{aligned} &\arg\min_{x'} \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T (x' - x_t) \\ =&\arg\min_{d} \sqrt{n}\sigma_t \left[\nabla_{x_t} L(\hat{x}_0(x_t), y)\right]^T d. \end{aligned} \tag{28}$$

Obviously, when $d = -\nabla_{x_t} L(\hat{x}_0(x_t), y)/||\nabla_{x_t} L(\hat{x}_0(x_t), y)||_2$, the optimization problem gets the minimal value $-\sqrt{n}\sigma_t$. Therefore, $x^*_{t-1} = x' = \mu_\theta(x_t) - \sqrt{n}\sigma_t \nabla_{x_t} L(\hat{x}_0(x_t), y)/||\nabla_{x_t} L(\hat{x}_0(x_t), y)||_2$.

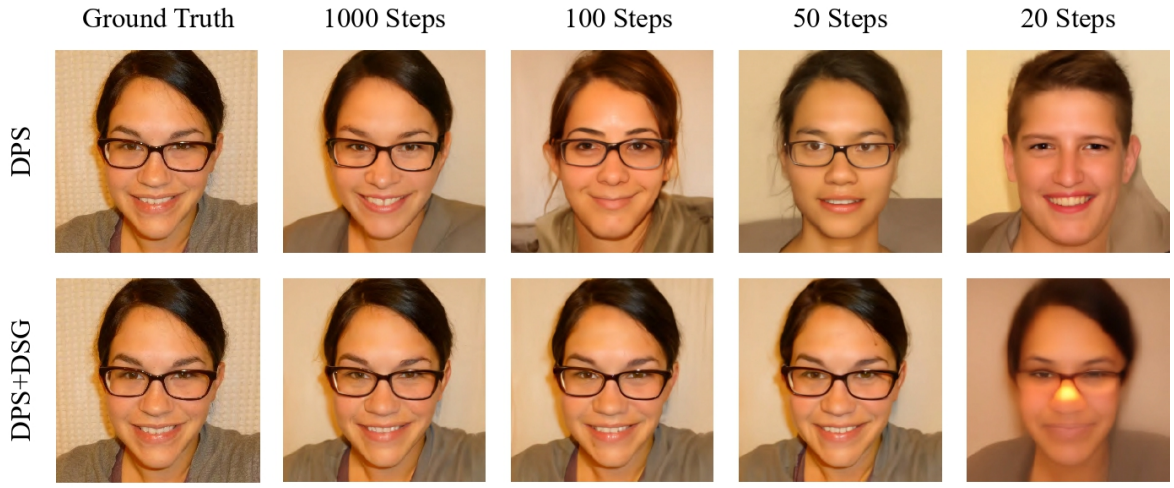

Figure 9. Comparison between DPS and DPS+DSG in Super-resolution task using different denoising steps.

B. Ablation Study

B.1. Hyperparameter Analysis

Our model has two main hyperparameters: guidance rate g_r and interval i. The guidance rate represents the weight of guidance in Equation 14. When it is zero, the denoising process is equivalent to unconditional generation. When it is set to 1, then the path of generation is determined. In practice, we find it better to choose values between 0.05 and 0.2 because a certain level of random noise is a trade-off between unconditional sampling diversity and better alignment. The interval

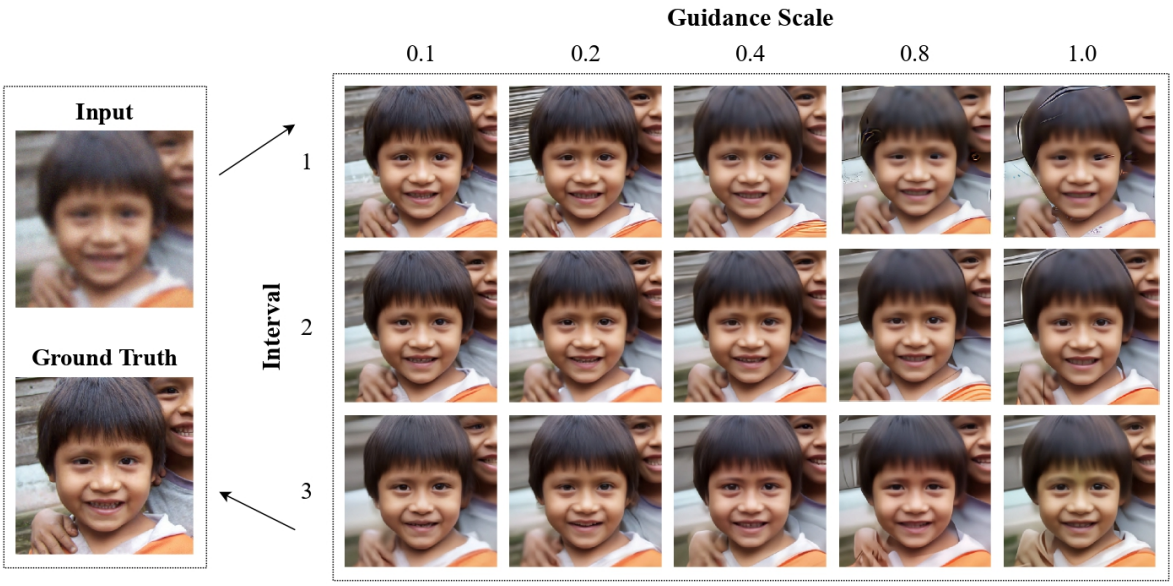

Figure 10. Different guidance rate using DSG using 100 DDIM steps

means that we apply guidance at a fixed interval and do unconditional sampling when not applying guidance. It can also increase the diversity of DSG and usually be used when the denoising steps are large, e.g. denoising steps larger than 100.

The hyperparameters we used for the linear inverse problem in FFHQ are shown in Table 4. For the linear inverse problem in ImageNet, we use $g_r = 0.2, i = 5$ for Inpainting, $g_r = 0.1, i = 10$ for Super-resolution, and $g_r = 0.1, i = 5$ for Gaussian deblurring. For Style Guidance, Text-Style Guidance and Text-Segmentation Guidance, we set $g_r = 0.1, i = 1$. For FaceID Guidance, we set the $g_r = 0.05, i = 1$.

From Table 4, we can observe that when the number of denoising steps is limited(≤ 100), simply setting the interval to 1 and the guidance rate to a small value ([0.05, 0.2]) can yield a satisfactory result. When denoising steps are significantly large (e.g., $T = 1000$), a large interval (larger than 5) can be set to increase the diversity of our method and decrease the number of guidance, thus accelerating the inference time while enhancing the quality of the generated image.

It is also worth noticing that while other training-free methods require careful tuning of the step size with different loss functions, the step sizes of DSG are adaptive and independent of the loss function. Moreover, the similar hyperparameter settings of DSG **can be applied to almost all tasks**, which reduces the cost for hyperparameter searching.

B.2. Ablation Study with Different Denoising Steps Compared with DPS

When conducting the ablation study of different denoising steps in the linear inverse problem, we carefully tune the step size γ of DPS (Chung et al., 2023) to $10\times$ than the original setting with denoising steps=100 and $20\times$ with denoising steps=20, 50. Further results are shown in Figure 9.

B.3. Ablation Study on Projecting DPS into Spherical Gaussian

To further demonstrate the advantages of Spherical Gaussian constraint, we first consider the single denoising step from x_t to x_{t-1}, DPS calculates x_{t-1} by estimating the additional correction step:

$$x_{t-1} = \underbrace{DDIM(x_t, \epsilon_\theta(x_t, t))}_{\text{sampling step}} - \underbrace{\gamma \nabla_{x_t} L(\hat{x}_0(x_t), y)}_{\text{correction step}}. \tag{29}$$

However, when γ is pretty large. we project x_{t-1} obtained by DPS onto the hyperplane $S^n_{\mu_\theta(x_t), \sqrt{n}\sigma}$ to force the spherical Gaussian constraint, which is called DPS+PDSG. The projection process can be represented as:

$$d_p = x_{t-1} - \mu_\theta(x_t), \tag{30}$$

$$x_{t-1}^{p} = \mu_\theta(x_t) + \sqrt{n}\sigma_t d_p / ||d_p||, \tag{31}$$

where x_{t-1}^{p} denates the projection point, which is the closest point in $S^n_{\mu_\theta(x_t),\sqrt{n}\sigma}$ w.r.t. x_{t-1}. Since a large step size γ will cause DPS to fall off the manifold (Figure 9), the operation in Equation 30, 31 can project the x_{t-1} obtained by DPS back into the intermediate manifold $\mathcal{M}_{t-1}$ while allowing for larger step sizes.

B.4. Ablation Study on Denoising Process

We compare our DSG with DPS in the Inpainting task in 1000 denoising steps and show the intermediate noisy image x_t in Figure 11. The results demonstrate that our guidance is more effective and can expedite the restoration of the overall appearance of the image.

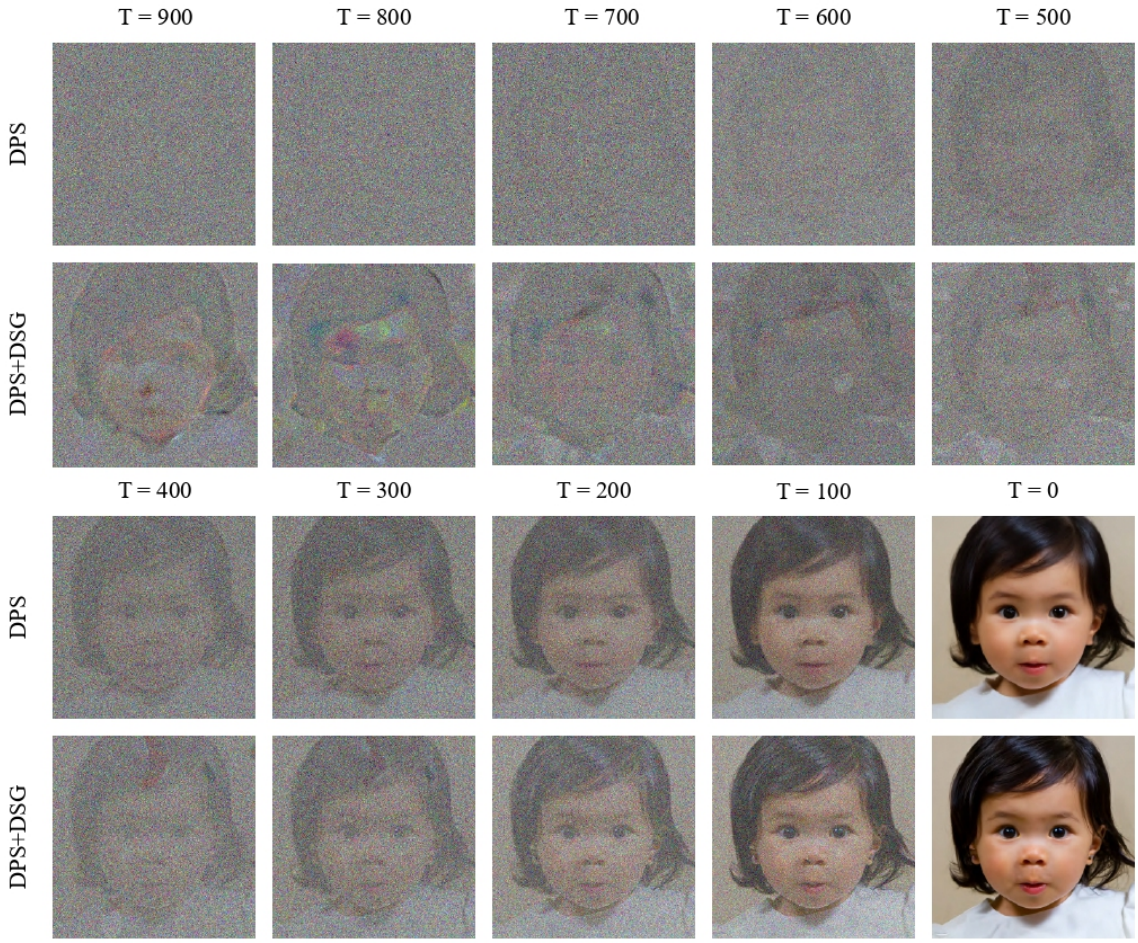

Figure 11. Qualitative results of the denoising process.

C. Additional Experimental Details and Results

In this section, to further demonstrate the applicability and strength of our DSG, we will provide additional experimental details and results.

C.1. More Experimental Details and Results for FaceID Guidance

In FaceID Guidance, we choose Freedom (Yu et al., 2023), LGD (Song et al., 2023) as the baselines. For Freedom, we follow the hyperparameter settings proposed in the original paper. For LGD, we set the number of Monte Carlo samples to 3, covariance to $0.5\sqrt{1-\alpha_t}$, and step size to $100\sqrt{\alpha_t}$. As shown in Table 3, Figure 6 and Figure 12, our DSG shows the

SOTA performance according to these qualitative and quantitative results.

Figure 12. The qualitative results in FaceID Guidance using a diffusion model pre-trained from CelebA-HQ256*256.

C.2. More Experimental Details and Results for Text-Style Guidance

In Text-Style Guidance, we follow the experiment setting of Freedom (Yu et al., 2023) and choose Freedom, LGD, and MPGD (Yu et al., 2023; Song et al., 2023; He et al., 2024) as the baseline methods for comparisons. For the hyperparameter setting, we follow the original setting of Freedom and MPGD. For LGD, we set the number of Monte Carlo samples to 2, covariance to $0.1\sqrt{1-\alpha_t}$, and step size same as Freedom. The qualitative results are shown in Figure 13.

C.3. More Experimental Details and Results for the Inverse Problems in ImageNet

We validate the performance of DSG in inverse problems (Inpainting, Super-resolution, and Gaussian deblurring) in the 1k validation set of ImageNet256*256 using the same setting in Sec 5.1. We choose DPS, LGD, and DDNM as the baseline. For DPS and DDNM, we follow the hyperparameter setting of the original paper. For LGD, we set the number of Monte Carlo samples to 10, covariance to $\sigma_t/\sqrt{1+\sigma_t^2}$, and the same step size as DPS. As shown in Table 5, Figure 14, Figure 15, and Figure 16, our DSG outperforms other baselines in ImageNet256*256.

It's worth noting that DDNM (Wang et al., 2022) is **primarily applicable to linear inverse problems** and cannot be directly applied to nonlinear cases due to its direct access to the noise scale of measurement, the forward operator, and its pseudo-inverse. However, despite these differences, DSG still offers superior performance compared to DDNM.

Table 5. Quantitative results in Linear Inverse Problem in ImageNet 256*256

Methods	Inpainting			Super resolution			Gaussian deblurring		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
DPS (Chung et al., 2023)	0.828	28.57	0.189	0.635	23.73	0.317	0.471	19.89	0.402
LGD (Song et al., 2023)	0.825	28.11	0.191	0.633	23.13	0.318	0.367	16.77	0.474
DDNM (Wang et al., 2022)	0.875	28.76	0.125	-	-	-	-	-	-
DPS+DSG(Ours)	**0.879**	**29.20**	**0.116**	**0.672**	**23.86**	**0.282**	**0.644**	**23.25**	**0.282**

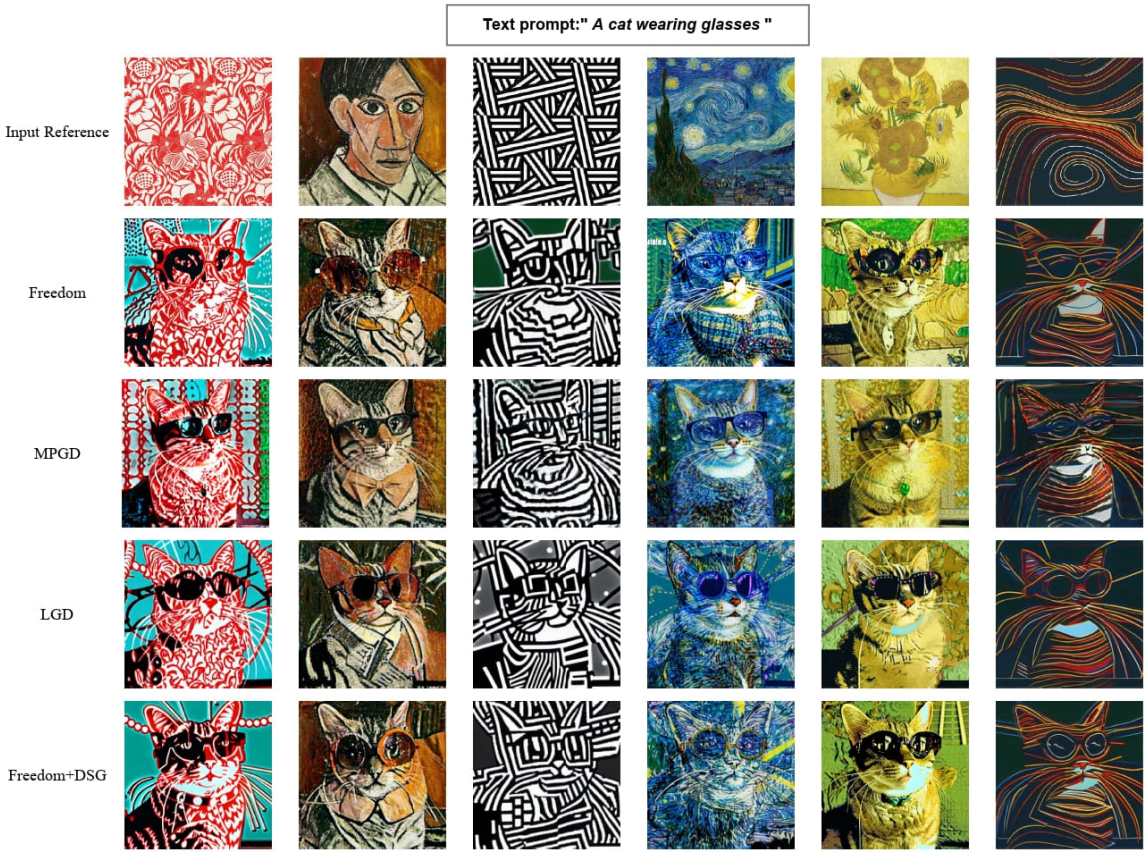

Figure 13. The qualitative results in Text-Style Guidance using Stable Diffusion.

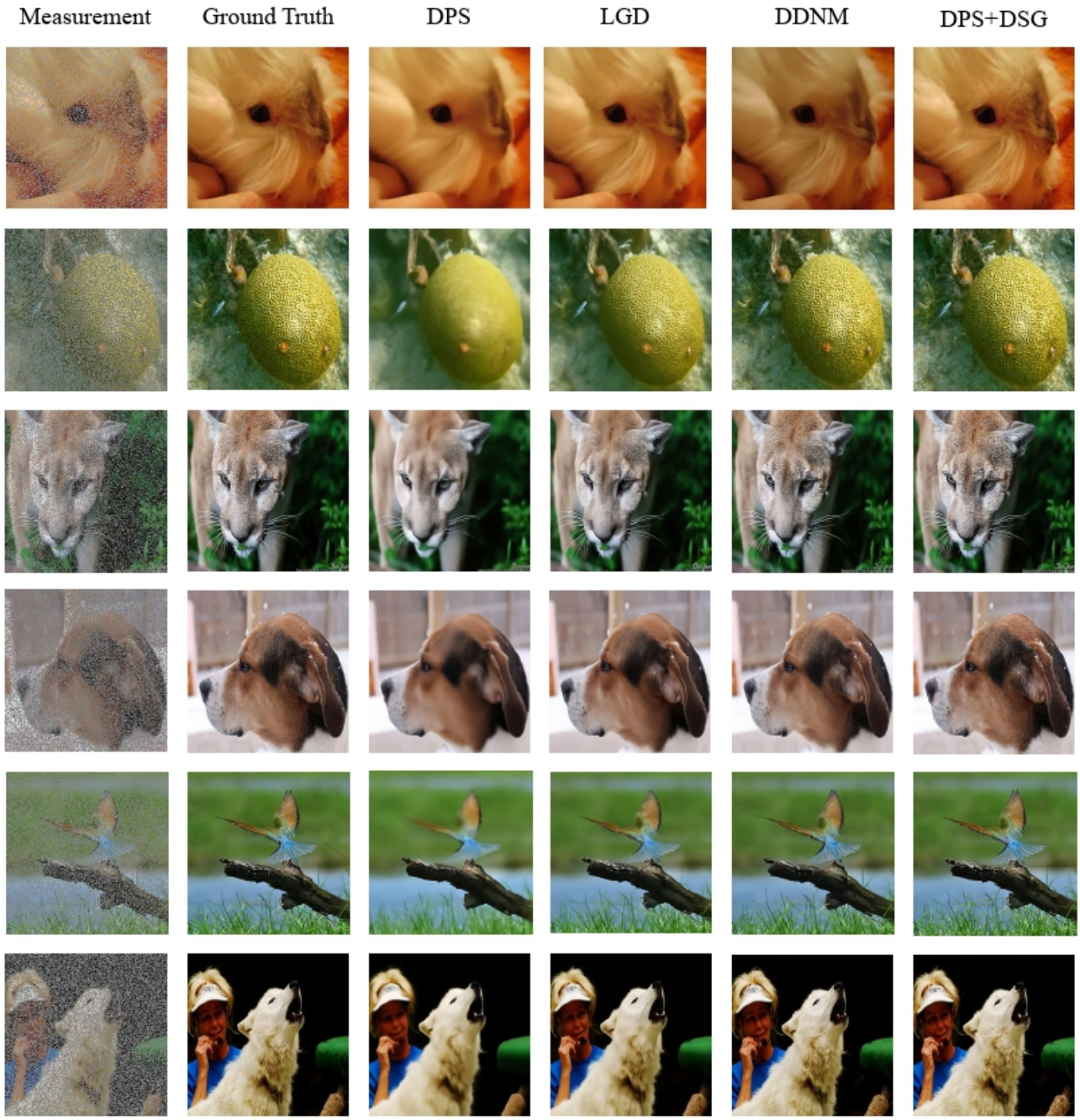

Figure 14. Additional qualitative results of Inpainting in Imagenet256*256.

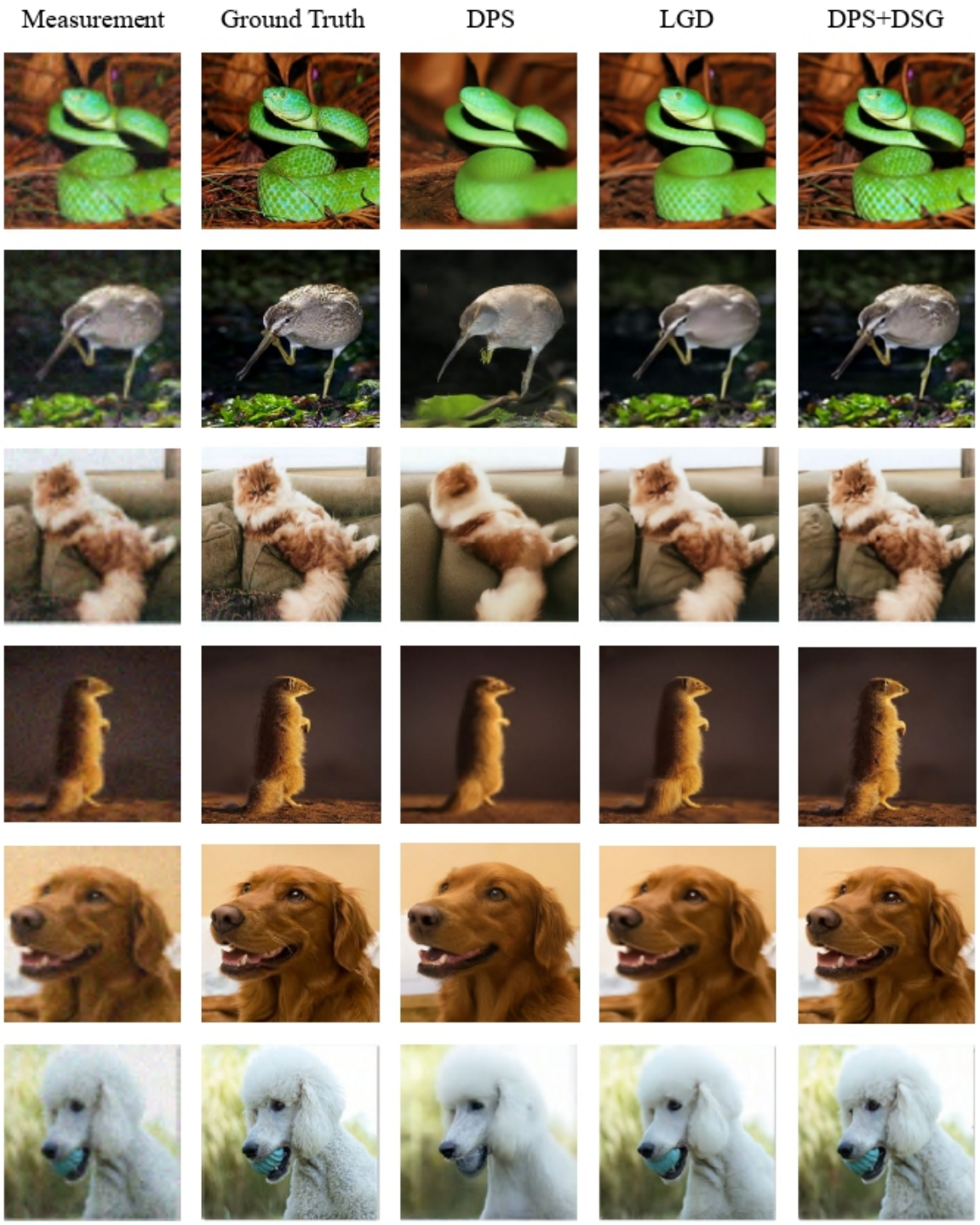

Figure 15. Additional qualitative results of Super-resolution in Imagenet256*256.

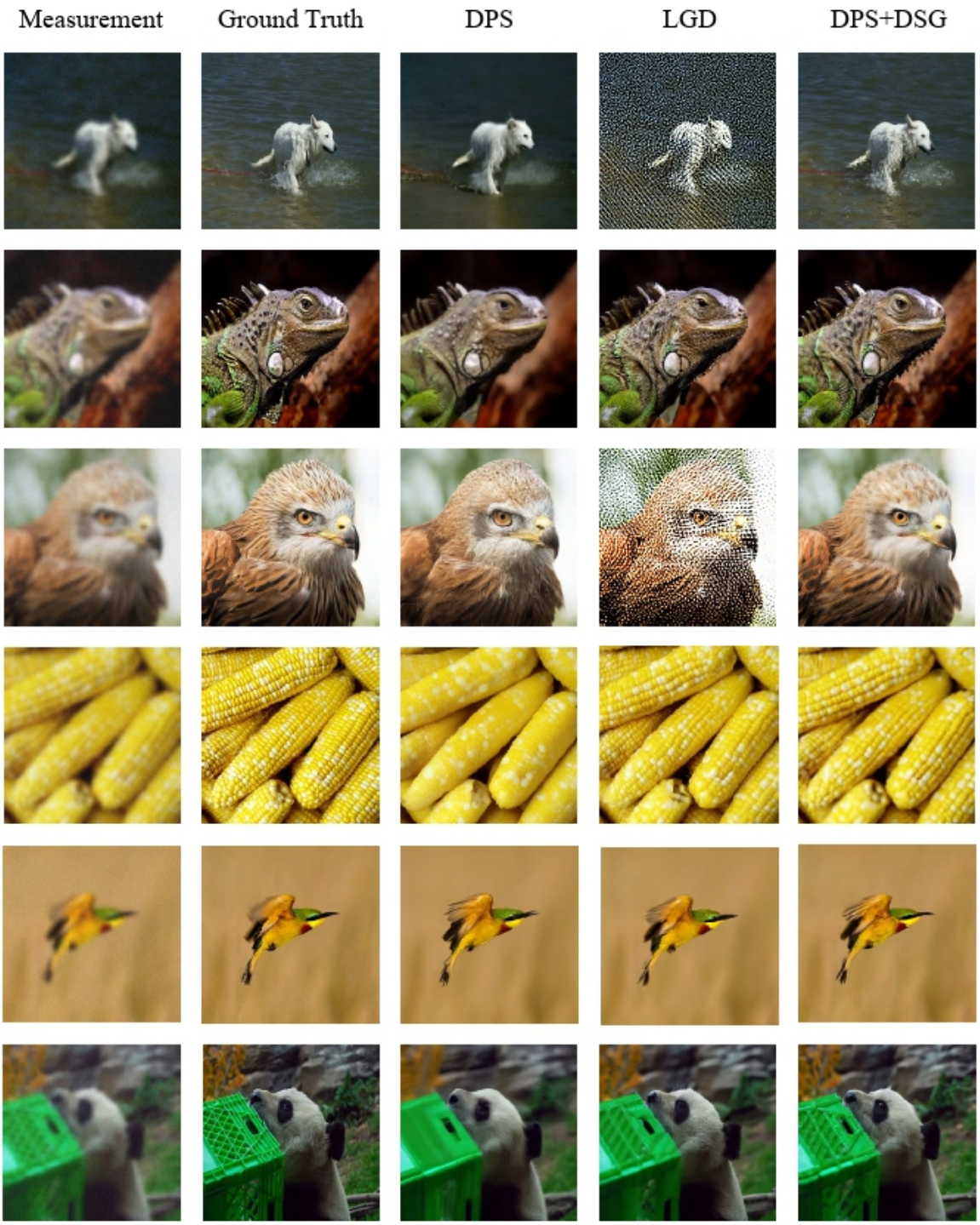

Figure 16. Additional qualitative results of Gaussian-deblurring in Imagenet256*256.

D. Additional Qualitative Results

We provided additional qualitative results to demonstrate that DSG can plug in other training-free methods while improving their performance.

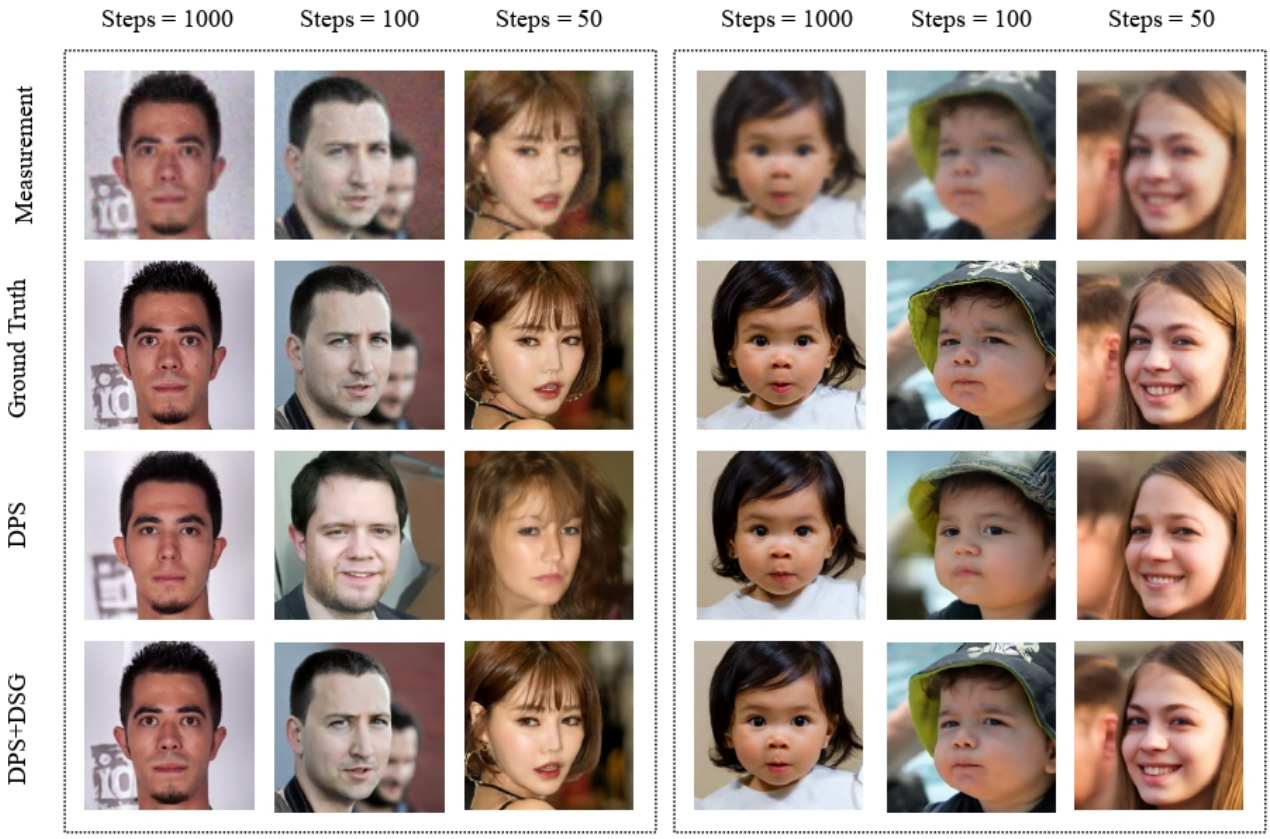

Figure 17. Extra qualitative examples of Super-resolution (left) and Gaussian deblurring (right) in the FFHQ dataset.

Figure 18. Enlargement of the qualitative results using DPS+DSG compared to DPS.

Figure 19. Additional qualitative results in Style Guidance using pre-trained Stable Diffusion.

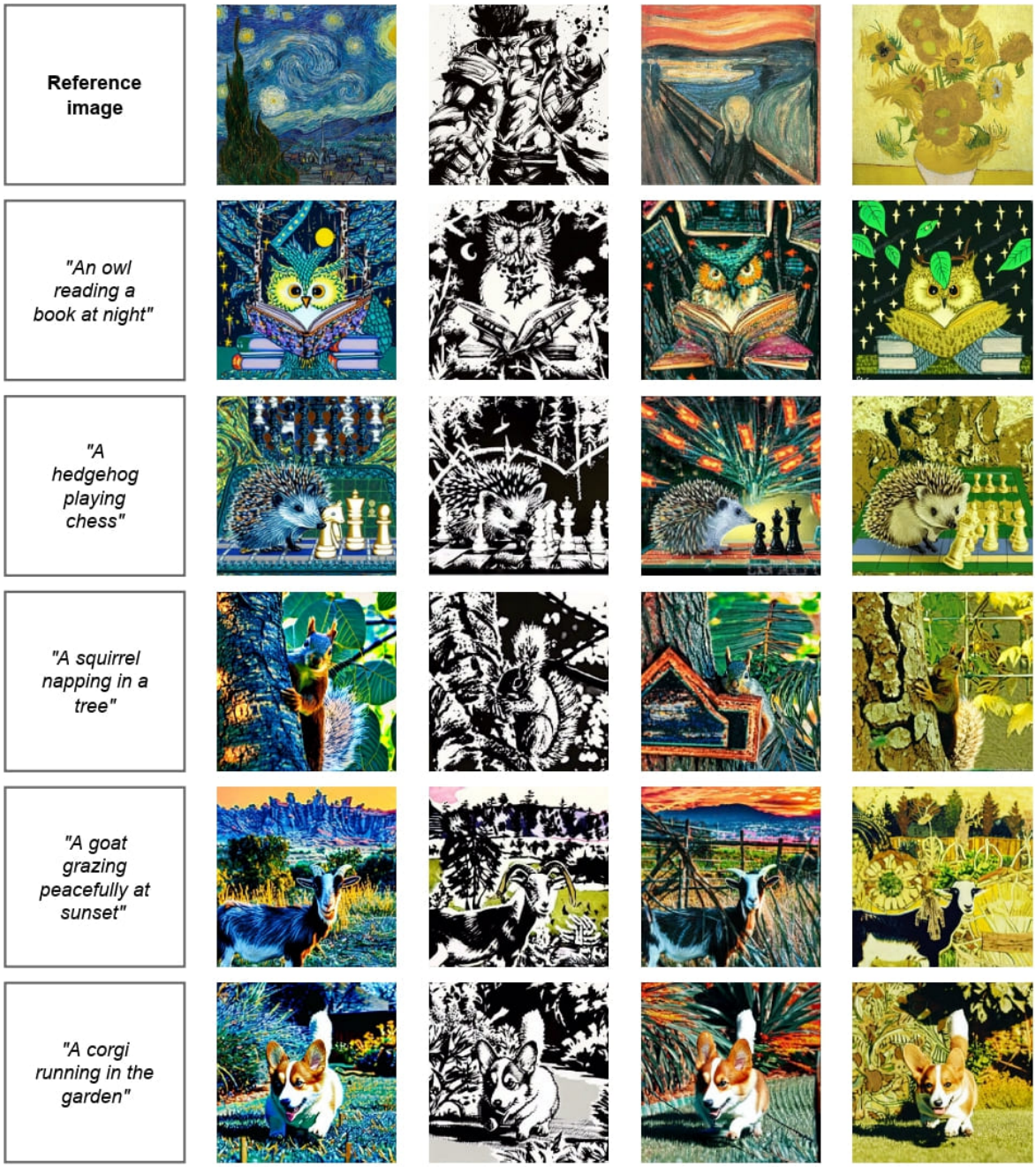

Figure 20. Additional qualitative results of DSG in Text-Style Guidance using pre-trained Stable Diffusion.