

Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance

Donghoon Ahn^{*1} Hyoungwon Cho^{*1} Jaewon Min¹
 Wooseok Jang¹ Jungwoo Kim¹ SeonHwa Kim¹
 Hyun Hee Park² Kyong Hwan Jin^{†1} Seungryong Kim^{†1}

¹Korea University

²Samsung Electronics

<https://ku-cvlab.github.io/Perturbed-Attention-Guidance>

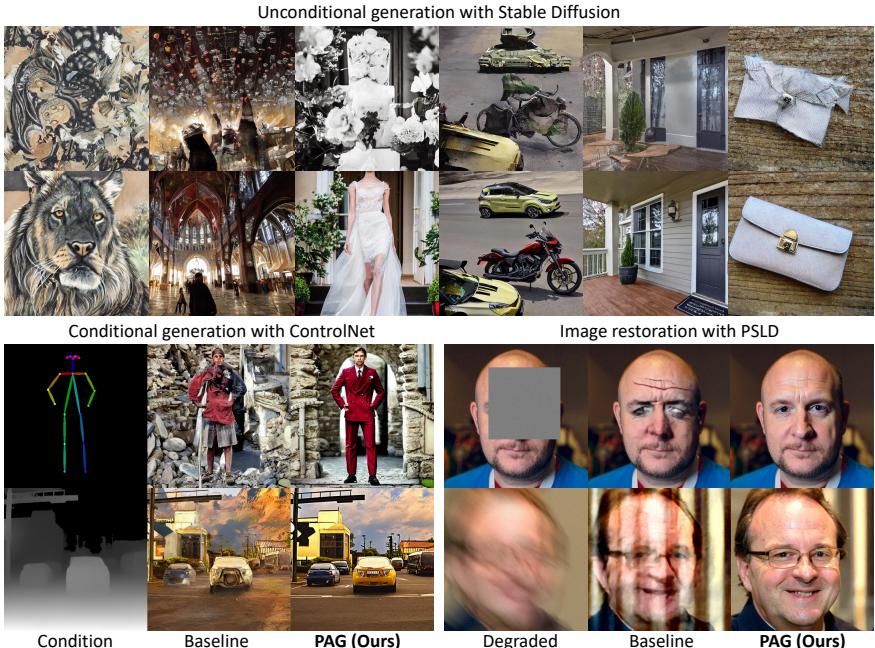


Fig. 1: Qualitative comparisons between unguided (baseline) and perturbed-attention-guided (PAG) diffusion samples. Without any *external conditions*, e.g., class labels or text prompts, or *additional training*, our PAG dramatically elevates the quality of diffusion samples even in unconditional generation, where classifier-free guidance (CFG) [19] is inapplicable. Our guidance can also enhance the baseline performance in various downstream tasks such as ControlNet [58] with empty prompt and image restoration such as inpainting and deblurring [6, 44].

Abstract. Recent studies have demonstrated that diffusion models are capable of generating high-quality samples, but their quality heavily depends on sampling guidance techniques, such as classifier guidance (CG)

^{*}: Equal contribution

[†]: Co-corresponding author

and classifier-free guidance (CFG). These techniques are often not applicable in unconditional generation or in various downstream tasks such as image restoration. In this paper, we propose a novel sampling guidance, called **Perturbed-Attention Guidance (PAG)**, which improves diffusion sample quality across both unconditional and conditional settings, achieving this without requiring additional training or the integration of external modules. PAG is designed to progressively enhance the structure of samples throughout the denoising process. It involves generating intermediate samples with degraded structure by substituting selected self-attention maps in diffusion U-Net with an identity matrix, by considering the self-attention mechanisms' ability to capture structural information, and guiding the denoising process away from these degraded samples. In both ADM and Stable Diffusion, PAG surprisingly improves sample quality in conditional and even unconditional scenarios. Moreover, PAG significantly improves the baseline performance in various downstream tasks where existing guidances such as CG or CFG cannot be fully utilized, including ControlNet with empty prompts and image restoration such as inpainting and deblurring.

1 Introduction

Diffusion models [18, 41, 47, 49, 50] have become prominent in image generation, showcasing their capability to produce high-fidelity and diverse samples. Sampling guidance techniques such as classifier guidance (CG) [10] or classifier-free guidance (CFG) [19] play crucial roles in directing diffusion models to generate higher-quality images. Without such techniques, as depicted in Fig. 1 and Fig. 2, diffusion models often produce lower-quality images, typically exhibiting collapsed structures in synthesized images. Despite their widespread use, these guidance approaches have drawbacks: they require additional training or the integration of external modules, often diminish the diversity of the output samples, and are unavailable in unconditional generation.

Meanwhile, unconditional generation offers significant practical advantages. It aids in understanding the fundamental principles of data creation and its underlying structures [5, 28]. Further, progress in unconditional techniques often enhances conditional generation. Importantly, it eliminates the need for potentially costly and complex human annotations such as class labels, text, segmentation maps, etc., which can be a major hurdle in tasks where accurate labeling is difficult, *e.g.*, modeling molecular structures [28]. Finally, unconditional generative models offer powerful general priors, as seen in their use for inverse problems solvers [6, 7, 25, 43, 44, 50, 56]. However, the unavailability of CG [10] or CFG [19] may lead to sub-optimal performance.

Recognizing the importance of unconditional generation, we propose a novel sampling guidance, called **Perturbed-Attention Guidance (PAG)**, that improves diffusion sample quality across both unconditional and conditional settings, without requiring additional training or the integration of external modules. Our approach leverages an implicit discriminator to distinguish between

desirable samples and undesirable samples. Leveraging the capability of self-attention maps in diffusion U-Net to capture structural information [2, 16, 32, 52, 53], we generate undesirable samples by substituting the diffusion model’s self-attention map with an identity matrix, and guide the denoising process away from these degraded samples. These undesirable samples serve to steer the denoising trajectory away from the structural collapse that is commonly observed in unguided generation.

Extensive experiments validate the effectiveness of our guidance method. Applied to ADM [10], it significantly improves sample quality in both conditional and unconditional settings. We also observe remarkable enhancements when applied to the widely-used Stable Diffusion [41] both qualitatively and quantitatively. In addition, combining PAG with conventional guidance methods such as CFG [19] leads to further improvement. Finally, our guidance significantly improves the performance of diffusion models in various downstream tasks like inverse problems [6, 44], ControlNet [58] with empty prompts, where the lack of conditions renders CFG [19] unusable.

2 Related Work

Diffusion models. Diffusion models (DMs) [47, 49, 50] have established a significant benchmark in image generation, achieving outstanding results in both sample quality and the estimation of sample distributions. The synergy between the inductive biases of image-like data and DMs, particularly when coupled with a U-Net backbone [18], has resulted in high-quality synthesis. Despite their impressive performance, these models face challenges of slow inference times and high training costs. In DDIM [48], the sampling speed has been enhanced by applying a non-Markovian process to the diffusion process. Latent diffusion models (LDMs) [41] operate in a compressed latent space, effectively striking a balance between computational efficiency and synthesis quality.

Sampling guidance for diffusion models. The surge in diffusion model research is largely attributed to advancements in sampling guidance techniques [10, 19]. Classifier guidance (CG) [10] increases fidelity at the expense of diversity by adding the gradient of a classifier which has to be trained. Classifier-free guidance (CFG) [19] models an implicit classifier to have a similar effect as CG. Self-attention guidance (SAG) [20] improves sample quality within an unconditional framework. It utilizes adversarial blurring on generating image samples to obscure crucial information and then guides the sampling process by leveraging the noise predicted with the blurred sample and the original one. In addition, various guidance methods have been proposed, focusing on conditioning [30] or image editing [3, 12].

Self-attention. Transformer [54] architecture shows that self-attention mechanism enables the model to dynamically weigh different words within a sequence, capturing complex dependencies and semantic relationships across the input. This capability has led to its success not only in natural language processing [9, 54] but also in computer vision [11, 40]. In diffusion models, self-attention

was effectively integrated into the U-Net [42] architecture in DDPM [18] and further utilized in DiT [36], leading to enhanced performance. Due to its rich representation, self-attention layer has been employed across a variety of applications, including personalization, image editing, and video generation [4, 31, 32, 35, 39].

3 Preliminaries

Diffusion models. In diffusion models [10, 18, 19, 50], random noise $\epsilon \sim \mathcal{N}(0, I)$ is added during forward path to an image x_0 to produce a noisy image x_t at an arbitrary timestep t :

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ according to a variance schedule β_1, \dots, β_T . A denoising network ϵ_θ is learned to predict ϵ by optimizing an objective

$$\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon \sim \mathcal{N}(0, I)} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right], \quad (2)$$

for uniformly sampled $t \in \{1, \dots, T\}$. Diffusion models often utilize U-Net [42] architecture for the denoising network [18, 41, 48].

During sampling, the model produces denoised image x_{t-1} from x_t at each timestep t based on the noise estimation $\epsilon_\theta(x_t, t)$ as follows:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \quad (3)$$

where $z \sim \mathcal{N}(0, I)$ and σ_t^2 is set to β_t . Starting with randomly sampled noise $x_T \sim \mathcal{N}(0, I)$, the process is applied iteratively to generate a clean image x_0 . For the sake of simplicity, throughout the remainder of this paper, we adopt the notation $\epsilon_\theta(x_t)$ to represent $\epsilon_\theta(x_t, t)$. Note that noise estimation of the diffusion model can be considered as $\epsilon_\theta(x_t) \approx -\sigma_t \nabla_{x_t} \log p(x_t)$ [10, 19, 49, 50], where $p(x_t)$ denotes the distribution of x_t .

In addition, using the reparameterization trick, it is possible to obtain the intermediate prediction of x_0 at a given timestep t as

$$\hat{x}_0 = (x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)) / \sqrt{\bar{\alpha}_t}. \quad (4)$$

Classifier-free guidance. To enhance the generation towards arbitrary class label c , CG [10] introduces a new sampling distribution $\tilde{p}_\theta(x_t|c)$ composed with both $p_\theta(x_t|c)$ and the classifier distribution $p_\theta(c|x_t)$, which is expressed as

$$\tilde{p}_\theta(x_t|c) \propto p_\theta(x_t|c) p_\theta(c|x_t)^s, \quad (5)$$

where s is the scale parameter. It turns out that sampling from this distribution with $s > 0$ leads the model to generate saturated samples with high probabilities for the input class labels, resulting in increased quality but decreased sample diversity [10].

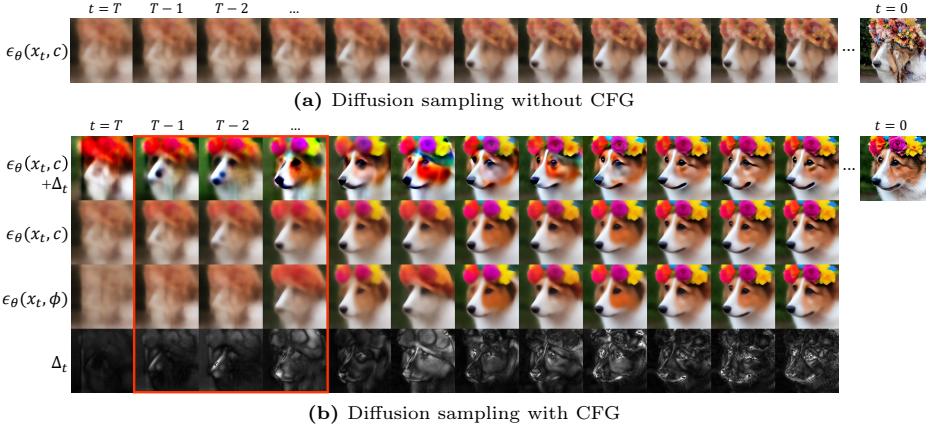


Fig. 2: Visualization of reverse process w/o and w/CFG [19]. To visualize predicted epsilon, we first convert it into \hat{x}_0 following Equation 4. Then, for the guidance signal $\Delta_t = \epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$, we apply an absolute value function and calculate the mean across all channels. **(a)** Without CFG [19], the diffusion models generate samples with collapsed structures. **(b)** With CFG [19], the diffusion models generate samples that are well-aligned to the prompt. The red rectangles highlight the distinction between *conditional* ($\epsilon_\theta(x_t, c)$) and *unconditional* ($\epsilon_\theta(x_t, \phi)$) predictions. Without a prompt, the diffusion model lacks guidance on what to generate in the early stages, often leading to the omission of salient features such as eyes and nose, and thus adding Δ_t amplifies features that are relevant to the prompt. Here the prompt “*a corgi with flower crown*” is used.

CG, however, has a drawback in that it requires a pretrained classifier for noisy images of each timestep. To address this issue, CFG [19] modifies the classifier distribution $p_\theta(c|x_t)$ by combining the conditional distribution $p_\theta(x_t|c)$ and the unconditional distribution $p_\theta(x_t)$:

$$\begin{aligned} \tilde{p}_\theta(x_t|c) &\propto p_\theta(x_t|c)p_\theta(c|x_t)^s = p_\theta(x_t|c) \left[\frac{p_\theta(x_t|c)p_\theta(c)}{p_\theta(x_t)} \right]^s \\ &= p_\theta(x_t|c)^{1+s} p_\theta(x_t)^{-s}. \end{aligned} \quad (6)$$

Then the score of new conditional distribution $\tilde{p}_\theta(x_t|c)$ would be $\nabla_{x_t} \log \tilde{p}_\theta(x_t|c) = (1+s)\epsilon^*(x_t, c) - s\epsilon^*(x_t)$, where ϵ^* denotes true score. By approximating this score using conditional and unconditional score estimates, we have

$$\begin{aligned} \tilde{\epsilon}_\theta(x_t, c) &= (1+s)\epsilon_\theta(x_t, c) - s\epsilon_\theta(x_t) \\ &= \epsilon_\theta(x_t, c) + s(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)) = \epsilon_\theta(x_t, c) + s\Delta_t. \end{aligned} \quad (7)$$

In practice, $\epsilon_\theta(x_t, c)$ and $\epsilon_\theta(x_t)$ are parameterized by a single neural network, which is jointly trained for both conditional and unconditional generation by assigning a null token ϕ as the class label for the unconditional model, such that $\epsilon_\theta(x_t) \approx \epsilon_\theta(x_t, \phi)$. The guidance signal $\Delta_t = \epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$ acts as the gradient of the implicit classifier, producing images that closely adhere to condition

c. In Fig. 2, we visualize Δ_t across timesteps and provide an explanation of its role in enhancing sample quality. A more detailed exploration of CFG’s workings is available in the Appendix E.1.

4 PAG: Perturbed-Attention Guidance

4.1 Self-rectifying sampling with implicit discriminator

Recently, it has been shown that the sampling guidance of diffusion models can be generalized as the gradient of energy function, for instance, which can be a negative class probability of classifier [10], negative CLIP similarity score [33], any type of time-independent energy [3], the distance between extracted signal such as pose and edges and reference signal [30] or any energy function which takes the noisy sample [12].

In this work, we introduce an implicit discriminator denoted \mathcal{D} that differentiates *desirable* samples following real data distribution from *undesirable* ones during the diffusion process. Similar to CFG [19] where the implicit classifier guides samples to be more closely aligned with the given class label, the implicit discriminator \mathcal{D} guides samples towards the desirable distribution and away from the undesirable distribution. By applying Bayes’ rule, we first define the implicit discriminator as

$$\mathcal{D}(x_t) = \frac{p(y|x_t)}{p(\hat{y}|x_t)} = \frac{p(y)p(x_t|y)}{p(\hat{y})p(x_t|\hat{y})}, \quad (8)$$

where y and \hat{y} denote the imaginary labels for desirable sample and undesirable sample, respectively.

Then similar to WGAN [1, 57], we set the generator loss of the implicit discriminator as our energy function, $\mathcal{L}_{\mathcal{G}}$, and compute its derivative as

$$\begin{aligned} \nabla_{x_t} \mathcal{L}_{\mathcal{G}} &= \nabla_{x_t} [-\log \mathcal{D}(x_t)] \\ &= \nabla_{x_t} \left[-\log \frac{p(y)p(x_t|y)}{p(\hat{y})p(x_t|\hat{y})} \right] = \nabla_{x_t} \left[-\log \frac{p(x_t|y)}{p(x_t|\hat{y})} \right] \\ &= -\nabla_{x_t} (\log p(x_t|y) - \log p(x_t|\hat{y})). \end{aligned} \quad (9)$$

Then, using Eq. 9, we define a new diffusion sampling such that

$$\begin{aligned} \tilde{\epsilon}_{\theta}(x_t) &= \epsilon_{\theta}(x_t) + s\sigma_t \nabla_{x_t} \mathcal{L}_{\mathcal{G}} \\ &= \epsilon_{\theta}(x_t) - s\sigma_t \nabla_{x_t} (\log p(x_t|y) - \log p(x_t|\hat{y})) \\ &= \epsilon_{\theta}(x_t) + s(\epsilon_{\theta}(x_t) - \hat{\epsilon}_{\theta}(x_t)) = \epsilon_{\theta}(x_t) + s\hat{\Delta}_t. \end{aligned} \quad (10)$$

Since diffusion models already learned the desired distribution, we use the pre-trained score estimation network $\epsilon_{\theta}(x_t)$ as an approximation of $-\sigma_t \nabla_{x_t} \log p(x_t|y)$. For the score with undesirable label \hat{y} , we approximate it by *perturbing* the forward pass of pretrained network which we denote $\hat{\epsilon}_{\theta}(x_t)$. Note that $\hat{\epsilon}_{\theta}(x_t)$ can embody any form of perturbation during the epsilon prediction process, including perturbations applied to the input [20] or internal representations, or both.

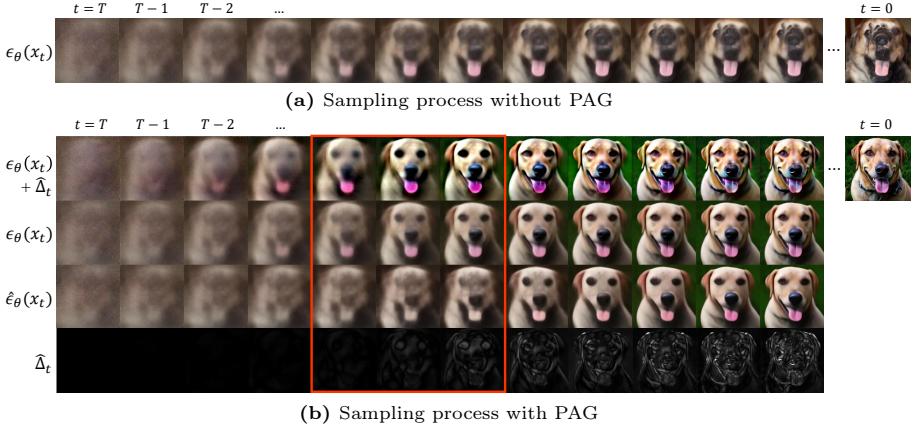


Fig. 3: Visualization of sampling process w/o and w/ PAG. To visualize predicted epsilon, we first convert it into \hat{x}_0 following Equation 4. Then, for the guidance signal $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$, we apply an absolute value function and calculate the mean across all channels. **(a)** Without guidance, the diffusion models generate samples with collapsed structures. **(b)** With our PAG, the diffusion models generate improved samples. The red rectangles highlight the distinction between the *original* ($\epsilon_\theta(x_t)$) and *perturbed* ($\hat{\epsilon}_\theta(x_t)$) predictions. With perturbed self-attention, the diffusion model lacks an understanding of the global structure, often leading to the omission of salient features such as eyes, nose, and tongue, and thus adding $\hat{\Delta}_t$ amplifies features that should be visible when global structures are successfully modeled.

Connections to CFG. The formulation in Eq. 10 resembles CFG [19]. Indeed, it is noteworthy that CFG can be considered a particular instance within our broader formulation. First, Eq. 10 can also be defined in class-conditional diffusion models such that

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + s(\epsilon_\theta(x_t, c) - \hat{\epsilon}_\theta(x_t, c)). \quad (11)$$

In CFG, $\hat{\epsilon}_\theta(x_t, c)$ is implemented by dropping the class label, resulting in $\epsilon_\theta(x_t, \phi)$, which in our terminology can be described as a *perturbed* forward pass. In this paper, we extend the concept of the *perturbed* forward pass to be more applicable even to the unconditional diffusion models.

4.2 Perturbing self-attention of U-Net diffusion model

In our framework, the perturbation strategies to implement $\hat{\epsilon}_\theta(x_t)$ can be arbitrarily chosen. However, perturbing the input image or the condition directly can cause the out-of-distribution problem, lead the diffusion model to create incorrect guidance signals, and steer the diffusion sampling towards the erroneous direction. To overcome this, CFG [19] explicitly trains an unconditional model. In addition, SAG [20] employs partial blurring to minimize deviation, but without careful selection of hyperparameters, it often deviates from the desired trajectory. This behavior is illustrated in Fig. 45 in Appendix E.3.

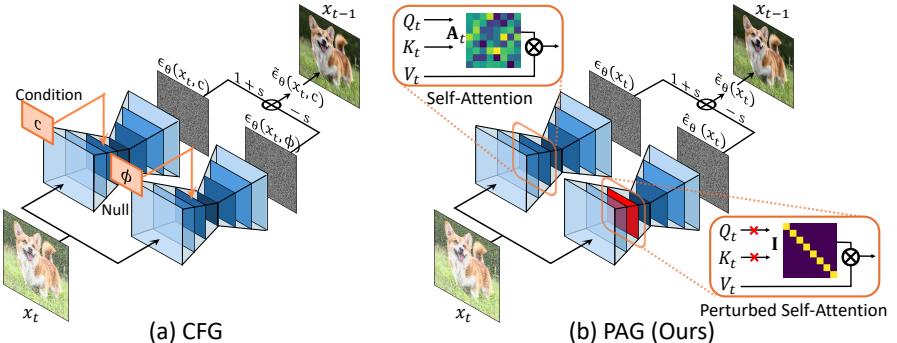


Fig. 4: Conceptual comparison between CFG [19] and PAG. CFG [19] employs jointly trained unconditional model as the *undesirable* path, whereas PAG utilizes perturbed self-attention for the same purpose. \mathbf{A}_t corresponds to the self-attention map $\text{Softmax}(Q_t K_t^T / \sqrt{d})$. In PAG, we perturb this by replacing with an identity matrix \mathbf{I} .

On the other hand, some studies have explored manipulating cross-attention and self-attention maps of the diffusion models for various tasks [4, 16, 26, 39, 46]. They show that modifying the attention maps has minimal impact on the model’s ability to generate plausible outputs. To design the perturbation strategy, we focus on the self-attention mechanism of denoising U-Net [42] to be applicable for both conditional and unconditional models.

Another criterion for selecting perturbations involves determining which aspects of the samples should be improved during the sampling process. As illustrated in the top row of Fig. 1 and Fig. 2, images generated by diffusion models without guidance often exhibit collapsed structures. To address this, the desired guidance should steer the denoising trajectory away from the sample exhibiting a collapsed structure, akin to the manner in which the null prompt in CFG is employed to strengthen class conditioning. Recently, several studies [2, 16, 32, 52, 53] demonstrate that the attention map contains structural information or semantic correspondence between patches. Thus, perturbing the self-attention map can generate a sample with a collapsed structure. We visualize the perturbed epsilon prediction in Fig. 3 in the same manner as in Fig. 2. Notably, within the red box in Fig. 3 (b), it can be seen that the generated samples have collapsed structures compared to the original prediction, while preserving the overall appearance of the original sample, attributable to the attention map’s robustness to manipulation.

Perturbed self-attention. Recent studies [2, 16, 52, 53] have shown that the self-attention module in diffusion U-Net [42] has two paths that have different roles, the query-key similarities for *structure* and values for *appearance*. Specifically, in the self-attention module, we compute the query $Q_t \in \mathbb{R}^{(h \times w) \times d}$, key $K_t \in \mathbb{R}^{(h \times w) \times d}$, value $V_t \in \mathbb{R}^{(h \times w) \times d}$ at timestep t , where h , w , and d refer to the height, width, and channel dimensions, respectively. The resulting output

from this module is defined by:

$$\text{SA}(Q_t, K_t, V_t) = \underbrace{\text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d}}\right)}_{\text{structure}} \overbrace{V_t}^{\text{appearance}} = \mathbf{A}_t V_t, \quad (12)$$

where the *structure* part is commonly referred to as the self-attention map.

Motivated by this insight, we focus on perturbing only the self-attention map to minimize excessive deviation from the original sample. This perspective can also be understood from the viewpoint of addressing out-of-distribution (OOD) issues for neural network inputs. Directly perturbing the appearance component V_t may cause the subsequent multilayer perceptron (MLP) to encounter inputs that it has not previously seen. This leads to OOD issues for MLP, resulting in significantly distorted samples. We will discuss this further in the experiments.

However, a linear combination of value features, such as using an identity matrix as a self-attention map that maintains the value of each element, is more likely to remain within the domain than direct perturbations to V_t . Therefore, we only perturb the *structural* component, $\mathbf{A}_t = \text{Softmax}(Q_t K_t^T / \sqrt{d}) \in \mathbb{R}^{hw \times hw}$, to eliminate the structural information while preserving the appearance information. This simple approach of replacing the selected self-attention map with an identity matrix $\mathbf{I} \in \mathbb{R}^{hw \times hw}$ can be defined as

$$\text{PSA}(Q_t, K_t, V_t) = \mathbf{I} V_t = V_t, \quad (13)$$

where we call perturbed self-attention (PSA). More ablation studies on perturbing a self-attention map can be found in 5.5 and the Appendix D.1.

By using SA and PSA module, we implement $\epsilon_\theta(x_t)$ and $\hat{\epsilon}_\theta(x_t)$, respectively. Fig. 4 illustrates the overall pipeline of our method, dubbed Perturbed-Attention Guidance (**PAG**). The input image x_t is fed into $\epsilon_\theta(\cdot)$ and $\hat{\epsilon}_\theta(\cdot)$ and the output of the two networks are linearly combined to get the final noise prediction $\tilde{\epsilon}_\theta(x_t)$ as in Eq. 10. The pseudo-code is provided in Alg. 1.

Algorithm 1 Sampling with PAG

Model(x_t), **Model'**(x_t) :
Diffusion model with self-attention and
perturbed self-attention (PSA), respectively.
 s : guidance scale, Σ_t : variance
 $x_T \sim \mathcal{N}(0, I)$
for t in $T, T-1, \dots, 1$ **do**
 $\epsilon_t \leftarrow \text{Model}(x_t)$, $\hat{\epsilon}_t \leftarrow \text{Model}'(x_t)$
 $\tilde{\epsilon}_t \leftarrow \epsilon_t + s(\epsilon_t - \hat{\epsilon}_t)$ ▷ Eq. 10
 $x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\tilde{\epsilon}_t), \Sigma_t\right)$ ▷ Eq. 3
end for
return x_0

4.3 Analysis on PAG

In this section, we delve into the reasons behind the effectiveness of our guidance method. Fig. 3 illustrates the diffusion model's sampling process utilizing PAG. Each row, apart from the last, showcases \hat{x}_0 at each timestep, employing the original epsilon prediction $\epsilon_\theta(x_t)$, the perturbed epsilon prediction $\hat{\epsilon}_\theta(x_t)$, and the guided epsilon $\tilde{\epsilon}_\theta(x_t)$. The final row in (b) presents the guidance signal $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$.

Table 1: Quantitative results on ADM [10]. The best values are in bold.

Model	Guidance	FID ↓	IS ↑	Precision ↑	Recall ↑
ImageNet 256×256 Unconditional	✗	26.21	39.70	0.61	0.63
	SAG	20.08	45.56	0.68	0.59
	PAG	16.23	88.53	0.82	0.51
ImageNet 256×256 Conditional	✗	10.94	100.98	0.69	0.63
	SAG	9.41	104.79	0.70	0.62
	PAG	6.32	338.02	0.51	0.82

This illustration demonstrates how our guidance term provides semantic cues. In the section marked by the red rectangle in Fig. 3, it is apparent that the perturbed prediction (row 3 in (b)) is missing salient features like eyes, nose, and tongue. This omission occurs because the perturbed prediction $\hat{\epsilon}_\theta(x_t)$ lacks an understanding of the global structure, which inhibits its ability to generate these essential features. As a result, the difference $\hat{\Delta}_t$ tends to focus on these salient points (row 4 in (b)). By adding this difference $\hat{\Delta}_t$ to the original prediction $\epsilon_\theta(x_t)$, the structure of the sample is strengthened, demonstrated in the first row of (b) in Fig. 3. It is also worth noting that, as the timestep approaches zero, $\hat{\Delta}_t$ progressively captures finer details, as illustrated in Fig. 3 and Fig. 42 in the Appendix. This suggests that PAG effectively guides samples across all timesteps towards a well-defined shape, transitioning from coarse structures to intricate details. Further analysis is available in Appendix E.1 and E.2.

We further illustrate a similar visualization with CFG [19] in Stable Diffusion in Fig. 2, showing that CFG constructs an *undesirable* sampling path through unconditional generation to enhance class conditioning.

5 Experiments

5.1 Experimental and Implementation Details

All experiments are conducted on NVIDIA GeForce RTX 3090 GPU and NVIDIA RTX A6000 GPU for image sampling. Our work is based on the pretrained models, ADM¹ [10] and Stable Diffusion² [41]. We accessed all necessary weights from their publicly available repositories and adopted the same evaluation metrics utilized in ADM [10], including FID [17], IS [45], and Improved Precision and Recall [27]. For additional details on the experimental setup, please refer to Appendix A. Detailed implementation details such as the selected self-attention layers can be found at Appendix A.

¹ <https://github.com/openai/guided-diffusion>

² <https://huggingface.co/runwayml/stable-diffusion-v1-5>

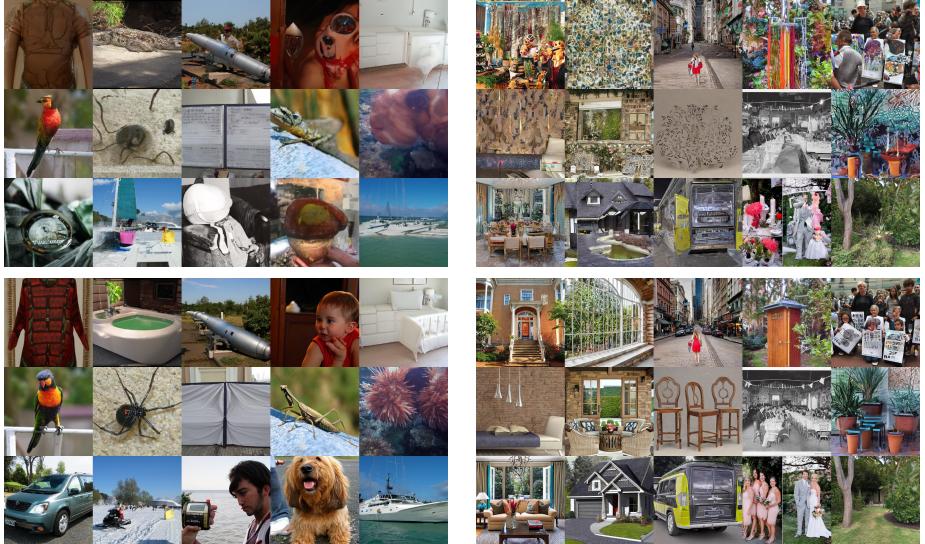


Fig. 5: Uncurated samples in unconditional generation w/o and w/ PAG. Figures display randomly selected samples from (Left) ADM [10] and (Right) Stable Diffusion [41]. Each set of images shows sampling without (Top) and with (Bottom) PAG. Samples guided by PAG appear more realistic and demonstrate semantically coherent structures in both ADM [10] and SD [41].

5.2 Pixel-Level Diffusion Models

With pretrained ADM [10], we generates 50K samples on ImageNet [8] 256×256 to evaluate metrics. In Table 1, we compare ADM [10] with SAG [20] and PAG in both conditional and unconditional generation. Table 1 shows that ADM [10] with PAG outperforms the others with large margin in FID [17], IS [45]. The contrastive patterns of Improved Recall and Precision [27] in unconditional and conditional generation in Table 1 are attributed to the trade-off between fidelity and diversity [10, 19, 20]. Despite this trade-off, the uncurated samples illustrated in Fig. 5 exhibit significant enhancements in quality, demonstrating PAG’s capability to rectify the diffusion sampling path leveraging perturbed self-attention. A qualitative comparison with SAG [20] is also presented in Fig. 6. For further exploration, additional samples generated using both the unconditional and conditional models of ADM [10] are available in Appendix B.1.

5.3 Latent Diffusion Models

Unconditional generation on Stable Diffusion. We further explored the application of our guidance to Stable Diffusion [41]. In the “Unconditional” part of Table 2, we compared the baseline without PAG to that with PAG for unconditional generation without prompts. The use of PAG resulted in improved FID [17] and IS [45]. Samples from Stable Diffusion’s unconditional generation



Fig. 6: Qualitative comparison between SAG [20] and PAG. Images are sampled from the ImageNet 256×256 unconditional model using the same seed sequence. Compared to samples guided by SAG, those guided by PAG exhibit significantly improved semantic structures with artifacts removed.

Table 2: Quantitative results on Stable Diffusion [41]. The results were obtained using Stable Diffusion v1.5. Sampling was conducted for each with 30K images, and the results were measured accordingly. For text-to-image tasks, 30k prompts were randomly selected from the MS-COCO 2014 validation set [29].

Type	Condition	PAG	CFG	FID ↓	IS ↑
Unconditional	X	X	-	53.13	16.26
		✓	-	47.57	21.38
Text-to-Image	✓	X	X	25.20	22.97
		X	✓	15.00	40.43
		✓	X	10.08	33.02
		✓	✓	8.73	36.99

with and without PAG are presented in the right column of Fig. 5 and in the top row of Fig. 1. Without PAG, the majority of images tend to exhibit semantically unusual structures or lower quality. In contrast, the application of PAG leads to the generation of geometrically coherent objects or scenes, significantly enhancing the visual quality of the samples compared to the baseline.

Table 3: Diversity comparison in samples generated by CFG [19] and PAG.

	IS ↑	LPIPS ↑
CFG	1.82	0.64
PAG	2.32	0.68

Text-to-image synthesis on Stable Diffusion. Results for text-to-image generation using prompts are presented in the “Text-to-Image” part of Table 2. In this case, since CFG [19] can be utilized, we conducted sampling in four different scenarios: without applying guidance as a



Fig. 7: Qualitative comparison between CFG [19] and CFG + PAG. Compared to using CFG alone, incorporating PAG alongside CFG noticeably improves the semantic coherence of the structures within the samples. This combination effectively rectifies errors in existing samples, such as adding a missing eye to a cat or eliminating extra legs from a zebra.

baseline, using CFG, using PAG, and combining both guidance methods with an appropriate scale.

Interestingly, combining PAG and CFG [19] with an appropriate scale leads to a significant improvement in the FID of the generated images. Fig. 7 offers a qualitative comparison between samples produced using solely CFG and those generated with both guidance methods. The synergy of CFG’s effectiveness in aligning images with text prompts and PAG’s enhancement of structural information culminates in visually more appealing images when these methods are applied together. Further analysis on the complementarity between PAG and CFG is provided in the Appendix E.2.

To examine the trade-off between sample quality and diversity when using CFG, we initially define per-prompt diversity as “*the capacity to generate a variety of samples for a given prompt*”. In text-to-image synthesis, this involves generating multiple images from different latents for a single prompt, forming a batch of generated samples. Assessing metrics on such a batch may not effectively measure per-prompt diversity. Thus, to compare the per-prompt diversity of CFG and PAG, we conduct samplings using various latents for a single prompt. For this comparison, the Inception Score (IS) [45] is calculated over 1000 generated samples, and the LPIPS [59] metric is averaged across pairwise comparisons of 100 samples (yielding 4950 pairs). The values presented in Table 3 are averages from experiments conducted on 20 prompts, chosen not by selection but by using the first 20 prompts based on the IDs from the MS-COCO 2014 validation set [29]. Further samples from Stable Diffusion are available in Appendix B.2 for additional reference.

Table 4: Quantitative results of PSLD [44] on FFHQ [23] 256×256 1K validation set.

Method	Box Inpainting		SR (8×)		Gaussian Deblur		Motion Deblur	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
PSLD	43.11	0.167	42.98	0.360	41.53	0.221	93.39	0.450
PSLD + PAG (Ours)	21.13	0.149	38.57	0.354	37.08	0.343	40.26	0.397



Fig. 8: Qualitative results of PSLD [44] with our PAG on FFHQ [23] dataset. Left Top: Box inpainting. Left Bottom: Super-resolution (×8). Right Top: Gaussian deblur. Right Bottom: Motion deblur. Using PAG leads to the removal of artifacts and blurriness, resulting in more realistic restorations.

5.4 Downstream Tasks

Inverse problems. Inverse problem is one of the major tasks in the unconditional generation, which aims to restore x from the noisy measurement $y = \mathcal{A}(x) + n$, where $\mathcal{A}(\cdot)$ denotes measurement operator (*e.g.*, Gaussian blur) and n represents a vector of noise. In this task, where text prompts are not available, our guidance can operate properly to improve sample quality without prompts, whereas it is challenging to utilize existing guidance methods that require prompts. We test PAG using a subset of FFHQ [23] 256×256 on PSLD [44] which leverages DPS [6] and LDM [41] to solve linear inverse problems. More details about experimental settings are provided in Appendix A.

Table 4 shows the quantitative results of PSLD with PAG on box inpainting, super-resolution (×8), gaussian deblur, and motion deblur. The performance of PSLD with PAG outperforms all of the tasks in FID [17], and mostly in LPIPS [59]. Fig. 8 highlights a considerable improvement in the quality of restored samples using PAG, with a notable reduction of artifacts present in the original method. In addition, PAG can be adopted to any other restoration model based on diffusion models, which can be shown through further experiments in Appendix C.

ControlNet. ControlNet [58], a method for introducing spatial conditioning controls in pretrained text-to-image diffusion models, sometimes struggles to produce high-quality samples under unconditional generation scenarios, particularly when the spatial control signal is sparse, such as pose conditions. However, as demonstrated in Fig. 9, PAG enhances sample quality in these instances. This



Fig. 9: ControlNet [58] sample images conditioned by pose and depth without text prompt. Samples guided by PAG appear more realistic, exhibiting fewer artifacts and semantically coherent structure.

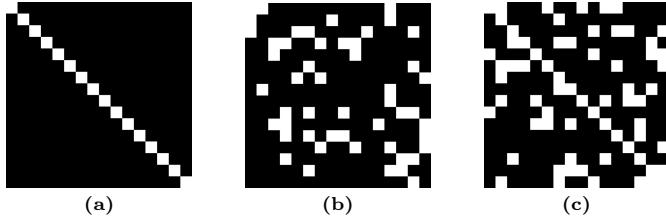


Fig. 10: Ablations study of self-attention map masking strategy. For the evaluation of FID [17], we sample 5K images from ADM [10] ImageNet [8] 256×256 unconditional model for each method. Black entries indicate the masked (set to $-\infty$) elements of the $Q_t K_t^T / \sqrt{d}$ component in Eq. 12 before the Softmax operation is applied. (a) Replacing attention map with identity matrix. **FID: 32.34**, (b) Random masking (ratio: 0.25). **FID: 40.20**, (c) Random masking off-diagonal entries (ratio: 0.25). **FID: 39.49**.

enables the generation of plausible samples conditioned solely on spatial information without the need for specific prompts, making it useful for crafting training datasets tailored to specific goals and allows artists to test diverse, imaginative works without relying on detailed prompts.

5.5 Ablation Studies

Self-attention perturbation strategy. We tested various methods for perturbing the self-attention map. The illustration of these approaches and qualitative results are presented in Fig. 10, including (a) random masking, (b) replacing the attention map with an identity matrix, and (c) applying random masking to off-diagonal entries only. Results for additional perturbations can be found in the Appendix D.1.



Fig. 11: Effects of PAG scales. With the increase in guidance scale s , images gain more semantically coherent structures and exhibit fewer artifacts, enhancing their overall quality. Note that overly large guidance scale can lead to smoother textures and slight saturation in the images, similarly to CFG [19]. Users can select the appropriate guidance scale according to their needs.

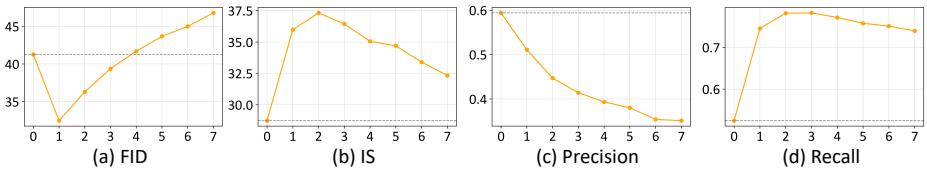


Fig. 12: Ablation study of the guidance scale.

Guidance scale. We conduct experiments to investigate the performance difference based on guidance scale. Using scales set from 0.0 to 7.0 with intervals of 1.0, we sampled 5K images with ADM [10] and measured FID [17], IS [45], Precision, and Recall metrics [27] for these images. The results can be seen in the graph in Fig. 12. PAG showed the best FID at a guidance scale of 1.0 and the best IS at a guidance scale of 2.0.

Additionally, we conduct a qualitative comparison of the guidance scale for unconditional generation using Stable Diffusion [41]. In Fig. 11, it can be observed that as the guidance scale increases from 0.0, the structure of the sampled images improves, leading to more natural images with fewer artifacts.

Computational cost. PAG, like CFG, can parallelize the two denoising passes in Fig. 4 by duplicating the input of the Diffusion U-Net and making a batch. As a result, the computational cost is nearly identical to that of CFG, and details on time and memory consumption are provided in the Appendix A.6.

6 Conclusion

In this work, we proposed a novel guidance method, termed Perturbed-Attention Guidance (**PAG**), which leverages structural perturbation for improved image

generation. Starting with an elucidation of how CFG [19] refines sample realism, by replacing the diffusion U-Net’s self-attention map with an identity matrix, we effectively guide the generation process away from structural degradation. Crucially, PAG achieves superior sample quality in both conditional and unconditional settings, requiring no additional training or external modules. Furthermore, we demonstrate the versatility of PAG by showing its effectiveness in downstream tasks such as image restoration. We believe that our exploration enriches the understanding of sampling guidance methods and diffusion models, and illuminates the applicability of unconditional diffusion models, liberating diffusion models from reliance on text prompts and CFG.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) [6](#)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022) [3](#), [8](#), [52](#)
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 843–852 (2023) [3](#), [6](#)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023) [4](#), [8](#)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) [2](#)
6. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687 (2022) [1](#), [2](#), [3](#), [14](#), [21](#), [23](#), [35](#), [37](#), [38](#), [39](#), [40](#)
7. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. Advances in Neural Information Processing Systems **35**, 25683–25696 (2022) [2](#)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., 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) [11](#), [15](#), [21](#), [23](#), [35](#), [36](#), [37](#), [38](#), [39](#), [45](#), [47](#), [49](#)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) [3](#)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) [2](#), [3](#), [4](#), [6](#), [10](#), [11](#), [15](#), [16](#), [21](#), [22](#), [23](#), [25](#), [26](#), [27](#), [28](#), [37](#), [45](#), [46](#), [47](#), [48](#), [50](#), [55](#)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020) [3](#)

12. Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems* **36** (2024) [3](#), [6](#)
13. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. <https://github.com/threestudio-project/threestudio> (2023) [43](#)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) [51](#)
15. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023) [51](#)
16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) [3](#), [8](#)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017) [10](#), [11](#), [14](#), [15](#), [16](#), [21](#), [37](#), [49](#)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020) [2](#), [3](#), [4](#), [37](#), [47](#)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [10](#), [11](#), [12](#), [13](#), [16](#), [17](#), [22](#), [24](#), [40](#), [43](#), [50](#), [51](#), [53](#), [54](#)
20. Hong, S., Lee, G., Jang, W., Kim, S.: Improving sample quality of diffusion models using self-attention guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7462–7471 (2023) [3](#), [6](#), [7](#), [11](#), [12](#), [21](#), [51](#), [54](#), [55](#)
21. Ignatov, A., Timofte, R., et al.: Pirm challenge on perceptual image enhancement on smartphones: report. In: European Conference on Computer Vision (ECCV) Workshops (January 2019) [40](#)
22. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. arXiv preprint arXiv:2401.09603 (2023) [21](#)
23. Karras, T., Laine, S., 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) [14](#), [23](#), [32](#), [33](#), [34](#), [37](#), [40](#)
24. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023) [51](#)
25. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. *Advances in Neural Information Processing Systems* **35**, 23593–23606 (2022) [2](#), [35](#)
26. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023) [8](#)
27. Kynkäanniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems* **32** (2019) [10](#), [11](#), [16](#)
28. Li, T., Katabi, D., He, K.: Self-conditioned image generation via generating representations. arXiv:2312.03701 (2023) [2](#)

29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) [12](#), [13](#)
30. Luo, G., Darrell, T., Wang, O., Goldman, D.B., Holynski, A.: Readout guidance: Learning control from diffusion features. arXiv preprint arXiv:2312.02150 (2023) [3](#), [6](#)
31. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023) [4](#)
32. Nam, J., Kim, H., Lee, D., Jin, S., Kim, S., Chang, S.: Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. arXiv preprint arXiv:2402.09812 (2024) [3](#), [4](#), [8](#)
33. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) [6](#)
34. Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. Advances in Neural Information Processing Systems **36** (2024) [52](#)
35. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. arXiv preprint arXiv:2303.11306 (2023) [4](#)
36. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022) [4](#)
37. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers> (2022) [22](#)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) [21](#), [43](#)
39. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023) [4](#), [8](#)
40. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) [3](#)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 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) [2](#), [3](#), [4](#), [10](#), [11](#), [12](#), [14](#), [16](#), [21](#), [29](#), [30](#), [31](#), [40](#), [41](#), [42](#), [47](#), [48](#)
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015) [4](#), [8](#), [47](#)
43. Rout, L., Chen, Y., Kumar, A., Caramanis, C., Shakkottai, S., Chu, W.S.: Beyond first-order tweedie: Solving inverse problems using latent diffusion. arXiv preprint arXiv:2312.00852 (2023) [2](#)
44. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. Advances in Neural Information Processing Systems **36** (2024) [1](#), [2](#), [3](#), [14](#), [23](#), [32](#), [33](#), [34](#), [35](#), [36](#)

45. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016) [10](#), [11](#), [13](#), [16](#)
46. Simsar, E., Tonioni, A., Xian, Y., Hofmann, T., Tombari, F.: Lime: Localized image editing via attention regularization in diffusion models. arXiv preprint arXiv:2312.09256 (2023) [8](#)
47. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) [2](#), [3](#)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) [3](#), [4](#), [22](#), [23](#), [47](#), [49](#)
49. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019) [2](#), [3](#), [4](#)
50. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) [2](#), [3](#), [4](#)
51. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**(1), 1929–1958 (2014) [51](#)
52. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) [3](#), [8](#)
53. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plugand-play diffusion features for text-driven image-toimage translation. arXiv preprint arXiv:2211.12572 (2022) [3](#), [8](#)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [3](#)
55. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: International conference on machine learning. pp. 1058–1066. PMLR (2013) [51](#)
56. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022) [2](#)
57. Wu, J., Huang, Z., Thoma, J., Acharya, D., Van Gool, L.: Wasserstein divergence for gans. In: Proceedings of the European conference on computer vision (ECCV). pp. 653–668 (2018) [6](#)
58. Zhang, L., Rao, A., 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) [1](#), [3](#), [14](#), [15](#), [23](#)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) [13](#), [14](#), [37](#)

Appendix

In the following, we provide detailed information on the implementation of all experiments (Sec. A), along with a broader range of qualitative results from samples enhanced by the Perturbed-Attention Guidance (PAG), which includes human evaluations and results from downstream tasks (Sec. B). Additionally, we highlight intriguing applications where PAG proves beneficial, such as DPS [6], the Stable Diffusion [41] super-resolution/inpaint pipeline, and text-to-3D [38] (Sec. C). We also present ablation studies focusing on perturbation methods and layer selection (Sec. D). Finally, a comprehensive analysis of CFG and PAG, including the dynamics of using CFG and PAG concurrently, is provided (Sec. E). Discussion on limitations is also included (Sec. F).

A Implementation Details

In this section, we provide detailed descriptions of the implementation and hyperparameter settings for all experiments in the paper.

A.1 Experiments on ADM

Quantitative results. For the main quantitative result presented in the main paper involving the ADM [10] ImageNet [8] 256×256 conditional and unconditional models, we utilized the official GitHub repository³ of ADM along with its publicly available pretrained weights. Our work builds upon the SAG [20] repository⁴, which is derived from the ADM official repository, to ensure precise comparison. We configured the PAG scale $s = 1.0$ and defined the perturbation to the self-attention mechanism as substituting $\text{Softmax}(Q_t K_t^T / \sqrt{d}) \in \mathbb{R}^{hw \times hw}$ with an identity matrix $\mathbf{I} \in \mathbb{R}^{hw \times hw}$. Here, Q_t , and K_t represent the query and key at timestep t and h , w , and d refer to the height, width, and channel dimensions, respectively. The specific layers for applying perturbed self-attention are as follows: `input_blocks.14.1`, `input_blocks.16.1`, `input_blocks.17.1`, `middle_block.1` for unconditional models and `input_blocks.14.1` for conditional models. We follow the same evaluation protocol as SAG [20], utilizing the DDPM sampler with 250 steps and employing the same evaluation code as provided by the official repository of ADM.

Qualitative results. For the qualitative results in the main paper, we configured the PAG scale $s = 3.0$. This choice of a higher s value stems from our observations in the ablation study on guidance scale. It shows that although sample quality improves with an increasing guidance scale the FID [17] score worsens. This may be due to the misalignment between FID and human perception [22]. Consequently, we increase the guidance scale to prioritize perceived quality improvement. We applied the same identity matrix substitution and the same layers for perturbed self-attention as in the quantitative experiments.

³ <https://github.com/openai/guided-diffusion>

⁴ <https://github.com/KU-CVLAB/Self-Attention-Guidance>

Visualization of diffusion sampling path. For the visualization of the reverse process in the Fig. 3, we obtain $\hat{\Delta}_t$ by calculating the absolute value of each channel, computing the channel-wise mean, and clipping outlier values to enhance clarity. The hyperparameters are consistent with those in the qualitative results with ADM [10].

A.2 Experiments on Stable Diffusion

Quantitative results. For all the quantitative experiments, we utilized Stable Diffusion v1.5⁵ implemented based on the pipeline provided by the Diffusers [37]. For the PAG guidance scale, $s = 2.0$ is used for unconditional generation, while $s = 2.5$ is used for text-to-image synthesis. In text-to-image synthesis, CFG [19] was set to the most commonly used value of $w = 7.5$, and for experiments combining CFG and PAG, $w = 2.0$ and $s = 1.5$ were employed. For the diversity comparison in the main paper, $s = 4.5$ and $w = 7.5$ were used respectively. In all experiments, perturbed self-attention was applied to the middle layer `mid_block.attentions.0.transformer_blocks.0.attn1` of the U-Net, and sample images were generated through DDIM [48] 50 step sampling method.

Qualitative results. Stable Diffusion v1.5 is used for all qualitative generation results. For the main qualitative results, PAG guidance scale $s = 4.5$ is used. Also, for CFG experiments, CFG guidance scale $w = 7.5$ was applied, and for the CFG+PAG experiment, $w = 6.0$ and $s = 1.5$ were used. We used DDIM sampling [48] with 200 steps for the teaser (Fig. 1), 50 steps for the main figure (Fig. 5), and 25 steps for comparison between CFG and CFG + PAG (Fig. 7). Perturbed self-attention was applied to the middle layer `mid_block.attentions.0.transformer_blocks.0.attn1` of the U-Net in all cases.

Visualization of diffusion sampling path. For the visualization experiment of reverse process in the main figure (Fig. 2), CFG [19] scale $w = 7.5$ is used, and perturbed self-attention was applied to the middle layer `mid_block.attentions.0.-transformer_blocks.0.attn1`, representing the initial 12 steps of DDIM 25 step sampling.

Combination of CFG and PAG. To apply CFG [19] and PAG together in text-to-image synthesis, we produced $\tilde{\epsilon}_\theta(x_t, c)$ using the following equation:

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + w(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)) + s(\epsilon_\theta(x_t, c) - \hat{\epsilon}_\theta(x_t, c)), \quad (14)$$

where w and s are guidance scale. These estimations involve adding the deltas of CFG and PAG, each weighted by each guidance scale w and s . To achieve this, we computed three estimations, $\epsilon_\theta(x_t, c)$, $\epsilon_\theta(x_t, \phi)$, and $\hat{\epsilon}_\theta(x_t, c)$ simultaneously, in the denoising U-Net.

⁵ <https://huggingface.co/runwayml/stable-diffusion-v1-5>

A.3 Experiments with PSLD

We use Stable Diffusion v1.5 used in PSLD [44]. The measurement operators for inverse problems are from DPS [6], as used in PSLD [44]. PSLD [44] leverages the loss term of DPS [6] and further implements the gluing objective to enhance fidelity, multiplied with step size η and γ respectively for updating gradients. $\eta = 1.0$ and $\gamma = 0.1$ are used in experiments of PSLD [44] without PAG as same as PSLD [44]. Practically, we find that it is better to use unconditional score $\epsilon_\theta(z_t)$ instead of guided score $\tilde{\epsilon}_\theta(z_t)$ when predicting \hat{z}_0 to update gradients. Furthermore, we conduct more experiments with ImageNet [8] dataset, which are provided in Sec. B.3. All experiments with PSLD [44] use DDIM [48] sampling and all hyperparameters with PAG are in Table 5. Perturbed self-attention is applied to the same layer, `input_block.8.1.transformer_blocks.0.attn1`, for both FFHQ [23] and ImageNet [8] dataset.

FFHQ				ImageNet			
Inpaint	SR×8	Gauss	Motion	Inpaint	SR×8	Gauss	Motion
η	0.15	0.7	0.1	0.15	0.5	0.7	0.1
γ	0.015	0.07	0.01	0.015	0.05	0.07	0.01
s	4.0	4.0	5.0	4.0	4.0	4.0	5.0

Table 5: Hyperparameters for PSLD [44] with PAG on FFHQ [23] dataset and ImageNet [8] dataset. Here, η and γ are the step size for gradients of PSLD [44] and s is the scale for PAG from Eq. 10 of main paper.

A.4 Experiments with ControlNet

For the ControlNet [58] experiment in Fig. A, Stable Diffusion v1.5 was utilized, implemented based on the ControlNet pipeline from Diffusers. For pose conditional generation, PAG guidance scale 2.5 is used, while for depth conditional generation, 1.0 was employed. Sampling was conducted using the DDIM 50 steps method, and perturbed self-attention was applied to the middle layer `mid_block.attentions.0.transformer_blocks.0.attn1` of the U-Net.

A.5 Ablation Study

For the ablation study on the guidance scale and perturbation strategy, we generated 5k images using the ADM [10] ImageNet 256×256 unconditional model with DDIM 25 step sampling and applied perturbed self-attention to the `input.13` layer. In the guidance scale ablation, identity matrix replacement was used consistently across other qualitative and quantitative experiments. For qualitative results with varying guidance scales on Stable Diffusion v1.5 (Fig. 12),

Table 6: Comparison of computational costs in Stable Diffusion.

	GPU Memory ↓	Sampling Speed ↑
No Guidance	3,147 MB	19.16 iter/s
CFG [19]	3,193 MB	12.67 iter/s
PAG	3,193 MB	12.68 iter/s

DDIM 50-step sampling was utilized with perturbed self-attention applied to `mid_block.attentions.0.transformer_blocks.0.attn1`, aligning with the approach used for Stable Diffusion qualitative samples in the bottom right of the main qualitative figure.

A.6 Computational Cost

We measured the computational costs for sampling without guidance, using CFG, and using PAG in Stable Diffusion. We utilized one NVIDIA GeForce RTX 3090 GPU and conducted sampling with one batch. Firstly, we measured GPU memory usage, which appeared to be nearly identical across all three scenarios. Next, we measured the iteration speed in the denoising U-Net, showing that both CFG and PAG exhibited similar sampling speeds, albeit slightly slower when compared to not using guidance..

B Additional Qualitative Results

B.1 ADM Results

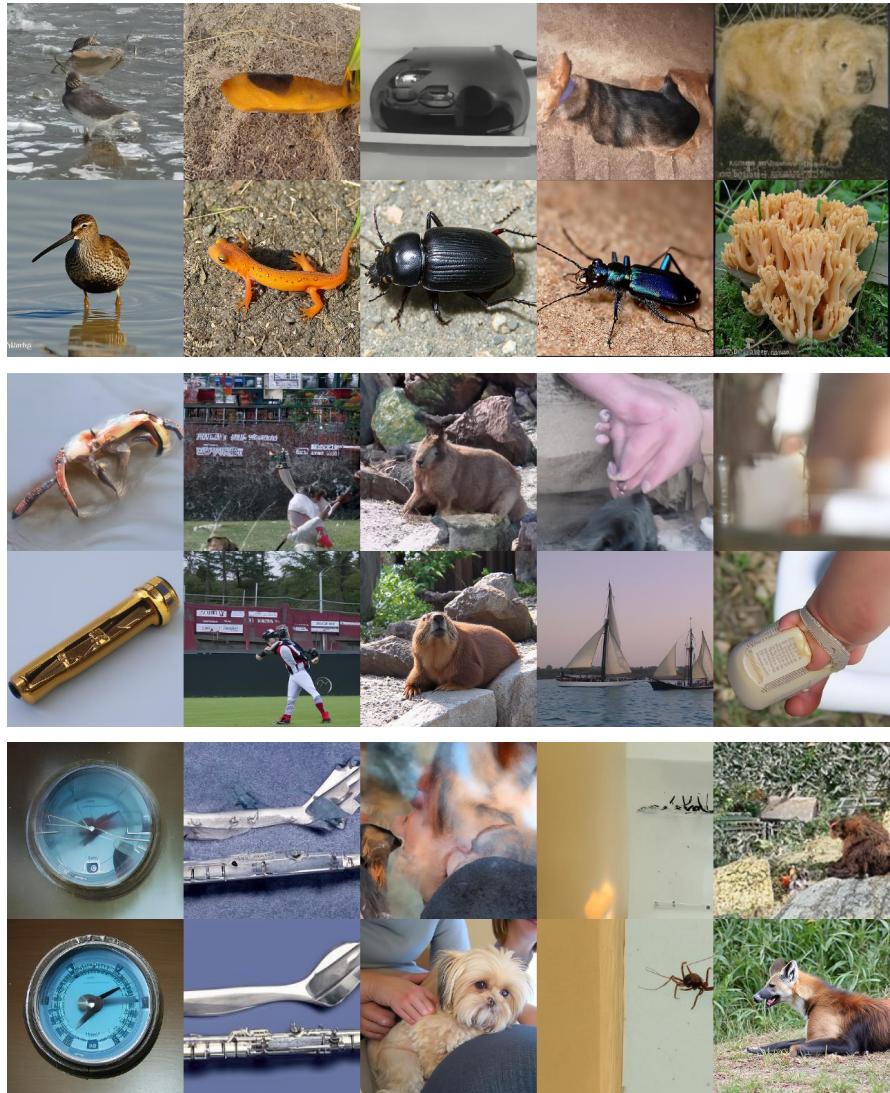


Fig. 13: Uncurated samples from ADM [10] ImageNet 256 *unconditional* model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: i13,i14,i16,m1.

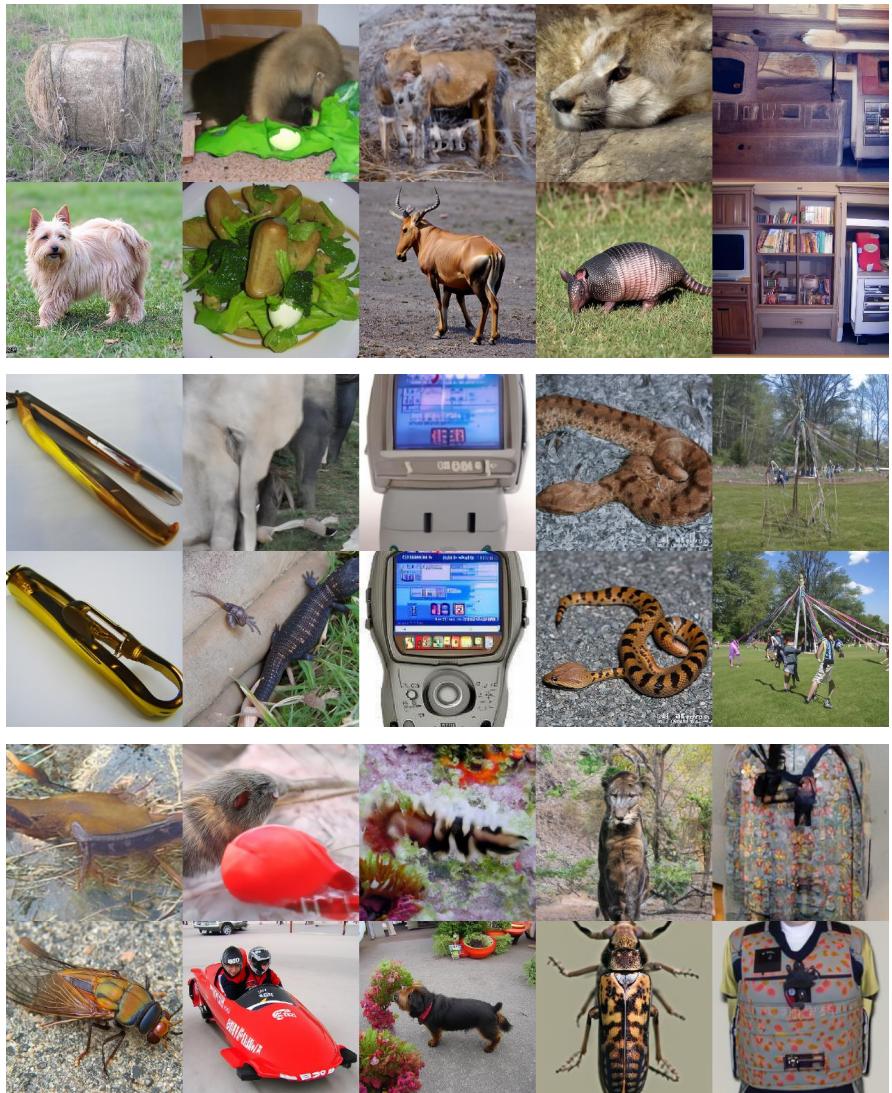


Fig. 14: Uncurated samples from ADM [10] ImageNet 256 unconditional model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: i13,i14,i16,m1.

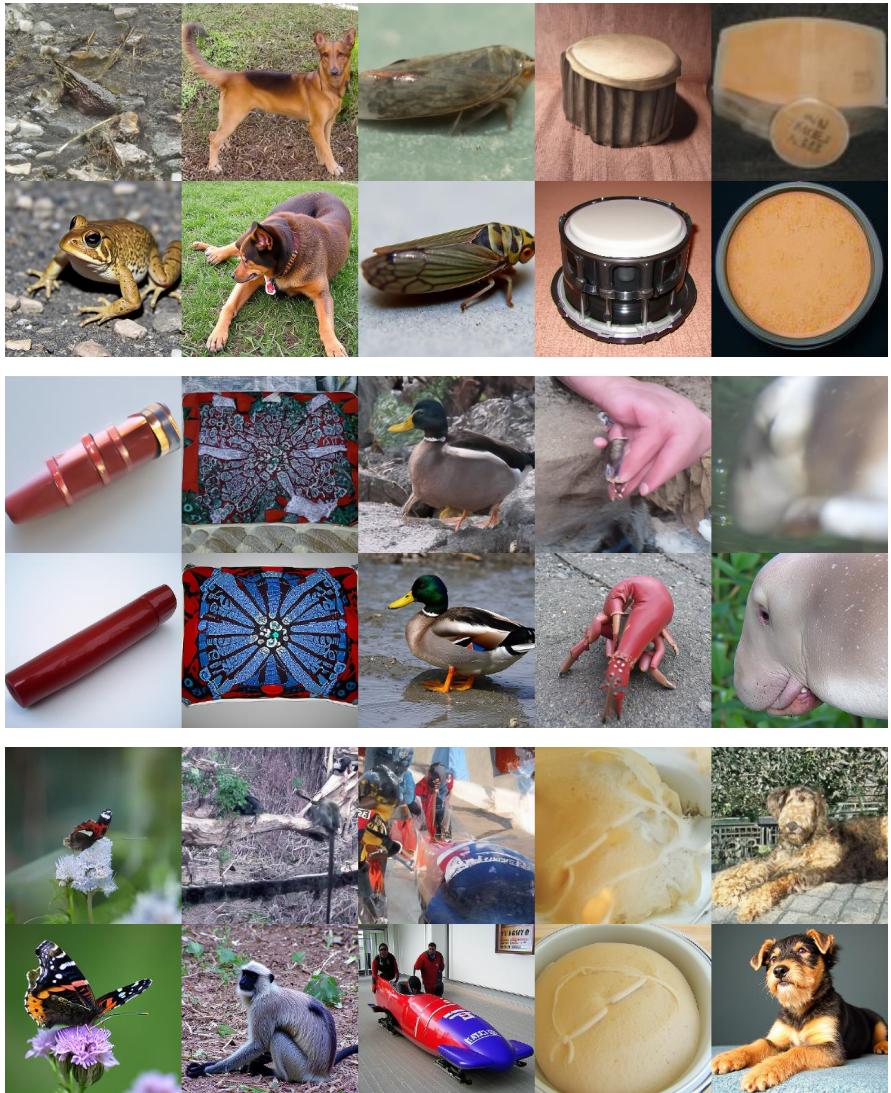


Fig. 15: Uncurated samples from ADM [10] ImageNet 256 conditional model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: i13,i14,i16,m1.



Fig. 16: Uncurated samples from ADM [10] ImageNet 256 *conditional* model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: i13,i14,i16,m1.

B.2 Stable Diffusion Results

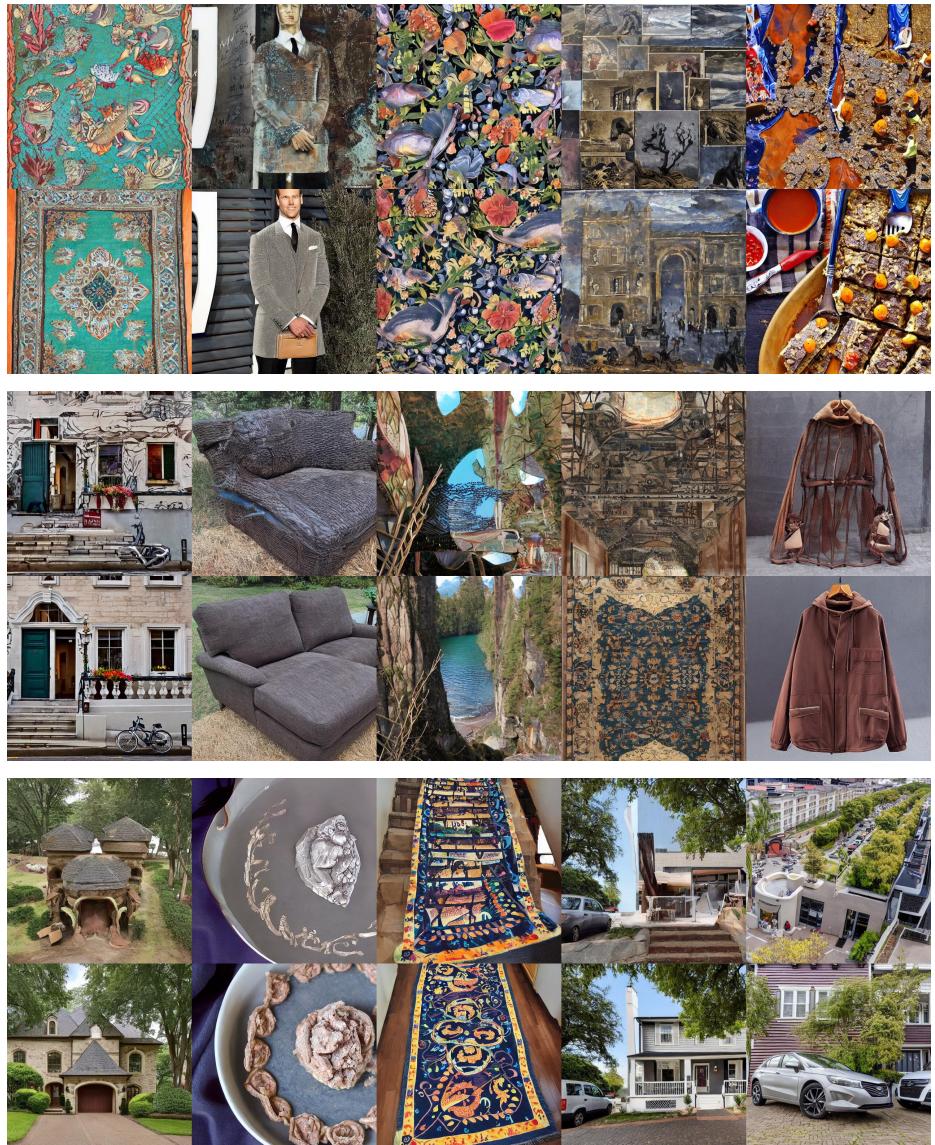


Fig. 17: Uncurated samples from SD [41] in *unconditional* generation w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 5.0$ and perturbed layer `mid_block.attentions.0.-transformer_blocks.0.attn1` are used.

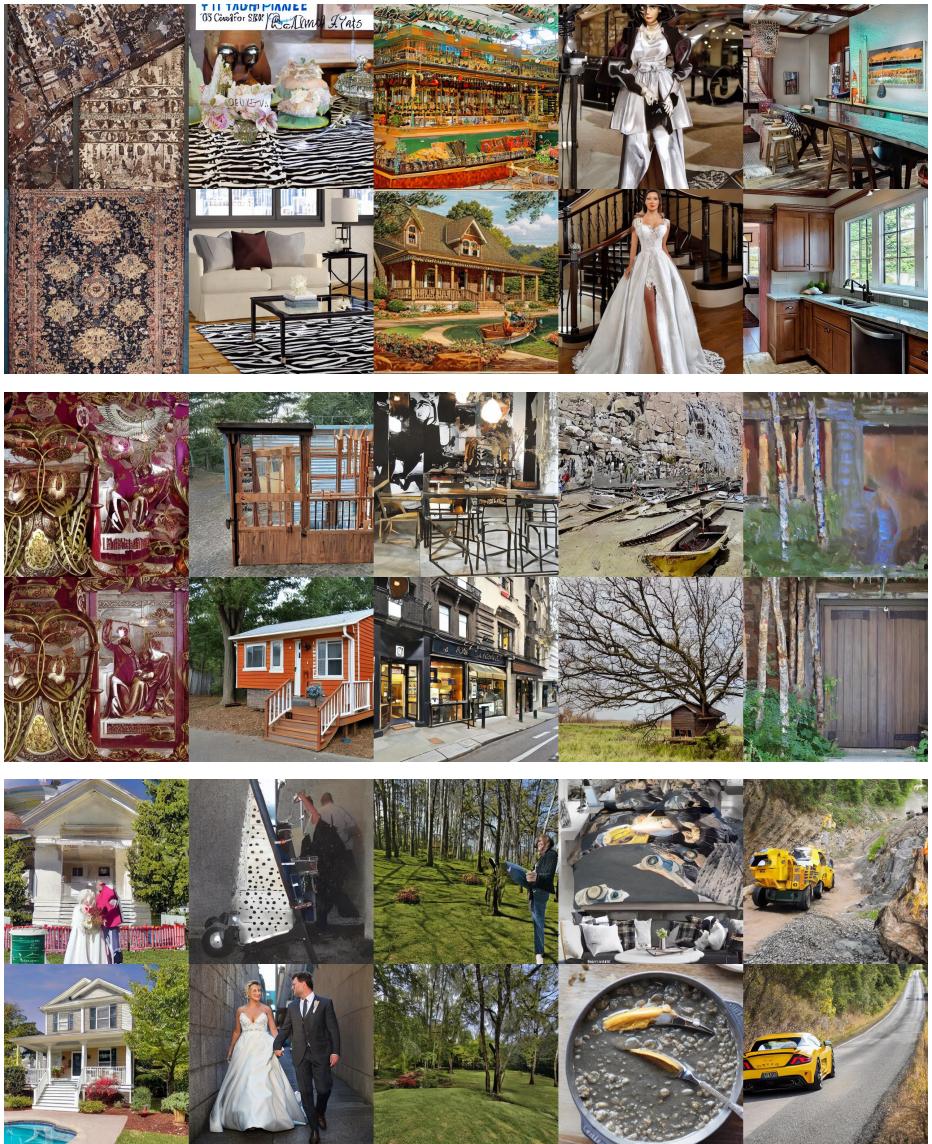


Fig. 18: Uncurated samples from SD [41] in *unconditional* generation w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 5.0$ and perturbed layer `mid_block.attentions.0.-transformer_blocks.0.attn1` are used.



Fig. 19: Uncurated samples from SD [41] in *unconditional* generation w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 5.0$ and perturbed layer `mid_block.attentions.0.-transformer_blocks.0.attn1` are used.

B.3 PSLD Results

FFHQ. Since we use Stable Diffusion v1.5, we upsample inputs to 512×512 as PSLD [44] does. Then, the outputs are downsampled to 256×256 for evaluation. Further qualitative results are provided in Fig. 20 21 22 23.

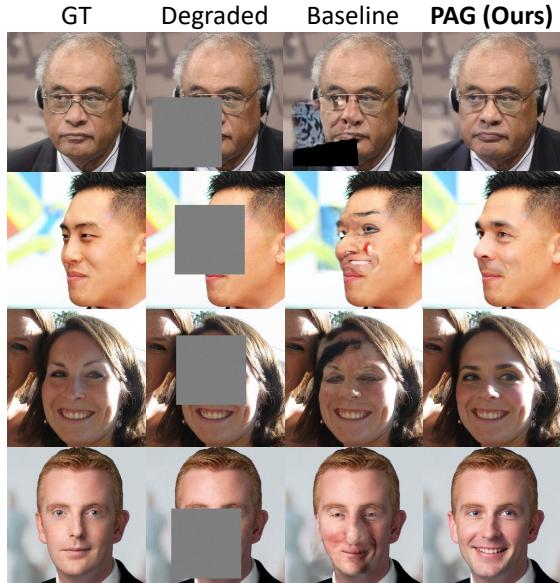


Fig. 20: Box inpainting results of PSLD [44] with PAG on FFHQ [23] dataset.

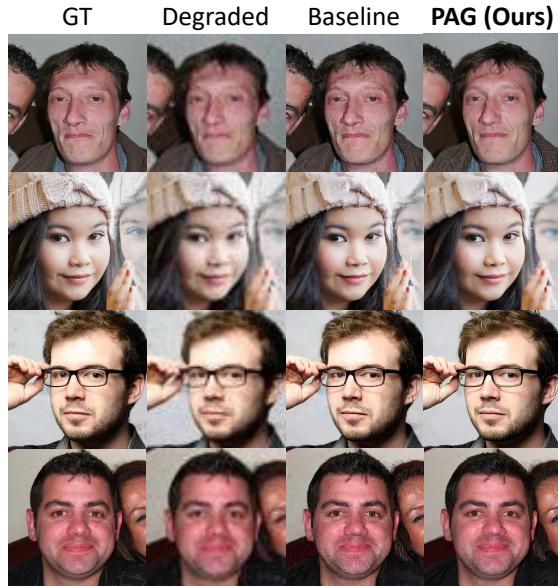


Fig. 21: Super-resolution ($\times 8$) results of PSLD [44] with PAG on FFHQ [23] dataset.

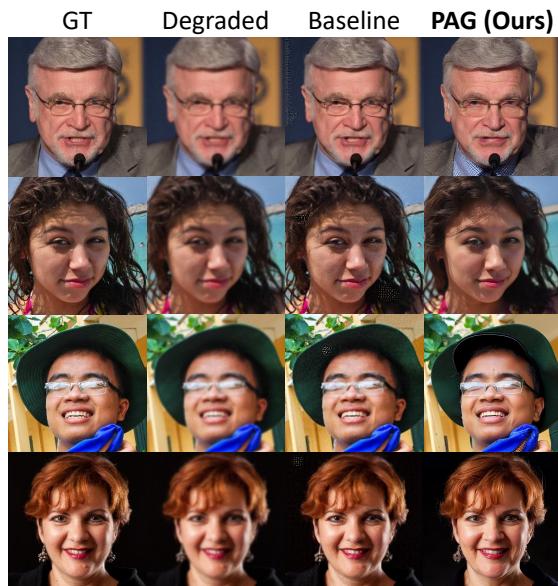


Fig. 22: Gaussian deblur results of PSLD [44] with PAG on FFHQ [23] dataset.



Fig. 23: Motion deblur results of PSLD [44] with PAG on FFHQ [23] dataset.

ImageNet. We use 1K ImageNet [8] 256×256 dataset which is used in [6, 25, 44]. Qualitative results shows that PAG properly improves sample quality with more various classes of images, as provided in 24 25 26 27.

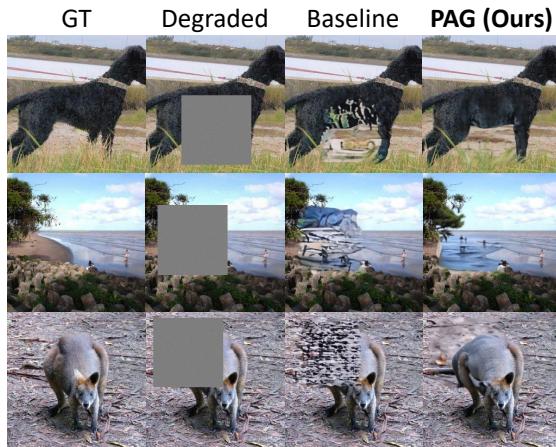


Fig. 24: Box inpainting results of PSLD [44] with PAG on ImageNet [8] dataset.

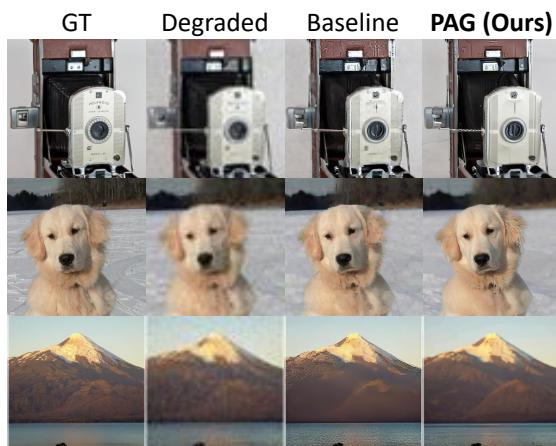


Fig. 25: Super-resolution($\times 8$) results of PSLD [44] with PAG on ImageNet [8] dataset.



Fig. 26: Gaussian deblur results of PSLD [44] with PAG on ImageNet [8] dataset.

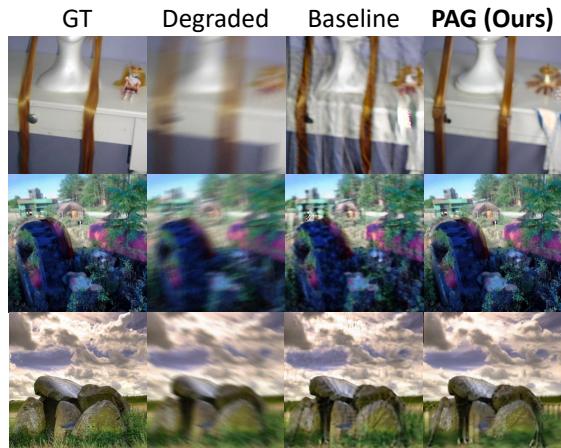


Fig. 27: Motion deblur results of PSLD [44] with PAG on ImageNet [8] dataset.

C Additional Applications

C.1 Diffusion Posterior Sampling

We conduct additional experiments on another diffusion restoration model, DPS [6], which is based on ADM [10]. DPS [6] updates the gradient of the loss term to perform sampling from the posterior distribution [6]. The detailed hyperparameters for all the DPS [6] experiments are presented in Table 7. Here, we only use the unconditional score $\epsilon_\theta(z_t)$ for predicting \hat{z}_0 , consistent with PSLD [6] experiments.

	η	FFHQ				ImageNet			
		Inpaint	SR×8	Gauss	Motion	Inpaint	SR×8	Gauss	Motion
DPS	η	1.0	1.0	1.0	1.0	1.0	1.0	0.4	0.6
DPS + PAG (Ours)	η	1.0	1.0	1.0	1.0	1.0	1.0	0.4	1.0
	s	1.0	1.0	1.0	1.0	2.0	2.0	1.0	2.0
	layer	input9.1				input9.1 middle.1 output2.1			

Table 7: Hyperparameters for DPS [6] w/o and w/ **PAG** on FFHQ [23] dataset and ImageNet [8] dataset. η is the step size for updating gradients of DPS [6] and s is the scale for PAG from Eq. 10 of main paper.

All experiments with DPS [6] use DDPM [18] sampling. Quantitative results on 1K 256 are provided in Table 8. PAG outperforms baseline on FID [17], except for super-resolution($\times 8$), where FID [17] is comparable. This result may be attributed to the point that sampling images with hard degradations can be regarded as generation rather than restoration, which emphasizes the importance of FID [17] over LPIPS [59]. Additional qualitative results are in Fig. 28 29 30 31 for ImageNet [8] dataset and Fig. 32 for FFHQ [23] dataset.

Table 8: Quantitative results of DPS [6] on FFHQ [23] 256×256 1K validation set [23].

Method	Box Inpainting		SR (8×)		Gaussian Deblur		Motion Deblur	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
DPS	33.12	0.168	34.00	0.320	44.05	0.257	39.92	0.242
DPS + PAG (Ours)	26.74	0.212	34.05	0.327	29.42	0.259	30.57	0.283

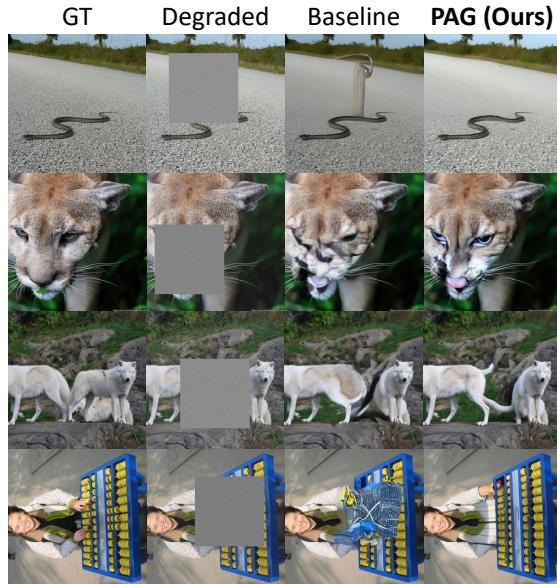


Fig. 28: Box inpainting results of DPS [6] with PAG on ImageNet [8] dataset.

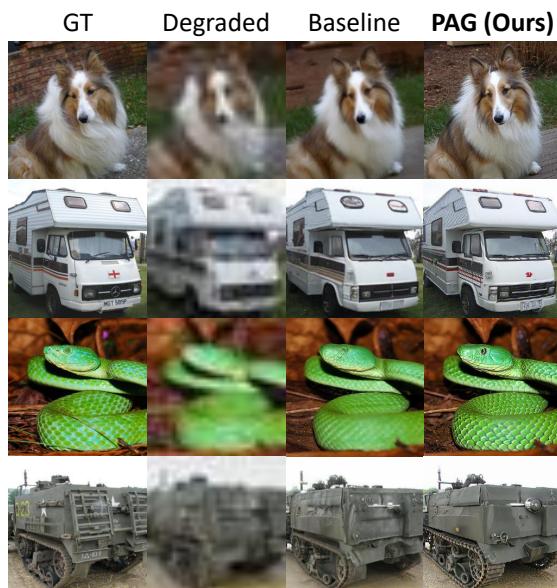


Fig. 29: Super-resolution($\times 8$) results of DPS [6] with PAG on ImageNet [8] dataset.

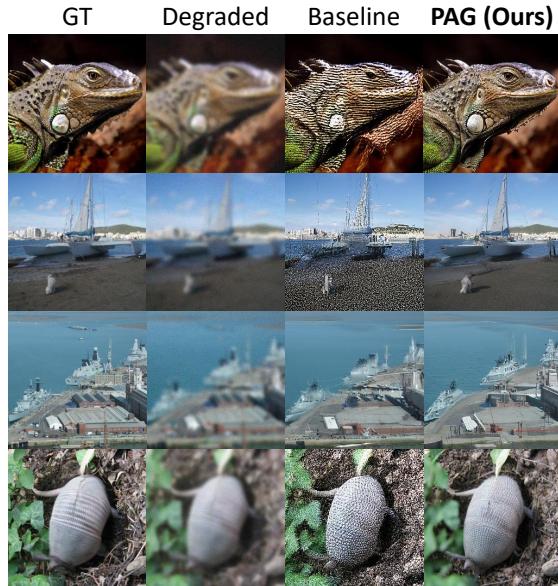


Fig. 30: Gaussian deblur results of DPS [6] with PAG on ImageNet [8] dataset.

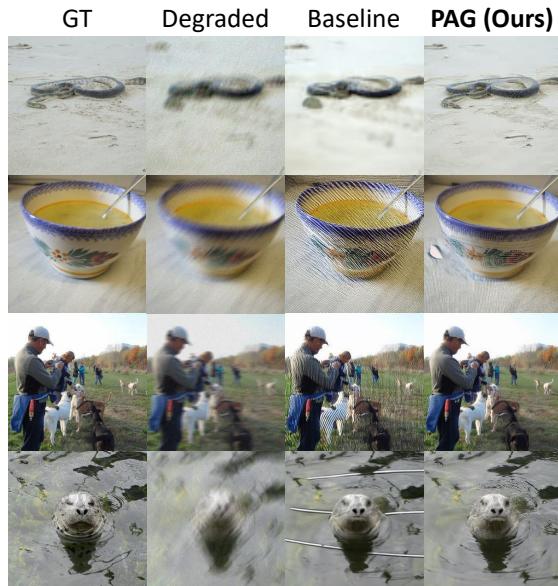


Fig. 31: Motion deblur results of DPS [6] with PAG on ImageNet [8] dataset.

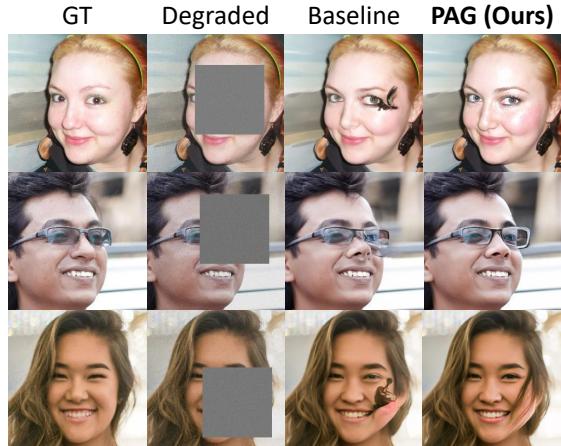


Fig. 32: Box inpainting results of DPS [6] with PAG on FFHQ [23] dataset.

C.2 Stable Diffusion Super-Resolution and Inpainting

Stable Diffusion [41] extends beyond the text-to-image pipeline to also support tasks requiring image input, such as super-resolution⁶ and inpainting⁷. The model also requires text input alongside image input to leverage CFG [19], yet there are instances where input prompts do not fit. For example, in a landscape photo, it may be more intuitive to specify only the area to be removed (such as a person in the background, shadows, or lens artifacts) and naturally fill it to match the surroundings, rather than providing a text prompt describing the entire content of the current image. Similarly, for super-resolution, it is more natural to input the image alone without having to describe it entirely in text, especially for real images. While synthetic images may have an associated creation prompt, real images do not, making it challenging to provide suitable text prompts. In contrast, PAG does not require text prompts, providing a natural way to enhance the quality of results in such pipelines. Fig. 33 and 34 present the outcomes of applying PAG to the Stable Diffusion super-resolution and inpainting pipelines, where the use of PAG produces sharper and more realistic results compared to those without it, offering a much more natural approach for these tasks. We select a subset of the DIV2K [21] dataset downsampled by a factor of 2 using bicubic interpolation and then center cropped to adjust the images to a resolution of 512×512 .

⁶ <https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler>

⁷ <https://huggingface.co/runwayml/stable-diffusion-inpainting>



Fig. 33: Comparison of Stable Diffusion [41] super-resolution results between w/o and w/ PAG. PAG applies guidance that enables the model to upscale images to high-quality renditions with clearer edges and finer details, even when using an empty prompt (3rd row). The guidance scales employed, from left to right, are sequentially 3.0, 2.0, and 1.0. The model upscaled a 256×256 input image to 512×512 .

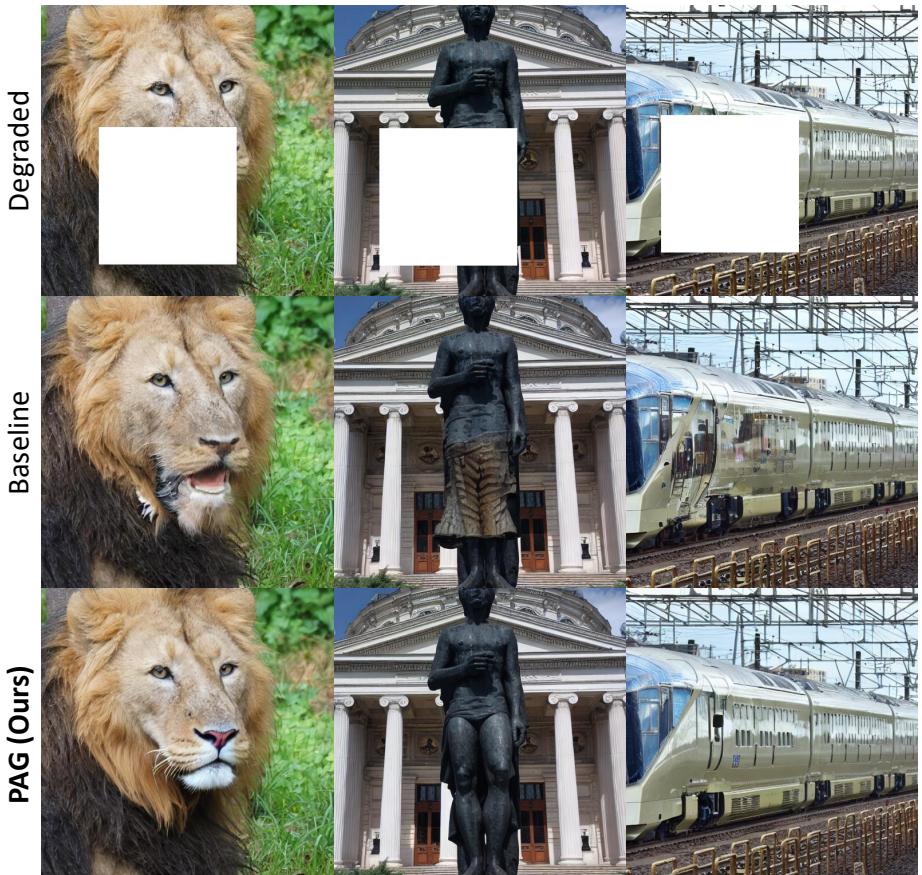


Fig. 34: Comparison of Stable Diffusion [41] inpainting results between w/o and w/ PAG. PAG aids the model in inpainting images, improving their realism and diminishing artifacts, without the necessity for a prompt (3rd row). The guidance scale of 1.5 is employed for all.

C.3 Text-to-3D

We integrated PAG with CFG for text-to-3D generation, utilizing the Dreamfusion [38] implementation provided by Threeestudio [13] due to the unavailability of official code. We employed a scale of 100 for both CFG and PAG. As seen in Fig. 35, combining CFG with PAG yields results with enhanced details and textures compared to using CFG alone.

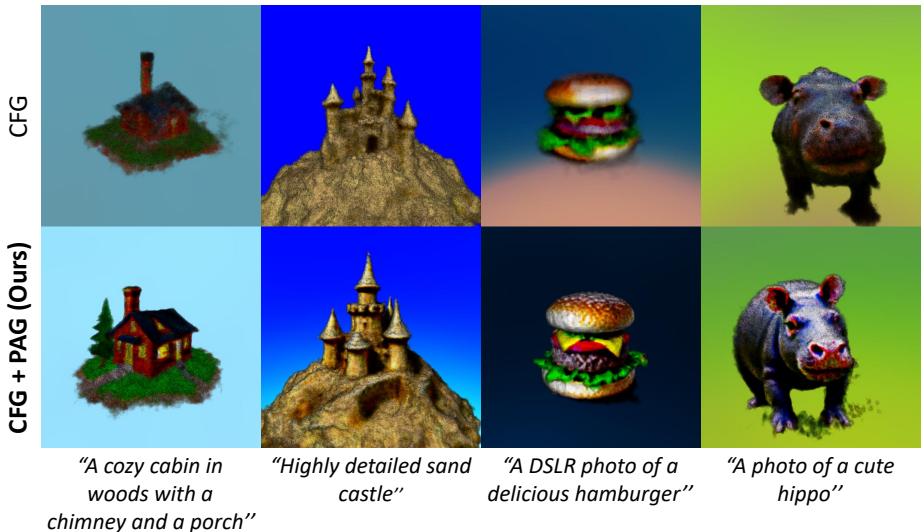


Fig. 35: Comparison of text-to-3D results between CFG [19], and CFG with PAG.

C.4 Human Evaluation

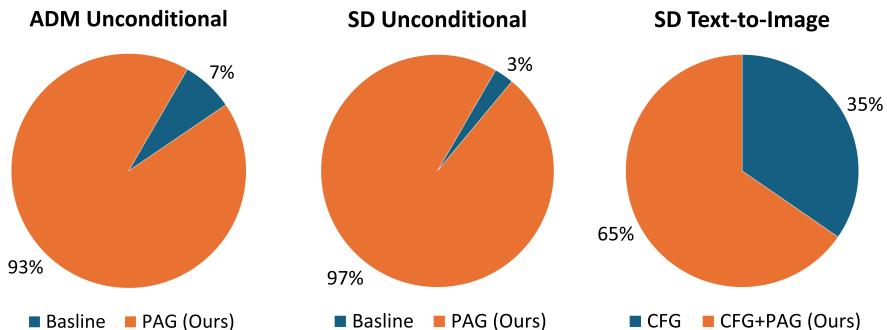


Fig. 36: The results of the user study.

A user study (Fig. 36) conducts to evaluate the quality of samples in ADM unconditional, Stable Diffusion unconditional, and Stable Diffusion text-to-image synthesis models. In the cases of unconditional generation, participants are presented with sets of four images sampled both with and without PAG and ask to identify the higher quality samples. For text-to-image synthesis, participants compare sets of four images generated using only CFG against those using both CFG and PAG. Each task comprises 10 questions, resulting in a total of 30 questions evaluated by 60 participants. The results show that the majority of unconditional generation questions prefer samples generated with PAG. Similarly, in the text-to-image synthesis task, samples generated with both CFG and PAG are frequently rated as higher quality.

Table 9: Ablation study on perturbations. We sampled 5K images from the ADM [10] ImageNet [8] 256×256 unconditional model. Perturbations are applied to the same layer (`input.13`) and the same guidance scale ($s = 1.0$) is used.

Perturbation strategy	FID \downarrow
Random Mask	40.20
Random Mask (off-diag)	39.49
Additive Noise	62.83
Gaussian Blur	35.48
Identity Matrix	32.34

D Ablation Studies

D.1 Perturbation on Self-Attention Maps

We explored various self-attention perturbation techniques that modify the structure part, $\text{Softmax}(Q_t K_t^T / \sqrt{d}) \in \mathbb{R}^{hw \times hw}$ in Eq. 12. These methods include replacing the attention map with an identity matrix, applying random masking, and selectively masking off-diagonal entries, as discussed in Sec. 5.5. We also tried additional perturbations, including applying Gaussian blur to the self-attention map and adding Gaussian noise to it. The quantitative results are detailed in Table. 9. The qualitative outcomes are depicted in Fig. 37, illustrating that substituting the self-attention map with an identity matrix enhances image realism by minimizing artifacts and making the objects’ structure semantically plausible. For additive noise, we use $\sigma = 0.1$ for Gaussian noise, and for Gaussian blur, we apply a blur kernel with a kernel size of 5 and a blur sigma of 1.0.

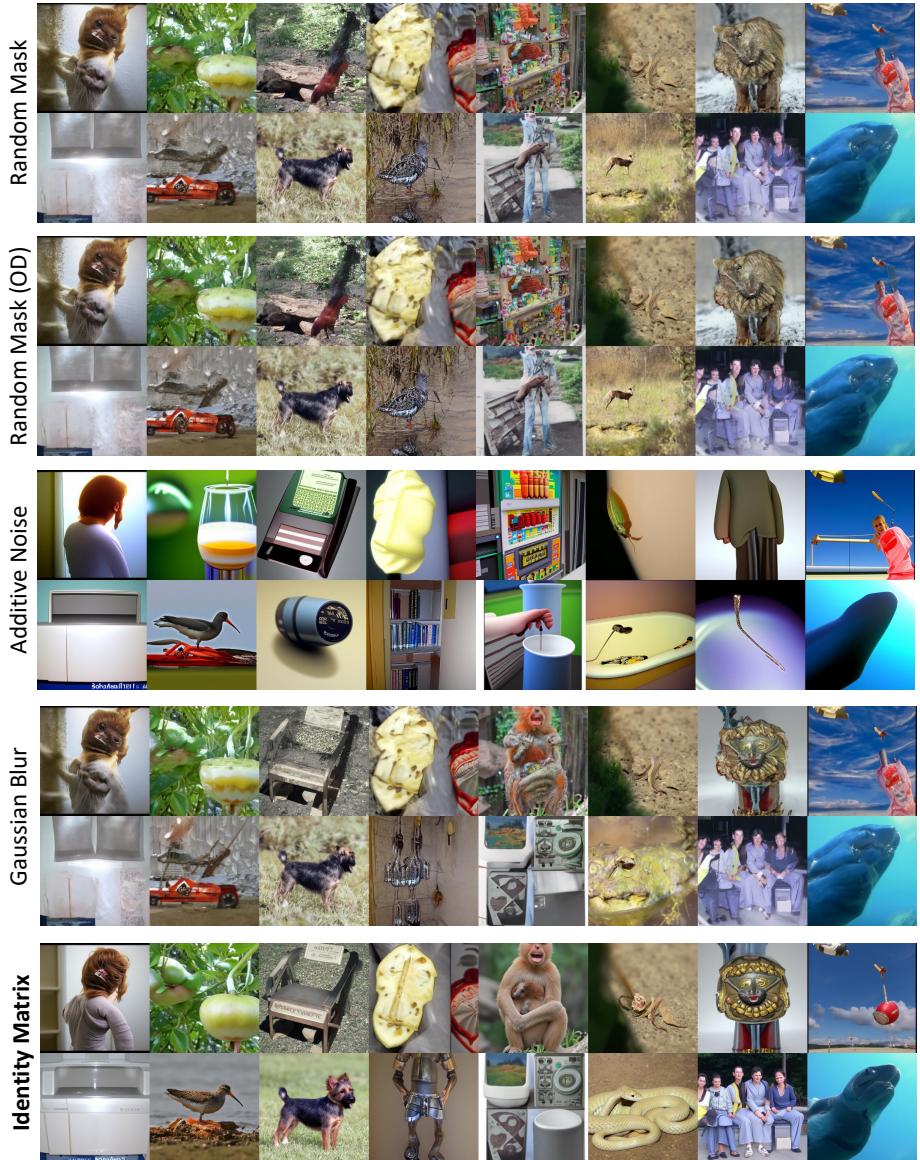


Fig. 37: Uncurated samples from ADM [10] with different perturbations on the self-attention map. Random Mask (OD) means masking on off-diagonal entries of the self-attention map. Note that all samples are not curated and use same layer to perturb (input.13) and same guidance scale ($s = 1.0$). The results clearly show that samples with identity matrix replacement generate plausible structures and semantics. In contrast, other perturbations often result in over-smoothed textures (additive noise) or introduce artifacts (other perturbations).

D.2 Layer Selection

We conduct an ablation study to determine the optimal layers for perturbing the self-attention map with the outcomes presented in Fig. 38. The experiments include both ADM [10] ImageNet 256×256 unconditional model and Stable Diffusion [41]. Observations indicate that perturbations applied to deeper layers generally yield relatively better outcomes compared to those applied to shallower layers of U-Net [42]. We apply perturbations to all combinations of the top-6 layers (`input.14`, `input.16`, `input.17`, `middle.1`, `output.2`), as ranked by FID, and present the results in Fig. 39 and Table 10. Some combinations show improved results for the ADM unconditional model but do not yield the same improvements in the case of Stable Diffusion [41]. Additionally, although experiments involving the random selection of layers for each timestep were conducted, we discover that selecting fixed layers across timesteps yields better outcomes.

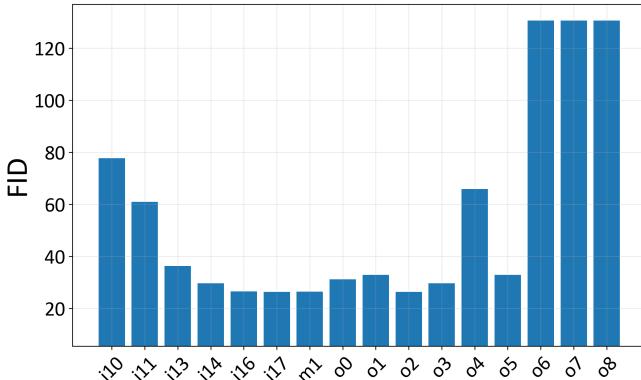


Fig. 38: Ablation study on which layer to apply perturbation with ADM [10].

Fig. 38 visualizes the FID scores obtained by perturbing each layer of the ADM [10] ImageNet [8] 256×256 unconditional models. A guidance scale of $s = 1.0$ is employed. FID scores are calculated using 5K image samples. Note that outlier values (`o6`, `o7`, `o8`) are clipped. It can be seen that perturbations on deeper layers, particularly near the bottleneck layer of U-Net, tend to show relatively better performance than those on shallower layers. The ablation results through DDIM [48] 25 step sampling are as follows, and in the case of sampling 5K images with DDPM [18] 250 step sampling, the layer we selected on Table. 1 shows the highest performance.

Fig. 40 shows the FID results from generating 5k samples using Stable Diffusion with PAG guidance scale $s = 2.5$ and DDIM 25 step sampling. We applied perturbation to different layers: “d0” represents the outermost encoder layer, “u8” is the outermost decoder layer, and “m0” is the mid-block. The best performance was achieved when perturbation was applied to the mid-block “m0”.

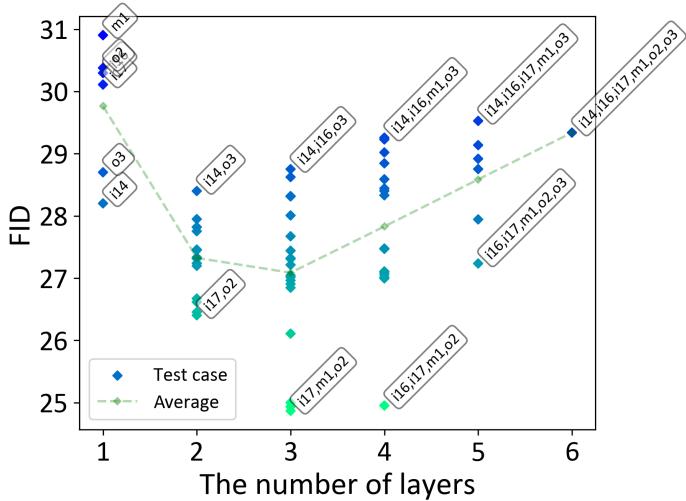


Fig. 39: Ablation study on layer combination for perturbed self-attention application in ADM [10]. Each data point represents the FID obtained when perturbed self-attention (PSA) is applied to the corresponding combination of layers. The annotations of points represent the combined layers. The green dashed line denotes the average FID across all combinations for a given number of layers involved. This analysis reveals that applying PSA to multiple layers can enhance sample quality to a certain extent. However, this trend does not hold for Stable Diffusion [41], indicating that the effectiveness of layer-wise perturbation varies across different diffusion models.

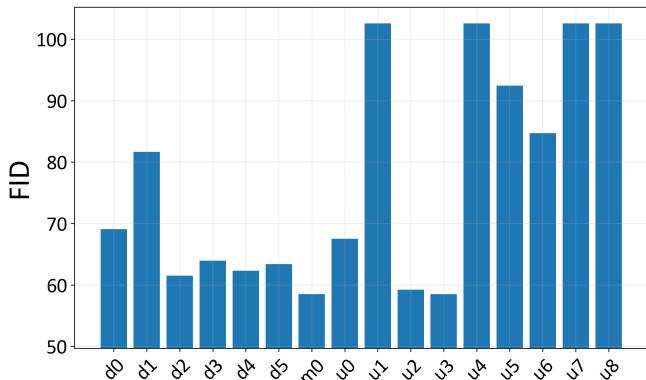


Fig. 40: Ablation study on which layer to apply perturbation with Stable Diffusion [41].

Table 10: Layer ablation on ADM. We evaluate the FID [17] of 5K samples from ImageNet [8] 256×256 unconditional model using DDIM [48] 25 step sampling.

# layers	layers	FID ↓
1	i14	28.20
	i16	30.30
	i17	30.11
	m1	30.90
	o2	30.38
	o3	28.70
2	i14 i16	27.95
	i14 i17	27.82
	i14 m1	27.82
	i14 o2	26.62
	i14 o3	28.40
	i16 i17	27.20
	i16 m1	27.31
	i16 o2	26.45
	i16 o3	27.45
	i17 m1	27.33
	i17 o2	26.40
	i17 o3	27.24
	m1 o2	26.67
	m1 o3	27.46
3	o2 o3	27.75
	i14 i16 i17	28.33
	i14 i16 m1	28.01
	i14 i16 o2	27.02
	i14 i16 o3	28.75
	i14 i17 m1	27.67
	i14 i17 o2	26.85
	i14 i17 o3	28.62
	i14 m1 o2	26.91
	i14 m1 o3	28.31
	i14 o2 o3	28.32
	i16 i17 m1	26.11
	i16 i17 o2	25.00
	i16 i17 o3	27.04
	i16 m1 o2	24.93
	i16 m1 o3	27.21
	i16 o2 o3	27.31
4	i17 m1 o2	24.87
	i17 m1 o3	26.96
	i17 o2 o3	27.44
	m1 o2 o3	27.32
	i14 i16 i17 m1	28.44
	i14 i16 i17 o2	27.47
	i14 i16 i17 o3	29.23
	i14 i16 m1 o2	27.48
	i14 i16 m1 o3	29.26
5	i14 i16 o2 o3	28.59
	i14 i17 m1 o2	27.09
	i14 i17 m1 o3	29.02
	i14 i17 o2 o3	28.84
	i14 m1 o2 o3	28.40
	i16 i17 m1 o2	24.95
	i16 i17 m1 o3	27.11
	i16 i17 o2 o3	27.01
	i16 m1 o2 o3	27.06
	i17 m1 o2 o3	27.00
6	i14 i16 i17 m1 o2 o2	27.94
	i14 i16 i17 m1 o3	29.53
	i14 i16 i17 o2 o3	29.14
	i14 i16 m1 o2 o3	28.92
	i14 i17 m1 o2 o3	28.75
	i16 i17 m1 o2 o3	27.24
6	i14 i16 i17 m1 o2 o3	29.34

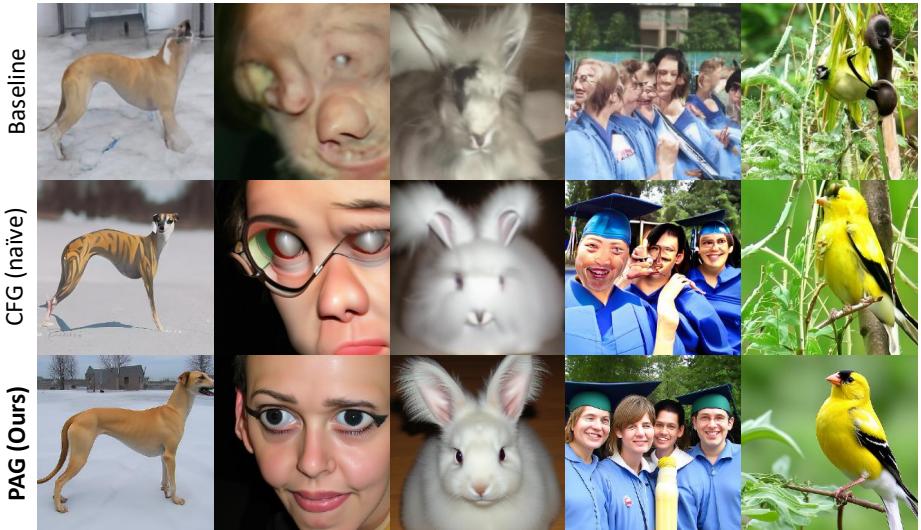


Fig. 41: Samples Using CFG with Separately Trained Models. We implement CFG [19] by employing separately trained ADM [10] ImageNet 256×256 conditional and unconditional models. Compared to samples with PAG in row 3, samples with naïve CFG in row 2 show inferior image quality. This suggests that when the conditional prediction $\epsilon_{\theta_1}(x_t, c)$ and $\epsilon_{\theta_2}(x_t)$ do not align, the guiding signal becomes ineffective, resulting in low-quality samples, where θ_1 and θ_2 are parameters from the conditional and unconditional models, respectively. Here, a guidance scale of 3.0 is employed for both naïve CFG and PAG, using the same seed and latent.

E Discussion

E.1 Further Analysis on CFG and PAG

CFG with separately trained models. As mentioned in the Sec. 3 in the main paper, the guidance term in CFG [19] originates from the gradient of the *implicit classifier* derived from Bayes’ rule. Therefore, in principle, CFG can be implemented by training the conditional and unconditional models separately. However, the authors implemented it using a single neural network by assigning a null token as the class label for the unconditional model. They mentioned, “It would certainly be possible to train separate models instead of jointly training them together,” suggesting it as an option during design. But in practice, this is not the case. We discover that as can be seen in Fig. 41, implementing CFG with separately trained conditional and unconditional models does not work properly (2nd row). This implies that CFG enhances image quality not merely by trading diversity but operates by some other key factor. The secret may be that as analyzed in Fig. 2 of main paper, predicting a sample missing salient features (such as eyes and nose) from the original conditional prediction and then adding the difference to reinforce those salient features. In other words, simply subtracting the unconditional generation made by a separate model does

not suffice for its operation, highlighting the utility of our **PAG**. While CFG creates a *perturbed* path missing salient features at the additional cost of training an unconditional model jointly, perturbed self-attention (PSA) in our **PAG** can produce predictions missing such salient features without any additional training or external model, simply by manipulating the self-attention map of U-Net. Especially when compared to SAG [20] and other perturbations (perturbation ablations and Sec. D.1), PSA can be considered an efficient and effective method.

Connections to delta denoising score. According to prior works [15, 24] that use diffusion models for score distillation sampling (SDS), the term $\hat{\epsilon}_\theta$ in our guidance framework can be interpreted as the model’s inherent *bias*. Delta denoising score [15] suggests that when conducting SDS, the gradient term contains *bias*, and by subtracting this from another gradient obtained with similar prompts, structures, and the same noise, one can eliminate the shared noisy components. From this perspective, CFG [19] and PAG can be interpreted as the removal of noisy components, which make locally aligned structures, in the diffusion model’s epsilon prediction as class label dropping and attention perturbing, respectively. This perspective underscores the importance of carefully calibrating perturbations to avoid significant deviations from the original sample. SAG [20] has shown a tendency to produce samples that diverge excessively from the original sample, due to aggressive perturbation applied directly to the model’s input, leading to out-of-distribution (OOD) samples and high hyperparameter sensitivity. Scale-wise qualitative result illustrates that PAG exhibits lower sensitivity to scale adjustments, attributed to the strategic perturbed self-attention approach, which preserves appearance information of the original sample. For a comprehensive comparison, see Sec. E.3.

Additional training for stability. Although our carefully designed perturbed self-attention (PSA, e.g., self-attention map replacement with identity) method effectively mitigates the out-of-distribution (OOD) issue without additional training, incorporating training can further improve its ability to address the OOD problem and enhance its robustness to hyperparameter settings.

Similar to various self-supervised learning or augmentation techniques [14, 51, 55] that achieve comparable results with augmented inputs/models to those with original inputs/models, $\hat{\epsilon}_\theta$ can be trained with PSA to produce more stable samples that maintain appearance and lack structural information. This improvement can be achieved by introducing a switching input to control the on/off status of self-attention map usage and fine-tuning the model while providing this switching input. Compared to training a new unconditional model for CFG [19], fine-tuning the model incurs lower computational costs. Furthermore, unlike CFG, which entangles sample quality and diversity, PAG with trained $\hat{\epsilon}_\theta$ enhances sample quality without compromising diversity. We leave this exploration as future work.

E.2 Complementarity between CFG and PAG

Recent research [2, 34] has shed light on the *temporal dynamics* of text-to-image diffusion models during their sampling process. The analysis, focusing on the model’s self-attention and cross-attention maps under different noise conditions, demonstrates a transition in the model’s operational focus from text to the pixels being generated. Initially, at the beginning of the sampling process where the network’s input is close to random noise, the model significantly relies on the text prompt for direction in the sampling. However, as the process continues, there’s a noticeable shift towards leveraging visual features for image denoising, showing higher activation of self-attention map, with the model gradually paying less attention to the text prompt. This shift is logical; in the early stage, the model relies on the prompt for cues on what to denoise in the image. As the denoising process progresses and the images take shape, the model shifts focus to refine these emerging visual details.



Fig. 42: Visualization of $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$ during reverse process with PAG. text-to-image generation using PAG with a prompt “*a fancy sports car*”.

This phenomenon can also be observed in PAG and CFG contexts. Fig. 42 visualizes $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$ during sampling with PAG. As mentioned earlier, since the self-attention map is not highly activated in the early stages of the diffusion sampling process, the difference $\hat{\Delta}_t$ between the predicted epsilon with self-attention map dropped and the original predicted epsilon appears weak initially. As the sampling process progresses and the image starts to take shape, the activation of the self-attention map gradually strengthens, leading to an increasing $\hat{\Delta}_t$ observable over time.



Fig. 43: Visualization of $\Delta_t = \epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$ during reverse process with CFG. text-to-image generation using CFG with a prompt “*a fancy sports car*”.

In contrast, CFG exhibits a different behavior. Fig. 43 displays the timestep-wise predicted epsilon difference $\Delta_t = \epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$ during sampling with CFG. As previously discussed, in the initial stages of generation, the diffusion model predominantly relies on the prompt to create images, leading to high activation in the cross-attention map. CFG creates a perturbed path using a null prompt for the prompt, which can be understood as applying perturbation to the cross-attention. (Indeed, we observe that making cross-attention map zero yield effects somewhat similar to CFG, though these effects were suboptimal.) Therefore, a high Δ_t is observed in the early stages of the sampling process, where the model focuses on the prompt, and the difference diminishes later on.

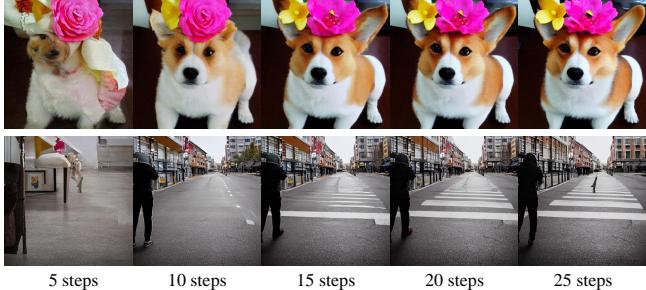


Fig. 44: Early stopping of CFG [19]. The process involves a total of 25 steps. The prompts used are “*A corgi with a flower crown*” (top) and “*A person walking on the street*” (bottom).

The impact of CFG being primarily in the early stages of the generation process can be validated through another observation. Fig. 44 shows the results when CFG is applied only in the first 5 steps, then in the first 10, 15, 20, and throughout all 25 steps of a 25 step generation process. It can be observed that applying CFG for the initial 60% of the total steps (15 steps) yields results comparable to those achieved when CFG is utilized for the full 25 steps.

Compared with CFG, PAG continues to influence throughout the mid to late stages of the timestep, offering highly detailed guidance, especially in the latter half as can be seen in Fig. 42. This indicates that PAG continues to provide a positive signal even in the later stages.

Therefore, using CFG and PAG together can guide the image towards better quality across the entire sampling process. This approach effectively utilizes both the self-attention and cross-attention maps, resulting in effective guidance throughout the entire timestep. Indeed, qualitative comparison between CFG and CFG + PAG, and Stable Diffusion quantitative results demonstrate that combining CFG and PAG yields superior outcomes compared to employing CFG alone. We also present a human evaluation of samples utilizing CFG versus CFG + PAG in Fig. 36.

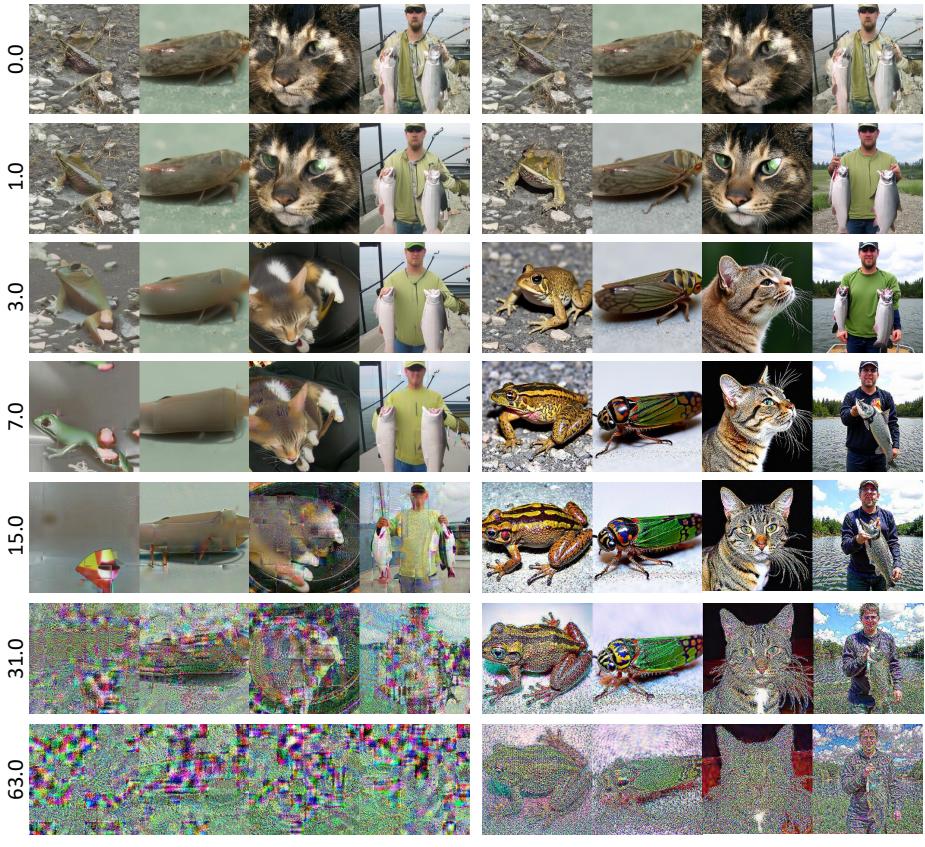
E.3 Comparison with SAG

In this section, we summarize the differences between SAG [20] and PAG, focusing on their formulation, stability, speed, and effectiveness. SAG emerged as an initial method for enhancing guidance in unconditional generation within diffusion models.

Generalizability. Both SAG and PAG aim to generalize guidance, albeit through distinct formulations. SAG proposes an imaginary regressor p_{im} to predict h_t given x_t , where h_t represents a generalized condition including external condition or internal information of x_t or both, and \bar{x}_t is a perturbed sample missing h_t from x_t . For instance, blur guidance in SAG leverages $\bar{x}_t = \tilde{x}_t$ and $h_t = x_t - \tilde{x}_t$, where \tilde{x}_t represents a sample with the high-frequency components of the original sample x_t removed. Specifically, \hat{x}_0 is derived from x_t by Eq. 4 and subsequently blurred using a Gaussian filter G_σ (expressed as $\tilde{x}_0 = \hat{x}_0 * G_\sigma$, with $*$ denoting a convolution operation), and then diffused back by incorporating the noise $\epsilon_\theta(x_t)$. The guided sampling can be formulated as $\tilde{\epsilon}(\bar{x}_t, h_t) = \epsilon_\theta(\bar{x}_t, h_t) - s\sigma_t \nabla_{\bar{x}_t} \log p_{\text{im}}(h_t | \bar{x}_t)$. However, this approach does not allow for perturbations on the model’s internal representation, whereas SAG can be considered a specific instance within our broader framework, as $\hat{\epsilon}_\theta(\cdot)$ could represent any perturbation process, including the adversarial blurring used by SAG.

Hyperparameter count and sensitivity to guidance scale. Since SAG utilizes Gaussian blur, it requires the setting of multiple hyperparameters. Hyperparameters related to blur include the blur kernel size and the σ of the blur kernel. Additionally, determining the area for adversarial blurring necessitates selecting the layer from which to extract the self-attention map and specifying a threshold value. In contrast, PAG simply requires the selection of the layer to which perturbed self-attention will be applied. Additionally, as seen in Fig. 45, SAG is sensitive to the guidance scale. The figure shows that as the guidance scale increases, the boundaries of the adversarial mask area become visible, and high-frequency artifacts appear. Therefore, SAG cannot use a large guidance scale, which is a significant drawback considering that stronger guidance often results in greater improvements in image quality. Indeed, considering CFG employs a large scale of around 7.5, this limitation is significantly notable. In contrast, PAG maintains the plausibility of object shapes and enhances details even at relatively high guidance scales.

Inference speed. SAG requires the extraction of self-attention maps during its first forward pass and the application of blur to the model input for adversarial blurring. Our method can be implemented to handle PSA and regular self-attention within the same batch, allowing guidance to be applied with a single evaluation of the denoising neural network, similar to CFG [19]. Therefore, if the GPU can perform concurrent computations swiftly, PAG could theoretically



(a) SAG

(b) PAG

Fig. 45: Comparison of samples with SAG [20] and PAG for different guidance scales. Samples are generated by ADM [10] conditional ImageNet 256×256 model, showcasing the impact of incrementally increasing the guidance scale from 0.0 to 63.0, from the top to the bottom of the figure. (a): Samples generated with a high guidance scale using SAG exhibit artifacts and over-smoothness due to excessive perturbation, specifically blurring on the input, with the outlines of the blur mask clearly visible. (b): Compared to SAG, samples generated with higher scale PAG display high-quality results, characterized by well-structured shape and high detail. Within each group, from left to right, the classes are *bell toad*, *leafhopper*, *tabby cat*, and *silver salmon*.

be up to more than twice as fast as SAG. We discuss the results of comparing the speed of PAG, implemented in this manner, with CFG and SAG in Sec. A.6.

F Limitation and Future Works

Although PAG demonstrates effectiveness across various tasks, it shares certain limitations with CFG. Notably, at high guidance scales, results can exhibit over-saturation. This highlights the need for careful calibration of the guidance scale to balance quality improvement with potential visual artifacts. Additionally, PAG requires two forward paths for each generation step. Future research could explore techniques to reduce this computational overhead or develop alternative guidance mechanisms with lower resource requirements.