Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance

Donghoon Ahn*[1], Hyoungwon Cho*[1], Jaewon Min[1], Wooseok Jang[1], Jungwoo Kim[1], SeonHwa Kim[1], Hyun Hee Park[2], Kyong Hwan Jin†[1], and Seungryong Kim†[1]

[1] Korea University
[2] Samsung Electronics

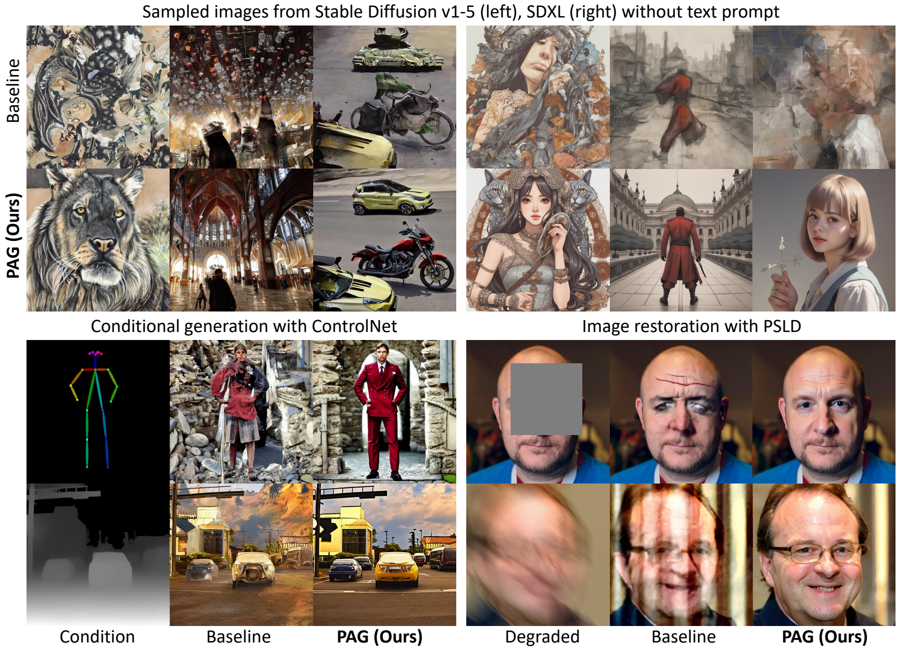

Sampled images from Stable Diffusion v1-5 (left), SDXL (right) without text prompt
Baseline
PAG (Ours)
Conditional generation with ControlNet
Image restoration with PSLD
Condition
Baseline
PAG (Ours)
Degraded
Baseline
PAG (Ours)

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) [14] is inapplicable. Our guidance can also enhance the baseline performance in various downstream tasks such as ControlNet [41] with empty prompt and inverse problems such as inpainting and deblurring [6, 30].

Abstract. Recent studies have demonstrated that diffusion models can generate high-quality samples, but their quality heavily depends on sampling guidance techniques, such as classifier guidance (CG) and classifier-free guidance (CFG). These techniques are often not applicable in unconditional generation or various downstream tasks such as the solving

$*$: Equal contribution

†: Co-corresponding author

inverse problems. In this paper, we propose 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 progressively enhances the structure of samples throughout the denoising process by generating intermediate samples with degraded structures and guiding the denoising process away from these degraded samples. These degraded samples are created by substituting selected self-attention maps in the diffusion U-Net, which capture structural information between image patches, with an identity matrix. In both ADM and Stable Diffusion, PAG surprisingly improves sample quality in conditional and even unconditional generation. Moreover, PAG significantly enhances baseline performance in various downstream tasks where existing guidance methods such as CG or CFG cannot be fully utilized, including ControlNet with empty prompts and solving inverse problems such as inpainting and deblurring. To the best of our knowledge, this is the first approach to apply guidance in solving inverse problems using diffusion models.

1 Introduction

Diffusion models [13, 27, 33, 35, 36] have gained prominence in image generation, demonstrating their capability to produce high-fidelity and diverse samples. Sampling guidance techniques, such as classifier guidance (CG) [9] and classifier-free guidance (CFG) [14], are crucial for directing diffusion models to generate higher-quality images. Without these techniques, as shown in Fig. 1 and Fig. 2, diffusion models often produce lower-quality images, typically exhibiting collapsed structures. Despite their widespread use, these guidance methods have drawbacks: they require additional training or the integration of external modules, often reduce 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 [5,20]. Furthermore, advancements in unconditional techniques often enhance conditional generation. Importantly, it eliminates the need for potentially costly and complex human annotations such as class labels, text, and segmentation maps, which can be a major hurdle in tasks where accurate labeling is difficult, such as modeling molecular structures [20]. Finally, unconditional generative models provide powerful general priors, as evidenced by their use in solving inverse problems [6,7,17,29,30,36,39]. However, the unavailability of CG [9] or CFG [14] can lead to sub-optimal performance.

Recognizing the importance of unconditional generation, we propose a novel sampling guidance method called **Perturbed-Attention Guidance (PAG)**. PAG improves diffusion sample quality in 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 and undesirable samples. By utilizing the capability of self-attention maps in the diffusion U-Net to capture structural information [2,11,23,37,38], 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 help steer the denoising trajectory away from the structural collapse commonly observed in unguided generation.

Extensive experiments validate the effectiveness of our guidance method. Applied to ADM [9], it exceptionally improves sample quality in both conditional and unconditional settings. We also observe remarkable enhancements, both qualitatively and quantitatively, when applied to the widely-used Stable Diffusion [27]. Additionally, combining PAG with conventional guidance methods such as CFG [14] leads to further improvements. Finally, our guidance profoundly enhances the performance of diffusion models in various downstream tasks, such as inverse problems [6,30] and ControlNet [41] with empty prompts, where the lack of conditions renders CFG [14] unusable. Notably, we have opened new avenues for fully leveraging the generative capabilities of diffusion models in solving inverse problems.

2 Related Work

Diffusion models. Diffusion models (DMs) [33,35,36] have set a high benchmark in image generation, achieving remarkable results in both sample quality and distribution estimation. DDIM [34] improves sampling speed by applying a non-Markovian process. Latent diffusion models (LDMs) [27] operate in a compressed latent space, balancing 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 [9, 14]. Classifier guidance (CG) [9] increases fidelity at the expense of diversity by adding the gradient of a pre-trained classifier. Classifier-free guidance (CFG) [14] models an implicit classifier to achieve similar effects as CG. Self-attention guidance (SAG) [15] enhances sample quality in an unconditional framework by using adversarial blurring to obscure crucial information and then guiding the sampling process with noise predicted from both blurred and original samples. Additionally, various guidance methods focus on conditioning [22] or image editing [3,10].

3 Preliminaries

Diffusion models. In diffusion models [9,13,14,36], 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:

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

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ according to a variance schedule $\beta_1, ..., \beta_t$. A denoising network ϵ_θ is learned to predict ϵ by optimizing an objective

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

for uniformly sampled $t \in \{1, ..., T\}$.

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:

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

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 $\epsilon_\theta(x_t) \approx -\sigma_t \nabla_{x_t} \log p(x_t)$ [9, 14, 35, 36], 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

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

Classifier-free guidance. To enhance the generation towards arbitrary class label c, CG [9] 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

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

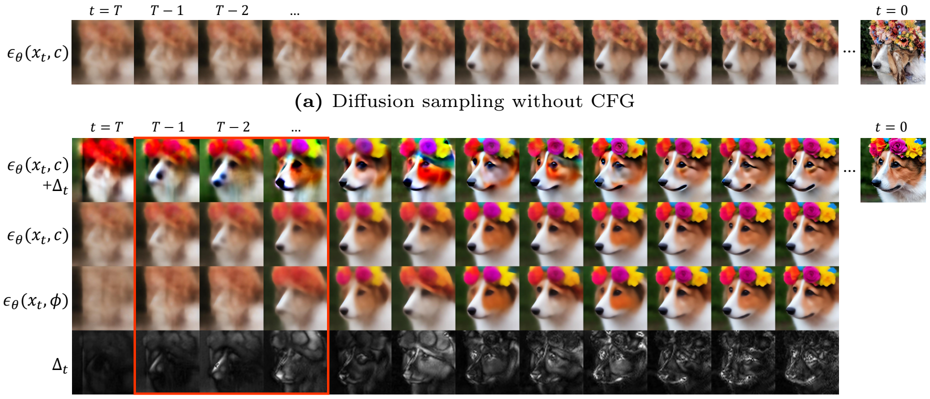
where s is the scale parameter. It turns out that sampling from this distribution with $s > 0$ leads the model to generate saturated samples with high probabilities for the input class labels, resulting in increased quality but decreased sample diversity [9].

CG, however, has a drawback in that it requires a pretrained classifier for noisy images of each timestep. To address this issue, CFG [14] modifies the classifier distribution $p_\theta(c|x_t)$ by combining the conditional distribution $p_\theta(x_t|c)$ and the unconditional distribution $p_\theta(x_t)$:

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

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

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


$t = T$
$T - 1$
$T - 2$
...
$t = 0$
$\epsilon_\theta(x_t, c)$
$\epsilon_\theta(x_t, c)$
$+\Delta_t$
$\epsilon_\theta(x_t, c)$
$\epsilon_\theta(x_t, \phi)$
Δ_t

(a) Diffusion sampling without CFG

(b) Diffusion sampling with CFG

Fig. 2: Visualization of reverse process w/o and w/CFG [14]. To visualize the predicted epsilon, we first convert it into $\hat{x}_0$ following Eq. 4. 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. We use the same latent and seed for both cases. **(a)** Without CFG, diffusion models generate samples with collapsed structures. **(b)** With CFG, 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 prompt, diffusion models lack 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 relevant to the prompt. Here the prompt *"a corgi with flower crown"* is used.

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 explain its role in enhancing sample quality. A more detailed exploration of CFG's workings is available in the Appendix E.2.

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 the energy function, for instance, which can be a negative class probability of classifier [9], negative CLIP similarity score [24], any type of time-independent energy [3], the distance between extracted signal such as pose and edges and reference signal [22] or any energy function which takes the noisy sample [10].

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 [14] 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

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

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

Then similar to WGAN [1, 40], we set the generator loss of the implicit discriminator as our energy function, $\mathcal{L}_\mathcal{G}$, and compute its derivative as

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

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

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

Since diffusion models already learned the desired distribution, we use the pretrained 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 [15] or internal representations, or both.

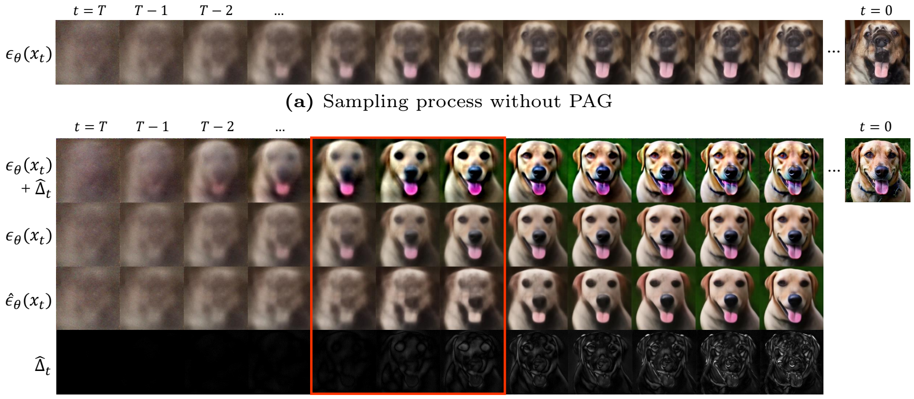
Connections to CFG. The formulation in Eq. 10 resembles CFG [14]. Indeed, it is noteworthy that CFG can be considered a particular instance within our broader formulation. First, Eq. 10 can also be defined in class-conditional diffusion models such that

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

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



(a) Sampling process without PAG

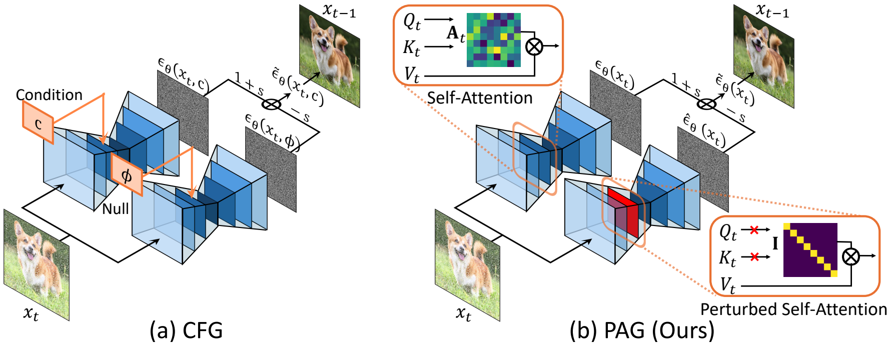
(b) Sampling process with PAG

Fig. 3: Visualization of sampling process w/o and w/ PAG. To visualize predicted epsilon, we first convert it into $\hat{x}_0$ following Eq. 4. 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. We use the same latent and seed for both cases. **(a)** Without guidance, diffusion models generate samples with collapsed structures. **(b)** With our PAG, 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. Adding $\hat{\Delta}_t$ thus enhances features that can only be accurately rendered with global structure information.

can cause the out-of-distribution problem, lead the diffusion model to create incorrect guidance signals, and steer the diffusion sampling toward the erroneous direction. To overcome this, CFG [14] explicitly trains an unconditional model. In addition, SAG [15] employs partial blurring to minimize deviation, but without careful selection of hyperparameters, it often deviates from the desired trajectory. This behavior is illustrated in Appendix E.4.

On the other hand, some studies have explored manipulating cross-attention and self-attention maps of the diffusion models for various tasks [4,11,18,26,32]. They show that modifying the attention maps has minimal impact on the model's ability to generate plausible outputs. We target the self-attention mechanism to design a perturbation strategy applicable to 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 how the null prompt in CFG is employed to strengthen class conditioning. Recently, several studies [2,11,23,37,38] demonstrate that the attention map contains structural information or semantic


Condition
c
ϕ
Null
x_t
x_{t-1}
$\epsilon_\theta(x_t, c)$
$1+s$
$-s$
$\tilde{\epsilon}_\theta(x_t, c)$
$\epsilon_\theta(x_t, \phi)$
Q_t
K_t
V_t
$\mathbf{A}_t$
Self-Attention
$\epsilon_\theta(x_t)$
$\hat{\epsilon}_\theta(x_t)$
$\tilde{\epsilon}_\theta(x_t)$
$\mathbf{I}$
Perturbed Self-Attention

(a) CFG

(b) PAG (Ours)

Fig. 4: Conceptual comparison between CFG [14] and PAG. CFG [14] 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 $\mathrm{Softmax}(Q_t K_t^T/\sqrt{d})$. In PAG, we perturb this by replacing with an identity matrix $\mathbf{I}$.

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, 11, 37, 38] have shown that the self-attention module in diffusion U-Net [28] 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:

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

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

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

where we call perturbed self-attention (PSA). More ablation studies on perturbing a self-attention map can be found in the Appendix D.2.

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

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

4.3 Analysis on PAG

In this section, we explore why our guidance method is effective. Fig. 3 shows the sampling process using PAG, with each row (except the last) depicting $\hat{x}_0$ at each timestep using 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 last row in (b) shows the guidance signal $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$. This figure highlights how our guidance term provides semantic cues. The red rectangle in Fig. 3 shows that the perturbed prediction (row 3 in (b)) misses key features like eyes, nose, and tongue due to a lack of global structure understanding. The difference $\hat{\Delta}_t$ focuses on these missing features (row 4 in (b)). Adding $\hat{\Delta}_t$ to the original prediction $\epsilon_\theta(x_t)$ strengthens the sample's structure, as shown in the first row of (b) in Fig. 3.

More analysis is in Appendix E.2 and E.3. We also visualize CFG [14] in Stable Diffusion in Fig. 2, showing how CFG uses an undesirable sampling path in the unconditional generation to enhance class conditioning. We also discuss the theoretical explanation for why replacing the attention map with an identity matrix is highly effective in Appendix E.1, drawing on the recent connection between the transformer's self-attention and Hopfield networks.

5 Experiments

5.1 Experimental and Implementation Details

Our work utilizes pretrained models, including ADM [9], Stable Diffusion 1.5 [27], and SDXL [25]. We accessed all necessary weights from their publicly available

Table 1: Quantitative results on ADM [9]. 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**

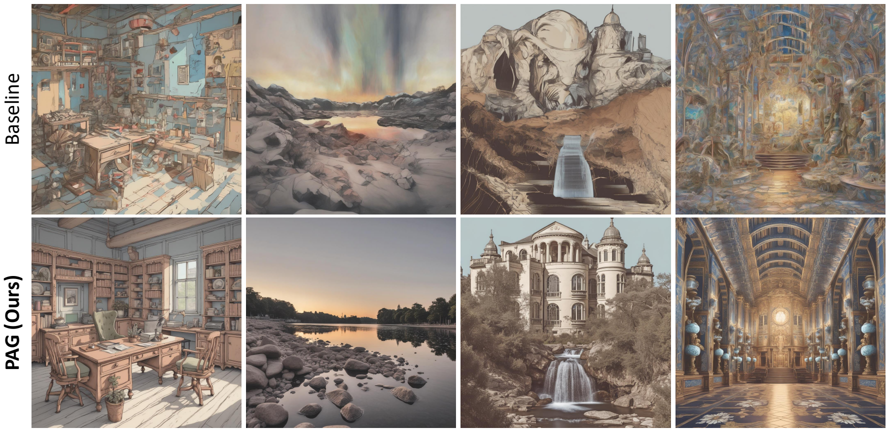


Fig. 5: Unconditional generation samples w/o and w/ PAG. Figures display sampled images from Stable Diffusion [27]. Each set of images shows sampling without (**Top**) and with (**Bottom**) PAG. Samples guided by PAG appear high perceptual quality and demonstrate semantically coherent structures.

repositories and used the same evaluation metrics as in ADM [9]. For additional experimental details, please refer to Appendix A.

5.2 Pixel-Level Diffusion Models

With pretrained ADM [9], we generates 50K samples on ImageNet [8] 256×256 to evaluate metrics. In Table 1, we compare ADM [9] with SAG [15] and PAG in both conditional and unconditional generation. Table 1 shows that ADM [9] with PAG outperforms the others with large margin in FID [12], IS [31]. The contrastive patterns of Improved Recall and Precision [19] in unconditional and conditional generation in Table 1 are attributed to the trade-off between fidelity and diversity [9,14,15]. 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 [15] is also presented in Fig. 6. For further exploration, additional samples from ADM [9] are available in Appendix B.1.

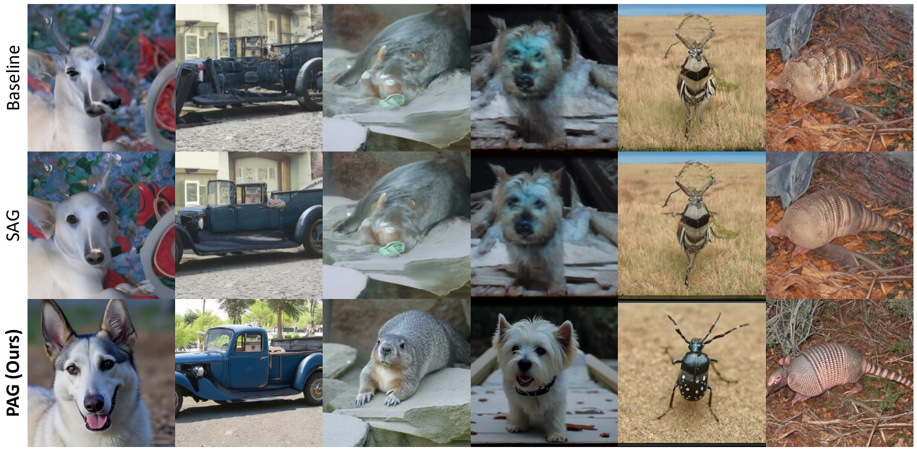


Fig. 6: Qualitative comparison between SAG [15] and PAG. Images are sampled from the ImageNet 256×256 unconditional model using the same seed sequence. Compared to samples guided by SAG, those guided by PAG exhibit significantly improved semantic structures with artifacts removed.

Table 2: Quantitative results on Stable Diffusion [27]. 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 [21].

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

5.3 Latent Diffusion Models

Unconditional generation on Stable Diffusion. We further explored the application of our guidance to Stable Diffusion [27]. 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 [12] and IS [31]. Samples from Stable Diffusion's unconditional generation 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.

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.

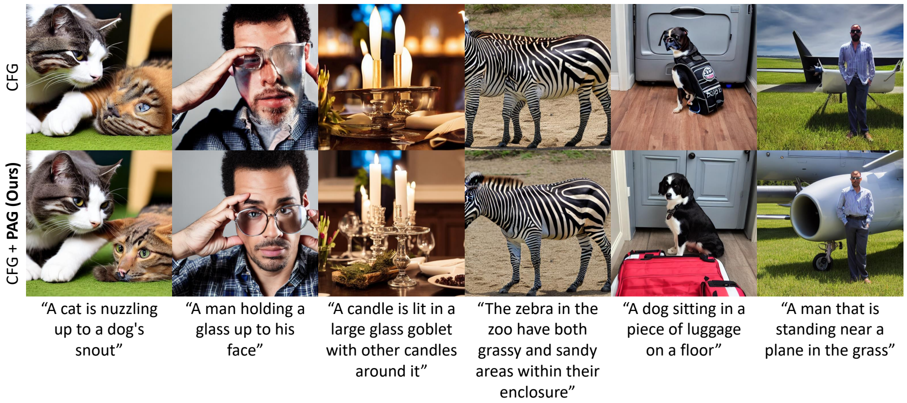

CFG
CFG + PAG (Ours)
"A cat is nuzzling up to a dog's snout"
"A man holding a glass up to his face"
"A candle is lit in a large glass goblet with other candles around it"
"The zebra in the zoo have both grassy and sandy areas within their enclosure"
"A dog sitting in a piece of luggage on a floor"
"A man that is standing near a plane in the grass"

Fig. 7: Qualitative comparison between CFG [14] 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.

In this case, since CFG [14] can be utilized, we conducted sampling in four different scenarios: without applying guidance as a baseline, using CFG, using PAG, and combining both guidance methods with an appropriate scale.

Interestingly, combining PAG and CFG [14] 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.3.

Table 3: Diversity comparison in samples generated by CFG [14] and PAG.

	IS ↑	LPIPS ↑
CFG	1.82	0.64
PAG	**2.32**	**0.68**

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) [31] is calculated over 1000 generated samples, and the LPIPS [42] 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

Table 4: Quantitative results of PSLD [30] on FFHQ [16] 256×256 1K validation set.

	Box Inpainting		SR (8×)		Gaussian Deblur		Motion Deblur	
Method	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**

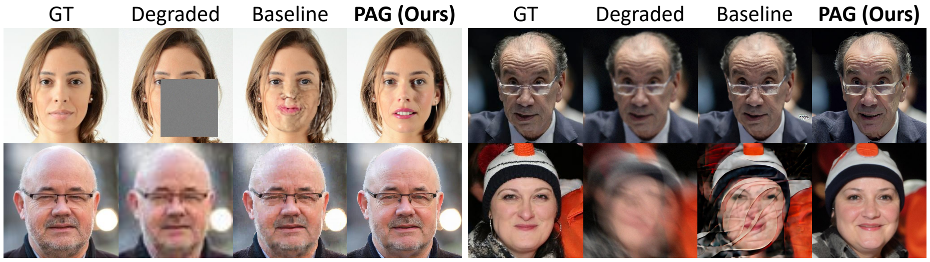


Fig. 8: Qualitative results of PSLD [30] with our PAG on FFHQ [16] dataset. Left Top: Box inpainting. **Left Bottom:** Super-resolution (×8). **Right Top:** Gaussian deblur. **Right Bottom:** Motion deblur. Using PAG leads to the removal of artifacts and blurriness, resulting in more realistic restorations.

2014 validation set [21]. Further samples from Stable Diffusion are available in Appendix B.2 for additional reference.

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, PAG 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 [16] 256×256 on PSLD [30] which leverages DPS [6] and LDM [27] 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 [12], and mostly in LPIPS [42]. 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. Importantly, PAG can be adopted to any other restoration model based on diffusion models, shown in Appendix C.

ControlNet. ControlNet [41], 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,

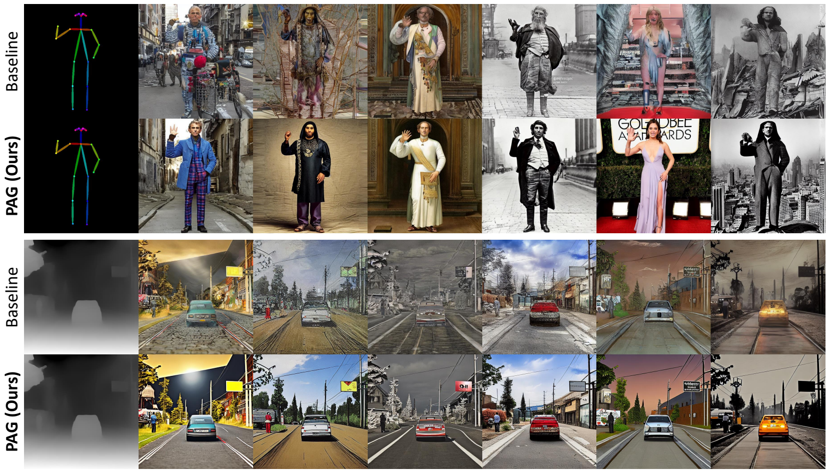


Fig. 9: ControlNet [41] sample images conditioned by pose and depth without text prompt. Samples guided by PAG appear more realistic, exhibiting fewer artifacts and semantically coherent structure.

as demonstrated in Fig. 9, PAG enhances sample quality in these instances. This 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 allowing artists to test diverse, imaginative works without relying on detailed prompts.

5.5 Ablation Studies

We provide ablations studies on self-attention perturbation strategy and effects of guidance scales on qualitative and quantitative results on Appendix D.

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 [14] 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.

Acknowledgements

This research was supported by the MSIT, Korea (IITP-2024-2020-0-01819, RS-2023-00227592), Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, RS-2024-00333068) and National Research Foundation of Korea (RS-2024-00346597). Thank you to Susung Hong for providing feedback on our research and manuscript.

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, 7, 8
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, 5
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) 7
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, 13
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) 10
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 2, 3, 4, 5, 9, 10
10. 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, 5
11. 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, 7, 8
12. 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, 13
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 3

14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14
15. 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, 10, 11
16. 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) 13
17. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022) 2
18. 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) 7
19. Kynkäänniemi, 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
20. Li, T., Katabi, D., He, K.: Self-conditioned image generation via generating representations. arXiv:2312.03701 (2023) 2
21. 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) 11, 13
22. 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, 5
23. 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, 7
24. 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) 5
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 9
26. 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) 7
27. 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, 9, 10, 11, 13
28. 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) 8
29. 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
30. 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, 13

31. 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, 12
32. 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) 7
33. 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
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 3
35. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 2, 3, 4
36. 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
37. 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, 7, 8
38. 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, 7, 8
39. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022) 2
40. 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
41. 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, 13, 14
42. 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) 12, 13