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 [37] super-resolution/inpaint pipeline, and text-to-3D [34] (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 [9] ImageNet [8] 256×256 conditional and unconditional models, we utilized the official GitHub repository[3] of ADM along with its publicly available pretrained weights. Our work builds upon the SAG [18] repository[4], 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 [18], 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 [15] score worsens. This may be due to the misalignment between FID and human perception [20]. 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 [9].

A.2 Experiments on Stable Diffusion

Quantitative results. For all the quantitative experiments, we utilized Stable Diffusion v1-5[5] implemented based on the pipeline provided by the Diffusers [32]. 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 [17] 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 [44] 50 step sampling method.

Qualitative results. Stable Diffusion v1-5 and SDXL[6] are 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 [44] 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 [17] 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 [17] 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
[6] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

A.3 Experiments with PSLD

We use Stable Diffusion v1.5 used in PSLD [40]. The measurement operators for inverse problems are from DPS [6], as used in PSLD [40]. PSLD [40] 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 [40] without PAG as same as PSLD [40]. Practically, we find that it is better to use unconditional score $\epsilon_\theta(\boldsymbol{z}_t)$ instead of guided score $\tilde{\epsilon}_\theta(\boldsymbol{z}_t)$ when predicting $\hat{\boldsymbol{z}}_0$ to update gradients. Furthermore, we conduct more experiments with ImageNet [8] dataset, which are provided in Sec. B.3. All experiments with PSLD [40] use DDIM [44] 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 [21] 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	0.3
γ	0.015	0.07	0.01	0.015	0.05	0.07	0.01	0.03
s	4.0	4.0	5.0	4.0	4.0	4.0	5.0	5.0

Table 5: Hyperparameters for PSLD [40] with PAG on FFHQ [21] dataset and ImageNet [8] dataset. Here, η and γ are the step size for gradients of PSLD [40] and s is the scale for PAG from Eq. 10 of main paper.

A.4 Experiments with ControlNet

For the ControlNet [53] 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 [9] 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. 34),

Table 6: Comparison of computational costs in Stable Diffusion.

	GPU Memory ↓	Sampling Speed ↑
No Guidance	3,147 MB	19.16 iter/s
CFG [17]	**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

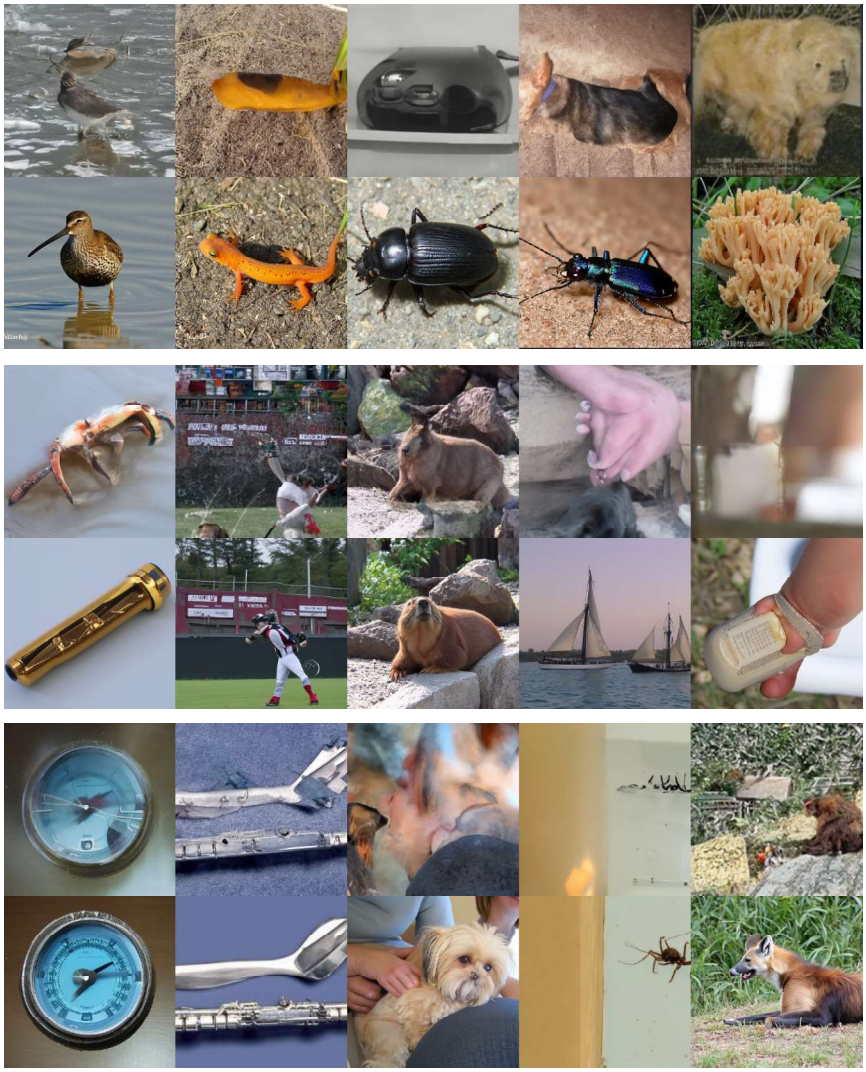

Fig. 10: Uncurated samples from ADM [9] ImageNet 256 *unconditional* model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: `i13`,`i14`,`i16`,`m1`.

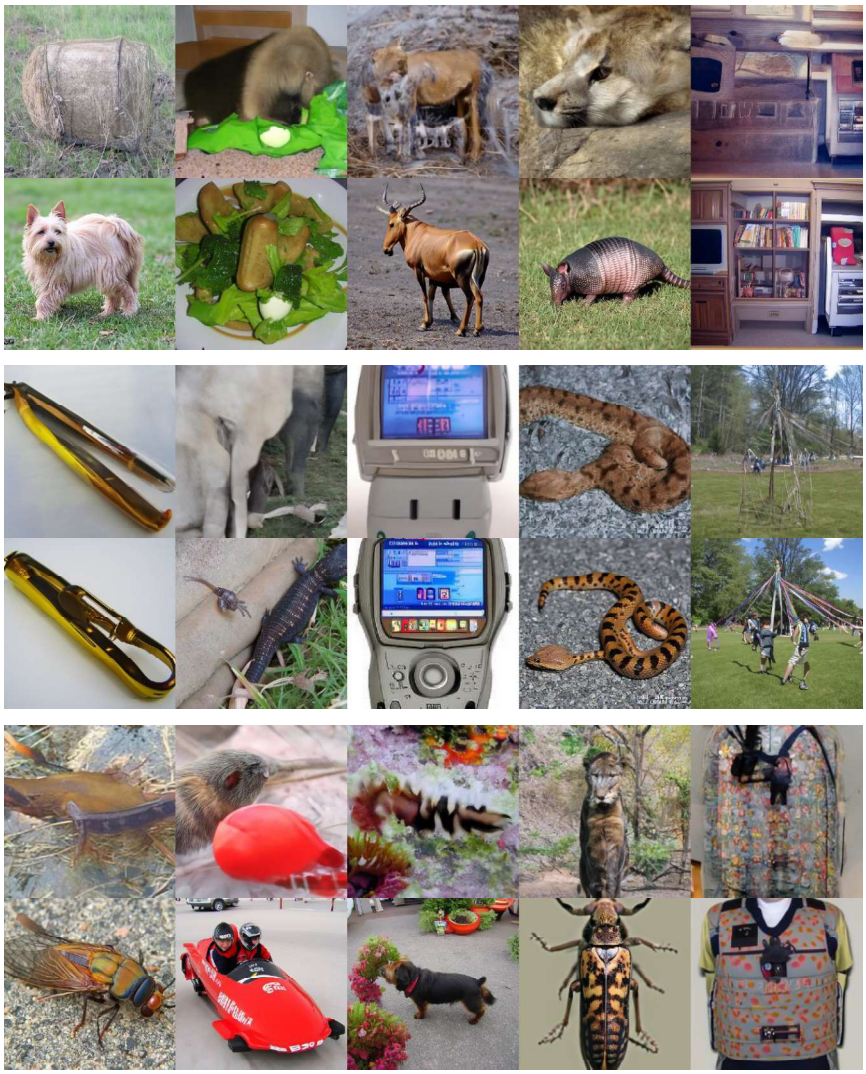

Fig. 11: Uncurated samples from ADM [9] ImageNet 256 *unconditional* model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: `i13`,`i14`,`i16`,`m1`.

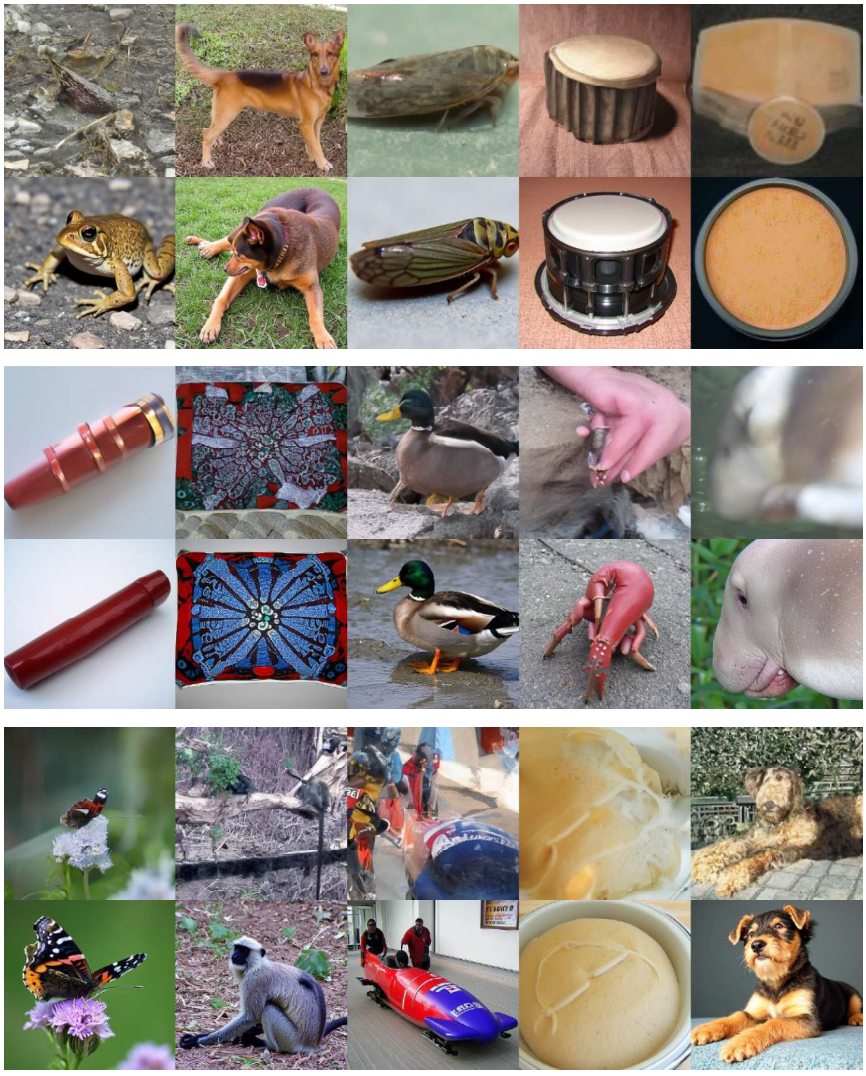

Fig. 12: Uncurated samples from ADM [9] ImageNet 256 *conditional* model w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 3.0$ is used and perturbed layers are following: `i13`,`i14`,`i16`,`m1`.

Fig. 13: Uncurated samples from ADM [9] 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. 14: Uncurated samples from SD [37] 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.

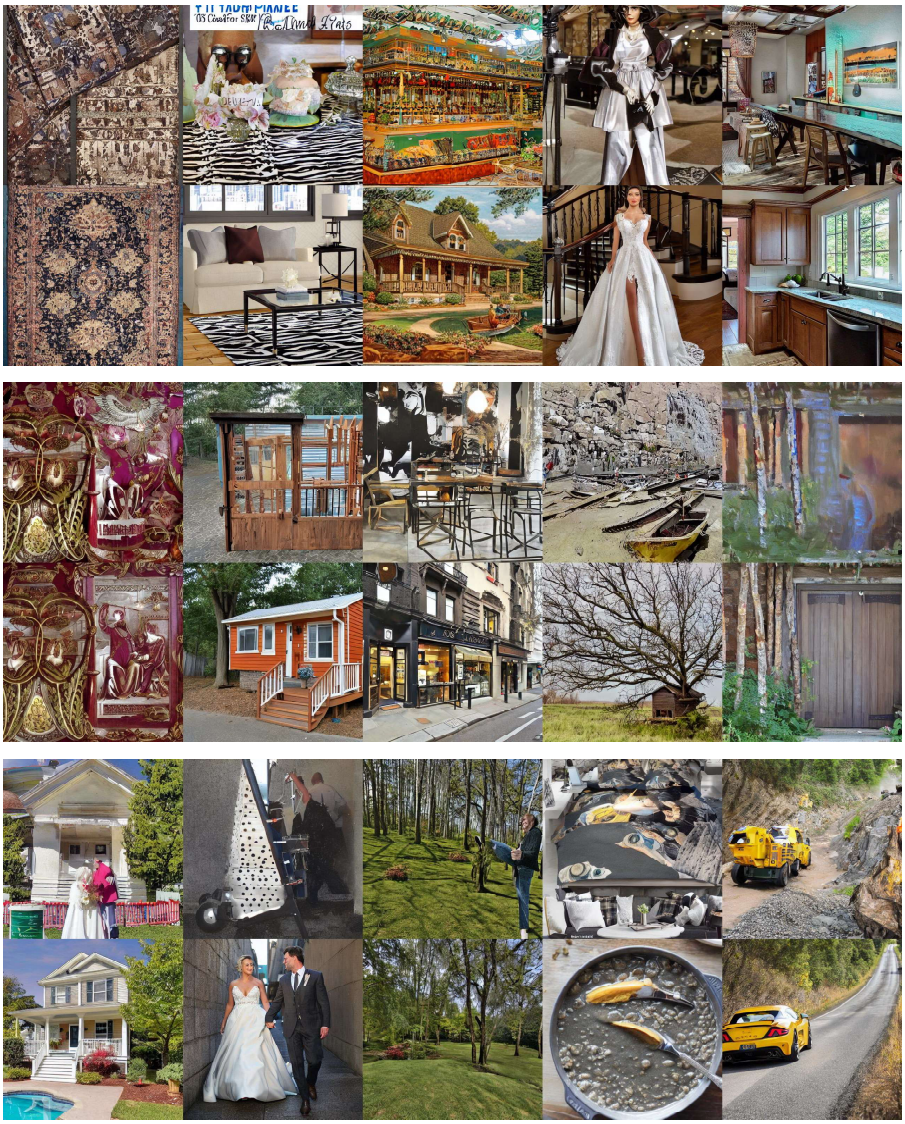

Fig. 15: Uncurated samples from SD [37] in *unconditional* generation w/o and w/ PAG. In each image set, the images in the top row are samples without using guidance, and the images in the bottom row are samples using PAG. PAG guidance scale $s = 5.0$ and perturbed layer `mid_block.attentions.0.transformer_blocks.0.attn1` are used.

Fig. 16: Uncurated samples from SD [37] 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 upsamle inputs to 512×512 as PSLD [40] does. Then, the outputs are downsampled to 256×256 for evaluation. Further qualitative results are provided in Fig. 17 18 19 20.

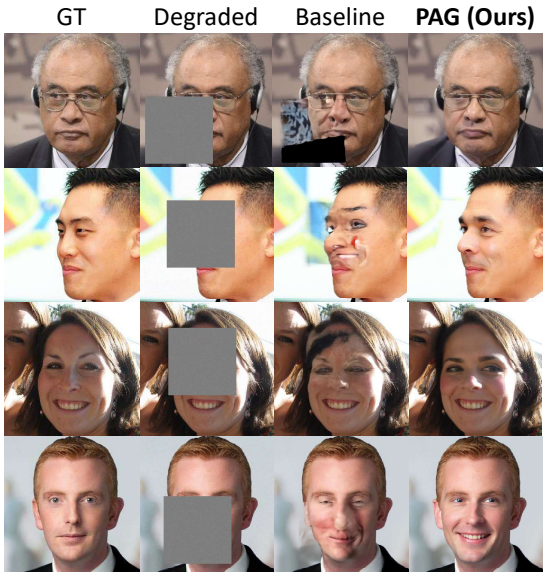

Fig. 17: Box inpainting results of PSLD [40] with PAG on FFHQ [21] dataset.

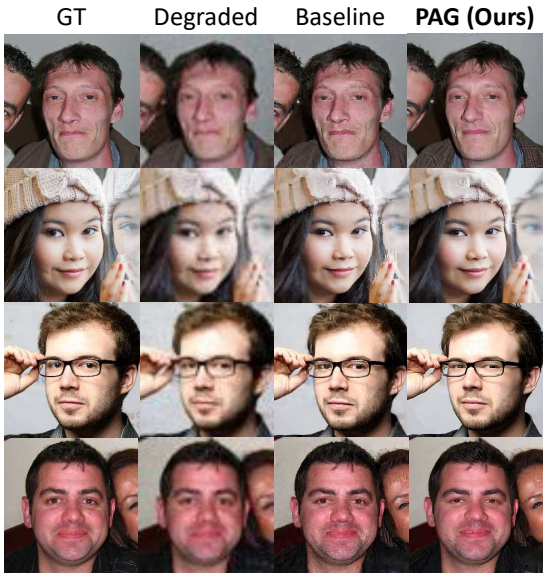

Fig. 18: Super-resolution (×8) results of PSLD [40] with PAG on FFHQ [21] dataset.

Fig. 19: Gaussian deblur results of PSLD [40] with PAG on FFHQ [21] dataset.

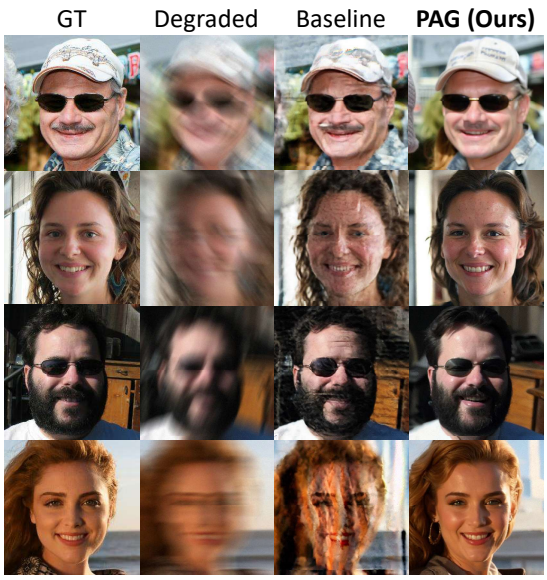

Fig. 20: Motion deblur results of PSLD [40] with PAG on FFHQ [21] dataset.

ImageNet. We use 1K ImageNet [8] 256×256 dataset which is used in [6,23,40]. Qualitative results shows that PAG properly improves sample quality with more various classes of images, as provided in Fig. 21 22 23 24.

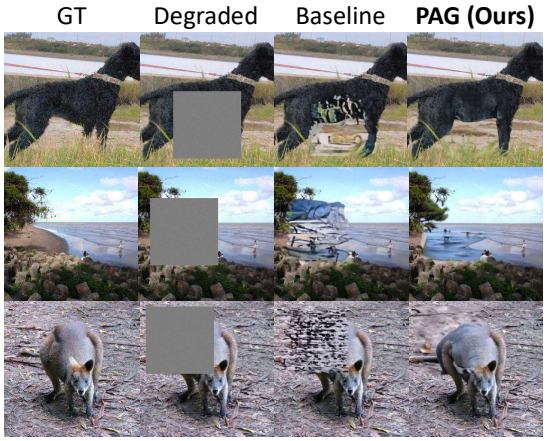

Fig. 21: Box inpainting results of PSLD [40] with PAG on ImageNet [8] dataset.

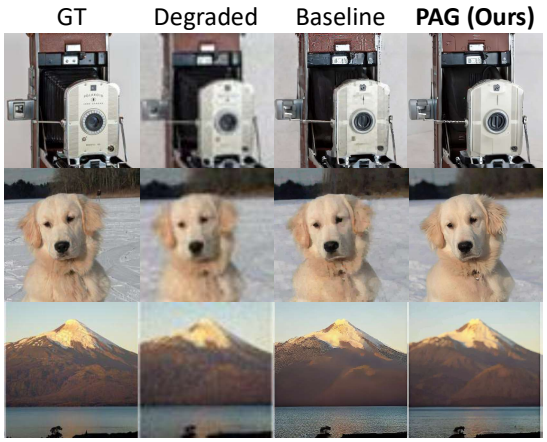

Fig. 22: Super-resolution(×8) results of PSLD [40] with PAG on ImageNet [8] dataset.

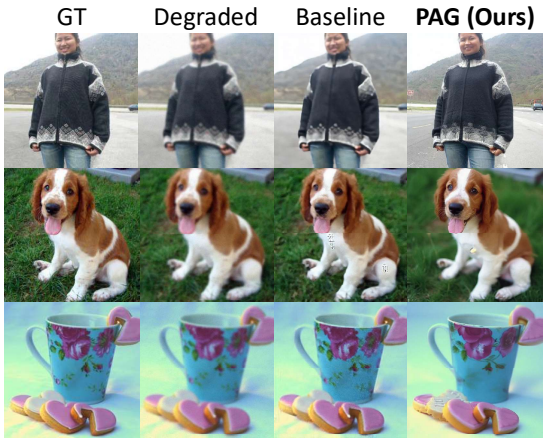

Fig. 23: Gaussian deblur results of PSLD [40] with PAG on ImageNet [8] dataset.

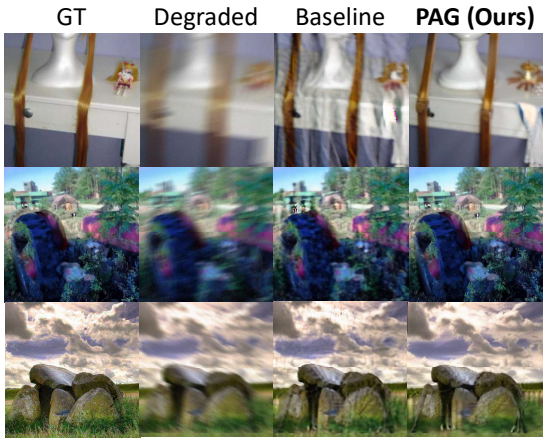

Fig. 24: Motion deblur results of PSLD [40] with PAG on ImageNet [8] dataset.

C Additional Applications

C.1 Diffusion Posterior Sampling

We conduct additional experiments on another diffusion restoration model, DPS [6], which is based on ADM [9]. 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 $\epsilon_\theta(\boldsymbol{z}_t)$ for predicting $\hat{\boldsymbol{z}}_0$, consistent with PSLD [6] experiments.

		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 [21] 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 [16] sampling. Quantitative results on 1K 256 are provided in Table 8. PAG outperforms baseline on FID [15], except for super-resolution(×8), where FID [15] 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 [15] over LPIPS [54]. Additional qualitative results are in Fig. 25 26 27 28 for ImageNet [8] dataset and Fig. 29 for FFHQ [21] dataset.

Table 8: Quantitative results of DPS [6] on FFHQ [21] 256×256 1K validation set [21].

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

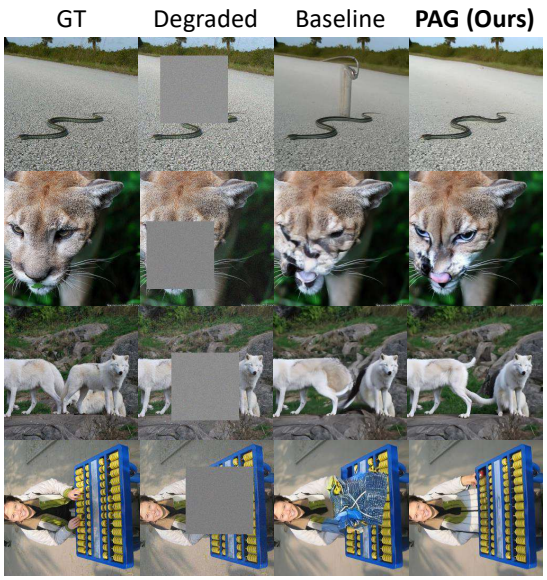

Fig. 25: Box inpainting results of DPS [6] with PAG on ImageNet [8] dataset.

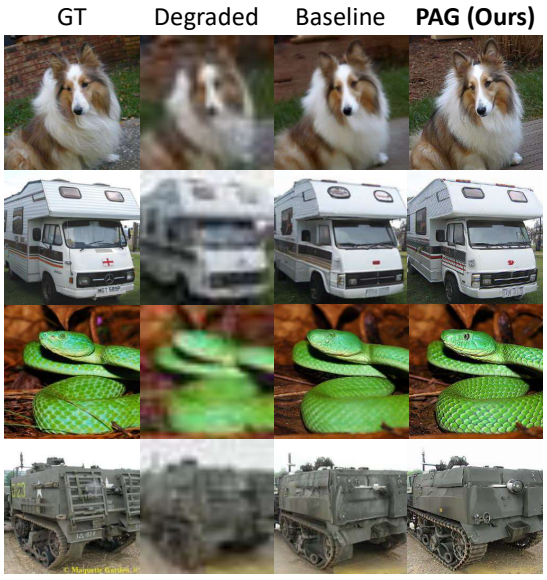

Fig. 26: Super-resolution(×8) results of DPS [6] with PAG on ImageNet [8] dataset.

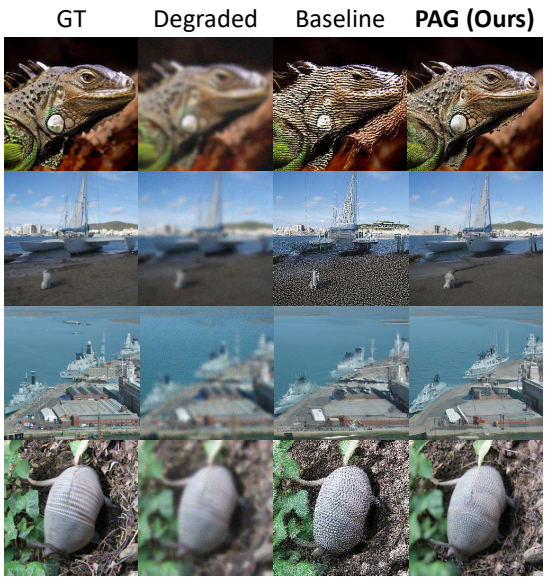

Fig. 27: Gaussian deblur results of DPS [6] with PAG on ImageNet [8] dataset.

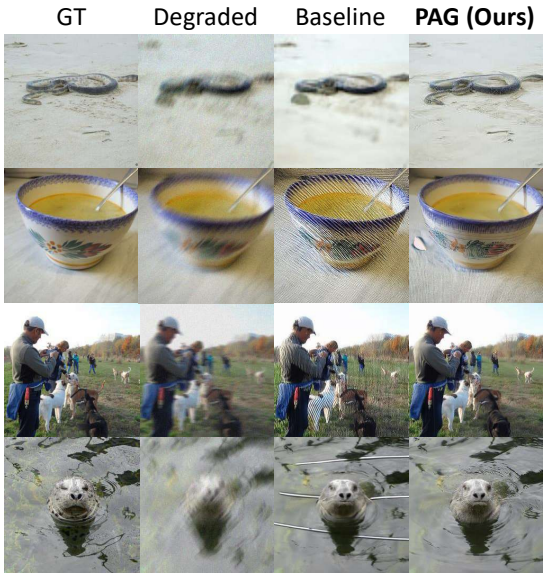

Fig. 28: Motion deblur results of DPS [6] with PAG on ImageNet [8] dataset.

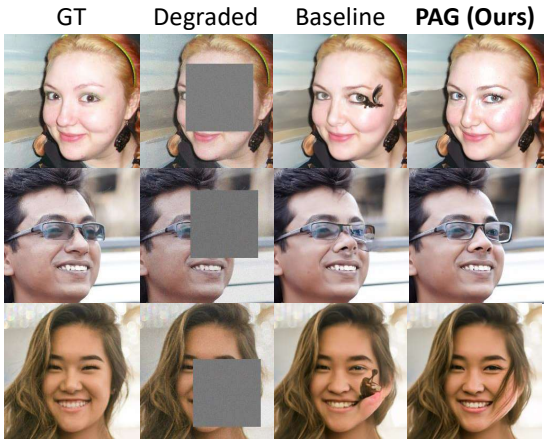

Fig. 29: Box inpainting results of DPS [6] with PAG on FFHQ [21] dataset.

C.2 Stable Diffusion Super-Resolution and Inpainting

Stable Diffusion [37] extends beyond the text-to-image pipeline to also support tasks requiring image input, such as super-resolution[7] and inpainting[8]. The model also requires text input alongside image input to leverage CFG [17], 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. 30 and 31 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 [19] dataset downscaled by a factor of 2 using bicubic interpolation and then center cropped to adjust the images to a resolution of 512×512.

[7] https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler
[8] https://huggingface.co/runwayml/stable-diffusion-inpainting

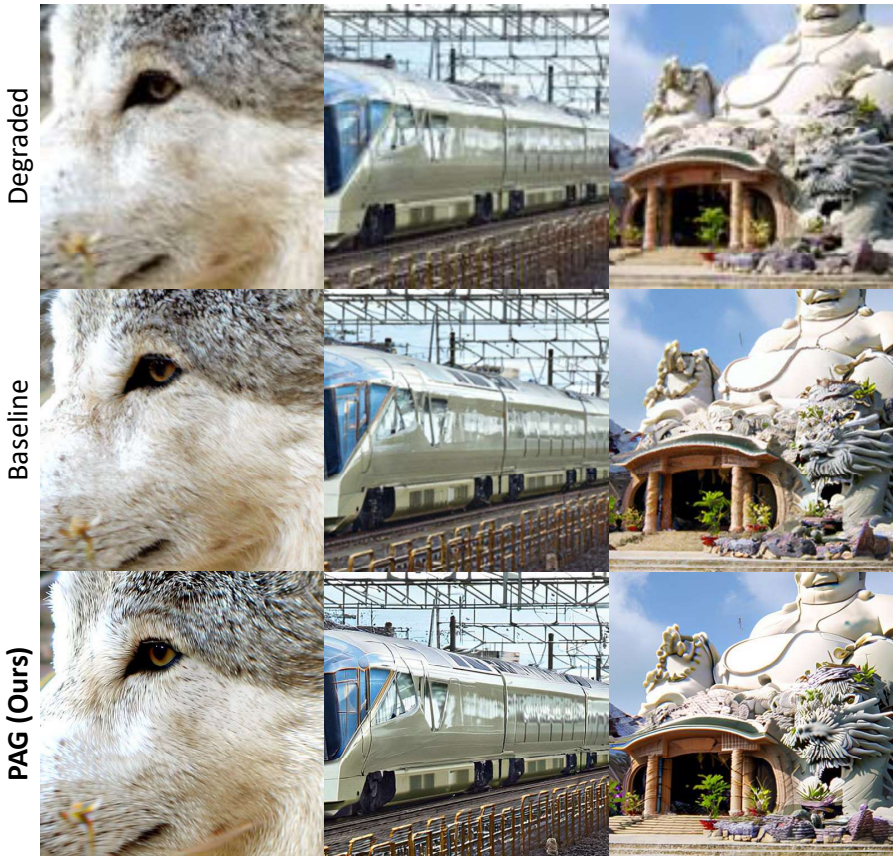

Fig. 30: Comparison of Stable Diffusion [37] super-resolution results between w/o and w/ PAG. PAG applies guidance that enables the model to upscale images to high-quality renditions with clearer edges and finer details, even when using an empty prompt (3rd row). The guidance scales employed, from left to right, are sequentially 3.0, 2.0, and 1.0. The model upscaled a 256×256 input image to 512×512.

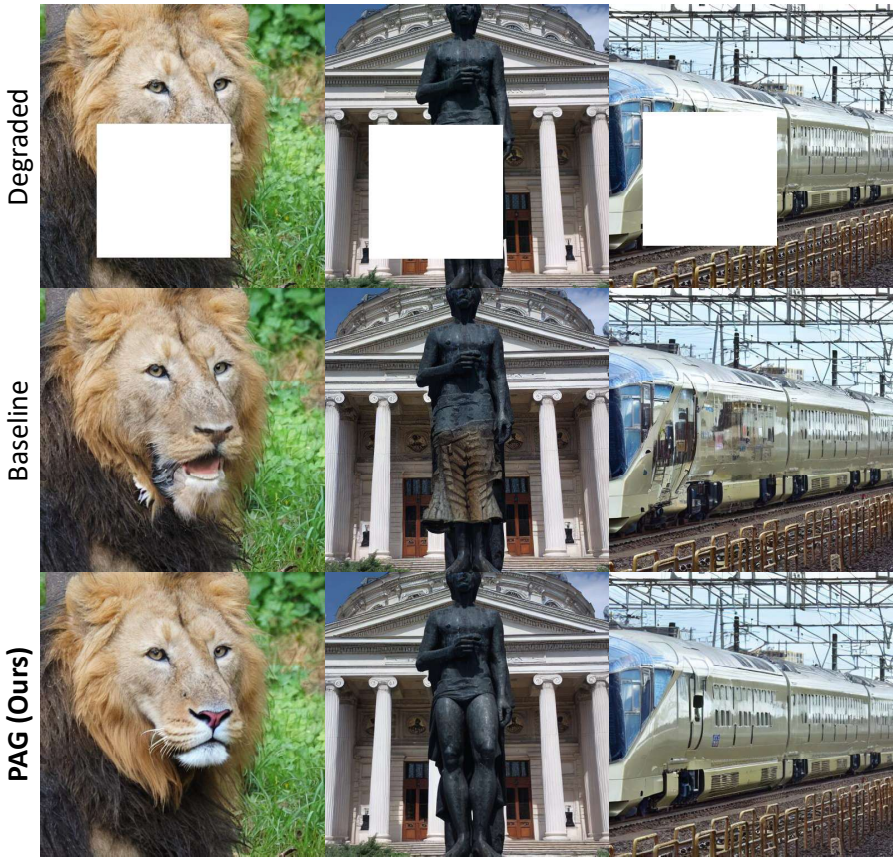

Fig. 31: Comparison of Stable Diffusion [37] inpainting results between w/o and w/ PAG. PAG aids the model in inpainting images, improving their realism and diminishing artifacts, without the necessity for a prompt (3rd row). The guidance scale of 1.5 is employed for all.

C.3 Text-to-3D

We integrated PAG with CFG for text-to-3D generation, utilizing the Dreamfusion [34] implementation provided by Threestudio [11] due to the unavailability of official code. We employed a scale of 100 for both CFG and PAG. As seen in Fig. 32, combining CFG with PAG yields results with enhanced details and textures compared to using CFG alone.

Fig. 32: Comparison of text-to-3D results between CFG [17], and CFG with PAG.

C.4 Human Evaluation

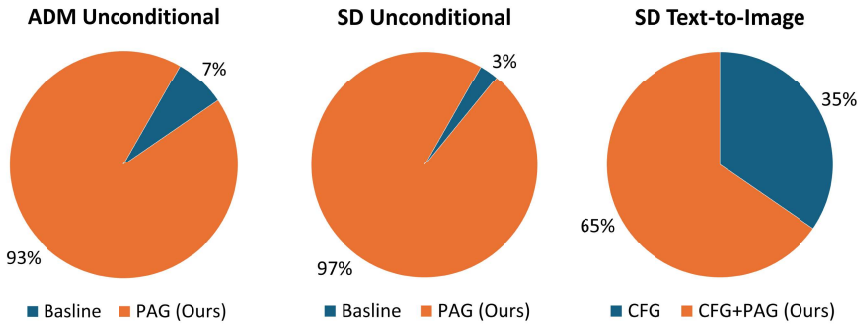

Fig. 33: The results of the user study.

A user study (Fig. 33) 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.

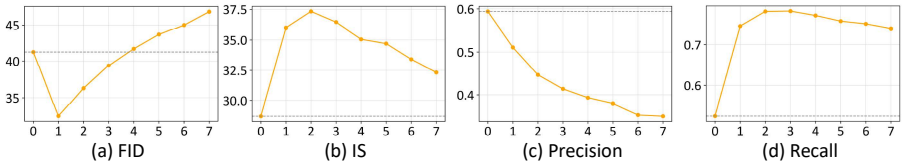

Fig. 34: Quantitative Analysis of Guidance Scale.

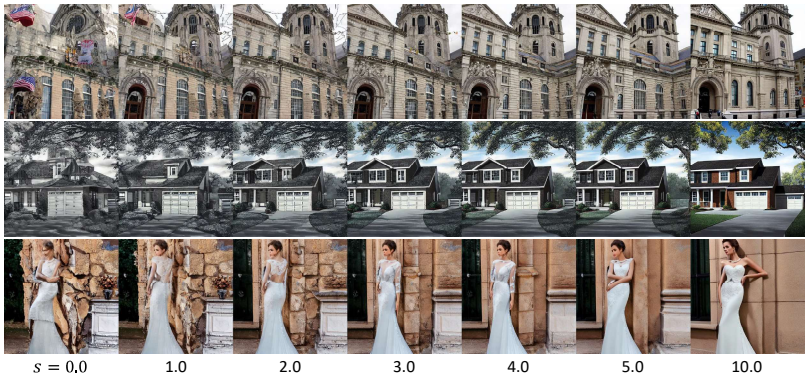

Fig. 35: Effect of Guidance Scale on Image Quality. Increasing the guidance scale s results in images with more semantically coherent structures and fewer artifacts, thereby improving their overall quality. However, an excessively large guidance scale can lead to smoother textures and slight saturation in the images, similar to the effects observed with CFG [17].

D Ablation Studies

D.1 Guidance scale

We conduct experiments to investigate the performance difference based on the guidance scale. Using scales set from 0.0 to 7.0 with intervals of 1.0, we sampled 5K images with ADM [9] and measured FID [15], IS [41], Precision, and Recall metrics [25] for these images. The results can be seen in the graph in Fig. 34. 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 [37]. In Fig. 35, 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.

D.2 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 illustrated in Fig. 36. 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. 36: Visualization of self-attention map masking strategy. For the evaluation of FID [15], we sample 5K images from ADM [9] ImageNet [8] 256×256 unconditional model for each method. Black entries indicate the masked (set to $-\infty$) elements of the self-attention map $\mathbf{A_t}$ 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**.

Table 9: Ablation study on perturbations. We sampled 5K images from the ADM [9] 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 ↓
Random Mask	40.20
Random Mask (off-diag)	39.49
Additive Noise	62.83
Gaussian Blur	35.48
Identity Matrix	**32.34**

Fig. 37: Uncurated samples from ADM [9] 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.3 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 [9] ImageNet 256×256 unconditional model and Stable Diffusion [37]. Observations indicate that perturbations applied to deeper layers generally yield relatively better outcomes compared to those applied to shallower layers of U-Net [38]. 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 [37]. 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.

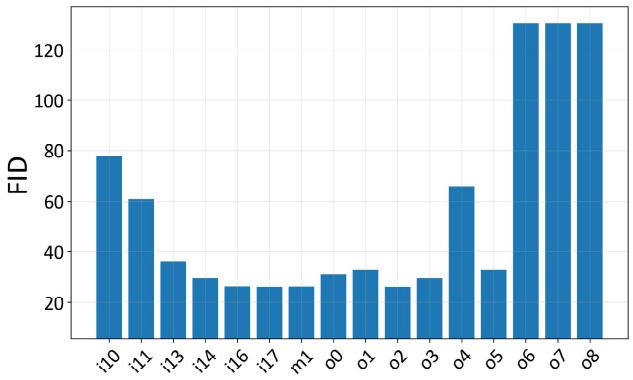

Fig. 38: Ablation study on which layer to apply perturbation with ADM [9].

Fig. 38 visualizes the FID scores obtained by perturbing each layer of the ADM [9] 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 [44] 25 step sampling are as follows, and in the case of sampling 5K images with DDPM [16] 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".

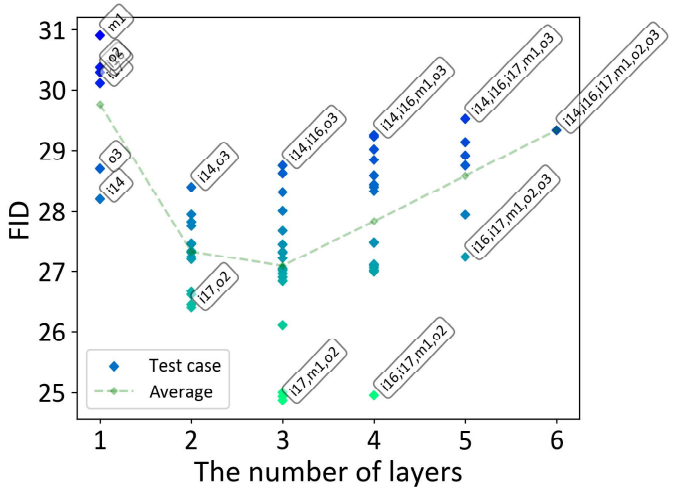

Fig. 39: Ablation study on layer combination for perturbed self-attention application in ADM [9]. Each data point represents the FID obtained when perturbed self-attention (PSA) is applied to the corresponding combination of layers. The annotations of points represent the combined layers. The green dashed 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 [37], indicating that the effectiveness of layer-wise perturbation varies across different diffusion models.

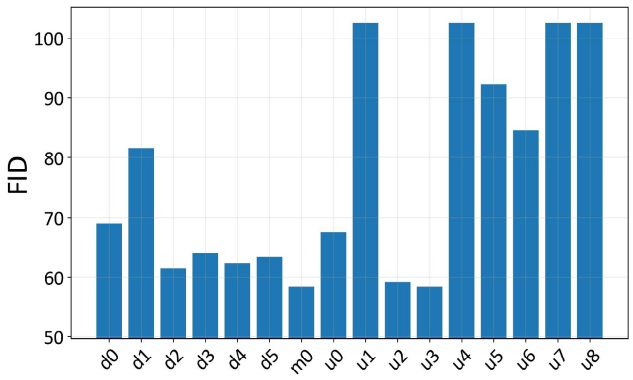

Fig. 40: Ablation study on which layer to apply perturbation with Stable Diffusion [37].

Table 10: Layer ablation on ADM. We evaluate the FID [15] of 5K samples from ImageNet [8] 256×256 unconditional model using DDIM [44] 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
	o2 o3	27.75
3	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
	i17 m1 o2	**24.87**
	i17 m1 o3	26.96
	i17 o2 o3	27.44
	m1 o2 o3	27.32
4	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
	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
5	i14 i16 i17 m1 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

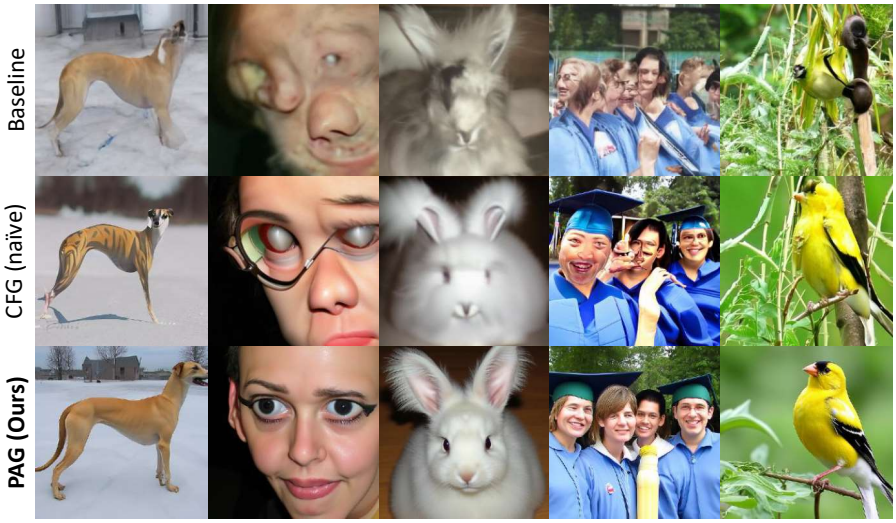

Fig. 41: Samples Using CFG with Separately Trained Models. We implement CFG [17] by employing separately trained ADM [9] ImageNet 256×256 conditional and unconditional models. Compared to samples with PAG in row 3, samples with naïve CFG in row 2 show inferior image quality. This suggests that when the 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 Theoretical Insights on Using Identity Matrix as Perturbation

Several studies have sought to establish its theoretical foundation, with a promising approach being its interpretation through pattern storage and retrieval behavior within the Energy-Based Model (EBM) framework. Based on this, we will explain why replacing the identity matrix works. Hopfield networks are associative memories that retrieve the pattern most similar to the input. They model an energy landscape with basins of attraction around desired patterns. [36] generalizes the energy function for continuous embeddings and demonstrates that the proposed update rule ensures global convergence: $\boldsymbol{\Xi}_{n+1} = \boldsymbol{X}\ \text{softmax}\left(\boldsymbol{X}^T \boldsymbol{\Xi}_n\right)$.

As shown in [36], this implicit energy minimization equation is closely linked to the self-attention forward pass of transformers by mapping X to K and Ξ to Q via projection matrices and introducing W_v for the key contents: $\mathbf{Q}^{\text{new}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$. This connection provides an insightful theoretical foundation for the attention mechanism. It suggests that the transformer's attention mechanism operates as an inner-loop optimization step minimizing the energy function determined by queries, keys, and values.

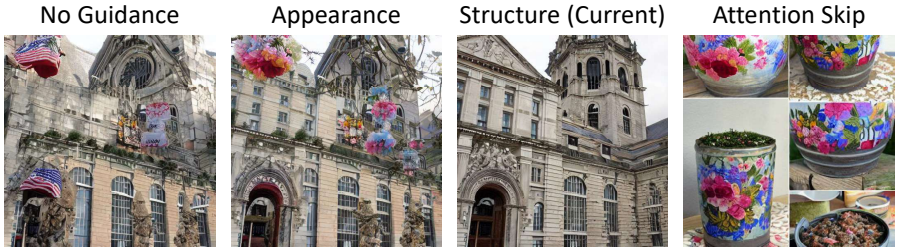

Fig. 42: Sampled images with various perturbations. Instead of replacing *structure* component Softmax ($Q_t K_t^T / \sqrt{d}$) with an identity matrix (Structure (Current)), we test perturbation on *appearance* component, replacing all tokens in V_t with its spatial average (Appearance). We also completely skip self-attention by ignoring self-attention which is learned in a residual manner (Attention Skip).

Thus, the forward pass of self-attention corresponds to *pattern retrieval* in the Hopfield network, and the backward pass updates the projection matrices (W_q, W_k, W_v) to reduce the final loss, implicitly learning to map the inputs to *useful patterns*. Specifically, W_q and W_k learn to model the relationships between inputs, while W_v learns the content patterns to be aggregated. Through pattern matching, self-attention effectively captures contextual relationships within the input.

We suggest replacing Softmax$(Q_t K_t^T / \sqrt{d})$ with an identity matrix to remove these relationships, resulting in an *undesirable* distribution in terms of structure. However, we keep the learned *content patterns* by passing value features, preserving local texture to make distribution *in-domain*. If we ignore both context and content patterns (value), it results in completely different images (Fig. 42 'Attention Skip'). Thus, identity-attention is a method intended to **maximize the use of learned self-attention patterns** to create an *in-domain* but *undesirable* distribution.

In summary, we selectively use learned intermediate representations to model the undesirable in terms of structure but still in-domain distribution by keeping the appearance information. There might be better perturbations but we observed that identity matrix replacement is an effective method, theoretically and empirically. We leave it for future works to find better perturbations for different tasks and models.

E.2 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** [17] 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 [18] and other perturbations (perturbation ablations and Sec. D.2), PSA can be considered an efficient and effective method.

Connections to delta denoising score. According to prior works [13, 22] 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 [13] 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 [17] 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 [18] 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.4.

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 [12, 47, 50] 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 sta-

ble 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 [17], 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.3 Complementarity between CFG and PAG

Recent research [2,31] 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.

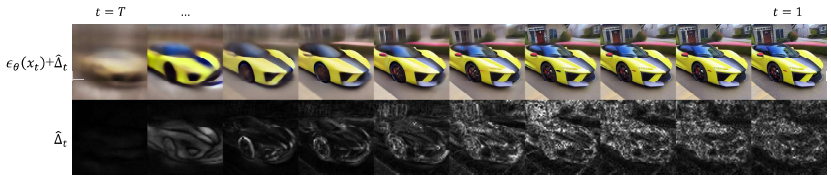

Fig. 43: Visualization of $\hat{\Delta}_t = \epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)$ during reverse process with PAG. text-to-image generation using PAG with a prompt *"a fancy sports car"*.

This phenomenon can also be observed in PAG and CFG contexts. Fig. 43 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.

In contrast, CFG exhibits a different behavior. Fig. 44 displays the timestep-wise predicted epsilon difference $\Delta_t = \epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)$ during sampling with

Fig. 44: 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"*.

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.

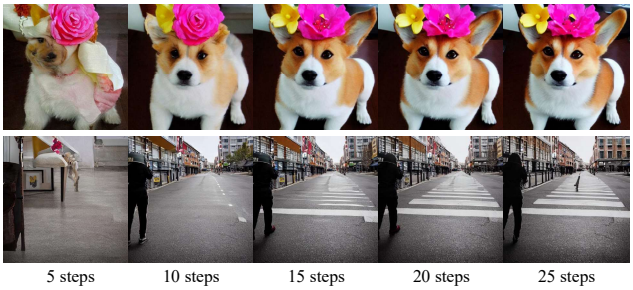

Fig. 45: Early stopping of CFG [17]. The process involves a total of 25 steps. The prompts used are *"A corgi with a flower crown"* (top) and *"A person walking on the street"* (bottom).

The impact of CFG being primarily in the early stages of the generation process can be validated through another observation. Fig. 45 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. 43. 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. 33.

E.4 Comparison with SAG

In this section, we summarize the differences between SAG [18] 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. 46, 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

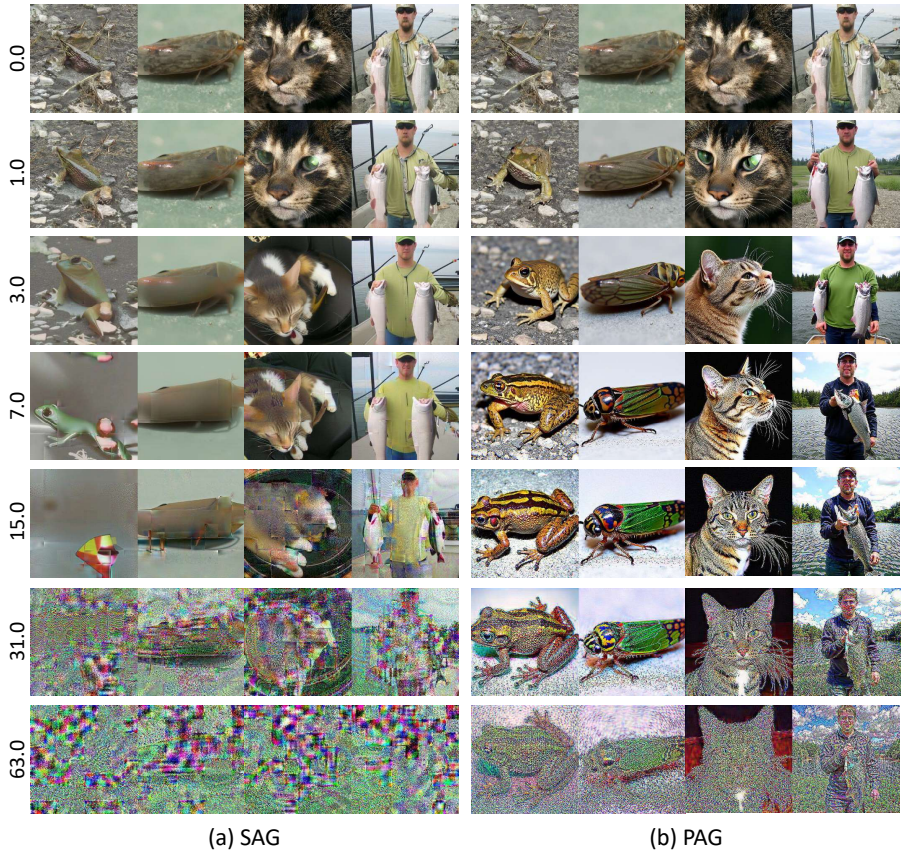

Fig. 46: Comparison of samples with SAG [18] and PAG for different guidance scales. Samples are generated by ADM [9] conditional ImageNet 256×256 model, showcasing the impact of incrementally increasing the guidance scale from 0.0 to 63.0, from the top to the bottom of the figure. **(a)**: Samples generated 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*.

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 [17]. Therefore, if the GPU can perform concurrent computations swiftly, PAG could theoretically 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.