Rethinking the Spatial Inconsistency in Classifier-Free Diffusion Guidance

Supplementary Material

1. Deriving Equation 11

In this section, we provide a derivation for Equation 11 based on one assumption that may be not particularly strict, i.e., *for any denoising step t, the semantic units, corresponding to token set $\{w_1, ..., w_L\}$, with masks $\{m_{t,1}, ..., m_{t,L}\}$ are independent of each other*. Along this line, we can derive:

$$\begin{aligned}
p(w_i|x_t) =& \ p(w_i|\sum_{j=1}^{L} m_{t,j} \odot x_t) \\
=& \frac{\prod_{j=1}^{L} p(m_{t,j} \odot x_t|w_i)p(w_i)}{\prod_{j=1}^{L} p(m_{t,j} \odot x_t)} \\
=& \frac{p(m_{t,i} \odot x_t|w_i)p(w_i) \prod_{j=1, j\neq i}^{L} p(m_{t,j} \odot x_t)}{\prod_{j=1}^{L} p(m_{t,j} \odot x_t)} \\
=& \frac{p(m_{t,i} \odot x_t|w_i)p(w_i)}{p(m_{t,i} \odot x_t)} \\
=& \ p(w_i|m_{t,i} \odot x_t).
\end{aligned}$$

Then, we can deduce Equation 11 as follows:

$$\begin{aligned}
p(c|x_t) &= \prod_{i=1}^{L} p(w_i|x_t) \\
&= \prod_{i=1}^{L} p(w_i|m_{t,i} \odot x_t). \\
\nabla_{x_t} \log p(w_i|m_{t,i} \odot x_t) &= \nabla_{m_{t,i} \odot x_t} \log p(w_i|m_{t,i} \odot x_t) \\
&= \nabla_{m_{t,i} \odot x_t} \log p(w_i|x_t) \\
&= \nabla_{m_{t,i} \odot x_t} \log p(c|x_t) \\
&= m_{t,i} \odot \nabla_{x_t} \log p(c|x_t).
\end{aligned}$$

Note that the prior assumption may not be strict in practice. However, it is intuitive that the patches among different semantic regions are more independent than those in the same patches. Meanwhile, based on the segmentation examples in Figure 3 and our experimental results, we believe that it is beneficial to segment the latent image and customize guidance degrees for different semantic regions.

2. More Experimental Details

Benchmark Models. In our experiment, we involve three special diffusion models as the benchmarks, which are all publicly accessible:

- Stable Diffusion v1.5 (**SD-v1.5**), a diffusion model in the latent space of powerful pre-trained autoencoders [1], which use the CLIP [2] as the text encoder and output images with the resolution 512x512.
- Stable Diffusion v2.1 (**SD-v2.1**), a variant of SD-v1.5 with more model size [2], which can output images with the resolution 768×768.
- DeepFloyd IF (**IF**), is a diffusion model in the pixel image space [3], which is constructed using multiple diffusion models with T5XXL as the text encoder. In particular, we use the first two stages of the middle-scale version, i.e., IF-I-M-v1.0 and IF-II-M-v1.0, which produce the 64×64 resolution image and boost them into 256× 256 resolution, respectively.

Quantitative Metric. Two qualitative metrics based on the MSCOCO validation dataset are used:

- **FID-30K**, where the FID score is computed on the 30K generated images with prompts selected from the validation set and the corresponding original images.
- **CLIP Score**, where 5K captions are selected randomly for guiding image synthesis, and CLIP-VIT-G-14 [4] is used to compute the similarity between the generated image and the corresponding caption.

In particular, our metric settings may be different from those in the official reports of the SD and IF models. It is somewhat weird that SD-v2.1 fails to outperform SD-v1.5 in our settings. Here, we also add another comparison on them based on a similar setting to their official report [5], i.e., where FID-10k and CLIP Score (CLIP-VIT-G-14) on MSCOCO dataset are used with the 50-step DDIM sampler. The results are shown in Figure 1. We can find that our S-CFG strategy also outperforms the original CFG strategy.

3. Analysis on the Efficiency

Here, we provide an additional analysis of the time cost of our S-CFG strategy. Specifically, we use DPMSolver++ with 50 steps as the sampler to generate images with different base models. All programs run on a single A100 GPU. Table 1 shows the average time cost for generating a sample in 10 runs. We can find only a tiny time cost has been required compared with the original CFG strategy.

4. More Ablation Analysis

Here, we provide an additional ablation analysis of the S-CFG on the diffusion model with multiple stages, such as DeepFloyd IF [3]. We try to respond to the question: *should*

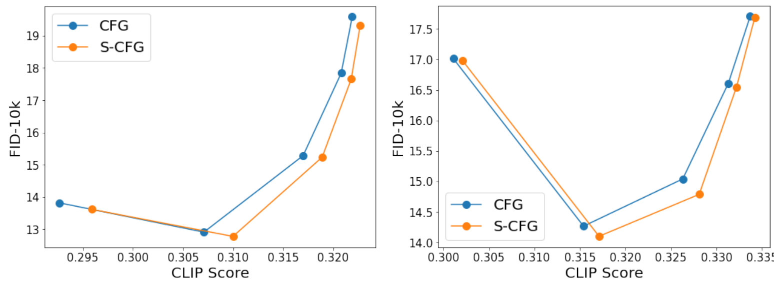
[1]https://huggingface.co/runwayml/stable-diffusion-v1-5

[2]https://huggingface.co/stabilityai/stable-diffusion-2-1

[3]https://huggingface.co/DeepFloyd/IF-I-M-v1.0

[4]https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s34B-b88K

[5]https://huggingface.co/stabilityai/stable-diffusion-2



(a) SD-v1.5 (b) SD-v2.1

Figure 1. **The trade-off curve of FID-10K VS CLIP Score with DDIM sampler.**

Table 1. **The analysis on the time cost.**

	CFG	S-CFG	improv.
SD-v1.5	2.773	2.848	2.70%
SD-v2.1	7.054	7.167	1.60%
IF	8.595	8.847	2.93%

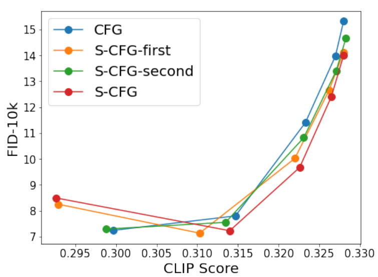


Figure 2. **The ablation analysis of the S-CFG on the diffusion model with multiple stages.**

the S-CFG strategy be used on all diffusion stages? Specifically, based on the IF model used in our paper, we compare the performance of three methods:

- **S-CFG-first**, where the S-CFG strategy is only used in the first diffusion model, i.e., IF-I-M-v1.0.
- **S-CFG-second**, where the S-CFG strategy is only used in the second diffusion model, i.e., IF-II-M-v1.0.
- **S-CFG**, where the S-CFG strategy is used in both two diffusion models.

In addition, the original CFG strategy is involved as a baseline. We use DPMSolver++ as the sampler with 50 steps and vary the parameter γ in [2.0, 3.0, 5.0, 7.5, 10.0]. The trade-off curve of FID-30k VS CLIP Score is shown in Figure 2. We can find that S-CFG tends to achieve the best trade-off between FID-30K and ClIP Score, while S-CFG-first and S-CFG-second perform similarly.

5. More Evaluation on Effectiveness

Recently, a new metric called T2I-CompBench [1] was introduced to evaluate diffusion models, which assesses image quality from 6 aspects and aligns with human preference better. Here, we provide another comparison based on

Table 2. Evaluation on T2I-CompBench, where the $\gamma = 7.5$.

Model	Attribute Binding			Object Relationship		Complex
	Shape	Color	Texture	Non-Spatial	Spatial	
SD-v1.5+CFG	0.3664	0.3761	0.4286	0.3109	0.111	0.2969
SD-v1.5+S-CFG	**0.3793**	**0.3879**	**0.4288**	**0.3111**	**0.1182**	**0.2993**
SD-v2.1+CFG	0.4518	0.549	0.5146	0.3096	0.1512	0.3154
SD-v2.1+S-CFG	**0.4558**	**0.5649**	**0.5333**	**0.3104**	**0.1567**	**0.3168**

this metric. The results in Table 2 show that SD-v2.1 outperforms SD-v1.5 significantly, and S-CFG performs better than CFG.

6. Detailed Table of Experiments

Here, we show the detailed tables for experiments in Figures 4 and 6. We can find that our S-CFG achieves the best performance on all settings, with the best FID-30K score and CLIP Score.

7. Additional Qualitative Samples

In this section, we present supplementary samples in Figure 3 generated by different base models with CFG and S-CFG. These additional samples further exhibit the superiority of S-CFG compared with the original CFG strategy.

References

[1] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[3] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd if, 2023. https://www.deepfloyd.ai/deepfloyd-if. 1

Table 3. **The trade-off curve of SD-v1.5**, where the best FID-30k and CLIP Score are highlighted.

	DDIM				DPMSolver++			
	CFG		S-CFG		CFG		S-CFG	
γ	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score
2.0	8.696	0.2948	8.656	0.2972	8.991	0.2954	9.023	0.2964
3.0	**7.904**	0.3097	**7.802**	0.3107	**7.760**	0.3091	**7.717**	0.3099
5.0	10.366	0.3184	10.069	0.3196	10.026	0.3182	9.757	0.3187
7.5	13.008	0.3217	12.620	0.3228	12.466	0.3223	12.059	0.3226
10.0	14.682	**0.3230**	14.101	**0.3231**	14.107	**0.3235**	13.694	**0.3236**

Table 4. **The trade-off curve of SD-v2.1**, where the best FID-30k and CLIP Score are highlighted.

	DDIM				DPMSolver++			
	CFG		S-CFG		CFG		S-CFG	
γ	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score
2.0	14.394	0.3053	13.892	0.3068	14.999	0.3040	14.864	0.3060
3.0	10.509	0.3191	10.227	0.3204	10.869	0.3187	10.797	0.3200
5.0	**10.429**	0.3286	**10.137**	0.3306	**10.241**	0.3291	**10.016**	0.3304
7.5	11.548	0.3331	11.278	0.3342	11.324	0.3339	10.944	0.3342
10.0	12.604	**0.3357**	12.371	**0.3359**	12.166	**0.3356**	11.833	**0.3359**

Table 5. **The trade-off curve of IF**, where the best FID-30k and CLIP Score are highlighted.

	DDIM				DPMSolver++			
	CFG		S-CFG		CFG		S-CFG	
γ	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score
2.0	**9.820**	0.3076	**9.309**	0.299	**7.242**	0.2997	8.494	0.2926
3.0	13.804	0.3195	10.864	0.3152	7.799	0.3147	**7.227**	0.314
5.0	17.267	0.3257	14.473	0.3259	11.396	0.3233	9.67	0.3226
7.5	18.532	0.329	16.621	0.3288	13.968	0.327	12.402	0.3265
10.0	19.029	**0.3296**	17.634	**0.3299**	15.31	**0.3280**	13.99	**0.3280**

Table 6. **The trade-off curve in the ablation analysis** , where the best FID-30k and CLIP Score are highlighted. The experiment is based on SD-v1.5 with 50-step DPMSolver++ Sampler.

	S-CFG-mean		S-CFG w/o sa		S-CFG-sa		S-CFG	
γ	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score	FID-30K	CLIP Score
2.0	10.703	0.2869	9.110	0.2963	9.063	0.2966	9.023	0.2964
3.0	**7.695**	0.3044	**7.811**	0.3089	**7.736**	0.3099	**7.717**	0.3099
5.0	8.813	0.3162	9.822	0.3185	9.755	0.3185	9.757	0.3187
7.5	11.204	0.3213	12.102	0.3222	12.083	0.3227	12.059	0.3226
10.0	12.838	**0.3233**	13.722	**0.3235**	13.690	**0.3235**	13.694	**0.3236**

Figure 3. **More samples generated by different base models with CFG (left) or S-CFG (right).**