Rethinking the Spatial Inconsistency in Classifier-Free Diffusion Guidance

Dazhong Shen[1], Guanglu Song[2], Zeyue Xue[3], Fu-Yun Wang[4], Yu Liu[1,2,*]
[1]Shanghai Artificial Intelligence Laboratory, [2]SenseTime Research,
[3]The University of Hong Kong, [4]The Chinese University of Hong Kong

Abstract

Classifier-Free Guidance (CFG) has been widely used in text-to-image diffusion models, where the CFG scale is introduced to control the strength of text guidance on the whole image space. However, we argue that a global CFG scale results in spatial inconsistency on varying semantic strengths and suboptimal image quality. To address this problem, we present a novel approach, Semantic-aware Classifier-Free Guidance (S-CFG), to customize the guidance degrees for different semantic units in text-to-image diffusion models. Specifically, we first design a training-free semantic segmentation method to partition the latent image into relatively independent semantic regions at each denoising step. In particular, the cross-attention map in the denoising U-net backbone is renormalized for assigning each patch to the corresponding token, while the self-attention map is used to complete the semantic regions. Then, to balance the amplification of diverse semantic units, we adaptively adjust the CFG scales across different semantic regions to rescale the text guidance degrees into a uniform level. Finally, extensive experiments demonstrate the superiority of S-CFG over the original CFG strategy on various text-to-image diffusion models, without requiring any extra training cost. our codes are available at https://github.com/SmilesDZgk/S-CFG.

1. Introduction

Recently, text-to-image generation has witnessed rapid development and various applications [29, 30, 32, 33, 45], where visually stunning images can be created by simply typing in a text prompt. In particular, after DDPM [7, 12] succeeded GANs [3, 8], diffusion models [39], such as Stable Diffusion [33] and DallE-3 [2], have emerged as the new state-of-the-art family for image-generative models.

The key feature of diffusion models is to approximate the true data distribution $p(x)$ by reversing the process of perturbing the data with noise progressively in a long iterative chain. To incorporate the text prompt c into the final generation, it is necessary to enhance the likelihood of c given the current latent image x_t at each reversed diffusion step t. Instead of training extra classifiers to model $p(c|x_t)$ at each diffusion step t [7], classifier-free guidance (CFG) [11] has recently been proposed to estimate both the classifier score $\nabla_{x_t} \log p(c|x_t)$ and the diffusion score $\nabla_{x_t} p(x_t)$ with the same neural model, such as U-net [34]. In particular, an empirical CFG scale is introduced to control the strength of the text guidance on the whole image space.

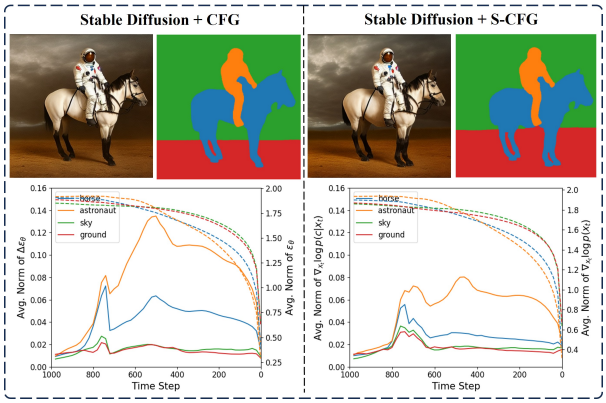


Figure 1. **A motivation example.** The first line shows images generated by Stable Diffusion with CFG and S-CFG, where the prompt is "*a photo of an astronaut riding a horse*" and the segmentation maps are manually labeled (Ground, Sky, Horse, Astronaut). The below line shows the average norm curves of the estimated classifier score $\nabla_{x_t} \log p(c|x_t)$ (solid line) and diffusion score $\nabla_{x_t} \log p(x_t)$ (dashed line) in each semantic region. The Y-axis scale unit is set as the dynamic variance parameter σ_t for better illustrations without damaging the conclusion.

However, we argue that *a global CFG scale results in spatial inconsistency on varying semantic strengths during the denoising process and suboptimal quality of the final image.* Figure 1 shows samples generated by Stable Diffusion [33]. The images can be segmented into four semantic regions corresponding to "astronaut", "horse", "sky" and "ground". To compare the guidance degrees assigned to different semantic units, the figures in the second line illustrate the average norm curves of the estimated classifier

*the corresponding author: liuyuisanai@gmail.com

score $\nabla_{x_t} \log p(c|x_t)$ and diffusion score $\nabla_{x_t} \log p(x_t)$ in each semantic region at any time step. as for the images with the original CFG strategy, we can find that the classifier score norm changes a lot on different semantic units, while the norms of diffusion scores seem to be closer. Intuitively, the larger classifier score implies a greater guidance degree received by the semantic unit. As a result, the final generative samples may exhibit spatial inconsistency in image qualities for different semantic units. For instance, the "astronaut" region, which consistently attains the highest score ratio, displays intricate and finely detailed structures that starkly contrast with the "sky" and "ground" regions.

Along this line, in contrast to the previous works, we propose to set customized CFG scales for different semantic regions of the latent image at each denoising step. In particular, we assume that the inter-patches in each semantic region serve a similar semantic concept and different regions are relatively independent. In this case, the classifier scores $\nabla_{x_t} \log p(c|x_t)$ can be approximately deduced into the combination of that conditioning on all independent semantic regions. Therefore, customized CFG scales can be safely involved for each semantic region, without the disruption of relative relations among interdependent patches. However, it is not trivial to conduct semantic segmentation on the latent image without accessing the final generated image. Meanwhile, determining the customized CFG scales to balance semantic units is another challenge.

To this end, in this paper, we propose a novel approach, called Semantic-aware Classifier-Free Guidance (S-CFG), to dynamically and customizedly control the text guidance degrees in text-to-image diffusion models. Specifically, when modeling the conditional distribution $p(x|c)$, diffusion models take c as another input with self-attention and cross-attention layers to mix up the image and text, which preserves the underlying semantic information. Along this line, we first design a training-free segmentation method for the latent images at each denoising step. In particular, the cross-attention map in the denoising U-net backbone is renormalized for assigning each patch to the corresponding token, while the self-attention map is used to complete the semantic regions. Then, to balance the amplification of diverse semantic information, we rescale the classifier score $\nabla_{x_t} \log p(c|x_t)$ across different semantic regions to a uniform level with the adaptive CFG scales. Finally, we conduct qualitative and quantitative analysis based on various diffusion models. The results demonstrate that S-CFG can outperform the original CFG strategy and obtain a robust improvement without any extra training cost. At first glance, the right part in Figure 1 demonstrates reduced disparities among the classifier score norms $\nabla_{x_t} \log p(c|x_t)$ of different semantic units in the image with S-CFG. As a result, more abundant clouds float in the "sky". The boundary between the "sky" and the "ground" is clearer.

2. Related Work

2.1. Image Diffusion Generative Models

Recently, diffusion models have emerged as an expressive and flexible family for image generation with remarkable image quality and various applications [1, 13, 17, 24, 29, 30, 33]. The general idea is to apply a forward diffusion process that adds tiny noise to the input data, then learn the reverse process with neural networks to gradually recover the original samples from the noisy data, step-by-step. Among them, Denoising Diffusion Probabilistic Model (DDPM) [12] is the representative baseline, which carefully designed the noise schedule on the pixel space during the forward process and the network architecture in the reverse process. As a result, diffusion models achieved better model coverage and training stability compared to GANs [3, 8, 15]. To further reduce computational costs, the subsequent study turned to combining DDPM and VAE [18, 31, 37] by applying diffusion models to the lower-dimensional latent space of a VAE trained on large-scale image datasets, such as Stable Diffusion [33]. In general, diffusion models suffer the downside of low inference speed compared to other generative models. However, this problem can be greatly alleviated by advanced sampling strategies, such as DDIM [40, 49], DPMSolver [22, 23], PNDM [16], Euler [16], and DEIS [48], which can perform 10X to 100X speedup compared to the original DDPM sampler. Here, we further explore a better way for image generation based on diffusion models.

2.2. Text-guided Generation

Recently, the text-guided generation in diffusion models has reached an unprecedented level, like DallE-3 [2]. This generative power stems from three aspects. First, to represent the unstructured text, expressive language embedding models are used to embed each token in the given text, such as CLIP [27] in Stable Diffusion [33], and T5 [28] in Imagen [36]. Second, to facilitate the interaction between text and image information, diffusion models typically enhance the network backbone, such as the U-net backbone [34], with the cross-attention mechanism. This mechanism involves utilizing the image embedding as the query and the key and value embeddings derived from the text. Third, Classifier-Free Guidance (CFG) [11] has recently been widely involved as a lightweight and robust technique to encourage text prompt adherence in generations. Instead of training extra classifiers [7, 21], CFG mixes the score estimates of the diffusion model with or without the conditional prompt. Some other works [14, 20] further separate a prompt into multiple concepts and generate an image by combining a set of diffusion models with each of them conditioning on a certain concept component. Here, we further emphasize the importance of varying CFG scales across different image semantic regions and design the semantic-ware CFG strategy to improve image quality.

2.3. Applications with Cross-Attention Maps

Cross-attention maps in the diffusion U-net Backbone are derived to represent the spatial relation between image patches and prompt tokens. They provide valuable semantic information for image segmentation and can contribute to various applications. For example, some works [5, 6, 44, 50] introduce layout control in image generation by minimizing the difference between the cross-attention-based semantic segmentation and the given layout conditions. Prompt2Prompt [10] achieves image editing by simply replacing, adding, or re-weighting cross-attention maps. Attend-and-Excite [4] improves the text alignment by optimizing the cross-attention maps during the inference process. Subsequent works further extend those ideas for image-to-image translation [26], text-driven image editing [9, 42], and compositional image generation [43]. In this paper, we further use cross-attention maps to improve image quality by segmenting latent images and customizing the guidance degrees of different semantic regions.

3. Preliminary

3.1. Diffusion Models

Given the image data space $\mathcal{X}$, diffusion models define a Markov Chain, known as the forward process, to corrupt the real data $x_0 \in \mathcal{X}$ by progressively adding Gaussian noise from time steps 0 to T:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\{\beta_t\}_{t=1:T}$ denotes the variance for each noise step, set as constant usually. Taking advantage of the properties of the Gaussian distribution, we can obtain x_t at an arbitrary time step t using the following closed form:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. x_T will degrade to standard Gaussian noise with $\overline{\alpha}_T \approx 0$.

The reverse denoising process aims to approximate the true posterior of each forward step via a time-dependent neural network parameterized by θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)\mathbf{I}), \tag{3}$$

which can be used to generate image $x_0 \sim p_\theta(x_0)$ by sampling Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ first and denoising step-by-step from x_{T-1} to x_0. In practice, to simplify the model training, $\sigma_\theta(x_t, t)$ is set as constant σ_t [7] and $\mu_\theta(x_t, t)$ is parameterized as follows:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{1-\overline{\alpha}_t}\epsilon_\theta(x_t, t)\right), \tag{4}$$

where the neural model ϵ_θ, such as U-net [34], is trained to predict the noise ϵ_t added in each forward step, which also mirrors the denoising score-matching, i.e, $\epsilon_\theta(x_t, t) \approx -\sigma_t\nabla_{x_t}\log p(x_t)$.

3.2. Classifier-free Guidance

The vanilla diffusion model described above is an unconditional generative model $p_\theta(x_0)$ to approximate the true data distribution $q(x_0)$. However, in practical scenarios, there is a growing demand to condition the generation on a label or text prompt c [46]. To address this requirement, classifier-guidance [7] incorporates an auxiliary classifier $p_\phi(c|x_t)$ to guide the sampling in each reverse denoising step, thereby increasing the likelihood of c given x_t. Specifically, the diffusion score is modified as follows:

$$\begin{aligned}\hat{\epsilon}_\theta(x_t.c, t) &= \epsilon_\theta(x_t, t) - \gamma\sigma_t\nabla_{x_t}\log p_\phi(c|x_t) \\ &\approx -\sigma_t\nabla_{x_t}\log(p(x_t)p_\phi^\gamma(c|x_t)),\end{aligned} \tag{5}$$

where γ is a scalar parameter to regulate the strength of the classifier guidance. While this method has demonstrated some performance improvements, training a robust classifier for all reverse steps, particularly for the highly noisy input at the initial step, poses a significant challenge and incurs additional training costs.

To avoid training a separate classifier model, classifier-free guidance [11] takes c as another input of the denoising neural network to model the conditional diffusion score, i.e., $\epsilon_\theta(x_t, c, t) \approx -\sigma_t\nabla_{x_t}\log p(x_t|c)$, while the unconditional score $\epsilon_\theta(x_t, t)$ is jointly estimated by randomly dropping the text prompt with a certain probability at each training iteration. Then the gradients for the classifier $p_\phi(c|x_t)$ can be estimated as:

$$\begin{aligned}\nabla_{x_t}\log p(c|x_t) &= \nabla_{x_t}\log p_\theta(x_t|y) - \nabla_{x_t}\log p_\theta(x_t) \\ &= -\frac{1}{\sigma_t}(\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)).\end{aligned} \tag{6}$$

Along this line, the corresponding diffusion score in Equation 5 can be derived as:

$$\hat{\epsilon}_\theta(x_t.c, t) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)), \tag{7}$$

where γ is also usually set as a global scalar parameter to control the guidance degree of the condition. However, in this paper, we argue that the CFG scale should be spatially adaptive, allowing for balancing the inconsistency of semantic strengths for diverse semantic units in the image.

4. Methods

In this section, we introduce the technical details of Semantic-aware Classifier-Free Guidance (S-CFG). where the overview of the framework is shown in Figure 2. At each denoising step in diffusion models, the current latent image is fed into the U-net backbone to estimate both diffusion score and conditional diffusion score without or with text prompt input. With the extracted attention maps, we can derive region masks for the relatively independent semantic units. In particular, the cross-attention map is renormalized for assigning each patch to the corresponding token, while the self-attention map is used to complete the

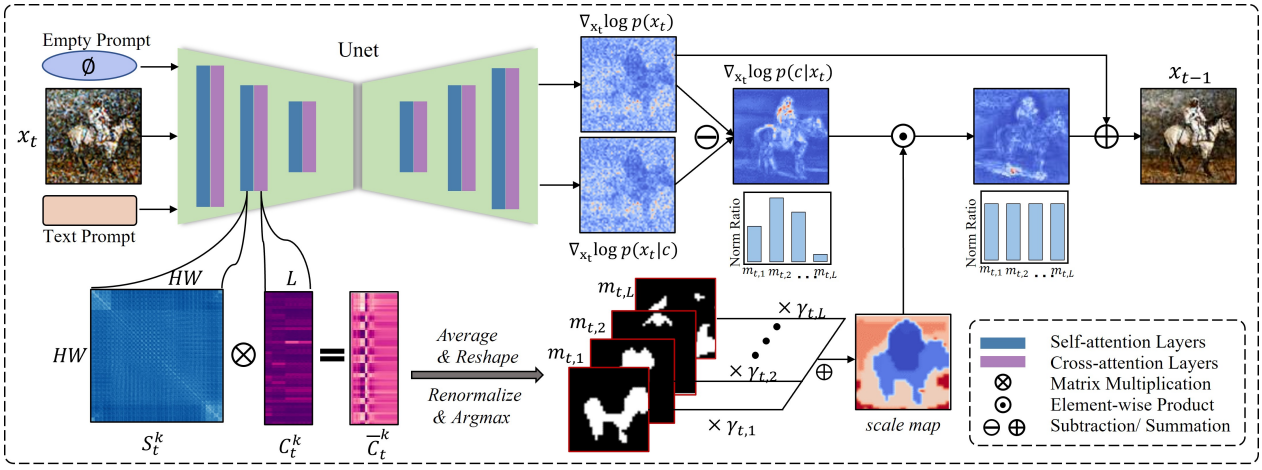


Figure 2. **The overall framework of our S-CFG method.** At each denoising step in diffusion models, the U-net backbone estimates both diffusion score $\nabla_{x_t} \log p(x_t)$ and conditional diffusion score $\nabla_{x_t} \log p(x_t|c)$ without or with text prompt input, which can further infer the classifier score $\nabla_{x_t} \log p(c|x_t)$. By extracting and exploiting self-attention map S_t^k and cross-attention map C_t^k in each attention layer of U-net, we can obtain the region masks $m_{t,i}$ for each prompt token i. With the goal of unifying the classifier score norm in different regions, the CFG scale map can be determined to control the semantic strengths spatially in the following step.

semantic regions. Then, to balance the amplification of diverse semantic information, we set adaptive CFG scales on diverse region masks and obtain the scale map to rescale their classifier scores into a uniform level.

4.1. Segmantic Map Generation

To customizedly control the amplification of diverse semantic units, we need to segment the latent image once using the CFG strategy defined in Equation 7, i.e., at each denoising step. However, this task is not trivial because the final image can not be accessed during the generation process. Fortunately, the attention layers in the U-net backbone have been reported to contain valuable semantic information for capturing relationships between image and text prompts [4, 41], which can be leveraged to efficiently extract semantic units.

Specifically, for most text-to-image diffusion models, the interaction between the text prompt and the generation image is performed using cross-attention mechanisms. In general, the denoising U-net network consists of self-attention layers followed by cross-attention layers at certain resolutions. For example, SD puts 16 self- and cross-attention layers at the resolution of 64, 32, 16, 8. In the k-th attention layer, a self-attention map $S_t^k \in \mathbb{R}^{HW \times HW}$ and a cross-attention map $C_t^k \in \mathbb{R}^{HW \times L}$ are calculated over linear projections of the intermediate image spatial feature $z_t^k \in \mathbb{R}^{HW \times C}$ or text embedding $e \in \mathbb{R}^{L \times D}$,

$$\begin{aligned} S_t^k &= \text{Softmax}\left(\frac{Q_s(z_t^k)K_s(z_t^k)^T}{\sqrt{d}}\right), \\ C_t^k &= \text{Softmax}\left(\frac{Q_c(z_t^k)K_c(e)^T}{\sqrt{d}}\right), \end{aligned} \tag{8}$$

where H and W are the current resolutions, L is the number of text tokens, C is the image feature channel, D is the token embedding dimension, and $Q_*(\cdot)$ and $K_*(\cdot)$ are linear projections with the dimension of output as d.

4.1.1 Cross-Attention-based Semantic Segmentation

Intuitively, at each denoising step t, each row in C_t^k defines the distribution over the text tokens, which is used to augment with the most relevant textual token for each patch. Therefore, a higher probability $C_t^k[s,i]$ indicates a closer relationship between the current patch s and the corresponding token w_i. Along this line, we propose to segment the latent image x_t as the set of regions masked by $\{m_{t,1}, ..., m_{t,L}\}$, which i-th masked region $m_{t,i} \in \{0,1\}^{HW}$ corresponds to the semantic token w_i.

Specifically, we first employ a fusion process to obtain the final cross-attention map $C_t \in \mathbb{R}^{HW \times L}$. This fusion involves averaging the cross-attention layers and heads with the smallest two resolutions, as these have been shown to contain the most substantial semantic information [10]. In particular, all attention maps are upsampled into the same size. Then, C_t is renormalized along the spatial dimension, and the argmax operation is applied on the token dimension to determine the activation of the current patch, denoted as:

$$\begin{aligned} \hat{C}_t[s,i] &= \frac{C_t[s,i]}{\sum_{s'=1}^{HW} C_t[s',i]}, \\ i_s &= \arg\max_i \hat{C}_t[s,i], \end{aligned} \tag{9}$$

where $\hat{C}_t[s,i]$ estimates the possibility assigned to the patch s for the token w_i. The corresponding region mask $m_{t,i}$ can be derived by setting the element in the patch set $\{s : i_s = i\}$ as 1, and 0 for others. Note that the renormalization in the above equation plays a crucial role in aligning the token with the image patch in our practice. Without the

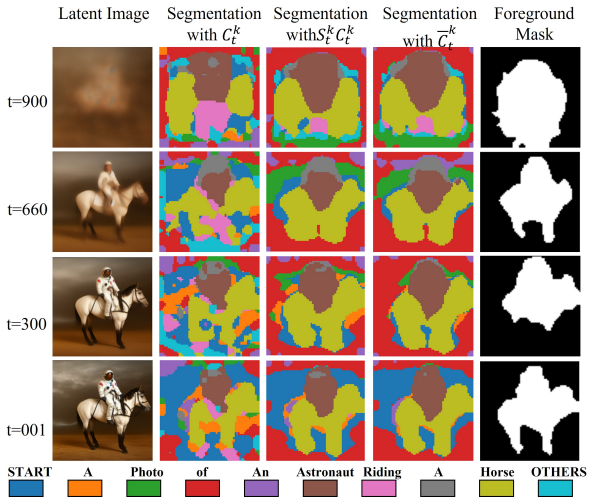


Figure 3. **The latent image segmentation based on attention maps at different denoising steps.** The first column shows the predicted image x_0 based on the current latent image x_t and noise estimation ϵ_θ with Equation 2. The following three columns show the semantic segmentation maps with different strategies. Regions labeled by different colors correspond to different tokens. The last column shows the foreground mask detected by our approach.

renormalization, C_t would tend to concentrate most of the attention on a single token, such as the START token, for all patches, damaging the semantic segmentation.

The second column in Figure 3 shows an example result of the above semantic segmentation, we can find that the semantic maps could successfully detect the rough locations of several important tokens, such as "astronaut" and "horse". However, it is worth noting that they often exhibit unclear object boundaries and may contain internal holes, particularly during the initial denoising steps. To alleviate this problem, we propose to refine and complete the semantic map with self-attention maps in the following section.

4.1.2 Self-Attention-based Segmentation Completion

Specifically, we follow [41] and refine each cross-attention map C_t^k by multiplying it with the corresponding self-attention maps at each attention layer. The hidden logic is rooted in the ability of self-attention maps to estimate the correlation between patches, enabling cross-attention to compensate for incomplete activation regions and perform region completion. Meanwhile, note that S_t^k can be interpreted as a transition matrix among all patches, where each element is nonnegative and the sum of each row equals 1. We can also enhance the region completion by transmitting semantic information among patches following the idea of feature propagation in graph [19]. Therefore, same as [51], we refine the cross-attention map C_t^k as follows:

$$\overline{C}_t^k = \frac{1}{R}\sum_{r=1}^{R}(S_t^k)^r C_t^k, \tag{10}$$

where R is a hyper-parameter and set as 4 in our experiments. Combining Eqaution 10, a refined version of cross-attention map, i..e, $\overline{C}_t$, would be computed, which would be put into Equation 9 for deriving refined segmentation masks. The fourth column in Figure 3 shows the corresponding results, where segmentation maps become better with clearer object boundaries and fewer internal holes, even better than the third column which sets $R = 1$.

4.2. Semantic-Aware CFG

At each denoising step t, given the semantic units with masks $\{m_{t,1}, ..., m_{t,M}\}$, we turn to design the semantic-aware CFG strategy to control the strength of each semantic unit separately. In particular, note that the image patches in the different semantic units usually have a more distant relationship than that among the same semantic unit. To simplify the discussion, we assume that *different semantic units are independent of each other at any time step*. Based on this assumption, we can derive the following expressions about the classifier $p(c|x_t)$:

$$\begin{aligned} p(c|x_t) &= \prod_{i=1}^{L} p(w_i|m_{t,i} \odot x_t), \\ \nabla_{x_t} \log p(w_i|m_{t,i} \odot x_t) &= m_{t,i} \odot \nabla_{x_t} \log p(c|x_t), \end{aligned} \tag{11}$$

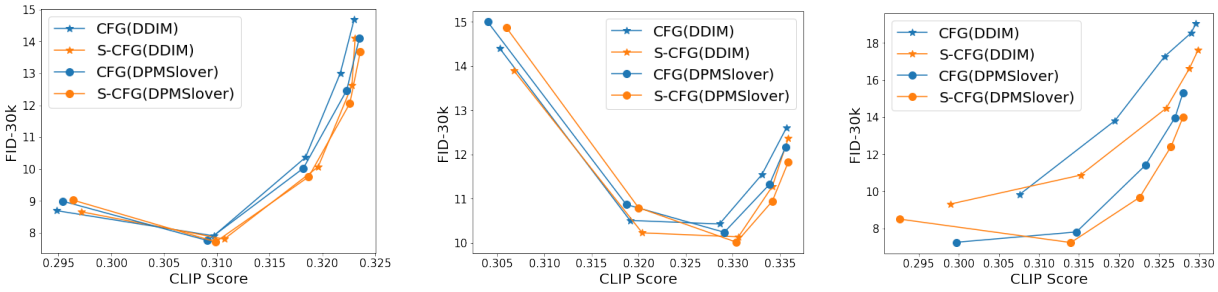
where $m_{t,i}$ is interpolated and reshaped to the same size as x_t and $\odot$ is the element-wise product. (The detailed derivation can be found in the Appendix.) Then, instead of using a single scalar to control the guidance degrees of all semantic units, like that in Equation 5 and 7, we define the composed diffusion score function as follows:

$$\begin{aligned} \hat{\epsilon}_\theta(x_t, c, t) &= \epsilon_\theta(x_t, t) \\ &+ \sum_{i=1}^{M} \gamma_{t,i} m_{t,i} \odot (\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)), \end{aligned} \tag{12}$$

where each term in the sum operation is the estimation of log-density for each semantic token w_i, and $\gamma_{t,i}$ is the scalar parameter to strengthen the corresponding semantic information. In particular, when all parameter $\gamma_{t,i}$ is set as the same as γ, the above equation reduces into the same as the original CFG strategy in Equation 7.

4.2.1 Adaptive CFG Scale $\gamma_{t,i}$

Here, we further propose an approach to adaptively set the CFG scale $\gamma_{t,i}$. The primary objective is to achieve a balanced amplification of diverse semantic units during each denoising step. To achieve this, an intuitive idea is to rescale the classifier scores in different semantic regions to a benchmark scale. This ensures that all semantic units undergo a comparable magnitude of change throughout the denoising



(a) SD-v1.5 (b) SD-v2.1 (c) DeepFloyd IF

Figure 4. **The qualitative evaluation results on the trade-off curve of FID-30K VS CLIP Score.**

process. Specifically, $\gamma_{t,i}$ is defined as follows:

$$\eta_t = \|\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)\|_2 \in \mathbb{R}^{HW},$$
$$\gamma_{t,i} = \gamma \frac{|m_{t,b} \odot \eta_t|}{|m_{t,i} \odot \eta_t|} \frac{|m_{t,i}|}{|m_{t,b}|}, \tag{13}$$

where $\|\cdot\|_2$ is the 2-norm operator of vectors used on the last dimension of a tensor, and $|\cdot|$ is the sum operator of a vector or matrix. γ is a hyper-parameter shared for all samples and time steps, like that in the original CFG strategy. In particular, the mask $m_{t,b} \in \{0,1\}^{HW}$ is introduced to assign the benchmarking region. For example, when setting $m_{t,b}$ as 1 for any patch, the average patch norm of the current latent image is the benchmark scale. Here we also introduce another benchmark region for better performance, i.e., the foreground region, such as the union of the regions of "astronaut" and "horse" in Figure 1.

Specifically, when estimating the unconditional score $\nabla_{x_t} \log p(x_t)$, an empty prompt $\emptyset$ is fed into the model, i.e, $\epsilon_\theta(x_t, \emptyset, t)$, where $\emptyset$ is usually represented as a list of padding tokens with a start token. Based on our approach in Section 4.1, we can detect the semantic region of the START token $m_{t,\text{START}}$, which effectively indicates the background area in our implementation (see the last column in Figure 3). Therefore, we can align the benchmarking region with the foreground region by setting:

$$m_{t,b} = 1 - m_{t,\text{START}}. \tag{14}$$

5. Experiments

Benchmark Models. We include two diffusion models as base models: Stable diffusion (SD) [33], which operates in the latent image space, and DeepFloyd IF (IF) [38], which operates in the image pixel space. Specifically, we consider two versions of SD: SD-v1.5 and SD-v2.1, which differ in terms of model sizes and generative qualities. For the IF model, we use the middle-scale version, IF-M, which is constructed using multiple diffusion models. To maintain simplicity, two model stages are used, where the base diffusion model produces low-resolution samples and an upscale diffusion model boosts them to a higher resolution. Both stages can benefit from the CFG or S-CFG strategy. Additionally, the IF model uses the T5XXL as the text encoder without using the start token. Therefore, instead of assigning the foreground region based on the start token, we set the benchmarking mask $m_{t,b}$ in Equation 13 as 1 for any patch. All three models are publicly accessible.

Meanwhile, two samplers are discussed for all three models, i.e., DDIM [41] and DPMSolver++ [23], which are both the most widely used in practice. Specifically, for DDIM, we follow [33] and set the number of sampling steps as 250 for SD models with the noise variance parameter as 0. Regarding the IF model, which employs learnable noise variance parameters, we adhere to the original noise settings and conduct DDIM sampling with 50 steps. As for DPM-Solver++, we set the number of sampling steps as 50.

5.1. Quantitative Evaluation

We compare the benchmark models with CFG and S-CFG on the MSCOCO 256× 256 dataset. Two qualitative metrics are used: 1) FID-30K: zero-shot Frechet Inception Distance with 30K images and the corresponding captions, which measures the quality and diversity of images. 2) CLIP Score [27]: which randomly selects 5K captions as prompts and uses the CLIP model to assess the alignments between the generated images and their corresponding text prompts. In particular, the trade-off between FID and CLIP scores has been widely reported with varying CFG scales [25]. Therefore, we present the trade-off curve across a range of the global scale $\gamma \in [2.0,\ 3.0,\ 5.0,\ 7.5,\ 10.0]$.

Based on the results presented in Figure 4, it is evident that our S-CFG strategy consistently outperforms the original CFG strategy across most experimental settings, where the trade-off curve of S-CFG consistently favors a position towards the bottom right of that of the original CFG strategy in each setting (See Appendix for a full detailed table). This phenomenon demonstrates the effectiveness and robustness of S-CFG, establishing its applicability in both latent image space and pixel space for diffusion models with different model sizes. In addition, we can find that the diffusion sampler may be crucial for the generative quality, specifi-

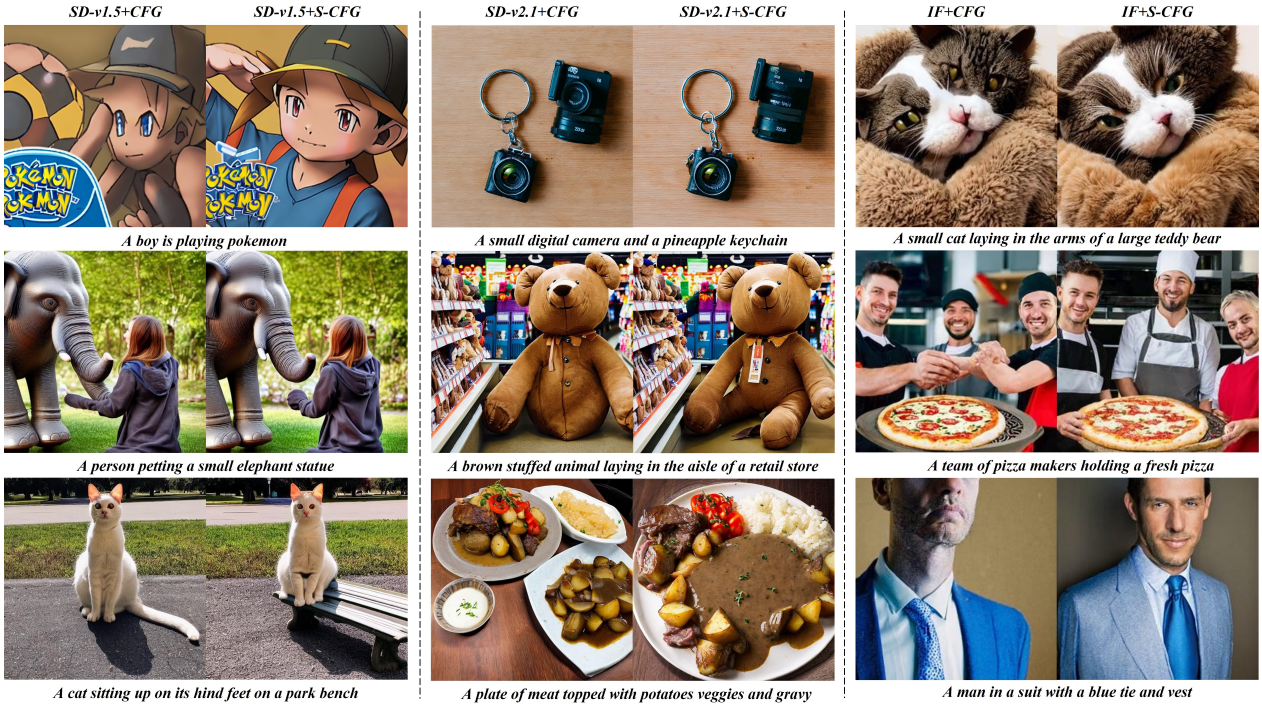


Figure 5. **Samples generated by different base models with CFG (left) or S-CFG (right).**

Table 1. **Human-level evaluation results.**

	Image Quality		Image-Text	
	CFG	S-CFG	CFG	S-CFG
SD-v1.5	26.78%	**73.22%**	23.20%	**76.80%**
SD-v2.1	28.16%	**71.84 %**	31.85%	**68.15%**
IF	32.39%	**67.61%**	29.17%	**70.83%**

cally for the pixel space model, i.e., IF, where a significant performance gap is observed for DDIM and DPMSolver++. However, S-CFG also achieve performance improvement.

5.2. Human-Level Evaluation

Here, 80 prompts are randomly selected from MSCOCO validation dataset for generative images with CFG and S-CFG. Then, we asked 5 participants to assess both the image quality and image-text alignment. Human raters are asked to select the superior respectively from the given two synthesized images, one from the original CFG strategy, and another from our S-CFG strategy. For fairness, we use the same random seed for generating both images. The voting results are summarised in Table 1. The majority of votes go to our S-CFG strategy for all base models, demonstrating superiority in both evaluated aspects.

5.3. Qualitative Evaluation

In Figure 5, we show some samples generated by different models with CFG and S-CFG. For fairness, we use the same setting and random seed for different strategies. The results exhibit a notable enhancement in the model's generative capacity from the aspects of semantic expressiveness and entity portrayal. For example, when given the prompt "A boy is playing Pokemon", S-CFG improves SD-v1.5 by ensuring the boy's appearance in a normal manner. In the case of "A person petting a small elephant statue", S-CFG eliminates the irregular elephant's trunk. Similar improvement in fine-grained structure completion can also be observed for SD-v2.1 and IF in the first two rows. Furthermore, for scenarios in the last rows, such as "A cat sitting ... on a park bench", "A plate of meat topped ..." and "A man in a suit with a blue tie ...", S-CFG helps models generate images that accurately represent the semantic descriptions.

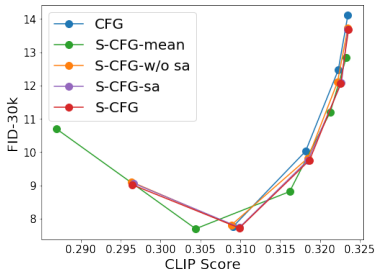


Figure 6. **The ablation analysis by evaluating the performance of different components in S-CFG.**

5.4. Ablation Analysis

Here, three variants of S-CFG are introduced: 1) S-CFG-mean sets the benchmarking mask $m_{t,b}$ as 1 for all patches. 2) S-CFG w/o sa is the variant without the segmentation completion based on self-attention maps. 3) S-CFG-sa is the variant with $R = 1$ in Equation 10.

Table 2. **Performance comparisons of ControlNet with CFG and S-CFG**, where the base model is SD-v1.5, the parameter $\gamma = 3.0$ and that sampler is DPMSolver++ with 50 steps.

	FID		CLIP Score	
	CFG	S-CFG	CFG	S-CFG
Canny	8.670	**8.382**	0.3006	**0.3019**
Segmentation	9.595	**9.549**	0.3004	**0.3017**

The results in Figure 6 based on SD-v1.5 demonstrate that all variants of S-CFG consistently outperform the original CFG strategy. This observation strongly supports our core idea of setting customized CFG scales for different semantic regions throughout the denoising process. In addition, when compared to other variants, S-CFG-mean exhibits increased performance instability and fails to achieve the optimal CLIP Score at the lowest FID score. It verifies the advantage of using the foreground region described in Equation 14 as the benchmarking region. Meanwhile, S-CFG w/o sa falls short in outperforming S-CFG-sa and S-CFG, albeit by a relatively small margin. This outcome highlights the effectiveness of self-attention-based segmentation completion. Furthermore, while S-CFG-sa and S-CFG demonstrate similar performance levels, Figure 3 shows that S-CFG exhibits superior segmentation capability, which should result in more accurate image generation. However, these improvements may not be fully captured by the current evaluation metrics.

5.5. Downstream tasks

Here, we extend the evaluations from foundational image generation to more specialized downstream tasks.

First, we incorporate S-CFG into ControlNet [47], which is a neural network architecture for adding various spatial conditioning controls to text-to-image diffusion models. Specifically, we utilize SD-v1.5 as the base model, incorporating image canny edge and image segmentation as the spatial conditions. Table 2 presents a performance comparison between CFG and S-CFG. The results demonstrate consistent improvement with the incorporation of S-CFG. Some examples are illustrated in Figure 7, showcasing notable improvements in image realism. Specifically, in the canny case of the duck toy, S-CFG enhances the structure of the duck's mouth and rectifies color imbalances around the tail. Likewise, in the segmentation case of the house, the ControlNet with CFG fails to synthesize the background sky, whereas S-CFG successfully addresses this issue.

We have also integrated S-CFG into DreamBooth [35], which enables the personalization of text-to-image diffusion models with specific subjects using only a few subject images. The examples presented in Figure 8 highlight the improvements in image quality and text-image alignment achieved by S-CFG. For instance, S-CFG enhances the appearance of the dog's mouth and brings the length of the toy's legs closer to the input images. Notably, in the second row, DreamBooth with CFG fails to align the image with the text prompt "river", whereas S-CFG succeeds.

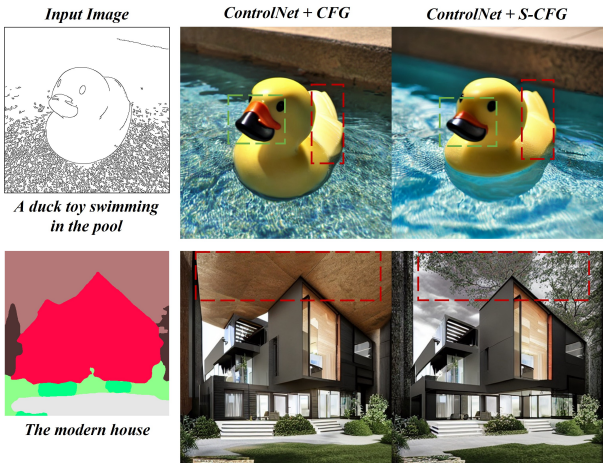


Figure 7. **Samples generated by ControlNet with CFG (middle) or S-CFG (right).**

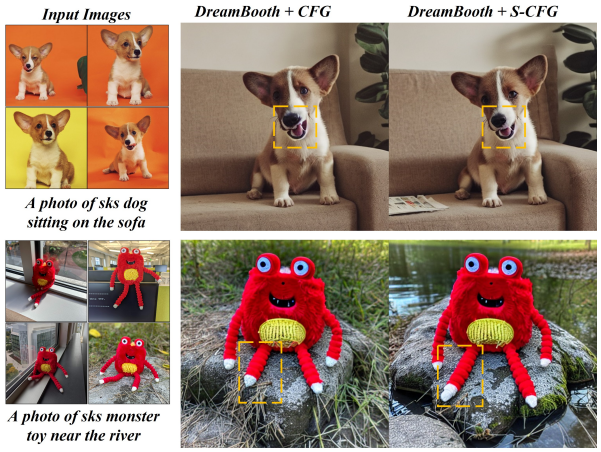


Figure 8. **Samples generated by DreamBooth with CFG (middle) or S-CFG (right).** The token "sks" represents the shared subject among the input images.

6. Conclusion

This paper argues that classifier-free guidance (CFG) in text-to-image diffusion models suffers from spatial inconsistency in semantic strengths and suboptimal image quality. To this end, we proposed Semantic-aware CFG (S-CFG), customizing the guidance degrees for different semantic units. Specifically, we first design a training-free semantic segmentation method to partition the latent image into relatively independent semantic regions at each denoising step. Then, the CFG scales across regions are adaptively adjusted to rescale the classifier scores into a uniform level. Experiments on multiple diffusion models demonstrated the superiority of S-CFG.

7. Acknowledgments

This research was supported by grants from the National Key R&D Program of China (No. 2022ZD0119302).

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 2

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. Improving image generation with better captions. *openai.com*, 2023. 1, 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 1, 2

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3, 4

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 3

[6] Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183, 2023. 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2

[9] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. *arXiv preprint arXiv:2312.10113*, 2023. 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 4

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 2, 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[13] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[14] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2

[21] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 2

[22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 2

[23] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2, 6

[24] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 6

[26] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[27] 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. 2, 6

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[31] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 1

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 6

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 1, 2, 3

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 8

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[37] Dazhong Shen, Chuan Qin, Chao Wang, Hengshu Zhu, Enhong Chen, and Hui Xiong. Regularizing variational autoencoder with diversity and uncertainty awareness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2964–2970. International Joint Conferences on Artificial Intelligence Organization, 2021. Main Track. 2

[38] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd if, 2023. https://www.deepfloyd.ai/deepfloyd-if. 6

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[41] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 4, 5, 6

[42] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *arXiv preprint arXiv:2309.15664*, 2023. 3

[43] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023. 3

[44] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3

[45] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 1

[46] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 3

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 8

[48] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2

[49] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022. 2

[50] Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023. 3

[51] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International conference on learning representations*, 2020. 5