Segmentation-Free Guidance for Text-to-Image Diffusion Models

Kambiz Azarian, Debasmit Das, Qiqi Hou, Fatih Porikli
Qualcomm AI Research*
{kambiza, debadas, qhou, fporikli}@qti.qualcomm.com

Figure 1. Segmentation-free vs. classifier-free guidance. Same amount of computations. Best viewed on a computer. Prompts: (a) "a cute Maltese white dog next to a cat," (b) "ultra realistic, predator, male, fangs, goth, tattoos, leather, fantasy, flesh, bone, body horror, intricate details, eerie, highly detailed, octane render, 8 k, art by artgerm and alphonse mucha and greg rutkowski," (c) "architectural drawing of a new town square for Cambridge England, big traditional museum with columns, fountain in middle, classical design, traditional design, trees," (d) "portrait of a kid," (e) "a beautiful ultradetailed painting of urbex building abandoned, nature, city, unfinished building architecture by april gornik, stormy darkacademia, archdaily, wallpaper, highly detailed, trending on artstation," and (f) "a girl hugging a Corgi on a pedestal."

Abstract

We introduce segmentation-free guidance, a novel method designed for text-to-image diffusion models like Stable Diffusion. Our method does not require retraining of the diffusion model. At no additional compute cost, it uses the diffusion model itself as an implied *segmentation network, hence named segmentation-free guidance, to dynamically adjust the negative prompt for each patch of the generated image, based on the patch's relevance to concepts in the prompt. We evaluate segmentation-free guidance both objectively, using FID, CLIP, IS, and PickScore, and subjectively, through human evaluators. For the subjective evaluation, we also propose a methodology for subsampling the prompts in a dataset like MS COCO-30K to keep the number of human evaluations manageable while ensuring that the selected subset is both representative in terms of content and fair in terms of model performance. The results demonstrate the superiority of our segmentation-free guidance to the widely used classifier-free method. Human evaluators preferred segmentation-free guidance over classifier-free* 60% *to* 19%*, with* 18% *of occasions showing a strong preference. Additionally, PickScore win-rate, a recently proposed metric mimicking human preference, also indicates a preference for our method over classifier-free.*

1. Introduction

Diffusion models are powerful generative models for creating visual content from textual prompts. Their success

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

stems from extensive training data and their ability to handle various modalities and signals, enabling diverse applications such as content editing, inpainting, and personalization.

Controlling a diffusion model can be achieved primarily in two ways - *conditioning* and *guidance*. When a diffusion model is conditioned, it is typically trained to accept a particular form of additional conditioning input, such as a text prompt, image edges, segmentation map, and class labels. However, adapting the model to a different condition often necessitates retraining from scratch. This reliance on expensive retraining poses challenges for end-users seeking to adopt and employ conditioning techniques to control diffusion models.

An alternative way to control a diffusion model is through a guidance mechanism. Unlike conditioning techniques, this approach does not rely on an external conditioning signal. Instead, it associates a guidance function with the diffusion model to fulfill a specific target criterion, which could be as simple as minimizing the CLIP distance between the generated image and the provided text description. When sampling an image, the reverse process iterations are steered in the direction of the guidance function's gradient, resulting in constrained image generation.

When comparing control techniques for diffusion models, guidance emerges as a more versatile approach. It treats the diffusion network as a foundational model which can accommodate different use cases. An earlier method in this domain involved classifier guidance [8], where an explicit classifier functioned as the guidance mechanism. This method utilized the classifier's gradients to drive the image generation process. However, classifier guidance has transitioned to classifier-free guidance [12], eliminating the need for an explicit classifier. In classifier-free guidance approaches, the network is trained to adapt class-label information and conditioning signals without relying on a fixed network architecture.

In this paper, we propose enhancing image generation quality beyond classifier-free guidance by introducing a novel and universal segmentation-free guidance approach. This methodology aims to improve image quality of diffusion models without necessitating costly retraining, architectural changes, or additional computing during inference.

Image generation using classifier-free guidance involves two forward passes of the diffusion network per iteration: one that uses conditional information and one that does not. The conditional information generally involves a (positive) text prompt describing different objects of interest in the generated image. For instance, in Fig. 1-a, the positive prompt is "a cute Maltese white dog next to a cat." The forward pass without conditional information is usually carried out by an empty (negative) prompt (i.e., ""). However, it is possible to employ non-empty negative prompts. This type of guidance allows the objects present in the positive, but not the negative, prompt to become more prominent. Nonetheless, the issue with having such negative prompts is that it interacts with the generated image globally.

Our objective is to dynamically adjust the negative prompt for each image patch. We examine attention maps within the diffusion model, specifically where it interacts with the text prompt embeddings. For each patch of the attention map, we aim to find the object in the positive prompt with the highest correlation. Subsequently, this selected object is excluded from the negative prompt interacting with that specific patch. Accordingly, the forward pass of the diffusion model carries out as if each patch cross-attends dynamically with a different negative prompt. Furthermore, the corresponding attention weight is adjusted to account for self-attention interactions. Since this proposed method of guidance does not involve any segmentation network as a guidance function, we term this method as *segmentation-free guidance*. Our method realizes local interaction between prompt embedding and feature patches while dynamically adjusting the negative prompts; thus, it produces better image generation quality, as shown in Fig. 1.

Our contributions can be summarized as follows:

- We introduce a novel mechanism named segmentation-free guidance that effectively adjusts the negative prompt for each patch of the generated image based on the category of the patch.
- We also propose an efficient subjective evaluation methodology that involves sub-sampling of prompts dataset for assessment. The chosen subset of prompts ensures the representation of dataset diversity while maintaining fairness in terms of model performance.
- Finally, we perform extensive evaluation on the MS-COCO datasets on which we show both qualitative and quantitative improvement.

2. Related Work

Our proposed work falls into the scope of controlled image generation using diffusion models. Controlled image generation can be broadly classified into conditional generation and guided generation. These are discussed as follows.

Conditional Generation These category of works generally require training diffusion models from scratch where conditional input can be of the form of prompts [2, 12, 19, 27, 29]. One of the most popular works [12] proposed use of classifier-free guidance with class labels as prompts. In this work, the diffusion model is trained such that the output is a linear combination between that of conditional and unconditional outputs. The authors of [2] trained a diffusion model, where it is enforced to solve linear inverse problems. This is realized

through a guidance function known as linear degradation operator. [19] used classifier-free guidance but extended it to descriptive phrases as prompts. Furthermore, the network was trained to enforce similarity between CLIP [21] representations of images and text. However, the major disadvantage of conditional generation methods is that the diffusion models need to be retrained and hence it is computationally intensive.

Guided Generation In this category, the diffusion model is kept frozen without any re-training. However, the sampling process for image generation is modified using gradients from a guidance function. There are prior works that studied guided image generation using various constraints and guidance functions [6–9, 14, 18, 28]. The most popular method in this category is classifier guidance [8]. In this method, a classifier is trained to distinguish images of different scales. The classifier is used as a guidance function, the gradients of which are used in the sampling process. Alternative methods include [28], where the guidance function is a linear operator. Since gradients of the linear operator are used, components of the images were generated in the null space of the linear operator. However, the use of null space does not naturally extend to non-linear guidance functions. In [6], the authors did an elaborate analyses of multiple simple non-linear guidance functions, e.g. non-linear blurring. The gradient of the non-linear function was calculated on expected denoised images and the sampling process was modified. Recently, [3] proposed a training-free universal guidance mechanism that can use guidance in the form of CLIP, segmentation map, face recognition, object location, style guidance to produce more controlled image generation.

In this paper, we consider a segmentation-free guidance mechanism, which does not require training from scratch. Furthermore, it does not require extra computation during image sampling compared to classifier-free guidance. Our method modulates the cross-attention weights pertaining to different categories in the prompt, which enhances the visual quality of the generated image.

3. Background

Gaussian diffusion models [13, 25, 26] are powerful generative methods for sampling $\mathbf{x}$, e.g., an image, according to $p(\mathbf{x})$, e.g., a dataset. They involve a T-step forward diffusion process that creates a Markov chain of ever more noisy latent representations

$$\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

with $\alpha_t^2 = 1/(1+e^{-\lambda_t})$, $\sigma_t^2 = 1-\alpha_t^2$, $\lambda_1 > \cdots > \lambda_T$ representing a log-SNR schedule, and a T-step reverse denoising process that starts by sampling Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and proceeds by sampling $\mathbf{z}_{t-1}$ according to

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mu_\theta(\mathbf{z}_t), \Sigma_t),$$

where $\mu_\theta(\mathbf{z}_t)$ is a function of diffusion model's output $\epsilon_\theta(\mathbf{z}_t)$ [12], which estimates ϵ in (1) by training on

$$\mathbb{E}_{\epsilon,t}[\|\epsilon_\theta(\mathbf{z}_t) - \epsilon\|_2^2].$$

ϵ and ϵ_θ are called the true and estimated scores, respectively.

Generative models have successfully been used in many applications, including text-to-image, where it is desired to generate an image consistent with a given prompt $\mathbf{c}$. Generating high quality images, however, requires using guidance methods [8]. Classifier-free guidance [12] is one such method which is inspired by an *implicit* classifier

$$\log p^i(\mathbf{c}|\mathbf{z}_t) = \log p(\mathbf{z}_t|\mathbf{c}) - \log p(\mathbf{z}_t) + \text{const.}, \tag{2}$$

with a gradient

$$\nabla_{\mathbf{z}_t} \log p^i(\mathbf{c}|\mathbf{z}_t) \propto -[\epsilon_\theta(\mathbf{z}_t, \mathbf{c}) - \epsilon_\theta(\mathbf{z}_t)]. \tag{3}$$

In classifier-free guidance, during sampling, diffusion model's output $\epsilon_\theta(\mathbf{z}_t, \mathbf{c})$ is steered in the direction of (3) to increase the implicit classifier's log likelihood of (2)

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) = (1+w)\epsilon_\theta(\mathbf{z}_t, \mathbf{c}) - w\epsilon_\theta(\mathbf{z}_t), \tag{4}$$

where w is the guidance strength. Note that $\epsilon_\theta(\mathbf{z}_t)$ is computed by applying an empty second, a.k.a., negative, prompt, i.e., $\epsilon_\theta(\mathbf{z}_t) = \epsilon_\theta(\mathbf{z}_t, \varnothing)$.

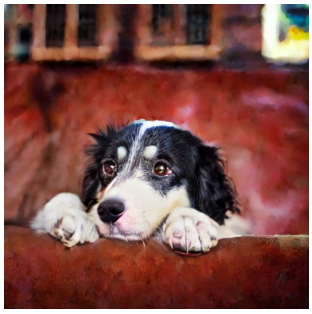
4. Segmentation-Free Guidance

To illustrate the motivation behind our method, we consider an example. All images in Figure 2 were generated using classifier-free guidance from the same seed and positive prompt, "A dog on a couch in an office." However, they differ in the negative prompts used. Figure 2a employs an empty prompt, while Figures 2b, 2c, and 2d omit the words "dog", "couch", and "office" from the negative prompt, respectively. As evident from the images, the regions corresponding to the omitted concepts exhibit enhanced detail. This raises the question: Can we enhance classifier-free guidance by dynamically adjusting the negative prompt for each patch of the generated image based on its semantic content? Segmentation-free guidance represents one such approach.

Next, we describe our segmentation-free guidance, explaining the motivation behind its different components. Let $\mathbf{c}_i$ denote the i-th CLIP embedding of prompt and $\mathbf{z}_p$ the p-th patch of $\mathbf{z}_t$. Let us define $\mathbf{z}_p$'s semantic as

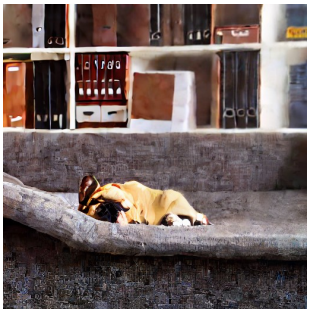
$$\mathbf{s}_p \triangleq \underset{\mathbf{c}_i}{\text{argmax}} \log p^i(\mathbf{c}_i|\mathbf{z}_p, \mathbf{c} - \{\mathbf{c}_i\}), \tag{5}$$

(a) Empty negative prompt, i.e., NP: ””



(b) NP: “A couch in an office” Dog becomes more prominent.

(c) NP: “A dog in an office” Couch becomes more prominent.



(d) NP: “A dog on a couch” Office becomes more prominent.

Figure 2. All images are generated using classifier-free guidance from the same seed and positive prompt, “A dog on a couch in an office,” but with different negative prompts (NP). As can be seen, the regions corresponding to the omitted concepts improve in detail.

where $\mathbf{c}-\{\mathbf{c}_i\}$ denotes prompt embeddings with $\mathbf{c}_i$ omitted. The Bayes rule can be used to write (compare to (2))

$$\log p^i(\mathbf{s}_p|\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_p\}) = \log p(\mathbf{z}_p|\mathbf{c}) - \log p(\mathbf{z}_p|\mathbf{c}-\{\mathbf{s}_p\}) + \text{const.} \quad (6)$$

Segmentation-free guidance enhances $\mathbf{z}_p$'s detail by steering diffusion model's output $\epsilon_\theta(\mathbf{z}_p, \mathbf{c})$ during sampling, in the direction that increases $\mathbf{s}_p$'s log-likelihood

$$\nabla_{\mathbf{z}_p} \log p^i(\mathbf{s}_p|\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_p\}) \propto -[\epsilon_\theta(\mathbf{z}_p, \mathbf{c}) - \epsilon_\theta(\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_p\})]. \quad (7)$$

In (7), diffusion model's output $\epsilon_\theta(\mathbf{z}_p, \mathbf{c})$ has been used to estimate the true score $\epsilon^*(\mathbf{z}_p, \mathbf{c})$. Since $\mathbf{s}_p$ is not known a priori, $\epsilon_\theta(\mathbf{z}_p, \mathbf{c}-\{\mathbf{c}_i\})$ has to be computed for all i, which makes (7) computationally prohibitive.

To overcome this limitation, we modify (7) to use $\mathbf{z}_p$'s local (i.e., layer) semantic

$$\nabla_{\mathbf{z}_p} \log p^i(\mathbf{s}_p|\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_p\}) \propto -[\epsilon_\theta(\mathbf{z}_p, \mathbf{c}) - \epsilon_\theta(\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_{pl}\}_{l=1}^L)]. \quad (8)$$

defined as (compare to (5))

$$\mathbf{s}_{pl} \triangleq \underset{\mathbf{c}_i, i>0}{\operatorname{argmax}} A(\mathbf{z}_{pl}, \mathbf{c}_i), \quad (9)$$

where l indexes diffusion model's cross-attention modules and $\mathbf{z}_{pl}$ denotes their input patch. $A(\mathbf{z}_{pl}, \mathbf{c}_i)$ denotes the computed cross-attention weight. As shown in Section 5, segmentation-free guidance offers a gain justifying use of $\mathbf{s}_{pl}$ instead of $\mathbf{s}_p$; However, as a motivation, we note that diffusion models like Stable Diffusion have an architecture similar to segmentation networks like Mask2Former [5], i.e., they are composed of a stack of transformer decoder layers. This suggests that $\mathbf{s}_{pl}$ could serve as a proxy for $\mathbf{s}_p$, though there is no guarantee of consistent semantics for a patch across layers, or for adjacent patches from the same layer. This is especially true during the first few iterations because of noise. Hence, we use classifier-free guidance for the first few iterations and switch to segmentation-free guidance only after, e.g., $t > t_s = 10$. We also note that in (9), $\mathbf{c}_0$, i.e., begin-of-sentence (BOS) embedding, has been excluded from designation as a patch's local semantic. This is because BOS's cross-attention weight is almost always significantly larger than those of others prompt tokens, while its projected value is negligible, i.e., BOS serves as a neutral text embedding.

Finally, we note that conditioning leakage, i.e., $\mathbf{s}_{pl}$ indirectly affecting $\epsilon_\theta(\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_{pl}\}_{l=1}^L)$, negatively affects guidance according to (8): First, we note that $\mathbf{c}-\{\mathbf{c}_i\}$ is different from the CLIP text encoding of a prompt that omits the i-th token. In the former the i-th token still substantially affects the subsequent embeddings $\{\mathbf{c}_j\}_{j>i}$. The second source of leakage are model's self-attention modules that allow a patch to be indirectly affected by conditioning applied to other patches. As a result of conditioning leakage, the gradients given by (8) are small and ineffective. Therefore we further modify it to

$$\nabla_{\mathbf{z}_p} \log p^i(\mathbf{s}_p|\mathbf{z}_p, \mathbf{c}-\{\mathbf{s}_p\}) \propto -[\epsilon_\theta(\mathbf{z}_p, \mathbf{c}) - \bar{\epsilon}_\theta(\mathbf{z}_p, \mathbf{c})], \quad (10)$$

where $\bar{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$ is computed identically to $\epsilon_\theta(\mathbf{z}_p, \mathbf{c})$, with the difference that in each cross-attention module, the $\mathbf{s}_{pl}$'s attention weight (c.f. (9)) is multiplied by $-a$, where a is a positive number, e.g., 10, called the segmentation-free scale. This effectively compensates for the conditioning leakage. Thus the segmentation-free modified score is

$$\tilde{\epsilon}(\mathbf{z}_t, \mathbf{c}) = (1+\bar{w})\epsilon(\mathbf{z}_t, \mathbf{c}) - \bar{w}\bar{\epsilon}(\mathbf{z}_t), \quad (11)$$

where $\bar{w}$ denotes segmentation-free guidance scale. We have found, through experiments, that a smaller value of $\bar{w} = 2.5$ (compared to $w = 7.5$) gives very good results. Note that segmentation-free guidance has almost the same computational complexity as classifier-free. Algorithm 1 gives the segmentation-free guidance method.

Algorithm 1: Segmentation-free guidance

Require: w: Classifier-free guidance strength (7.5)
Require: T: Total number of iterations (20)
Require: t_s: Classifier-free guidance iterations ($T/2$)
Require: a: Segmentation-free scale (10.0)
Require: $\bar{w}$: Segmentation-free guidance strength (2.5)
Require: $\mathbf{c}$: Prompt CLIP text embeddings
Require: $\lambda_1 > \cdots \lambda_T$: log-SNR schedule

$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \cdots, 1$ **do**
  if $t \geq T - t_s$
    # compute classifier-free score
    $\tilde{\epsilon}(\mathbf{z}_t, \mathbf{c}) = (1 + w)\epsilon(\mathbf{z}_t, \mathbf{c}) - w\epsilon(\mathbf{z}_t)$
  else
    # compute segmentation-free score
    $\tilde{\epsilon}(\mathbf{z}_t, \mathbf{c}) = (1 + \bar{w})\epsilon(\mathbf{z}_t, \mathbf{c}) - \bar{w}\bar{\epsilon}(\mathbf{z}_t)$
  # sample $\mathbf{z}_{t-1}$
  if $t > 1$
    $\mathbf{z}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{z}_t), \Sigma_t)$
  else
    $\mathbf{z}_0 = (\mathbf{z}_1 - \sigma_1\tilde{\epsilon_1})/\alpha_1$
return $\mathbf{x} = \mathbf{z}_0$

5. Experiments

In this section we report our experiments. We provide detailed analysis of a few cases in Sec. 5.1, give our quantitative results in Sec. 5.2 and finally report our qualitative results in Sec. 5.3. All images and results are generated using the Stable Diffusion v1.5 model with $T = 20$. Unless otherwise stated, classifier-free guidance strength is set to $w = 7.5$. We use a segmentation-free guidance strength of $\bar{w} = 2.5$ (c.f., Algorithm 1).

5.1. Case Studies

To provide insight into the nature of the improvement and the role of different parameters, we present a few case studies comparing images generated using the two guidance methods. We also include some examples showing segmentation-free method's limitations.

Fig. 3 shows images generated using classifier-free and segmentation-free guidance methods for the prompt "a cute Maltese white dog next to a cat". As Fig. 3a shows, under classifier-free guidance some of the dog's features, particularly its long white hair, interfere with cat's rendering, resulting in its unnatural appearance. Segmentation-free guidance improves details of the dog and the cat by adjusting their conditioning individually (3b).

Fig. 4 shows the effect of segmentation-free scale (parameter a in Algorithm 1). As evident from Fig. 4b, setting $a = 0$, which is equivalent to ignoring the most relevant prompt CLIP embedding, is ineffective. As explained in Sec. 4, this is due to conditioning leakage resulting from both, correlation among prompt text embeddings, and self-attention modules in the diffusion model. In contrast, as seen in Fig. 4d and 4e, very large values, e.g., $a > 20$, cause artifacts by over compensating for this leakage. In Sec. 5.3 we present human evaluation results showing $a = 10$ to provide the best results.

We next consider the effects of the number of classifier-free guidance iterations performed before switching to segmentation-free guidance (parameter t_s in Algorithm 1). Generally, reducing t_s improves a region's detail by customizing its conditioning earlier. However, switching too early may have some negative effects. For example, Fig. 5c shows images generated for the prompt "a girl hugging a Corgi on a pedestal". As Fig. 5b shows, segmentation-free guidance with $t_s = 10$ greatly enhances the image. However, reducing t_s to 1 in Fig. 5c causes compositional defects, i.e., the dog becomes too large. Fig. 6 shows another negative effect of using a small t_s. While segmentation-free guidance with $t_s = 10$ enhances the image in Fig. 6b, using $t_s = 5$ in Fig. 6c has caused prompt's "architectural drawing" directive to be ignored. We note, however, that using larger values of classifier-free guidance strength w generally requires using a smaller t_s, as argued next.

It is a well known fact that the performance of text-to-image diffusion models, e.g., Stable Diffusion, is very dependent on the quality of the text encoder used, e.g., CLIP [23]. For example Fig. 7 shows images generated from three different seeds for the prompt "A man in yellow shirt next to a woman in blue dress". Note that in Fig. 7b and 7c the colors are switched, while in Fig. 7a, there is no man. The first issue may either be due to "yellow shirt" coming before "woman" in the prompt, compared to "blue dress" that is coming after (CLIP is a unidirectional text encoder), or the fact that men wearing blue shirts appear more often in Stable Diffusion's training data than yellow ones. In either case segmentation-free guidance cannot rectify this issue as it relies on classifier-free guidance for the first few iterations. However, the second issue, i.e., Fig. 7a not showing a man, can be rectified by using a larger classifier-free guidance strength, i.e., $w = 12.5$ as shown in Fig. 8b. A larger w, however, suppresses unrelated noise more aggressively, which makes a moderate value of $t_s = 10$ ineffective as seen in Fig. 8c. A smaller $t_s = 5$ however provides a significant improvement, c.f., Fig. 8d. We show the noisy image resulting after applying $t_s = 5$ iterations of classifier-free guidance in Fig. 9b. This is the image that is used in the first iteration of segmentation-free guidance. As expected, the image is quite noisy, which shows the importance of relying on the diffusion model as an implicit segmentation network.

Finally we note that segmentation-free guidance improves on classifier-free method by identifying the most relevant concept from the prompt for each patch and excluding the rest from interfering. This bring us to the interesting case where morphing distinct concepts in prompt, such as

(a) Classifier-free guidance (b) Segmentation-free, $t_s = 5$

Figure 3. Effect of segmentation-free guidance. Prompt: "a cute Maltese white dog next to a cat". The long hair characteristic of the dog spills over to the cat under classifier-free guidance (3a). Segmentation-free guidance improves quality by adjusting conditioning for the dog and the cat individually (3b).

the "bat" and the "cat" in the prompt "hybrid of a bat and a cat" is actually the goal. As Fig. 10 shows, segmentation-free under-performs classifier-free guidance, and reducing t_s exacerbates the deficiency.

5.2. Quantitative Results

Evaluation metrics. We report our results on the MS-COCO dataset [17]. Following prior works [1, 22, 23], we utilize the first 30K captions from the val2014 subset for image generation (also known as MS-COCO-30K). A variety of evaluation metrics are adopted to measure the objective quality of these synthesized images, including FID [11] score, CLIP score [10], IS score [24], and PickScore [15] . We compute the FID score between the 30K generated images and 30K reference ground truth images following the official implementation, which leverages the `Inception-V3` model. In line with recent studies [1, 22, 23], we use `ViT-g-14` model for computing the CLIP score. We calculate the IS score using `Inception-V3` model. For PickScore, we follow the official implementation [16] (which uses `CLIP-ViT-H-14` model) and report the win-rate. As stated earlier, all images are generated using the Stable Diffusion v1.5 model with $T = 20$. Unless otherwise stated, classifier-free guidance strength is set to $w = 7.5$ and a segmentation-free guidance strength of $\bar{w} = 2.5$.

Tab. 1 gives the FID, CLIP and IS scores for classifier-free and segmentation-free guidance methods. As the table shows, segmentation-free guidance does not improve either FID or CLIP scores. There is increasing evidence [4, 16, 20], however, that such metrics may not reflect visual aesthetics. In fact [16], based on their large dataset of text-to-image prompts and human preferences, finds that FID is negatively correlated with subjective quality. [16] further leverages this dataset to train a scoring function, PickScore, which predicts human preferences well. Tab. 2 gives the PickScore win-rate for segmentation-free guidance against classifier-free. As the table shows, segmentation-free guidance, with scales $a = 5$ and 10, score PickScore win rates of 63.24% and 60.25% against classifier-free, respectively.

Method	FID↓	CLIP↑	IS↑
Class.-free	17.37	0.3044	37.51
Segm.-free, $a = 5$	19.47	0.2982	37.96
Segm.-free, $a = 10$	20.55	0.2961	25.59

Table 1. FID, CLIP and IS scores for classifier-free and segmentation-free guidance methods.

Method vs. Class.-free	PickScore win-rate
Segm.-free, $a = 5$	63.24%
Segm.-free, $a = 10$	60.25%

Table 2. PickScore win-rate for segmentation-free guidance against classifier-free.

5.3. Qualitative Results

To further demonstrate the performance of our segmentation-free guidance, we conduct a subjective evaluation study. We use the MS-COCO-30K validation set prompts to generate images using the two guidance methods and ask human evaluators to choose one of five options based on image quality and match to the prompts. The five options are "much better," "slightly better," "no preference," "slightly worse," and "much worse." Evaluating on the entire MS-COCO-30K dataset requires a large number of human evaluations, so we sample the dataset to form a smaller subset. This raises a few important considerations: the size of the subset, how to ensure that it adequately captures the dataset's diversity, and how to ensure that it represents the diffusion model's performance.

Next we present our sampling methodology. We evaluate the diversity of a random subset of prompts by measuring its Fréchet distance to the entire MS-COCO-30K validation set prompts. Note that computing this measure does not use any images, as it is the Fréchet distance between two distributions of prompts' CLIP text encodings. As Fig. 11 shows, a subset of 5K randomly sampled prompts adequately represents MS-COCO-30K's diversity, whereas smaller subsets, such as those with 150 samples, show a large Fréchet distance from the original set and therefore are not good representations. To further reduce the number of samples, we use the classifier-free guidance to generate images for the 5K subset and rank them based on their CLIP score. Then, we use the 90th, 50th, and 10th percentiles, representing the high-performing, middle-performing, and low-performing prompts, respectively, to form a final subset of 150 prompts for our human evaluations. These two steps allow us to sub-

(a) Classifier-free (b) Segm.-free, $a = 0$ (c) Segm.-free, $a = 10$ (d) Segm.-free, $a = 20$ (e) Segm.-free, $a = 50$

Figure 4. Effect of segmentation-free scale (a in Algorithm 1). Prompt: "portrait of a dog and a kid". $a = 0$ is ineffective due to conditioning leakage (4b), while large values $a > 20$ cause artifacts (4d, 4e). Subjective evaluations show that $a = 10$ gives the best results (4c).

(a) Classifier-free guidance (b) Segmentation-free, $t_s = 10$ (c) Segmentation-free, $t_s = 1$

Figure 5. Compositional effect of t_s. Prompt: "a girl hugging a Corgi on a pedestal". Switching too early from classifier-free to segmentation-free guidance improves local detail, but hurts the overall composition, i.e., too large a dog in (5c).

(a) Classifier-free guidance (b) Segmentation-free, $t_s = 10$ (c) Segmentation-free, $t_s = 5$

Figure 6. Effect of t_s in skipping aspects of prompt. Prompt: "architectural drawing of a new town square for Cambridge England, big traditional museum with columns, fountain in middle, classical design, traditional design, trees". Using a smaller t_s in (6c) improves the overall image quality with respect to (6b), but at the expense of ignoring the "architectural drawing" aspect of the prompt.

stantially reduce the resources needed for human evaluation while ensuring that the prompt subset is both diverse and fair. For example with 17 human evaluators each rating 30 pairs of images, we obtain an average of 1.7 evaluations per each of the 300 independent pairs from two tests.

Fig. 12 shows the subjective results of comparing segmentation-free to classifier-free guidance. Human evaluators favored segmentation-free guidance in 60% of cases (with a segmentation-free scale of 10), compared to 19% for classifier-free guidance. More significantly, human evaluators indicated a strong preference for segmentation-free guidance in 18% of cases, compared to 2% for classifier-free guidance. For visual comparisons, c.f., Fig. 1, where images for few user specified prompts have been generated.

(a) Seed 1 (b) Seed 2 (c) Seed 3

Figure 7. Effect of CLIP text encoding. All images are for the prompt "A man in yellow shirt next to a woman in blue dress" and generated from three different seeds. Note that in (7b, 7c) the colors are switched, while in (7a) there is no man. This may be due to "yellow shirt" coming before "woman" in the prompt, whereas "blue dress" comes after it (CLIP is a unidirectional text encoder), or the fact that men with blue shirts appear more often in Stable Diffusion's training data than yellow ones.

6. Conclusion

We presented segmentation-free guidance, a novel guidance method for text-to-image diffusion models like Stable Diffusion. Our method does not increase the computational load. It does not require retraining or fine-tuning. Segmentation-free guidance uses the diffusion model as an implied segmentation network to dynamically customize the negative prompt for each image patch, by focusing on the most relevant concept from the prompt. We evaluated segmentation-free guidance both objectively, using FID, CLIP, IS, and PickScore, and subjectively, through human evaluators. For the subjective evaluation, we proposed a methodology for reducing the number of prompts in a dataset like MS-COCO-30K to keep the number of human evaluations manageable while ensuring that the selected subset is both representative in terms of diversity and fair in terms of model performance. The results showed the superiority of our segmentation-free guidance to the

(a) Classifier-free, $w = 7.5$ (b) Classifier-free, $w = 12.5$ (c) Segm.-free, $t_s = 10$ (d) Segm.-free, $t_s = 5$

Figure 8. A larger w requires a smaller t_s. All images generated for the prompt "A man in yellow shirt next to a woman in blue dress". A larger $w = 12.5$ aligns 8b better with the prompt (a man appears, although colors are still switched). However the larger w means more aggressive suppression of unrelated noise, which makes a moderate $t_s = 10$ ineffective (8c). A smaller $t_s = 5$ fixes this issue in (8d).

(a) $t = 20 (= T)$ (b) $t = 5 (= t_s)$

Figure 9. Importance of using the diffusion model for implicit segmentation. (9b) is the intermediate image after $t_s = 5$ iterations of classifier-free guidance, and before applying segmentation-free guidance which results in (9a) at the end. As can be seen, the image is too noisy to be processed by any off-the-shelf network.

(a) Classifier-free guidance (b) Segmentation-free, $t_s = 10$ (c) Segmentation-free, $t_s = 5$

Figure 10. A limitation of segmentation-free guidance. Prompt: "hybrid of a bat and a cat". Segmentation-free guidance (10b) under-performs classifier-free (10a) when the goal is to morph distinct concepts into one. Reducing t_s (10c) exacerbates the deficit, as expected.

widely used classifier-free method. Human evaluators preferred segmentation-free guidance over classifier-free 60% to 19%, with 18% of occasions showing a strong preference. PickScore win-rate, a recently proposed metric mimicking human preference, indicated a preference for our method over classifier-free, too.

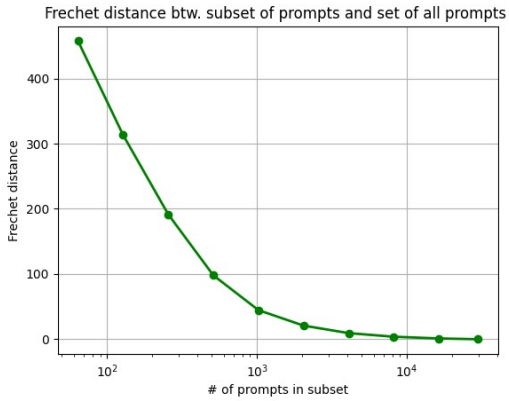


Figure 11. Fréchet distance between MS-COCO-30K valid set prompts and its subsets of various sizes. As can be seen, a subset of 5K randomly sampled prompts adequately represents MS-COCO-30K's diversity, whereas smaller subsets, e.g., 150 samples, show a large Fréchet distance and therefore are not good representations.

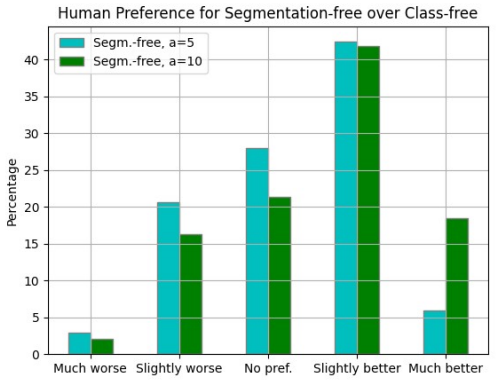


Figure 12. Preferences of human evaluators comparing segmentation-free to classifier-free guidance. Human evaluators preferred segmentation-free guidance $a = 10$ over classifier-free 60% to 19%, with 18% of occasions showing a strong preference.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 6

[2] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022. 2

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 3

[4] Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. A study on the evaluation of generative models, 2022. 6

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 4

[6] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3

[7] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

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

[9] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 3

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 3

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 3

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 6

[16] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 6

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[18] 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. 3

[19] Alexander Quinn 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. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2, 3

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 6

[21] 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. 3

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 6

[23] 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. 5, 6

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 3

[26] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019. 3

[27] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2

[28] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[29] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2