Improving image synthesis with diffusion-negative sampling

Alakh Desai[1] and Nuno Vasconcelos[1]

University of California San Diego, USA
{ahdesai,nuno}@ucsd.edu

Abstract. For image generation with diffusion models (DMs), a negative prompt $\mathbf{n}$ can be used to complement the text prompt $\mathbf{p}$, helping define properties not desired in the synthesized image. While this improves prompt adherence and image quality, finding good negative prompts is challenging. We argue that this is due to a semantic gap between humans and DMs, which makes good negative prompts for DMs appear unintuitive to humans. To bridge this gap, we propose a new *diffusion-negative prompting* (DNP) strategy. DNP is based on a new procedure to sample images that are least compliant with $\mathbf{p}$ under the distribution of the DM, denoted as *diffusion-negative sampling* (DNS). Given $\mathbf{p}$, one such image is sampled, which is then translated into natural language by the user or a captioning model, to produce the negative prompt $\mathbf{n}^*$. The pair $(\mathbf{p}, \mathbf{n}^*)$ is finally used to prompt the DM. DNS is straightforward to implement and requires no training. Experiments and human evaluations show that DNP performs well both quantitatively and qualitatively and can be easily combined with several DM variants.

Keywords: Image generation · Diffusion Models · Negative prompting

1 Introduction

Diffusion models (DMs) [19, 22, 24] have shown exquisite capacity to synthesize visually appealing images guided by textual prompts. However, they are not easy to control. While the synthesized images are usually impressive, various classes of prompts are known to be difficult, e.g. prompts involving humans [12], hands [17], or multiple objects and their interactions [3]. The synthesized image quality can often be unsatisfactory for such prompts, and prompt adherence is usually weak. This is illustrated in Figure 1, which shows an image synthesized by Stable Diffusion (SD) for a complex prompt $\mathbf{p}$. The problem can be mitigated by adding alternative conditioning inputs to the diffusion model, such as visual conditioning with sketches or layouts [31,32]. This, however, usually requires skilled users and can be labor-intensive. Ideally, it should be possible to improve consistency with text prompts alone. In this work, we explore negative prompting [16], which has been quite effective but is difficult to use. We argue that this difficulty stems from a *semantic gap* between the concept representations of a human user and the DM. We then introduce a new prompting technique, denoted as *diffusion-negative prompting* (DNP), to bridge this gap, as illustrated in Figure 1.

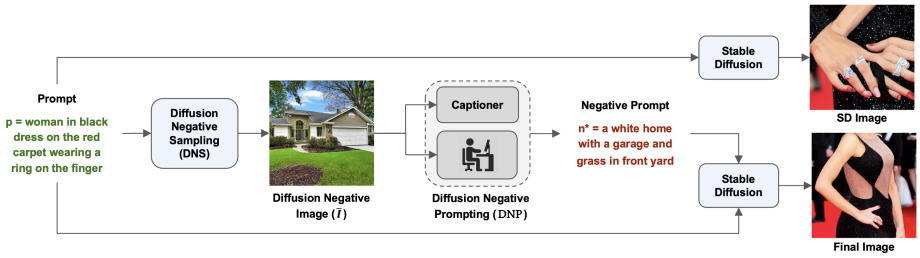
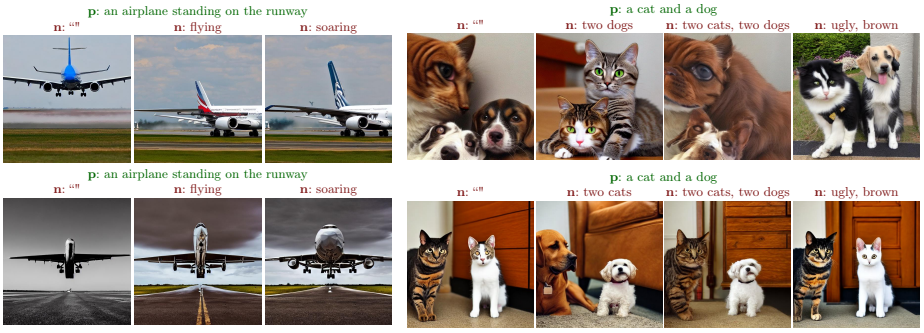


Fig. 1: `DNP` improves quality of synthesis for prompts $\mathbf{p}$ (green) of SD's images (top-right). A diffusion-negative image, $\bar{\mathcal{I}}$, is sampled using `DNS`, enabling the user to visualize the negation of $\mathbf{p}$ under DM's distribution. The user translates the $\bar{\mathcal{I}}$ into a negative prompt $\mathbf{n}^*$ (red), by a process denoted as `DNP`, and the DM is prompted with the pair $(\mathbf{p}, \mathbf{n}^*)$. This increases compliance and quality of the synthesized image (bottom-right). Replacing the user with a captioning model is denoted as `auto-DNP`.

Negative prompting complements the text prompt $\mathbf{p}$, e.g. *"an airplane standing on the runway"*, with negative prompts $\mathbf{n}$, e.g. *"flying"* or *"soaring"*. As illustrated in the top row of Figure 2a, this greatly improves compliance of the synthesized image with prompt $\mathbf{p}$ and can improve image quality. However, negative prompts are difficult to use, for two reasons. First, there are many prompts, e.g. see $\mathbf{p}$ = "a cat and a dog" in Figure 2b, without a clear negative. Second, even when an intuitive negative exists, e.g. "flying" vs. "standing", it is not necessarily a successful negative for DMs across seeds, as seen in the bottom row of Figure 2a. This raises the question: what is a good negative prompt for a DM? In general, $\mathbf{n}$ *is a good negative prompt if the images synthesized by prompt-pair* $(\mathbf{p}, \mathbf{n})$ *adhere more to* $\mathbf{p}$ *than those synthesized by* $\mathbf{p}$ *alone.* This, however, is a *DM-centric* definition of negative prompt. The problem is that the semantic representation of the DM is only a weak approximation to that used by a human. In general, DMs cannot replicate the human definition of concept negation. We refer to this problem as the *semantic gap* between DMs and humans. As a result, good negative prompts for a DM are frequently not intuitive for humans, i.e. what a human would consider a negative for $\mathbf{p}$. This makes it difficult for users to produce good negative prompts for DMs.

In this work, we address this problem by devising a strategy that enables humans to visualize the concepts that DMs consider negatives for prompt $\mathbf{p}$. This procedure is inspired by classifier-free guidance (CFG) [11]. While CFG uses a sampling guidance factor that increases the probability of images compliant with $\mathbf{p}$, we introduce a *diffusion-negative sampling* (`DNS`) guidance factor that encourages the sampling of the *diffusion-negative images*, $\bar{\mathcal{I}}$, least compliant with $\mathbf{p}$, under the DM's image distribution. This $\bar{\mathcal{I}}$ usually does not comply with the human understanding of the negation of concepts in $\mathbf{p}$. However, because it represents the DM's understanding of this negation, it can be shown to the user to overcome the semantic gap between them. The user can then produce a negative



(a) Semantic negatives: While effective, they do not guarantee compliance across seeds. Success (top), failure (bottom).

(b) Non-semantic negatives: A good negative prompt needs multiple iterations for every seed as they can be highly unintuitive.

Fig. 2: Example of negative prompting for both semantic and non-semantic scenarios. The positive prompt, $\mathbf{p}$ and negative prompts, $\mathbf{n}$ are on top of each image.

prompt $\mathbf{n}^*$ *in the language of the DM* by simply captioning $\bar{\mathcal{I}}$. Prompting the DM with the pair $(\mathbf{p}, \mathbf{n}^*)$ usually produces better images than with $\mathbf{p}$ alone.

Figure 1 shows how `DNP` can help when DMs produce poor-quality images. In response to the user prompt $\mathbf{p}$ = “woman in black dress on the red carpet wearing a ring on the finger”, SD generates the top image, with distorted hands and rings. The rest of the figure illustrates `DNP`. Given the prompt $\mathbf{p}$, the DM samples a negative image $\bar{\mathcal{I}}(\mathbf{p})$ using `DNS`. In the figure, $\bar{\mathcal{I}}$ contains the house with a garden. The user inspects the image and produces a *diffusion-negative prompt* $\mathbf{n}^*(\bar{\mathcal{I}})$ = “a white home with a garage and grass in the front yard” that reflects the content of this image. This is similar to captioning the image, and can also be done automatically by an image captioning model, resulting in what we denote as `auto-DNP`. Finally, the DM is prompted with the pair $(\mathbf{p}, \mathbf{n}^*)$ to produce an image that has much better compliance with $\mathbf{p}$ than the original.

Our experiments show that `DNP` improves prompt adherence and image quality, both in terms of quantitative metrics, like CLIP scores of image-prompt similarity, and subjective ratings by human evaluators. This is shown to hold both for vanilla DMs, like SD, and DMs optimized for solving specific problems, e.g. the *A&E* [3] model tailored to synthesize images containing multiple objects. Various ablations also show that `DNP` outperforms negative prompting solutions commonly used by practitioners, e.g. using a dictionary of universal negative prompts, random prompts, etc. It also provides much-needed transparency to the negation process. `auto-DNP` provides a fully automated implementation of `DNP`. While not as effective as `DNP`, it maintains all the benefits discussed above, avoids user captioning effort, and allows automatic comparisons to other negative prompting procedures. Finally, our experiments consistently demonstrate the existence of a *semantic gap* between DMs and humans. Although the negative images generated by `DNS` are frequently very unintuitive for humans, they help the DM significantly when translated into negative prompts.

Overall, this paper makes the following contributions. First, we hypothesize that DM image synthesis can be substantially enhanced by using negative prompts. However, these prompts are difficult for humans to generate, due to the *semantic gap* between humans and DMs, which leads to a different understanding of concept negation. Second, we introduce the `DNS` procedure to overcome this *semantic gap*, allowing humans to visualize the negative concept as understood by the DM. This enables negative prompting with `DNP`, as shown in Figure 1. Third, we present extensive experiments on various datasets, showing that `DNP` complements any existing DM or its variant. Fourth, we show that the process can be automated by using a pre-trained image captioner, leading to `auto-DNP`. Finally, we show that `DNP` is trivial to implement and requires no training.

2 Diffusion Models

Latent Diffusion: DMs combine a forward and a backward process for generative modeling. In the forward process, a sample is gradually corrupted by sequential application of small amounts of Gaussian noise. The backward process then sequentially denoises the sample using a learned neural network ϵ_θ. Latent DMs (LDMs) increase the efficiency of the diffusion process by operating on the low dimensional latent space $\mathcal{Z}$ of an autoencoder, implemented with a pre-trained encoder/decoder pair $(\mathcal{E}(.), \mathcal{D}(.))$. This transforms image $x \in \mathcal{X}$ into latent representation $z = \mathcal{E}(x) \in \mathcal{Z}$ and reconstructs the image with $x = \mathcal{D}(\mathcal{E}(z))$.
Training: In the forward process, a noisy version of the latent representation is obtained at each time step t with $z_t = \sqrt{\bar{\alpha}_t}z + \sqrt{1-\bar{\alpha}_t}\epsilon$. Here, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, where $\{\beta_1, ..., \beta_T\}$ is fixed according to the variance schedule and $\epsilon \sim \mathcal{N}(0, I)$. In the backward process, the denoising model ϵ_θ estimates the noise ϵ_t added to the noisy latent representation at each time step t. The denoising is conditioned by the embedding $\tau(\mathbf{p})$, of text prompt $\mathbf{p}$, where $\tau(.)$ is a text encoder. The denoising network parameters are learned by minimizing the loss

$$\mathcal{L} = \mathcal{E}_{t\sim\mathcal{U}(1,T),\epsilon_t\sim\mathcal{N}(0,I)}\Big[\|\epsilon_t - \epsilon_\theta(z_t; t, \tau(\mathbf{p}))\|^2\Big], \tag{1}$$

where $\mathcal{U}$ and $\mathcal{N}$ are the uniform and Gaussian distributions, respectively.
Inference: Given prompt $\mathbf{p}$, images are sampled by iteratively alternating between denoising and sampling, with

$$\hat{\epsilon}_p = \epsilon_\theta(z_t; t, \tau(\mathbf{p})), \quad z_{t-1} = \mathit{sample}(z_t, \hat{\epsilon}_p, t). \tag{2}$$

The sampling method, *sample*, varies with the DM. Popular approaches are DDPM [9] and DDIM [27]. This process is initialized with a noise seed $z_T \sim \mathcal{N}(0, I)$ and produces a latent image code z_0, which is finally passed to the decoder to obtain image $x_0 = \mathcal{D}(z_0)$.
Classifier-free Guidance: CFG [11] is a method to trade off prompt compliance and image diversity. It consists of training the DM with and without prompt $\mathbf{p}$, randomly setting $\mathbf{p}$ to the empty prompt ϕ="", with probability

$P_{uncond} = 0.1$. At inference, each denoising step uses a linear combination of the prompt-conditional ($\hat{\epsilon}_p$) and unconditional ($\hat{\epsilon}_\phi$) noise estimates

$$\hat{\epsilon} = \hat{\epsilon}_\phi + s(\hat{\epsilon}_p - \hat{\epsilon}_\phi), \tag{3}$$

where $\hat{\epsilon}_p = \epsilon_\theta(z_t; t, \tau(\mathbf{p}))$, $\hat{\epsilon}_\phi = \epsilon_\theta(z_t; t, \tau(\phi))$, and s is an hyper-parameter that controls the conditioning strength, known as the guidance-scale. The modified noise $\hat{\epsilon}$ is then utilized to update the latent representation.

Negative Prompting: Despite conditioning on text prompt $\mathbf{p}$, prompt adherence of CFG can be weak when $\mathbf{p}$ refers to complex scenes, such as that in Figure 1. Negative prompting [16] is useful as an extra conditioning input. Given negative prompt $\mathbf{n}$, the inference denoising steps are implemented with

$$\hat{\epsilon} = \hat{\epsilon}_\phi + s(\hat{\epsilon}_p - \hat{\epsilon}_n) \tag{4}$$

where

$$\hat{\epsilon}_n = \epsilon_\theta(z_t; t, \tau(\mathbf{n})). \tag{5}$$

Practitioners have shown that simply replacing the empty prompt ϕ of CFG with the negative prompt $\mathbf{n}$, i.e. using $\hat{\epsilon} = \hat{\epsilon}_n + s(\hat{\epsilon}_p - \hat{\epsilon}_n)$, yields similar results, albeit for a slightly different guidance scale s [1]. This has the benefit of computational efficiency, as it only requires the calculation of two noise vectors instead of three. Hence, this method has gained popularity and has become the default implementation of negative prompting in many T2I generation models. However, in what follows, we use the theoretically more grounded (4) to derive a procedure to sample *diffusion-negative* images ($\bar{\mathcal{I}}$).

3 Creating Good Negative Prompts

In this section, we introduce `DNP`.

3.1 Energy-based model interpretation

Sampling with a DM of network $\epsilon_\theta(z_t; t, \tau(\mathbf{p}))$ can be interpreted as sampling from probability distribution $p_\theta(z_t|\mathbf{p}) \propto e^{-E_\theta(z_t; t, \tau(\mathbf{p}))}$, where E_θ is an energy function, using the score function $\nabla_z \log P(z|\mathbf{p})$ and Longevin dynamics [28,30]. The DM network is trained to learn the score function, i.e.

$$\epsilon_\theta(z_t; t, \tau(\mathbf{p})) = -\nabla_{z_t} \log p_\theta(z_t|\tau(\mathbf{p})). \tag{6}$$

This motivates the CFG denoising step of (3), which corresponds to sampling from the distribution

$$p_{cfg}(z_t, \mathbf{p}) \propto p_\theta(z_t)\gamma^s_{cfg}(z_t, \mathbf{p}) \tag{7}$$

where

$$\gamma_{cfg}(z_t, \mathbf{p}) = \frac{p_\theta(z_t|\mathbf{p})}{p_\theta(z_t)} \propto p_\theta(\mathbf{p}|z_t) \tag{8}$$

is a guidance factor that increases the probability of codes z_t corresponding to images compliant with prompt $\mathbf{p}$. Similarly, the negative prompt denoising step of (4) corresponds to sampling from

$$p_{np}(z_t, \mathbf{p}, \mathbf{n}) \propto p_\theta(z_t)\gamma_{np}^s(z_t, \mathbf{p}, \mathbf{n}) \tag{9}$$

with guidance factor

$$\gamma_{np}(z_t, \mathbf{p}, \mathbf{n}) = \frac{p_\theta(z_t|\mathbf{p})}{p_\theta(z_t|\mathbf{n})} \propto \frac{p_\theta(\mathbf{p}|z_t)}{p_\theta(\mathbf{n}|z_t)}. \tag{10}$$

This increases the probability of sampling latent codes z_t of large odds ratio

$$o(z_t, \mathbf{p}, \mathbf{n}) = \frac{p_\theta(\mathbf{p}|z_t)}{p_\theta(\mathbf{n}|z_t)}. \tag{11}$$

Since the Bayes decision rule for deciding between $\mathbf{p}$ and $\mathbf{n}$ is to choose $\mathbf{p}$ when $o(z_t, \mathbf{p}, \mathbf{n}) \geq 1$ and $\mathbf{n}$ otherwise, it increases the probability that sampled codes comply with prompt $\mathbf{p}$ and not with prompt $\mathbf{n}$.

3.2 Limitations of Negative Prompting

Besides being theoretically well-motivated, negative prompting frequently improves image quality, as illustrated in the top row of Figure 2a. However, negative prompts are difficult to specify, because a natural negative frequently does not exist or does not produce the intended results. Figure 2b shows an example for prompt $\mathbf{p}$ = "a cat and a dog". Here, the user may need to attempt several negative prompts $\mathbf{n}$, requiring multiple iterations of image synthesis, as illustrated in the figure. Eventually, the user notices the insistence of the DM in brown objects and uses $\mathbf{n}$ = *"brown"*. This is complemented with *"ugly"*, which practitioners have identified as a good general-use negative prompt. To compound the problem, a negative prompt successful for one random seed can be unsuccessful for another seed, i.e. good negative prompts depend on *both* prompt $\mathbf{p}$ and seed z_T, as illustrated in Figure 2a. In summary, negative prompting is an art form and can be quite cumbersome, even when successful.

In practice, the human-provided negative prompt frequently fails to improve the quality of the image generated by the DM. We hypothesize that this is because the semantic representation of the DM is only a weak approximation to that of humans. Hence, there is no guarantee that, under the probability distribution modeled by the DM, there will be a large classification margin between $\mathbf{p}$ and the human-provided $\mathbf{n}$. In this case, negative prompting fails to produce samples of large odds ratio $o(z_t, \mathbf{p}, \mathbf{n})$ and the prompt pair $(\mathbf{p}, \mathbf{n})$ fails to induce the model to produce more prompt compliant images than those generated with CFG, i.e. the posterior probability $p_\theta(\mathbf{p}|z_t)$ of (8). In the extreme case where $p_\theta(\mathbf{p}|z_t) \approx p_\theta(\mathbf{n}|z_t)$ for the samples of high $p_\theta(\mathbf{p}|z_t)$, $o(z_t, \mathbf{p}, \mathbf{n}) \approx 1$ for those samples and (9) looses all sensitivity to prompt $\mathbf{p}$.

3.3 Diffusion-Negative Guidance

The discussion above suggests that a good negative prompt *for the DM* "grounds" the odds-ratio of (11), providing a reference for the prompt $\mathbf{p}$, which is defined in terms of the margin to negative $\mathbf{n}$. However, this grounding must happen under the *internal* representation of the DM, which is not necessarily that of humans. The goal is to encourage samples more likely to comply with given prompt $\mathbf{p}$ than the negative prompt, under the conceptual representation *of the model*. We formalize this goal by defining the optimal negative prompt $\mathbf{n}^*$ *for the DM* as the one that maximizes the steepness of the odds ratio with $\mathbf{p}$, i.e.

$$\mathbf{n}^*(\mathbf{p}) = \arg\max_{\mathbf{n}} \nabla_{z_t} \log o(z_t, \mathbf{p}, \mathbf{n}). \tag{12}$$

This is denoted as the *diffusion-negative* prompt for $\mathbf{p}$. Using (2), (5), and (11),

$$\nabla_{z_t} \log o(z_t, \mathbf{p}, \mathbf{n}) = \nabla_{z_t} \log \frac{p_\theta(\mathbf{p}|z_t)}{p_\theta(\mathbf{n}|z_t)} = \nabla_{z_t} \log \frac{p_\theta(z_t|\mathbf{p})}{p_\theta(z_t|\mathbf{n})} = \hat{\epsilon}_p - \hat{\epsilon}_n. \tag{13}$$

Hence, for a given z_t, a natural alternative definition of the *diffusion-negative prompt* is that which induces the noise vector maximally distant from $\hat{\epsilon}_p$, i.e.

$$\epsilon_{n^*} = \underset{n|\|\hat{\epsilon}_n\|^2=K}{\arg\max} \|\hat{\epsilon}_p - \hat{\epsilon}_n\|^2 \tag{14}$$

where K prevents the noise magnitude from becoming infinitely large. This is the prompt that induces the steeper log odds surface at z_t, encouraging the largest margin between the probabilities $p_\theta(z_t|\mathbf{p})$ and $p_\theta(z_t|\mathbf{n}^*)$ in the neighborhood of z_t. In Supplementary Section 1, we show that (14) has a solution

$$\hat{\epsilon}_{n^*} = -\sqrt{K}\frac{\hat{\epsilon}_p}{\|\hat{\epsilon}_p\|} \propto -\hat{\epsilon}_p. \tag{15}$$

Choosing $K = \|\hat{\epsilon}_p\|^2$, i.e. equal noise strength under the positive and negative conditions, results in $\hat{\epsilon}_{n^*} = -\hat{\epsilon}_p$.

3.4 Diffusion-Negative Sampling (DNS)

It follows from (6) that

$$\nabla_{z_t} \log p_\theta(z_t|\mathbf{n}^*) = -\nabla_{z_t} \log p_\theta(z_t|\mathbf{p}) \implies p_\theta(z_t|\mathbf{n}^*) \propto \frac{1}{p_\theta(z_t|\mathbf{p})}. \tag{16}$$

However, for most $p_\theta(z_t|\mathbf{n})$, (16) is an improper distribution [4]. This stems from the fact that a negative condition *must* be specified in context [5]. A general contextual constraint for DM-based sampling is that the negatively conditioned DM should still sample images from the image distribution $p_\theta(z_t)$. To account for this, we define the *diffusion-negative* guidance factor as

$$\gamma_{dn}(z_t, \mathbf{p}) = \frac{p_\theta(z_t)}{p_\theta(z_t|\mathbf{p})} \propto \frac{1}{p_\theta(\mathbf{p}|z_t)}, \tag{17}$$

which leads to the *diffusion-negative* distribution

$$p_{dn}(z_t, \mathbf{p}) \propto p_\theta(z_t)\gamma_{dn}^s(z_t, \mathbf{p}) \tag{18}$$

and inference denoising equation

$$\hat{\epsilon} = \hat{\epsilon}_\phi + s(\hat{\epsilon}_\phi - \hat{\epsilon}_p). \tag{19}$$

Sampling with this equation is denoted as *diffusion-negative sampling* (DNS). When compared to (10), the *diffusion-negative* guidance factor of (17) has two main differences. First, it does not require the semantic negative prompt $\mathbf{n}$, reflecting the fact that the definition of negative is fully based on the internal representation of the DM. Second, it replaces the odds ratio of (11) by the ratio between the image distribution $p_\theta(z_t)$ and the image distribution under the positive condition $\mathbf{p}$. This emphasizes sampling images *from the DM probability distribution with the lowest prompt conditional probability* $p_\theta(z_t|\mathbf{p})$. When compared to the guidance factor (8) of CFG, (17) simply flips the role of the image distributions $p_\theta(z_t)$ and $p_\theta(z_t|\mathbf{p})$. This provides a simple interpretation of *diffusion-negative* guidance as the "opposite" of CFG. Rather than sampling natural images that align best with the condition $\mathbf{p}$, it emphasizes sampling natural images that least comply with it.

The relation between the guidance factors of (17) and (10)

$$\gamma_{dn}(z_t, \mathbf{p}) = \frac{p_\theta(z_t)}{p_\theta(z_t|\mathbf{p})} = \gamma_{np}(z_t, \phi, \mathbf{p}) \tag{20}$$

makes the implementation of DNS trivial for any model that supports negative prompting. It suffices to prompt the latter with empty positive prompt ϕ and negative prompt $\mathbf{p}$. Hence, DNS *requires no retraining of the DM, nor any optimizations at inference.*

3.5 Diffusion-Negative Prompting (DNP)

DNP relies on DNS to implement the procedure of Figure 1. Given prompt $\mathbf{p}$, DNS is used to sample a *diffusion-negative* image, $\bar{\mathcal{I}}$, as discussed above. This image is then captioned by the user, to produce the DNP $\mathbf{n}^*$. The DM is finally prompted with the prompt pair $(\mathbf{p}, \mathbf{n}^*)$ to produce the final image. As illustrated in Figure 1, the captioning step can be performed by either the user or any captioning model. In the latter case, the process is denoted auto-DNP. auto-DNP has the advantage of fully automating DNP, making the entire process transparent to the user and reducing user effort. The price is a small increase in the failure rate due to captioning errors. In this paper, we use BLIP2 [14] as the captioning model to implement auto-DNP. However, auto-DNP can be used with any other captioner. We show results obtained with GPTv4 (*gpt-4-vision-preview*) in Supplementary Section 4. Figure 3 and the final column of Figure 4 illustrate how DNP produces images that either have higher quality or comply better with the prompt $\mathbf{p}$.

4 Experiments

In this section, we discuss the experimental evaluation of DNP.

4.1 Datasets

For a comprehensive dataset description see Supplementary Section 2.2
Existence and Attribute Binding (A&E) Dataset: This is the benchmark dataset introduced by [3]. This dataset is focused on entity neglect, which occurs when one or more entities are completely ignored by the generative model, and attribute assignment, which evaluates whether an attribute is associated with the correct object in the image. To evaluate these, each prompt in the dataset comprises two entities and their associated attributes. There are three categories of prompts: (1) "an animal and an animal"; (2) "a color in-animate object and an animal"; (3) "a color in-animate object and a color in-animate object". We sample 50 prompts from each category for 32 random seeds.
Human and Hand Generation (H&H) Dataset: Humans and hands occur in various shapes and forms, including gender, race, activity, and individual differences. Despite the size of the training datasets, this variation has posed a major challenge to DMs, which are known to synthesize images with extra or merged limbs and fingers. We curated a challenging benchmark dataset to test the effect of DNP on human and hand generation. The dataset consists of 45 human-based prompts such as "a man wearing a sombrero" and 25 hand-based prompts such as "hand with a ring on it". We run all prompts for 32 seeds.

4.2 Evaluation Metrics

Text-Image CLIP Score: [8] is a widely used metric of the similarity between a generated image and its text prompt. For the H&H dataset, we calculate the CLIP Score between each generated image and its corresponding prompt. For the A&E dataset, we follow the evaluation protocols established by [3], which measure CLIP Score at full prompt and entity level. For evaluating the *Full Prompt* CLIP Score, we calculate the score between the generated image and the full prompt. For the *Minimum Object* CLIP Score, we take the minimum score between the entities in the prompt, as the entity with the minimum CLIP Score represents the neglected entity for the given prompt. This allows us to measure the CLIP Score at two levels, thereby providing better insight into the presence of both entities in the A&E dataset prompts.
Inception Score (IS): [25] is a popular metric for assessing the quality and realism of images. It is the exponential of the average of the entropy of the label distribution predicted by the Inception v3 classifier model. In particular, we choose IS over FID as it does not require a dataset with real images.
Human Evaluation: While CLIP Score and IS provide an objective measure of correctness and quality, they are known for not being fully aligned with human aesthetics and preferences. To evaluate DNP in terms of human preferences, we used Amazon Mechanical Turk (AMT) to evaluate a subset of each dataset. We

Method	CLIP Score			
	Human	Hand	A&E	
			Min. Object	Full Prompt
SD+auto-DNP	0.330	0.323	0.250	0.344
SD+DNP	**0.331**	**0.324**	**0.253**	**0.345**

Table 1: Benchmarking DNP to auto-DNP with CLIP Score.

performed human evaluation on 100 prompt-seed pairs for the human dataset, 100 pairs for the Hands dataset, and 150 (50 per category) pairs for the A&E dataset. The evaluators were asked to compare the images using two criteria. 1) Adherence to the prompt, and 2) Quality (natural or realistic nature) of the image. MTurkers were tasked with choosing one image based the two criteria. They could also choose *no clear winner* if they did not prefer one over the other. More details can be found in the Supplementary Section 2.3.

4.3 Benchmarking DNP vs auto-DNP

The evaluation of DNP requires human users. PhD students, with no experience in visual language research, were shown an image and asked to produce a caption shorter than 60 words. Due to the prohibitive resources required to caption the image $\bar{\mathcal{I}}$ for all prompt-seed pairs of all the datasets, we restricted the evaluation of DNP to the prompt-seed pairs used for human evaluation. We compared the performance of DNP and auto-DNP on this set and used auto-DNP in the remaining experiments. Besides being cost-effective, this enables the replication of our experiments for comparison to other negative prompting methods. Table 1 compares the performance of DNP and auto-DNP, on the prompt-seed pairs used for the human evaluation[1]. For the rest of the paper, we use the notation DM+X, where DM is the diffusion model and X the negative prompting method.

While DNP achieves the best performance, auto-DNP provides almost identical results. This suggests that the BLIP2 model is quite effective at captioning the diffusion-negative negative images $\bar{\mathcal{I}}$.

4.4 Quantitative Results

We next discuss the results of quantitative experiments on image quality using auto-DNP. Various ablations are also presented in Supplementary Section 4.
A&E dataset: We first compare SD to SD+auto-DNP, on the A&E dataset in Table 2a. SD+auto-DNP achieves a *Min. Object* CLIP Score of 0.258, which is 6.6% higher than that of SD, illustrating that SD+auto-DNP generates images that align better with the prompts. Since CLIP scores practically range from [0, 0.4] [8], the full prompt CLIP Score of 0.346 of SD+auto-DNP suggests good alignment between prompt and image. Human evaluation underscores the true

[1] IS score cannot be computed in this case as the number of seeds per prompt is not large enough for IS to provide reliable results.

Method	CLIP Score		IS	Human Evaluation	
	Min. Object	Full Prompt		Correctness	Quality
No Clear Winner	-	-	-	16.88%	10.46%
Stable Diffusion (SD)	0.242	0.335	13.17	18.90%	25.73%
SD+`auto-DNP`	**0.258** (+6.61%)	**0.346** (+3.28%)	**13.35** (+1.37%)	**64.22%**	**63.81%**
No Clear Winner	-	-	-	16.95%	7.35%
Attend & Excite (*A&E*)	0.264	0.349	11.86	30.06%	33.12%
A&E+`auto-DNP`	**0.276** (+4.54%)	**0.362** (+3.72%)	**13.12** (+10.62%)	**52.99%**	**59.53%**

(a) Quantitative Results on the A&E dataset.

Dataset	Method	CLIP Score	IS	Human Evaluation	
				Correctness	Quality
Human Prompts	No Clear Winner	-	-	14.42%	5.10%
	Stable Diffusion (SD)	0.322	9.95	27.60%	29.0%
	SD+`auto-DNP`	**0.331** (+2.80%)	**10.13** (+1.80%)	**57.98%**	**65.80%**
Hand Prompts	No Clear Winner	-	-	19.89%	8.18%
	Stable Diffusion (SD)	0.309	12.04	20.45%	22.47%
	SD+`auto-DNP`	**0.321** (+3.88%))	**12.39** (+2.94%)	**59.66%**	**69.35%**

(b) Quantitative Results on the H&H dataset.

Table 2: Quantitative Results: comparing prompt adherence and image quality by CLIP Score, IS and Human Evaluation. The percentage gains are shown in (brackets)

strength of `DNP`, as human evaluators prefer images generated by SD+`auto-DNP` 3.4× more in terms of correctness and 2.5× more in terms of quality than those generated by SD. Note that SD+`auto-DNP` even outperforms the *A&E* model in IS and matches its CLIP Score for both full prompt and minimum object, despite not being specifically designed for multiple object generation.

To demonstrate the flexibility of the `DNP` approach, we also applied it to the *A&E* model, comparing *A&E*+`auto-DNP` to *A&E*. We observe a moderate increase in the CLIP Score since *A&E* is already quite efficient at generating multiple objects. *A&E*+`auto-DNP` achieves a *Min. Object* CLIP Score of 0.276, which is very close to the theoretical *upper bound* for *Min. Object* CLIP Score of 0.29 as defined by [3]. The 10.62% improvement in IS shows that *A&E* is more inclined to generate unrealistic images and the addition of `DNP` improves its realism. The human evaluation also reflects this, as around 53% of the evaluators prefer *A&E*+`auto-DNP` over the *A&E* images for correctness and 60% for quality, which is close to 1.8× the preference for *A&E*.

H&H dataset: Table 2b summarizes the ability of `DNP` to address the specific SD weaknesses of synthesizing humans and hands. Since both CLIP and IS may ignore distorted or maligned results when considering correctness or realism, they fail to give substantial insight into image quality (despite improvement). The human evaluation is much more important for this task. SD+`auto-DNP` outperforms SD by a substantial margin, as evaluators prefer SD+`auto-DNP` 3× more for correctness and 2× more for quality, for both human and hand prompts. Since evaluators were specifically tasked with checking the number and pose of limbs and fingers and the quality of the faces, their preference for images generated with `DNP` shows that the latter improves SD performance substantially.

Fig. 3: Images synthesized by SD (left) for prompt $\mathbf{p}$ vs. SD+`auto-DNP` (right) for prompt pair $(\mathbf{p}, \mathbf{n}^*)$, where $\mathbf{n}^*$ is the `DNP` estimated from the `DNS` image. `DNP` produces negative prompts that are not intuitive for humans but improve quality and adherence.

4.5 Qualitative Results

Figure 3 shows qualitative examples from the three datasets, comparing SD with SD+`auto-DNP`. There are more visual examples in Supplementary Section 5.

The H&H dataset results, e.g. the images synthesized for prompt "*hand with a ring on it*", illustrate how `auto-DNP` improves the quality of humans and hands in terms of the number of limbs, fingers, poses, etc. It also induces the model to create realistic images, rather than drawings or sketches. See, for example, the "*man in a sombrero*" and "*boy in the moonlight*" examples. We observe a major improvement in the quality of the humans generated by `auto-DNP` wherein the faces do not exhibit the uncanny, robotic, or distorted features that SD is inclined to create, instead having a natural warmth.

The results on the A&E dataset show that `auto-DNP` can correct for entity neglect while ensuring superior image quality. This is especially visible when we compare *A&E* with *A&E*+`auto-DNP` (see Figures 5,6,7 in Supplementary Section 5), which produces more realistic images. We posit that `auto-DNP`'s inclination towards generating more realistic images is due to setting the diffusion process onto a better Markov chain without disrupting it at each step, unlike *A&E*.

Finally, note that none of the negative prompts produced by `auto-DNP` in Figure 3 is intuitive for humans. This illustrates the semantic gap between humans and SD and the difficulty of producing good negative prompts manually.

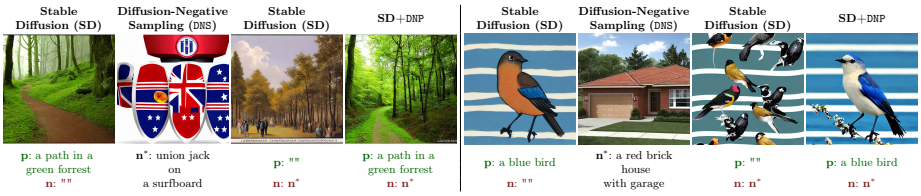


Fig. 4: Exploring Semantic Gap: From left to right: 1) images generated for prompt $\mathbf{p}$, 2) `DNS` images generated by the DM and resulting caption $\mathbf{n}^*$ by `DNP`, 3) images synthesized with $\mathbf{n}^*$ as the negative prompt alone, and 4) with prompt pair $(\mathbf{p}, \mathbf{n}^*)$.

4.6 Semantic Gap

Figure 4 explores the semantic gap between humans and DMs in more detail. From left to right, the first column shows images synthesized by SD with prompt $\mathbf{p}$. The second column shows examples of images sampled with `DNS`, i.e. (19). These can be seen as images that the model hallucinates as "negatives" of the prompt $\mathbf{p}$ since they usually do not have this interpretation for humans. The caption $\mathbf{n}^*$ is obtained the user. To confirm that the prompt mirrors the DM internal representation of the "negative" for the image in the left column, we prompted the model with an empty positive prompt $\mathbf{p} = \phi$ and $\mathbf{n}^*$ as a negative prompt, which produces the images in the third column of the figure. Note how these images comply with $\mathbf{p}$ even though $\mathbf{n}^*$ is non-intuitive for a human. Finally, the last column shows the result of prompting with the pair $(\mathbf{p}, \mathbf{n}^*)$. In all cases, the synthesized image is at least as good as that on the first column, and usually better, e.g. a greener forest or a more realistic bird.

Two aspects are worth noting. First, these examples purposefully uses prompts $\mathbf{p}$ for which SD works well. This proves that the "DM knows what it's doing," at least to the point of generating a sensible image (first column). However, its notion of negative is totally different from that of humans, demonstrating the semantic gap between the two. The practical result is that a totally unintuitive negative prompt is needed to produce good images. Figure 3 shows that `auto-DNP` can improve the quality and correctness of the synthesised image even when the DM fails. Second, because all derivations above are functions of z_t, the `DNS` procedure is valid for any seed z_T. This is an additional advantage over the human-centric negative prompting of Figure 2a, whose success is seed-dependent.

5 Related Work

Text-to-Image Diffusion: High-resolution T2I generation has advanced dramatically with the introduction of large-scale diffusion models, due to: (1) availability of large-scale text-image datasets [2, 26], (2) advances in various training and inference techniques [9–11, 27], and, (3) development of scalable model architectures [19, 20, 22, 24]. The most popular T2I models employ classifier-free guidance [11] to balance prompt compliance and image diversity using a guidance scale hyper-parameter, s. While prompt adherence improves by increasing

s, image quality deteriorates beyond a certain value. DNP helps ensure prompt adherence even for lower values of s, thereby maintaining image quality.

Structured Image Conditioning: Various image generation and editing improvements have been introduced since the advent of DMs. They range from finetuning the models for certain tasks to editing the noise at each iteration. Many methods require users to provide structured conditioned inputs like layouts [15, 18, 32], example images [13, 23, 29], and depth maps [31]. These can be hard to produce and require considerable user skill or additional computer vision modules. DNP allows us to control T2I with text alone while allowing it to be combined with most structured conditioning techniques as well.

Attention-based Methods: These methods attempt to tackle DM errors by editing attention maps at each diffusion step, changing the cross-attention between prompt tokens. [3] updates the noise latent with a loss that maximizes attention to each noun, while [21] first creates a linguistic binding between prompt attributes and nouns and then maximizes (minimizes) attention IOUs for mapped (non-mapped) tokens. [7] uses text formatting such as font style, size, color, and footnotes to allow for creative human input and uses region-specific guidance to combine the noise corresponding to each token. These methods require careful fine-tuning of a loss guidance hyper-parameter and can heavily alter the DM Markov chain, which can create highly saturated and low-quality images. While the proposed method also forces the DM to use a different Markov chain, it does not interfere with its intermediate steps through any loss updates.

Text based Methods: These methods try to explore and correct the semantic failures of the DM text encoder. [16] tries to decouple the text and combine noise latent later and [6] introduces consistency trees to split the prompt into noun phrases, for better cross-attention in the text-encoder. These methods fail to correct those errors which arise not from the text but from the DM itself.

Negative Prompts: Negative prompts [16], are quite effective at removing unwanted concepts from the synthesized image and have been incorporated into most T2I models. However, finding a good negative prompt requires trial and error. Also, the effect of negative prompts varies significantly with prompt and seed. The proposed method, DNP provides a solution to this.

6 Conclusion

Do diffusion models truly understand negation? In this work, we have hypothesized that the semantic gap between DMs and humans is responsible for the poor performance of negative prompting. We proposed DNP as a simple yet effective method for bridging this gap. This consists of sampling a negative diffusion image $\bar{\mathcal{I}}$, using a novel DNS procedure, and asking the user to translate it into a natural language negative prompt. This greatly improves the prompt adherence and quality of the generated images. DNP is universally applicable across diffusion models and can be combined with other methods. It highlights the difference between semantic and diffusion negation and leverages this difference to improve the performance of these models without additional training.

Acknowledgements

This work was partially funded by NSF award IIS-2303153 a gift from Qualcomm, and NVIDIA GPU donations. We also acknowledge and thank the use of the Nautilus platform for the experiments discussed above.

References

1. AUTOMATIC1111: Negative prompt: Stable diffusion webui, https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt
2. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models (2023)
4. Dawid, A.P., Stone, M., Zidek, J.V.: Marginalization paradoxes in bayesian and structural inference. Journal of the Royal Statistical Society Series B: Statistical Methodology **35**(2), 189–213 (1973)
5. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6637–6647. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/49856ed476ad01fcff881d57e161d73f-Paper.pdf
6. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis (2023)
7. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: IEEE International Conference on Computer Vision (ICCV) (2023)
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation (2022)
13. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
15. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
16. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)

17. Lu, W., Xu, Y., Zhang, J., Wang, C., Tao, D.: Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting (2023)
18. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023)
19. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
20. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
21. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment (2023)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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 (CVPR). pp. 22500–22510 (June 2023)
24. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
25. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans (2016)
26. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 25278–25294. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
28. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
29. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation (2023)
30. Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 681–688 (2011)
31. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
32. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)