# Guiding Diffusion Models with Adaptive Negative Sampling Without External Resources

Alakh Desai  Nuno Vasconcelos
University of California, San Diego
{ahdesai, nvasconcelos}@ucsd.edu


Figure 1. Images synthesized by SDXL (CFG) (Left) vs. SDXL+ANSWER (Right) for the prompt displayed below each image pair. ANSWER optimizes the negative guidance at each diffusion step to maximize image quality and prompt compliance.

## Abstract

*Diffusion models (DMs) have demonstrated an unparalleled ability to create diverse and high-fidelity images from text prompts. However, they are also well-known to vary substantially regarding both prompt adherence and quality. Negative prompting was introduced to improve prompt compliance by specifying what an image must not contain. Previous works have shown the existence of an ideal negative prompt that can maximize the odds of the positive prompt. In this work, we explore relations between negative prompting and classifier-free guidance (CFG) to develop a sampling procedure, Adaptive Negative Sampling Without External Resources (ANSWER), that accounts for both positive and negative conditions from a single prompt. This leverages the internal understanding of negation by the diffusion model to increase the odds of generating images faithful to the prompt. ANSWER is a training-free technique, applicable to any model that supports CFG, and allows for negative grounding of image concepts without an explicit negative prompts, which are lossy and incomplete. Experiments show that adding ANSWER to existing DMs outperforms the baselines on multiple benchmarks and is preferred by humans 2× more over the other methods.*

## 1. Introduction

Diffusion models (DMs), such as DALL-E 2 [22], Imagen [27], or SDXL [20], have shown a remarkable ability to generate compelling images from textual prompts, and are now considered state-of-the-art image synthesis. However, while the quality of the synthesized images is usually excellent, these methods still suffer from limited control over the image semantics. As shown in Figure 2, the generated images frequently fail to comply with the prompt (e.g. "two cats" instead of "a cat on the left of a dog"). This has motivated a literature on DMs that condition image synthesis on user-specified visual constraints, such as bounding boxes [17, 19, 39], sketches [33, 38], etc. While effective, these approaches have limitations of their own. First, it is difficult to devise a universal visual conditioning module. While bounding boxes can easily specify object locations, it is much less clear what visual user input can help a model

produce a scene that "exudes happiness" or "is artsy". Second, while visual specifications are suitable for professional image creators, they are too cumbersome for casual users.

An alternative is to modify the sampling process of the DM. A popular sampling technique is classifier-free guidance (CFG) [12]. This combines samples from a conditioned and an unconditioned network to improve the trade-off between image diversity and prompt compliance, under the control of a guidance scale parameter $s$. While increasing $s$ tends to improve prompt adherence, it is usually insufficient to guarantee compliance with complex prompts, e.g, involving spatial relations between multiple objects. In the example of Figure 2, even a high guidance scale ($s = 20$) creates an image of "two cats". This problem motivated various techniques that optimize the DM latent during the denoising process [5, 23] to maximize prompt compliance. However, this alters the Markov chain implemented by the model, creating a mismatch between training and inference that can easily result in low-quality images.

A less drastic manipulation of the sampling process is to use negative prompting [18]. This allows the user to specify an additional prompt of what the image *is not* about. Sampling techniques like classifier-free guidance (CFG) [12] are then modified to use positive and negative text prompts to improve prompt compliance. While negative prompting is a popular and successful mechanism for removing specific concepts, it is frequently difficult to specify negative prompts (e.g., the negative of the prompt of Figure 2 is unclear). This has motivated efforts to determine the negative prompt automatically. DNP [6] has shown that this can be done by sampling a negative image using the DM itself, which is then captioned to produce a negative prompt. This was shown to outperform latent optimization methods and can be quite successful, as illustrated in Figure 2, where DNP generates an image better aligned with the prompt (apart from the dog's extra leg) than CFG, even for a small guidance scale ($s = 6$). The authors of [6] observe a semantic mismatch between successful negative prompts generated in this way and the human definition of a negative. In Figure 2, the negative prompt used to create the DNP image is "a colorful pattern with flowers and leaves". This confirms the difficulty of generating good negative prompts.

While DNP is effective, it requires running the denoising process twice and captioning the negative image with an external visual-language model. This has three major limitations: (1) the added complexity of the external model, (2) the discrete approximation of the negative image by a text prompt that fails to capture many of its details, and (3) the fact that negation is only estimated once (first run of the DM), rather than iteratively throughout the diffusion process of the negatively prompted model (second run). We will show that the negation prompt can change significantly as the diffusion chain progresses.

Figure 2. Effect of guidance scale and different sampling methods on the prompt **p** = `a cat on the left of a dog`.

We propose a new *Adaptive Negative Sampling Without External Resources* (ANSWER) procedure, which addresses all these limitations without explicit user input or model retraining. We hypothesize that the capacity of negative prompting to alter the DM's Markov chain can be utilized in more powerful ways by continuously adapting the negative hypothesis during sampling, thus accounting for the changing nature of the image being denoised. Instead of producing a single negative prompt that conditions all denoising steps, we *dynamically apply negative guidance at each diffusion step* to estimate the negative noise that *matches* the noisy image synthesized at that step. As a result, the negation is performed *adaptively throughout the chain*. This makes ANSWER an enhanced DM sampling mechanism that, as illustrated in Figure 2, tends to generate images that better comply with the prompt than CFG and DNP. There is also no need for translation into a language prompt, eliminating the associated loss of information and the need for external captioning.

Overall, this work makes three main contributions:

- It is the first to investigate DM sampling schemes that involve a negative hypothesis produced adaptively to match the denoised image at each diffusion step. It is fully implemented in the DM latent space, thus avoiding costly and incomplete translation to text.
- We introduce the fully autonomous ANSWER sampling procedure, which requires no negative input by the DM user and no model retraining.
- Through extensive experiments on various datasets, we demonstrate that ANSWER can be seamlessly integrated into existing DMs, offering a practical solution to reduce the semantic gap between humans and DMs.

## 2. Preliminaries

### 2.1. Diffusion Models

A DM is a denoising model that synthesizes an image by sequentially denoising a noise code or *seed*. DM training involves a combination of a forward and a backward process. The forward process is a Markov chain that maps an image into a seed by gradually adding noise. The backward process performs step-by-step denoising, using a neural network $\epsilon_\theta$ that predicts the noise added at each forward step. Latent DMs are a popular class of DM that uses an encoder/decoder pair $(\mathcal{E}(.), \mathcal{D}(.))$ to produce a low-dimensional latent space $\mathcal{Z}$, where diffusion takes place. The forward and backward processes are as follows.

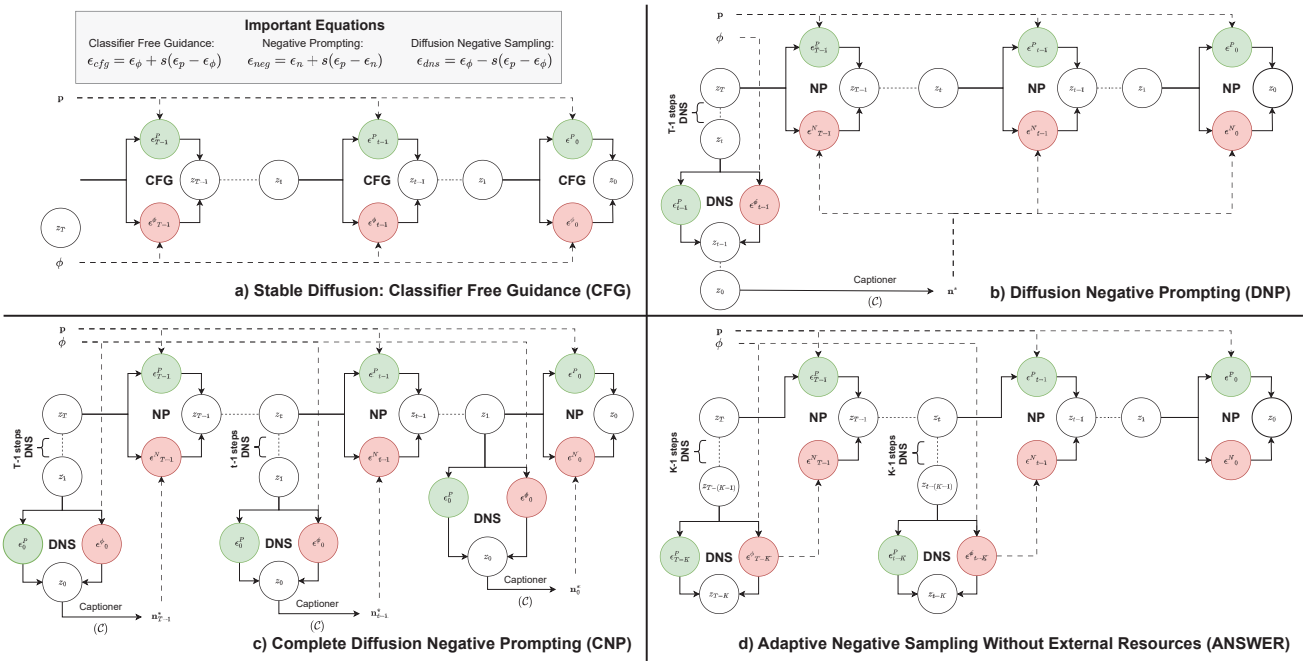


Figure 3. Sampling approaches discussed in this work. DNP, CNP, and ANSWER all rely on a DNS chain to generate the negative condition. While DNP, and CNP use an external captioning model $\mathcal{C}$ to produce negative text prompts, ANSWER performs the negative conditioning in the latent space of the DM. CNP runs a complete DNS chain at each diffusion iteration $t$, whereas ANSWER only requires $K$ DNS iterations.

**Forward Process:** A noisy latent representation is obtained at each time step $t$, with $z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. Here, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, with $\{\beta_1, ..., \beta_T\}$ fixed according to a variance schedule. During training, the network $\epsilon_\theta$ learns to predict the noise added ($\epsilon_t$) to the latent representation $z$ at each time step $t$. This prediction is conditioned by the embedding $\tau(\mathbf{p})$ of a text prompt $\mathbf{p}$, where $\tau(.)$ is a text encoder. The following loss is then minimized to learn the parameters $\theta$ of the denoising network

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

where $\mathcal{N}$ represents the Gaussian distribution.

**Backward Process:** Image generation is performed by iteratively alternating between denoising and sampling with the learned network $\epsilon_\theta$

$$\hat{\epsilon}_p = \epsilon_\theta(z_t; t, \tau(\mathbf{p})) \quad (2)$$

A sampling method [13, 14, 30] is then used to obtain $z_{t-1}$ from $z_t$ and $\hat{\epsilon}_p$. A noise seed initializes $z_T \sim \mathcal{N}(0, I)$ to produce a latent $z_0$. The final image is obtained by passing the latent to the decoder, $x_0 = \mathcal{D}(z_0)$.

**Energy-based model interpretation:** Langevin Dynamics [31, 34] enable an energy-based interpretation of DM sampling. If $E_\theta$ is an energy function, sampling from DM can be viewed as sampling from the probability distribution, $p_\theta(z_t|\mathbf{p}) \propto e^{-E_\theta(z_t;t,\tau(\mathbf{p}))}$ and training the DM to learn the noise $\epsilon_t$, can be viewed as learning the negative of the score function, $\nabla_{z_t} \log p_\theta(z_t|\tau(\mathbf{p})) = -\epsilon_\theta(z_t; t, \tau(\mathbf{p}))/\sqrt{1-\bar{\alpha}_t}$

## 2.2. Guidance

Using Eq. (2) in the DM backward process can produce images of weak compliance with prompt $\mathbf{p}$. Several enhancements have been proposed.

**Classifier-free Guidance (CFG) [12]:** provides a trade-off between diversity and compliance by training the model both with (90%) and without (10%) the prompt $\mathbf{p}$. A linear combination of prompt-conditional ($\hat{\epsilon}_p$) and unconditional ($\hat{\epsilon}_\phi$) noise estimates is used at inference time.

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

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 known as the guidance-scale that controls the conditioning strength. At each step, the noise $\hat{\epsilon}$ is used to denoise the latent. Under the energy-based interpretation, this is equivalent to sampling from

$$p_{cfg}(z_t, \mathbf{p}) \propto p_\theta(z_t) p_\theta(\mathbf{p}|z_t)^s. \quad (4)$$

While increasing $s$ can strengthen the conditioning process, it cannot offset an extremely low $p_\theta(\mathbf{p}|z_t)$. This is shown in Figure 2, where CFG fails even for high guidance values.

**Negative Prompting (NP) [18]:** is a popular tool in the DM user community to improve prompt compliance and image quality. It leverages a negative prompt $\mathbf{n}$ to provide additional guidance to the DM. Most systems [1] implement negative prompting in a slightly different manner than introduced by [18], replacing Eq. (3) by the denoising equation

$$\hat{\epsilon} = \hat{\epsilon}_n + s(\hat{\epsilon}_p - \hat{\epsilon}_n) \quad (5)$$

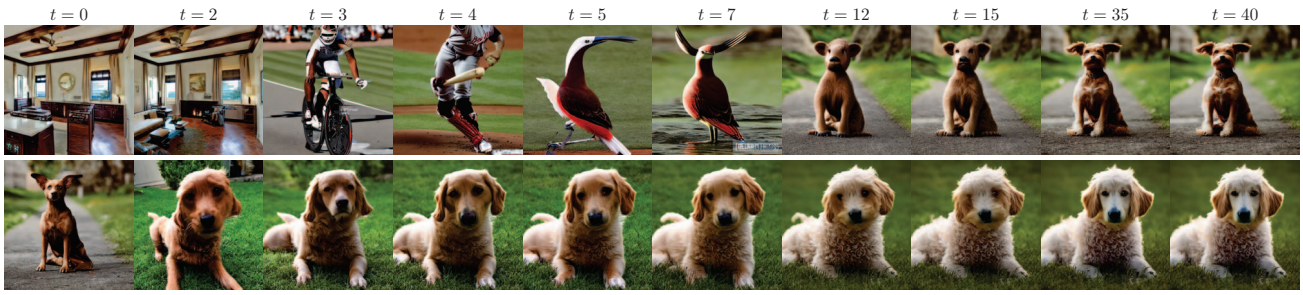


Figure 4. **Top:** Negative images produced by SD (T = 40 step) using DNS for the latent $z_t$ of the CFG chain, for the prompt $\mathbf{p}$ = photo of a dog, at different time steps $t$. **Bottom:** ANSWER images corresponding to $t$ ANSWER steps and remaining CFG steps.

where $\hat{\epsilon}_n = \epsilon_\theta(z_t; t, \tau(\mathbf{n}))$. Under the energy-based interpretation, this corresponds to sampling from

$$p_{np}(z_t, \mathbf{p}) \propto p_\theta(z_t) \left[\frac{p_\theta(\mathbf{p}|z_t)}{p_\theta(\mathbf{n}|z_t)}\right]^s. \quad (6)$$

**Diffusion-Negative Prompting (DNP) [6]:** seeks the optimal negative prompt. It notes that Eq. (6) encourages the sampling of latent codes $z_t$ of large odds ratio

$$o(z_t, \mathbf{p}, \mathbf{n}) = \frac{p_\theta(\mathbf{p}|z_t)}{p_\theta(\mathbf{n}|z_t)}. \quad (7)$$

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, this increases the probability that sampled codes comply with prompt $\mathbf{p}$ and not with prompt $\mathbf{n}$. It then defines 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.

$$\mathbf{n}^*(\mathbf{p}) = \arg\max_{\mathbf{n}} \nabla_{z_t} \log o(z_t, \mathbf{p}, \mathbf{n}). \quad (8)$$

The authors of [6] show that this leads to the *Diffusion Negative Sampling* (DNS) equation

$$\hat{\epsilon} = \hat{\epsilon}_\phi + s(\hat{\epsilon}_\phi - \hat{\epsilon}_p) \quad (9)$$

for sampling optimal negative images. DNP samples one such image and captions it to estimate the optimal negative prompt, $\mathbf{n}^*$. This is finally used in the NP equation of Eq. (5) to synthesize the desired image.

Both guidance scale and negative prompting can increase the odds ratio in Eq. (6). However, the difference between them can be seen by comparing Eq. (4) and Eq. (6). In Eq. (4), increasing the guidance scale does not work well when $p_\theta(\mathbf{p}|z_t)$ is very low due to the seed choice. Given the same seed, we can use Eq. (8) to estimate the "optimal" negative prompt that can maximize the odds ratio and increase $p_{np}(z_t, \mathbf{p})$. Another difference is observed in practice, where increasing the guidance scale beyond a certain threshold causes the noise to be outside the distribution of the trained DM and can, therefore, result in over-saturation or artifacts. This can be avoided with negative prompting because the noise distribution remains unchanged.

# 3. Proposed Method

## 3.1. Sampling Procedures

Figure 3 shows the different sampling procedures discussed in this work. CFG uses a single Markov chain conditioned by prompts $\mathbf{p}$ and $\phi$ to produce the denoised latent $z_0$. DNP uses two, a DNS chain to estimate $\mathbf{n}^*$ and another to generate the final image using $(\mathbf{p}, \mathbf{n}^*)$.

$$\mathbf{n}^*(\mathbf{p}) = \mathcal{C}(\text{DNS}_{k=T}(z_T, \mathbf{p})) \quad (10)$$

where $\text{DNS}_k$ denotes $k$ DNS steps and $\mathcal{C}$ is an external captioning model. The prompt $\mathbf{n}^*(\mathbf{p})$ is then used, instead of $\phi$, in the NP chain to generate the final image. This has a few limitations. First, $\mathcal{C}$ can be much larger than the DM, increasing complexity. Second, much of the information in the negative image synthesized by DNS is "lost in translation". A caption is, by definition, a limited representation of image content, and $\mathcal{C}$ may not even pay attention to the visual details that make the image a negative from the DM perspective. Finally, and most importantly, the fundamental DNP assumption that a single caption $\mathbf{n}^*(\mathbf{p})$ is optimal for every diffusion step $t$, in Eq. (5), is questionable. This can be observed simply by inspection of Eqs. (7) and (8). Since the latent $z_t$ varies with $t$, so will the optimal prompt $\mathbf{n}^*(\mathbf{p})$.

## 3.2. Complete diffusion-Negative Prompting (CNP)

In this work, we hypothesize that this assumption does not hold, i.e., DNS produces a different negative image for the latent $z_t$ of each step $t$ and, consequently,

$$\text{DNS}_{k=T}(z_T, \mathbf{p}) \neq \text{DNS}_{k=t}(z_t, \mathbf{p}) \;\; \forall\, t \in [0, T). \quad (11)$$

To test this hypothesis, we used the DNS chain of Eq. (9) to generate the diffusion-negative image $I_{n,t}$ for different latents $z_t$, sampled with CFG for a given prompt $\mathbf{p}$ and latent $z_0$. The top half of Figure 4 shows the image $I_{n,t}$ for various values of $t$. As the Markov chain of Eq. (3) progresses and $z_t$ changes, the corresponding diffusion-negative image $I_{n,t}$ and corresponding optimal negative prompt, $\mathbf{n}^*(\mathbf{p})$ change as well. For example, the negative image for "a photo of a dog" depicts a living room, cyclist, baseball player, birds, and dogs as $t$ varies. To reflect this dependence clearly, we

modify the notation from $\mathbf{n}^*(\mathbf{p})$ to $\mathbf{n}^*(\mathbf{p}, t)$. This is also more consistent with Eqs. (7) and (8) and leads to

$$\mathbf{n}^*(\mathbf{p}, t) = \mathcal{C}(\text{DNS}_{k=t}(z_t, \mathbf{p})) \tag{12}$$

Comparing with Eq. (10) shows that the prompt $\mathbf{n}^*$ of DNP can be written as $\mathbf{n}^*(\mathbf{p}, T)$ in the new notation and is optimal only for the first DNP step ($t = T$).

To ensure optimality at every step, $\mathbf{n}^*(\mathbf{p}, t)$ must be estimated for every $t \in (0, T]$. This is denoted *Complete diffusion-Negative Prompting* (CNP). As illustrated in Figure 3, it consists of implementing Eq. (12) instead of Eq. (10) by initializing the DNS chain at $z_t$ and running it for $t$ steps, at each step of the NP chain. The resulting negative prompt $\mathbf{n}^*(\mathbf{p}, t)$ is then used in the next step of the NP chain, ensuring that the negative prompt used at every step is optimal for that step.

**Limitations of CNP:** While CNP eliminates DNP's limitation of using a single negative for all steps, the added complexity is substantially larger because Eq. (12) requires using $\mathcal{C}$ at every time step $t$. As $t$ approaches 0 (end of the chain), only small details are edited and $\mathcal{C}$ may ignore the details that differentiate the positive and negative conditions. For example, as shown in the top half of Figure 4 for $\mathbf{p}$ = "a photo of a dog", as $t$ approaches 0, the diffusion negative image also contains a dog (low quality, distorted, etc.). The correct negative prompt should be $\mathbf{n}^*(\mathbf{p}, t)$ = "a distorted dog", to improve the image quality. However, captioning models caption it as "dog", which is detrimental. Beyond this, the complexity of CNP is $O(T^2)$, which is intractable for most applications. Hence, while better justified mathematically, in practice, CNP can be intractable and more prone to captioning errors.

### 3.3. ANSWER

To address these problems, we propose the *Adaptive Negative Sampling Without External Resources* (ANSWER) sampling procedure of Figure 3. This is a variant of CNP that eliminates the external captioning model. At step $t$, rather than producing a negative caption $\mathbf{n}^*(\mathbf{p}, t)$ and using it to condition the noise term $\hat{\epsilon}_n$ of Eq. (5), this noise term is replaced by an estimate produced by running a few ($K_t$) steps of the DNS chain, namely the negative noise produced by this chain at time $t - K_t$. The intuition is that the solution of Eq. (8) is actually the term $-\hat{\epsilon}_p$ of Eq. (9). The remaining $\hat{\epsilon}_\phi$ terms result from a normalization to produce a valid probability distribution and some arguments regarding the need to always define negative conditions in context [7] (see [6] for details). While for DNS that context is the distribution of natural images, associated with prompt $\phi$, the NP chain of Eq. (5) already establishes a context: the noise distribution $\hat{\epsilon}_p$ associated with the positive prompt $\mathbf{p}$. Note that if the DNS chain were run for a single iteration ($K_t = 1$) and the same denoising network used in the DNS and NP chains,

**Algorithm 1:** ANSWER

**Input**: DM $\epsilon_\theta$, text prompt $\mathbf{p}$, initial noise, $z \sim \mathcal{N}(0, I)$, number of timesteps $T$, number of DNS timesteps $K$, guidance scale $s$, negative guidance scale $s_n$, scheduler $Schdlr$

**for** $t \leftarrow T$ **to** 1 **do**
- $\epsilon_p^t \leftarrow \epsilon_\theta(z, \mathbf{p}, t), \epsilon_\phi^t \leftarrow \epsilon_\theta(z, \text{" "}, t)$ ;
- $K_t \leftarrow Schdlr(K, t)$ ;
- **if** $K_t > 0$ *and* $t > T/2$ **then**
  - $z^n \leftarrow z$ ;
  - **for** $i_n \leftarrow 0$ **to** $K_t - 1$ **do**
    - $t_n \leftarrow t$ - $i_n$ ;
    - $\epsilon_p^{t_n} \leftarrow \epsilon_\theta(z^n, \mathbf{p}, t_n)$
    - $\epsilon_\phi^{t_n} \leftarrow \epsilon_\theta(z^n, \text{" "}, t_n)$ ;
    - $\epsilon^{t_n} \leftarrow \epsilon_p^{t_n} + s_n(\epsilon_\phi^{t_n} - \epsilon_p^{t_n})$ // DNS;
    - $z^n \leftarrow Sample(z^n, \epsilon^{t_n}, t_n)$ ;
  - **end**
  - $\epsilon_n^t \leftarrow Normalize(\epsilon^{t_n}, t)$ ;
  - $\epsilon^t \leftarrow \epsilon_n^t + s(\epsilon_p^t - \epsilon_n^t)$ // NP;
- **end**
- **else**
  - $\epsilon^t \leftarrow \epsilon_\phi^t + s(\epsilon_p^t - \epsilon_\phi^t)$ // CFG;
- **end**
- $z \leftarrow Sample(z, \epsilon^t, t)$ ;

**end**

Eq. (5) would reduce to $\hat{\epsilon} = \hat{\epsilon}_\phi + (s^2 - s + 1)(\hat{\epsilon}_p - \hat{\epsilon}_\phi)$, which is equal to simply increasing the guidance scale of CFG. While this effectively increases sampling odds, $\epsilon_{n^*}$ remains $\epsilon_\phi$, and the Markov chain is not substantially changed. Using a small $K_t > 1$ allows $\epsilon_{n^*}$ to diverge from $\epsilon_\phi$ and further increase the odds ratio of Eq. (6). This also makes the sampling highly adaptive to the prompt $\mathbf{p}$ and can be implemented with no external resources, neither in the form of a captioning model nor a user-specified negative prompt.

An implementation of ANSWER is detailed in Algorithm 1. At diffusion step $t$, $K_t$ iterations of DNS sampling are implemented with Eq. (9). $K_t$ is the number of DNS timesteps and is determined by a scheduler, $Schdlr(K, t) = K \cdot \left(\frac{2t-T}{2t-T+20}\right) \cdot \left(\frac{T+20}{T}\right)$, where $T$ is the total number of timesteps and $t \in \{T, .., T/2\}$ is the current timestep. This gradual decay, starting at $K_T = K$, allows a smoother transition between negative and unconditional noise. The negative noise $\epsilon^{t_n}$ is normalized to match the noise statistics $\epsilon_\phi^t$ using $Normalize(\epsilon^{t_n}, t) = \left(\frac{\epsilon^{t_n} - \mu(\epsilon^{t_n})}{\sigma(\epsilon^{t_n})}\right) \cdot \sigma(\epsilon_\phi^t) + \mu(\epsilon_\phi^t)$. This is helpful because the noise distributions change with time step $t$, and the negative noise is $K_t$-steps ahead. The normalized noise $\epsilon_n^t$ is then passed on to the NP step of Eq. (5). Experimentally, this negative prompting step is more effective for early time steps $t$, where the image is still poorly formed but does not

| Dataset | Method | CLIP | IS | Image Reward | Pick Score | HPSv2 |
|---|---|---|---|---|---|---|
| Attend&Excite | SDXL (CFG) | 32.97 (27%) | 11.24 | 0.95 (26%) | 22.73 (33%) | 30.07 (24%) |
| | SDXL+DNP | 33.20 (21%) | **11.65** | 1.03 (25%) | 22.79 (26%) | 30.38 (26%) |
| | SDXL+ANSWER | **33.62 (52%)** | 11.15 | **1.20 (49%)** | **22.88 (41%)** | **30.92 (50%)** |
| Pick-a-Pic | SDXL (CFG) | 33.42 (27%) | 11.11 | 0.91 (24%) | 22.14 (29%) | 30.24 (24%) |
| | SDXL+DNP | 33.47 (21%) | **11.67** | 0.85 (25%) | 22.08 (23%) | 30.47 (25%) |
| | SDXL+ANSWER | **34.09 (52%)** | 11.28 | **1.00 (51%)** | **22.39 (48%)** | **30.95 (51%)** |
| DrawBench | SDXL (CFG) | 32.59 (28%) | 14.88 | 0.61 (28%) | 22.37 (32%) | 28.86 (24%) |
| | SDXL+DNP | 32.87 (22%) | **15.34** | 0.71 (26%) | 22.45 (24%) | 29.38 (30%) |
| | SDXL+ANSWER | **33.27 (50%)** | 15.19 | **0.79 (46%)** | **22.56 (44%)** | **29.45 (48%)** |
| PartiPrompts | SDXL (CFG) | 32.33 (28%) | 16.35 | 0.65 (26%) | 22.27 (33%) | 28.43 (23%) |
| | SDXL+DNP | 32.02 (22%) | 16.45 | 0.62 (26%) | 22.18 (23%) | 28.72 (31%) |
| | SDXL+ANSWER | **32.91 (50%)** | **16.55** | **0.81 (48%)** | **22.39 (44%)** | **28.97 (47%)** |

Table 1. Quantitative results on the Attend&Excite, Pick-a-Pic, DrawBench, and PartiPrompts datasets. %WinRates are shown in parentheses next to each metric. Best values are in **bold**.

| Dataset | Method | CLIP | IS | FID | Image Reward | Pick Score | HPSv2 |
|---|---|---|---|---|---|---|---|
| ImageNet | SDXL (CFG) | 30.22 | 27.26 | 69.8 | 0.21 | 21.24 | 25.63 |
| | SDXL+DNP | 30.19 | 34.55 | 70.11 | 0.26 | 21.30 | 26.80 |
| | SDXL+ANSWER | **30.64** | **40.83** | **67.67** | **0.48** | **21.46** | **26.87** |

Table 2. Quantitative Results on the Imagenet dataset. Best results are in **Bold**.

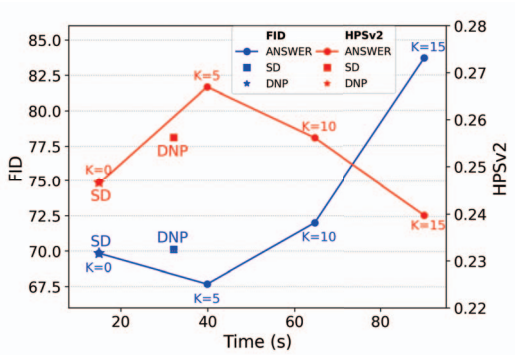


Figure 5. Inference time Pareto curves over K: FID (Left axis) and HPSv2 (Right axis)


Figure 6. Complex and imaginary scenarios: (Left: SDXL (CFG), Right: SDXL+ANSWER)

help in the later steps. For example, the bottom half of Figure 4 shows the images created by performing $t$ ANSWER steps followed by the $T - t$ CFG steps. Therefore, applying ANSWER beyond a certain number of steps in the NP chain does not increase performance significantly. In this example, the negative and positive chain converge around $t = 12$, and the positive image does not change much beyond that point. Additionally, latent editing literature [5, 23] observes that only cosmetic changes can be made in the final timesteps. Therefore, to save time and resources, we replace the NP chain of Eq. (5) with the standard CFG chain for $t < T/2$. Restricting to the early stages also reduces inference time of ANSWER. Stopping criteria can be researched, as different prompt/seed pairs may need more or fewer steps based on difficulty.

Note that, rather than operating in the text domain, ANSWER defines the negative prompt directly in noise space. This eliminates the "lost in translation" problem associated with captions. Since ANSWER does not estimate the negative prompt at each step, it does not need to complete the Markov chain and can use any $\epsilon^{t_n}$ from the DNS chain as a substitute for the negation. This allows choosing the number of steps $K$, thereby reducing the computational complexity of ANSWER to $O(KT)$, which is tractable for low values of $K$. Overall, ANSWER captures dynamic changes in $\mathbf{n}^*(\mathbf{p}, t)$ through each step without requiring full sampling to the final step. As a result, sampling with ANSWER is tractable, independent of external models, and estimates an adaptive negative noise for every step in the diffusion chain. Using the negative noise obtained from DNS in the diffusion process bridges the semantic gap more effectively. This also allows ANSWER to capitalize on the flexibility of DNS sampling to improve efficiency and effectiveness for longer or more complex prompts.

# 4. Experiments

We discuss the experiments we performed to evaluate ANSWER in this section. Comprehensive details about datasets and evaluation metrics with additional results can be found in the Supplementary.

**Datasets:** We evaluate ANSWER using five prominent datasets to assess its ability to generate images with comprehensive and diverse descriptions that extend beyond common training data. Specifically, we utilize the ImageNet (IM) [26], Attend & Excite (AE) [5], Pick-a-Pic (PK) [15], DrawBench (DB) [27], and PartiPrompts (P2) [37] datasets. These datasets test the model's ability to handle entity attributes and interactions, spatial relationships, text rendering, numeracy, and rare or imaginative scenarios.

**Evaluation metrics:** We use both human and automated evaluation metrics for the large-scale assessment of the generated images. The CLIP Score [10], FID [11] and Inception Score (IS) [28] are used as the traditional measures along with emerging popular Human Preference Metrics, such as HPSv2 [35], ImageReward [36] and PickScore [15]. Since, they are trained as preference metrics, not for absolute scoring, their absolute values are not highly informative of visual quality. Since, small metric deltas can correspond to large differences in visual quality, we also report win-rate percentages to offer further insights into generation quality.

## 4.1. Quantitative Results

We evaluate ANSWER on Stable Diffusion (SD) and Stable Diffusion XL (SDXL). Table 1 compares the performance of models using ANSWER to the respective baselines and their negative prompting with DNP, on the AE dataset, Pick-a-Pic, DrawBench, and PartiPrompts. SD+ANSWER outperforms both SD (CFG) and SD+DNP for all metrics except IS,

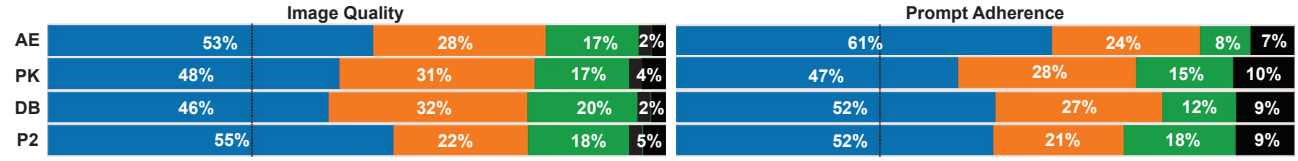


Figure 7. **Human evaluation using Amazon Mechanical Turk:** From left to right: SDXL+ANSWER is denoted by **blue**, SDXL+DNP denoted by **orange**, SDXL (CFG) is denoted by **green** and *'No Clear Winner'* is **black**.

under which it has slightly inferior performance compared to SD+DNP. SDXL (CFG) also mirrors this behavior where SDXL+ANSWER outperforms on all metrics except IS. This shows that SDXL+ANSWER generates images that are both prompt-compliant and natural.

Table 2 shows a similar comparison for the ImageNet dataset. For all datasets, SDXL+ANSWER outperforms SDXL (CFG) and SDXL+DNP across all metrics. To ensure that ANSWER improves correctness and quality without reducing the diversity of generated images, we also show FID on the ImageNet dataset in Table 2. Note that SDXL+ANSWER has a lower FID score (which implies high diversity) than SDXL (CFG) and SDXL+DNP while maintaining higher prompt adherence.

While these metrics show that sampling with ANSWER improves the quality and prompt adherence of the images generated by all models, they do not convey an intuitive understanding of the significance of these gains. This is mostly because relatively small differences in the metrics can correspond to significant improvements in image synthesis. To provide better insight into the performance of the methods, we show win rate of each method in parenthesis Table 1. SDXL+ANSWER wins $41-52\%$ of the time with the remaining split between SDXL (CFG) and SDXL+DNP. These results show SDXL+ANSWER is preferred $2\times$ over SDXL (CFG) or SDXL+DNP.

Human evaluations remain the gold standard for this assessment. For example, we have observed that adding ANSWER sampling nudges SDXL (CFG) toward more realistic images when the model is inclined to create paintings or cartoonish imagery. Generally, this makes images produced using ANSWER preferable to humans. Figure 7 presents the results of the AMT evaluation, showing that humans prefer the image generated using SDXL+ANSWER over that synthesized by SDXL (CFG) for both quality and adherence. SDXL+ANSWER is preferred $46-61\%$ of the time, SDXL+DNP is preferred $21-32\%$ of the time, and SDXL (CFG) is preferred $8-20\%$ of the time. Most *'No clear winner'* cases correspond to situations where the prompt was too easy (all methods created similar or good enough images) or too hard (all methods failed to generate a correct image). This corroborates HPSv2's 2-to-1 win rate of SDXL+ANSWER over the baselines. When the prompt is too hard, increasing the value of $K$ in ANSWER allows corrections that no amount of guidance scale increase in CFG can achieve. However, the results above do not capture this advantage since $K$ was kept fixed in these experiments.

### 4.2. Qualitative Results

Qualitatively, we show the impact of adding ANSWER sampling to SDXL (CFG) for sample prompts all datasets in Figure 1, where each image pair was synthesized with the same seed and guidance scale. These examples show the extent to which ANSWER produces better-quality images, illustrating how it is a general solution to many problems of current diffusion models. In particular, it shows that while SDXL (CFG) tends to generate paintings or cartoons for unnatural or complex prompts like "a confused grizzly bear", SDXL+ANSWER is more inclined towards realistic or life-like images. This is also evident in the prompts "a cat and a dog" and "a mouse and a pink bow".

We observe that SDXL+ANSWER can add missing objects with their corresponding attributes, for example, the dog in "a cat and a dog" and the glasses in "a white bench and purple glasses". SDXL+ANSWER can also help with one of diffusion's most well-known weaknesses, anatomy, e.g., generating more anatomically correct hands, with no extra or missing fingers, as shown in the images for "a girl diving into the pool" and "a diamond ring on a girl's hand". Finally, ANSWER sampling allows SDXL (CFG) to better focus on spatial relations such as "to the left of", text rendering, and numeracy.

Even for rare or imaginary prompts such as "Burger with wheels on a racetrack," shown in Figure 6, where SDXL shows extreme resistance towards generating this unreal scenario, maximizing the odds ratio using ANSWER forces compliance with the input prompt. For the prompt "Dog and Santa. Black and white Christmas trees in the background", ANSWER not only ensures the presence of a dog and Santa Claus, but also enforces the complex requirement of black and white Christmas trees.

## 5. Ablations

**Number of DNS steps:** The most important hyperparameter of ANSWER is $K$, which determines the number of DNS steps performed at each iteration of the NP chain. While longer steps allow bigger corrections to the chain, there is a trade-off. When $K$ is too large, the mismatch between the DNS and NP chains produces a noise-smoothing effect that results in blurred images. In Figure 5,

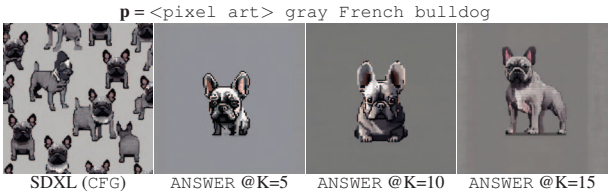

Figure 8. Effect of DNS hyperparameter $K$ on ANSWER.

we show how FID and HPSv2 vary with run time for different $K$ using the SDXL (CFG) model and the ImageNet dataset [26]. Lower values of $K$ are equivalent to SDXL. As $K$ increases, both HPSv2 and FID improve. However, performance drops when $K$ is too high because the negative chain progresses too far, creating a large distribution mismatch and disrupting the Markov chain. This can be seen in the example of Figure 8, which shows the effect of the correction introduced by ANSWER on SDXL (CFG) as $K$ increases. While small values of $K$ improve prompt adherence, the image quality decreases beyond a certain threshold (for $K = 15$ in this example). We empirically choose $K = 5$ in our implementation and use it for all experiments in the paper.

We note, however, that for individual prompts, other values of $K$ may achieve a better trade-off between prompt compliance and image quality. For simple prompts, SDXL suffices ($K = 0$). For complex prompts, ANSWER enforces prompt adherence when SDXL fails even with the highest guidance scale. Overall, the throughput of SDXL+ANSWER can be controlled by limiting $K$ to as little as required for prompt compliance. More research is needed to automate the selection of $K$.

**Effect of Normalization:** Figure 9 shows the effect of the mean and variance normalization performed in ANSWER (*Normalize* step of Algorithm 1). While ANSWER corrects the image and removes the front of a second car from the SDXL (CFG) image, it causes blurring when normalization is not used. This is because the DNS chain is $K$ steps ahead of the NP chain, which creates a mismatch in the noise statistics. Normalization corrects this and helps eliminate this blurring effect, as seen in Figure 9.

## 6. Related Work

**Diffusion Models:** There is now a plethora of T2I models based on scalable model architectures [20–22, 24, 27] trained on large datasets [4, 29]. Many models leverage CFG [12] to balance prompt compliance and diversity. This enables the synthesis of higher-quality images for more complex prompts. ANSWER will, in principle, benefit any model that leverages CFG.

**Visual Conditioning & Guidance:** Some works have introduced forms of conditioning other than text prompts, such as layouts [17, 19, 39], example images [3, 16, 25], sketches, or depth maps [38]. While these conditions add structure and constraints to diffusion models (DMs), they

Figure 9. Effect of normalization on ANSWER.

are cumbersome or require skill. In [32] they address this by training a chain of models to generate visual conditions using adapters. However, these methods require additional resources and/or training. ANSWER requires no external resources and is intuitive since only a text prompt is needed.

**Text Conditioning:** Several approaches focus on text-only conditioning. Works like [8, 18] aim to correct errors induced by the DM text encoder by splitting the prompt into its constituent tokens or noun phrases, while, [9] accepts Word document-like text formatting (e.g., font style, size, color, and footnotes) as input and applies the corresponding condition to each token based on the attention map. In [5, 23], they dynamically update latents by adjusting cross-attention between prompt tokens to enhance the presence and binding of attributes, respectively. These methods directly edit the latent at each step, which can significantly alter the Markov chain of the DM, frequently resulting in overly saturated and low-quality images. Recently, [6] has shown that negative prompting with DNP achieves better performance. Since ANSWER sampling uses NP sampling and updates latents based solely on DM-generated noise, it is much less prone to this problem.

**Negative Prompting:** The inference-time guidance introduced by [18] enables concept negation in diffusion models. Due to its effectiveness in filtering out unwanted elements from generated images, it has been widely embraced by most T2I models. However, crafting an effective negative prompt requires trial and error. To address this, [6] proposed DNP to learn a negative prompt optimized directly in diffusion space. Recently, [2] attempted to use corresponding text tokens as negatives by selectively removing them. Although effective, these techniques have limitations inherent to external captioning and latent manipulation by attention maps that do not affect ANSWER.

## 7. Conclusion

In this work, we hypothesize that the optimal negation for a diffusion model changes as the Markov chain proceeds. We then develop a sampling scheme, ANSWER, that uses adaptive negative sampling to correct the diffusion Markov chain without any external inputs or conditions. Our experiments show that updating the diffusion negative at each step generates images of higher quality and better prompt adherence. ANSWER requires no training or additional input and can be used with any model that supports classifier-free guidance.

## 8. Acknowledgements

This work was partially funded by NSF award IIS-2303153, NAIRR-240300, 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. 3

[2] Kambiz Azarian, Debasmit Das, Qiqi Hou, and Fatih Porikli. Segmentation-free guidance for text-to-image diffusion models, 2024. 8

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 8

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 8

[5] 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, 2023. 2, 6, 8

[6] Alakh Desai and Nuno Vasconcelos. Improving image synthesis with diffusion-negative sampling, 2024. 2, 4, 5, 8

[7] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In *Advances in Neural Information Processing Systems*, pages 6637–6647. Curran Associates, Inc., 2020. 5

[8] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. 8

[9] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 8

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 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. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 8

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 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] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 8

[17] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1, 8

[18] 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, 3, 8

[19] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 1, 8

[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. 1, 8

[21] 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.

[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. 1, 8

[23] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment, 2023. 2, 6, 8

[24] 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. 8

[25] 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. 2022. 8

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6, 8

[27] 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. 1, 6, 8

[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. 6

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294. Curran Associates, Inc., 2022. 8

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[32] Deepak Sridhar, Abhishek Peri, Rohith Rachala, and Nuno Vasconcelos. Adapting diffusion models for improved prompt compliance and controllable image synthesis, 2024. 8

[33] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models, 2022. 1

[34] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 3

[35] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. 6

[36] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 6

[37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 6

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 8

[39] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1, 8