# Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models

HILA CHEFER*, Tel-Aviv University, Israel
YUVAL ALALUF*, Tel-Aviv University, Israel
YAEL VINKER, Tel-Aviv University, Israel
LIOR WOLF, Tel-Aviv University, Israel
DANIEL COHEN-OR, Tel-Aviv University, Israel

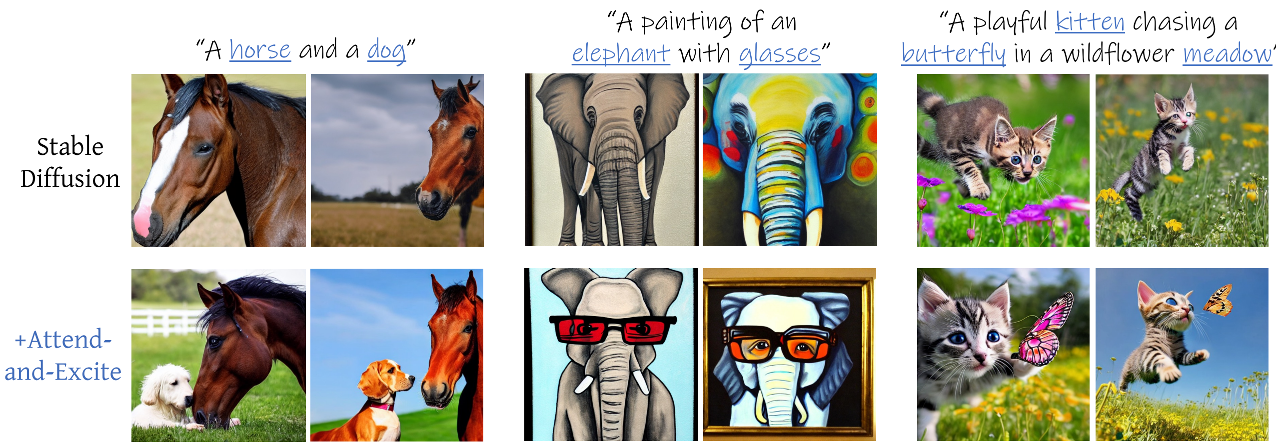


Fig. 1. Given a pre-trained text-to-image diffusion model (*e.g.*, Stable Diffusion [Rombach et al. 2022]) our method, Attend-and-Excite, guides the generative model to modify the cross-attention values during the image synthesis process to generate images that more faithfully depict the input text prompt. Stable Diffusion alone (top row) struggles to generate multiple objects (*e.g.*, a horse and a dog). However, by incorporating Attend-and-Excite (bottom row) to strengthen the subject tokens (marked in blue), we achieve images that are more semantically faithful with respect to the input text prompts.

Recent text-to-image generative models have demonstrated an unparalleled ability to generate diverse and creative imagery guided by a target text prompt. While revolutionary, current state-of-the-art diffusion models may still fail in generating images that fully convey the semantics in the given text prompt. We analyze the publicly available Stable Diffusion model and assess the existence of *catastrophic neglect*, where the model fails to generate one or more of the subjects from the input prompt. Moreover, we find that in some cases the model also fails to correctly bind attributes (*e.g.*, colors) to their corresponding subjects. To help mitigate these failure cases, we introduce the concept of *Generative Semantic Nursing (GSN)*, where we seek to intervene in the generative process on the fly during inference time to improve the faithfulness of the generated images. Using an attention-based formulation of GSN, dubbed *Attend-and-Excite*, we guide the model to refine the cross-attention units to *attend* to all subject tokens in the text prompt and strengthen — or *excite* — their activations, encouraging the model to generate all subjects described in the text prompt. We compare our approach to alternative approaches and demonstrate that it conveys the desired concepts more faithfully across a range of text prompts. Code is available at our project page: https://attendandexcite.github.io/Attend-and-Excite/.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Image processing**.

Additional Key Words and Phrases: Image Generation, Diffusion Models

**ACM Reference Format:**
Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.* 42, 4, Article 148 (August 2023), 10 pages. https://doi.org/10.1145/3592116

*Denotes equal contribution.

Authors' addresses: Hila Chefer, Tel-Aviv University, Israel, hilach70@gmail.com; Yuval Alaluf, Tel-Aviv University, Israel, yuvalalaluf@gmail.com; Yael Vinker, Tel-Aviv University, Israel, yael.vinker@mail.huji.ac.il; Lior Wolf, Tel-Aviv University, Israel, liorwolf@gmail.com; Daniel Cohen-Or, Tel-Aviv University, Israel, cohenor@gmail.com.


0730-0301/2023/8-ART148 $15.00
https://doi.org/10.1145/3592116

## 1 INTRODUCTION

Recent advancements in text-based image generation [Balaji et al. 2022; Gafni et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022] have demonstrated an unprecedented ability to generate diverse and creative imagery provided a free-form text prompt. However, it has been shown [Feng et al. 2022; Wang et al.

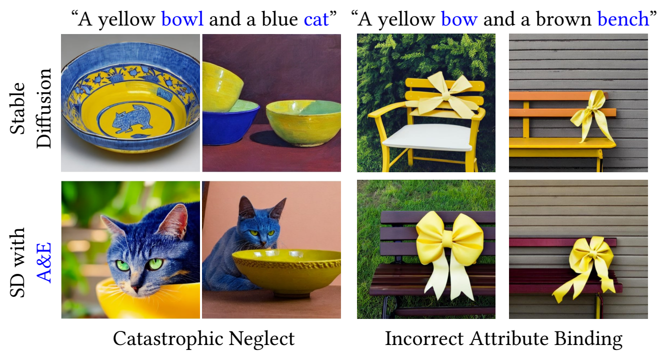


Fig. 2. Failure cases of Stable Diffusion (SD) [Rombach et al. 2022]. In the top row, we show examples of two failure settings: catastrophic neglect (left) and incorrect attribute binding (right). In the bottom row, we show images obtained when applying Attend-and-Excite over SD using the same seeds.

2022] that images produced by such models do not always faithfully reflect the semantic meaning of the target prompt.

We observe two key semantic issues in state-of-the-art text-based image generation models: (i) "catastrophic neglect", where one or more of the subjects of the prompt are not generated; and (ii) incorrect "attribute binding", where the model binds attributes to the wrong subjects or fails to bind them entirely. Examples of cases where the aforementioned issues arise can be found in the top row of Figure 2, which depicts images generated by the state-of-the-art Stable Diffusion model [Rombach et al. 2022]. In the left column, we provide an example of catastrophic neglect where the model fails to generate the blue cat, choosing to focus solely on generating the bowl. In the right column, we demonstrate incorrect attribute binding where the color "yellow" is incorrectly binded to the bench.

To mitigate these semantic issues, we introduce the concept of *"Generative Semantic Nursing" (GSN)*. In the GSN process, one slightly shifts the latent code at each timestep of the denoising process such that the latent is encouraged to better consider the semantic information passed from the input text prompt.

We propose a form of GSN dubbed *Attend-and-Excite*, which leverages the powerful cross-attention maps of a pre-trained diffusion model. The attention maps define a probability distribution over the text tokens for each image patch, which determines the dominant tokens in the patch. We observe that this text-image interaction is susceptible to neglect. Although each patch can attend freely to all text tokens, there is no mechanism to ensure that *all* tokens are attended to by some patch in the image. In cases where a subject token is not attended to, the corresponding subject will not be manifested in the output image.

Thus, intuitively, in order for a subject to be present in the generated image, the model should assign at least one image patch to the subject's token. Attend-and-Excite embodies this intuition by demanding that each subject token is dominant in some patch in the image. We carefully guide the latent at each denoising timestep and encourage the model to *attend* to all subject tokens and strengthen — or *excite* — their activations. Importantly, our approach is applied on the fly during inference time and requires no additional training or fine-tuning. We instead choose to preserve the strong semantics already learned by the pre-trained generative model and text encoder. Example generations with our approach applied over Stable Diffusion are shown in the bottom row of Figure 1.

As shall be demonstrated, although Attend-and-Excite explicitly tackles only the issue of catastrophic neglect, our solution implicitly encourages correct bindings between attributes and their subjects. This can be attributed to the connection between the two issues of catastrophic neglect and attribute binding. The embedding of the text, obtained by a pre-trained text encoder, links information between each subject and its corresponding attributes. For example, in the prompt "a yellow bowl and a blue cat", the token "cat" receives information from the token "blue" during the text encoding process. Therefore, mitigating catastrophic neglect over the cat should ideally result in enhancing the color attribute (*i.e.*, allowing for correct binding between "cat" and "blue").

We demonstrate Attend-and-Excite's superiority in generating semantically-faithful images over Stable Diffusion and alternative methods that explore similar semantic issues. We additionally analyze the cross-attention maps realized with and without Attend-and-Excite and demonstrate the importance of applying our method to mitigate catastrophic neglect, while enabling the use of cross-attention as a form of explanation for the generated content.

## 2 RELATED WORK

Early works studied text-guided image synthesis in the context of GANs [Tao et al. 2022; Xu et al. 2018; Ye et al. 2021; Zhang et al. 2021; Zhu et al. 2019]. More recently, impressive results were achieved with large-scale auto-regressive models [Ramesh et al. 2021; Yu et al. 2022] and diffusion models [Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022]. Yet, generating images that faithfully align with the input prompt is often difficult. To enforce heavier reliance on the text, classifier-free guidance [Ho and Salimans 2022; Nichol et al. 2021; Saharia et al. 2022] allows extrapolating text-driven gradients to better guide the generation by strengthening the reliance on the text. However, even when employing this technique, extensive prompt engineering is often required to achieve the expected result [Liu and Chilton 2022; Marcus et al. 2022; Wang et al. 2022; Witteveen and Andrews 2022].

To provide users with more control over the synthesis process, several works employ a segmentation map or spatial conditioning [Avrahami et al. 2022b; Gafni et al. 2022; Zhao et al. 2019]. In the context of image editing, while most methods are generally limited to global edits [Chefer et al. 2022a; Crowson et al. 2022; Gal et al. 2022b; Kwon and Ye 2022], several works introduce a user-provided mask to specify the region that should be altered [Avrahami et al. 2022a; Bau et al. 2021; Couairon et al. 2022; Nichol et al. 2021].

Another related line of work aims to introduce specific concepts to a pre-trained text-to-image model by learning to map a set of images to a "word" in the embedding space of the model [Gal et al. 2022a; Kumari et al. 2022; Ruiz et al. 2022]. Several works have also explored providing users with more control over the synthesis process solely through the use of the input text prompt [Brooks et al. 2022; Hertz et al. 2022; Kawar et al. 2022; Valevski et al. 2022].

Recently, two works have explored the semantic flaws of text-to-image models. First, Liu *et al.* [2022] propose Composable Diffusion models where an image is generated by composing multiple outputs of a pre-trained diffusion model. Each output is tasked with capturing different image components which are then joined using compositional operators to attain a unified image. Yet, we observe that this method often struggles in achieving realistic compositions of multiple objects (see Section 5). Moreover, the approach is limited to operating over conjunctions and negations of subjects.

Feng *et al.* [2022] propose StructureDiffusion which employs consistency trees or scene graphs to split the prompt into several noun phrases. An attention map is computed for each noun phrase and the output of the cross-attention unit is the average of all attention operations. In contrast, our Attend-and-Excite technique directly optimizes the noised latent, allowing us to synthesize images that vary significantly from those produced by Stable Diffusion. We find that results obtained by StructureDiffusion often resemble those produced by Stable Diffusion, falling short of achieving meaningful modifications that amend the semantic faults (see Section 5).

It should be noted that there are additional semantic issues in text-based image synthesis, *e.g.*, object relations and compositions. Addressing such issues may require additional models to determine the object relations [Ashual and Wolf 2019; Johnson et al. 2018]. However, this deviates from the scope of this work where we focus on inference-time guidance of a pre-trained generative model.

## 3 PRELIMINARIES

*Latent Diffusion Models.* We apply our method over the state-of-the-art Stable Diffusion model (SD) [Rombach et al. 2022]. Instead of operating in the image space, SD operates in the latent space of an autoencoder. First, an encoder $\mathcal{E}$ is trained to map a given image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$. A decoder $\mathcal{D}$ is then tasked with reconstructing the input image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

Given the trained autoencoder, a denoising diffusion probabilistic model (DDPM) [Ho et al. 2020; Sohl-Dickstein et al. 2015] operates over the learned latent space to produce a denoised version of an input latent $z_t$ at each timestep $t$. During the denoising process, the diffusion model can be conditioned on an additional input vector. In Stable Diffusion, this additional input is typically a text encoding produced by a pre-trained CLIP text encoder [Radford et al. 2021]. Given a conditioning prompt $y$, we denote the conditioning vector by $c(y)$. The DDPM model $\varepsilon_\theta$ is trained to minimize the loss,

$$\mathcal{L} = \mathbb{E}_{z \sim \mathcal{E}(x), y, \varepsilon \sim \mathcal{N}(0,1), t} \left[ ||\varepsilon - \varepsilon_\theta(z_t, t, c(y))||_2^2 \right]. \quad (1)$$

In words, at each timestep $t$, the denoising network $\varepsilon_\theta$ is tasked with correctly removing the noise $\varepsilon$ added to the latent code $z$, given the noised latent $z_t$, timestep $t$, and conditioning encoding $c(y)$. Here, $\varepsilon_\theta$ is a UNet network [Ronneberger et al. 2015] consisting of self-attention and cross-attention layers, discussed below.

At inference, a latent $z_T$ is sampled from $\mathcal{N}(0, 1)$ and is iteratively denoised to produce a latent $z_0$ using the DDPM. The denoised latent is then passed to the decoder to obtain the image $x' = \mathcal{D}(z_0)$.

*Text-Conditioning Via Cross-Attention.* Text guidance in Stable Diffusion is performed using the cross-attention mechanism. The denoising UNet network consists of self-attention layers followed by cross-attention layers at resolutions of 64, 32, 16, and 8.

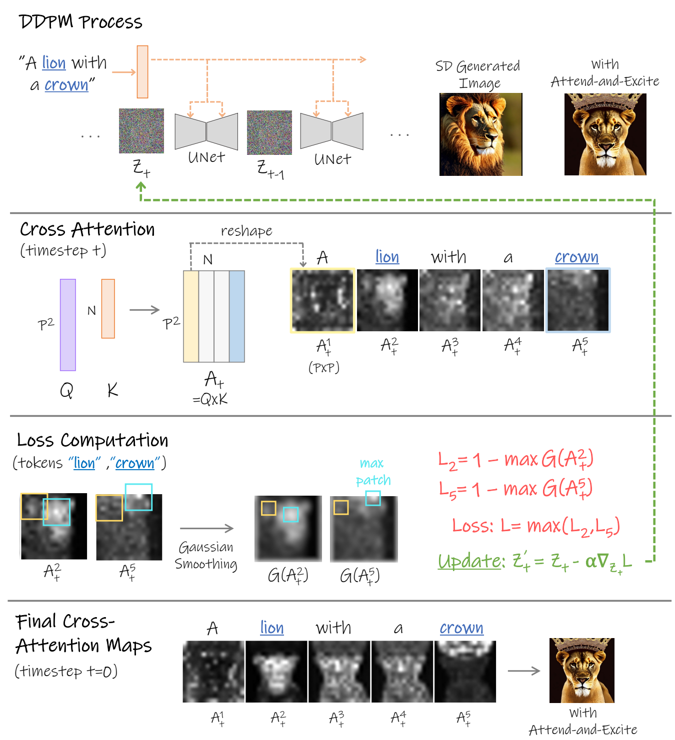


Fig. 3. Overview of Attend-and-Excite. Given a prompt (*e.g.* "A lion with a crown"), we extract the subject tokens (lion, crown), and their corresponding attention maps ($A_t^2, A_t^5$). We apply a Gaussian kernel on each attention map to obtain smoothed attention maps that consider the neighboring patches. Our optimization enhances the maximal activation for the most neglected token at timestep $t$ and updates the latent code $z_t$ accordingly. The final cross-attention maps at $t = 0$ are illustrated in the final row.

Denote by $P$ the spatial dimension of the intermediate feature map (*i.e.*, $P \in \{64, 32, 16, 8\}$), and by $N$ the number of text tokens in the prompt. An attention map $A_t \in \mathbb{R}^{P \times P \times N}$ is calculated over linear projections of the intermediate features ($Q$) and text embedding ($K$), as illustrated in the second row of Figure 3. $A_t$ defines a distribution over the text tokens for each spatial patch $(i, j)$. Specifically, $A_t[i, j, n]$ denotes the probability assigned to token $n$ for the $(i, j)$-th spatial patch of the intermediate feature map. Intuitively, this probability indicates the amount of information that will be passed from token $n$ to patch $(i, j)$. Note that the maximum value of each of the $P \times P$ cells is 1.

We operate over the $16 \times 16$ attention units since they have been shown to contain the most semantic information [Hertz et al. 2022].

## 4 ATTEND-AND-EXCITE

At the core of our method is the idea of *generative semantic nursing*, where we gradually shift the noised latent code at each timestep $t$ toward a more semantically-faithful generation. At each denoising step $t$, we consider the attention maps of the subject tokens in the prompt $\mathcal{P}$. Intuitively, for a subject to be present in the synthesized image, it should have a high influence on some patch in the image. As such, we define a loss objective that attempts to maximize the

**Algorithm 1** A Single Denoising Step using Attend-and-Excite

**Input:** A text prompt $\mathcal{P}$, a set of subject token indices $\mathcal{S}$, a timestep $t$, a set of iterations for refinement $\{t_1, \dots, t_k\}$, a set of thresholds $\{T_1, \dots, T_k\}$, and a trained Stable Diffusion model $SD$.
**Output:** A noised latent $z_{t-1}$ for the next timestep

1: $\_, A_t \leftarrow SD(z_t, \mathcal{P}, t)$
2: $A_t \leftarrow \text{Softmax}(A_t - \langle sot \rangle)$
3: **for** $s \in \mathcal{S}$ **do**
4: $\quad A_t^s \leftarrow A_t[:, :, s]$
5: $\quad A_t^s \leftarrow \text{Gaussian}(A_t^s)$
6: $\quad \mathcal{L}_s \leftarrow 1 - \max(A_t^s)$
7: **end for**
8: $\mathcal{L} \leftarrow \max_s(\mathcal{L}_s)$
9: $z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$
10: **if** $t \in \{t_1, \dots, t_k\}$ **then** ▷ If performing iterative refinement at $t$
11: $\quad$ **if** $\mathcal{L} > 1 - T_t$ **then**
12: $\quad\quad z_t \leftarrow z_t'$
13: $\quad\quad$ **Go to** Step 1
14: $\quad$ **end if**
15: **end if**
16: $z_{t-1}, \_ \leftarrow SD(z_t', \mathcal{P}, t)$
17: **Return** $z_{t-1}$

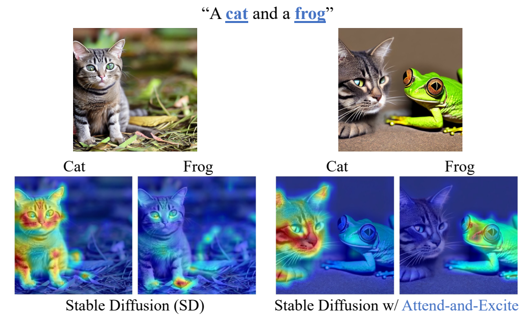


Fig. 4. Visualization of the cross-attention maps for each subject token with and without Attend-and-Excite over Stable Diffusion.

attention values for each subject token. We then update the noised latent at time $t$ according to the gradient of the computed loss. This encourages the latent at the next timestep to better incorporate all subject tokens in its representation. This manipulation occurs on the fly during inference (*i.e.*, no additional training is performed).

In the next sections, we discuss each of the steps presented in Algorithm 1 for a single denoising timestep $t$ as illustrated in Figure 3.

*Extracting the Cross-Attention Maps.* Given the input text prompt $\mathcal{P}$, we consider the set of all subject tokens (*e.g.*, nouns) $\mathcal{S} = \{s_1, ..., s_k\}$ present in $\mathcal{P}$. Our objective is to extract a spatial attention map for each token $s \in \mathcal{S}$, indicating the influence of the token $s$ on each image patch.

Given the noised latent $z_t$ at the current timestep, we perform a forward pass through the pre-trained UNet network using $z_t$ and $\mathcal{P}$ (Step 1 in Algorithm 1). We then consider the resulting cross-attention map obtained after averaging all $16 \times 16$ attention layers and heads. The resulting aggregated map $A_t$ contains $N$ spatial attention maps, one for each of the tokens of $\mathcal{P}$, *i.e.* $A_t \in \mathbb{R}^{16 \times 16 \times N}$.

The pre-trained CLIP text encoder prepends a specialized token $\langle sot \rangle$ to $\mathcal{P}$ indicating the start of the text. We note that Stable Diffusion learns to consistently assign a high attention value to the $\langle sot \rangle$ token in the token distribution defined in $A_t$. Since we are interested in enhancing the actual prompt tokens, we re-weigh the attention values by ignoring the attention of $\langle sot \rangle$ and performing a Softmax operation on the remaining tokens (Step 2 in Algorithm 1). After the Softmax operation, the $(i, j)$-th entry of the resulting matrix $A_t$ indicates the probability of each of the textual tokens being present in the corresponding image patch. We then extract the $16 \times 16$ normalized attention map for each subject token $s$ (Step 4).

*Obtaining Smooth Attention Maps.* Observe that the attention values $A_t^s$ calculated above may not fully reflect whether an object is generated in the resulting image. Specifically, a single patch with a high attention value could stem from partial information being passed from the token $s$. This may occur when the model does not generate the full subject, but rather a patch that resembles some part of the subject, *e.g.*, a silhouette that resembles an animal's body part. See the supplementary materials for such failure cases.

To avoid such adversarial solutions, we apply a Gaussian filter over $A_t^s$ in Step 5 of Algorithm 1. After doing so, the attention value of the maximally-activated patch is dependent on its neighboring patches since, after this operation, each patch is a linear combination of its neighboring patches in the original map.

*Performing On the Fly Optimization.* Intuitively, successfully generated subjects should have an image patch that significantly attends to their corresponding token. Our optimization objective embodies this intuition directly.

For each subject token in $S$, our optimization encourages the existence of at least one patch of $A_t^s$ with a high activation value. Therefore, we define the loss quantifying this desired behavior as

$$\mathcal{L} = \max_{s \in S} \mathcal{L}_s \qquad \text{where} \qquad \mathcal{L}_s = 1 - \max(A_t^s). \tag{2}$$

That is, the loss attempts to strengthen the activations of the most neglected subject token at the current timestep $t$. It should be noted that different timesteps may strengthen different tokens, encouraging all neglected subject tokens to be strengthened at some timestep.

Having computed our loss $\mathcal{L}$, we shift the current latent $z_t$ by

$$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}, \tag{3}$$

where $\alpha_t$ is a scalar defining the step size of the gradient update. Finally, we perform another forward pass through $SD$ using $z_t'$ to calculate $z_{t-1}$ for the next denoising step (Step 16 of Algorithm 1). The above update process is repeated for a subset of the timesteps $t = T, T-1, \dots, t_{end}$ where we set $T = 50$, following Stable Diffusion, and $t_{end} = 25$. This is based on the observation that the final timesteps do not alter the spatial locations of objects in the generated image.

*Iterative Latent Refinement.* So far, we have made a single latent update at each denoising timestep. However, if the attention values of a token do not reach a certain value in the early denoising stages, the corresponding object will not be generated. Therefore, we iteratively update $z_t$ until a pre-defined minimum attention value is achieved for *all* subject tokens. Yet, many updates of $z_t$ may lead to the latent becoming out-of-distribution, resulting in incoherent

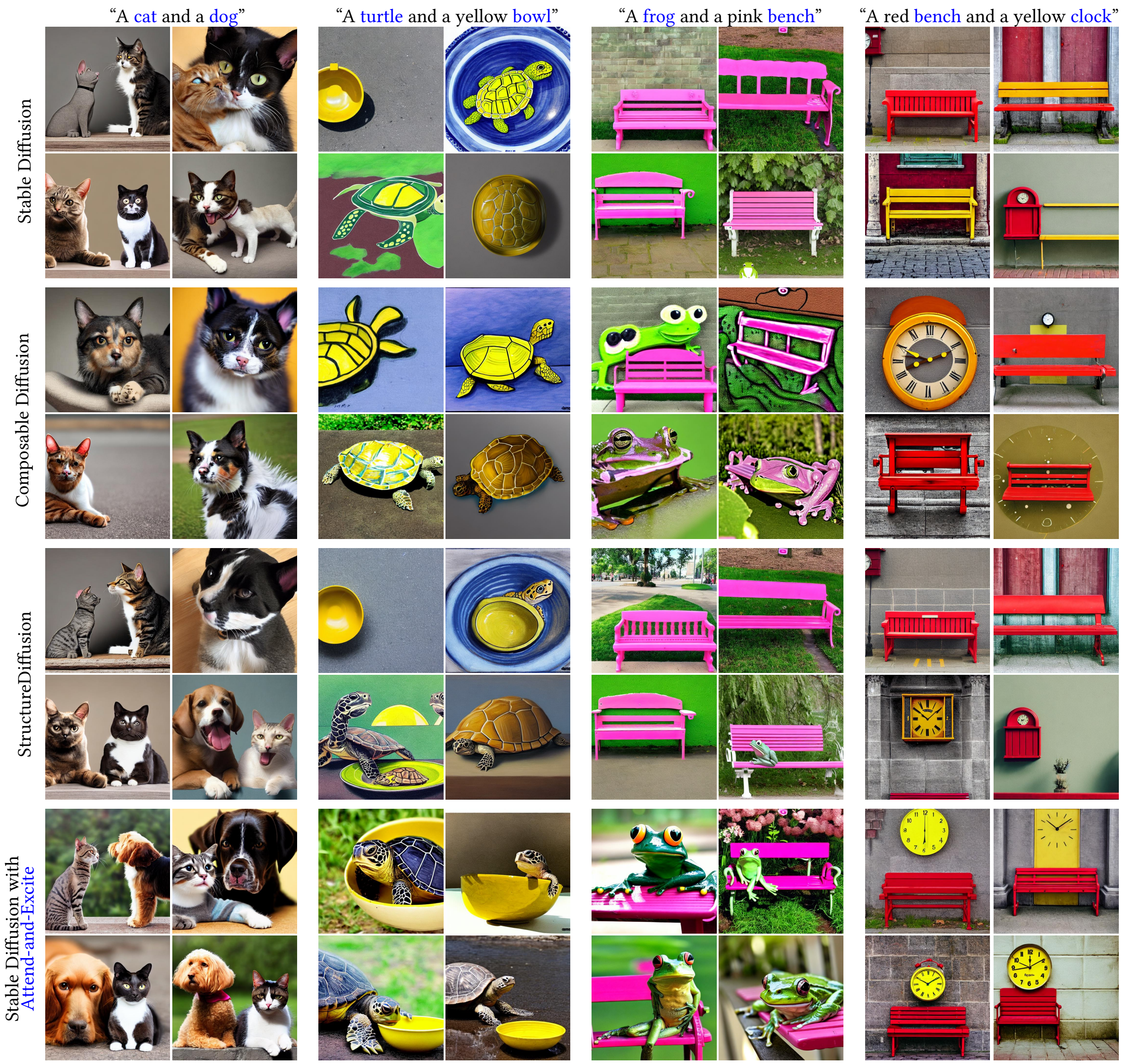


Fig. 5. Qualitative Comparison using prompts from our dataset. For each prompt, we show four images generated by each of the four considered methods where we use the same set of seeds across all approaches. The subject tokens optimized by Attend-and-Excite are highlighted in blue.

images. As such, this refinement is performed *gradually* across a *small* subset of timesteps.

Specifically, we demand that each subject token reaches a maximum attention value of at least 0.8. To do so gradually, we perform the iterative updates at various denoising steps (Steps 10-15 in Algorithm 1). We set the iterations to $t_1 = 0, t_2 = 10$, and $t_3 = 20$ with minimum required attention values of $T_1 = 0.05, T_2 = 0.5$, and $T_3 = 0.8$. This gradual refinement prevents $z_t$ from becoming out-of-distribution while encouraging more faithful generations.

*Obtaining Explainable Image Generators.* The extent to which attention can be used as an explanation has been widely explored [Abnar and Zuidema 2020; Chefer et al. 2021, 2022b]. In the context of text-based image generation, the cross-attention maps have been considered a natural explanation for the model [Hertz et al. 2022].

However, a direct result of catastrophic neglect is that the attention map corresponding to the neglected subject no longer faithfully represents the subject's localization in the generated image, as can be seen in the left column of Figure 4. While the cross-attention map for the cat is correctly localized, the map corresponding to the frog highlights irrelevant regions since a frog is not present. Thus, the cross-attention maps do not constitute viable explanations, as they are misleading and inaccurate. Conversely, as can be seen on the right of Figure 4, by mitigating neglect using Attend-and-Excite,


Fig. 6. Additional comparisons with Stable Diffusion using prompts describing complex scenes and multiple subject tokens. For each prompt, we show four generated images where we use the same set of seeds for both approaches. The subject tokens optimized by Attend-and-Excite are highlighted in blue.

both the cat and the frog are accurately localized in the attention maps, and the maps can now be considered a faithful explanation.

## 5 RESULTS

*Evaluation Setup.* As there are currently no openly-available datasets that analyze semantic issues in text-based image generation, we construct a new benchmark to evaluate all methods. To analyze the existence of catastrophic neglect, we construct prompts containing two subjects. Additionally, to test correct attribute binding, the prompts should contain a variety of attributes matched to the subject tokens. Specifically, we consider three types of text prompts: (i) "a [*animalA*] and a [*animalB*]", (ii) "a [*animal*] and a [*color*][*object*]", and (iii) "a [*colorA*][*objectA*] and a [*colorB* ][*objectB*]". To compose the prompts, we consider 12 animals and 12 object items with 11 colors, detailed in the supplementary materials. For each prompt containing a subject-color pair, we randomly select a color for the subject. This results in 66 Animal-Animal and Object-Object pairs and 144 Animal-Object pairs. For each prompt, we then generate 64 images using 64 random seeds applied across all methods.

For ease of evaluation, our prompts are constructed of conjunctions and color attributes. Yet, our method is not limited only to such cases and can be applied to a range of prompts with any number and type of subjects and attributes (see Figures 6 and 7 and the supplementary materials).

### 5.1 Qualitative Comparisons

In Figure 5, we present results using prompts from our dataset. As can be seen, Composable Diffusion [Liu et al. 2022] tends to generate images containing a mixture of the subjects. For example, for "A cat and a dog", the images tend to mix the cat's body with the dog's face and vice versa. This can similarly be seen in the prompt "A frog and a pink bench" where the images may contain a frog in the shape of a bench. For StructureDiffusion [Feng et al. 2022], the generated images tend to be very similar to those of Stable Diffusion, indicating that the approach fails to adequately address the semantic issues since it heavily relies on the inaccurate semantics captured by Stable Diffusion. Further, in the second and last column, the alternative methods either fail to generate all subjects or fail to correctly bind colors to each subject (*e.g.*, a blue bowl instead of a yellow bowl and

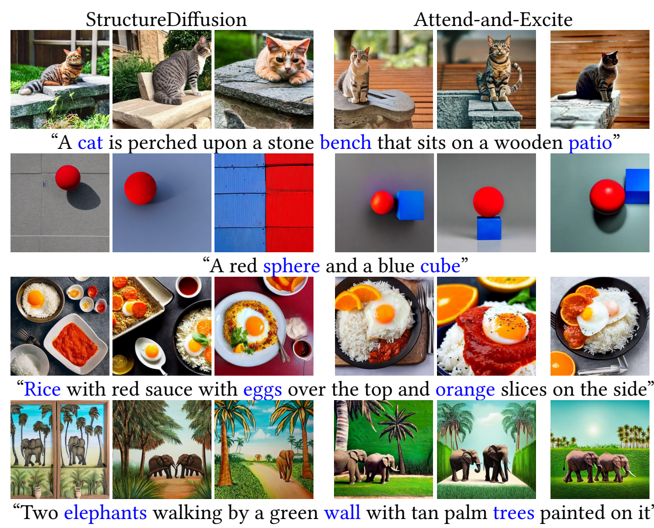


Fig. 7. Comparison with prompts appearing in Feng *et al.* [2022]. For each prompt, we apply the same set of random seeds across the two methods.

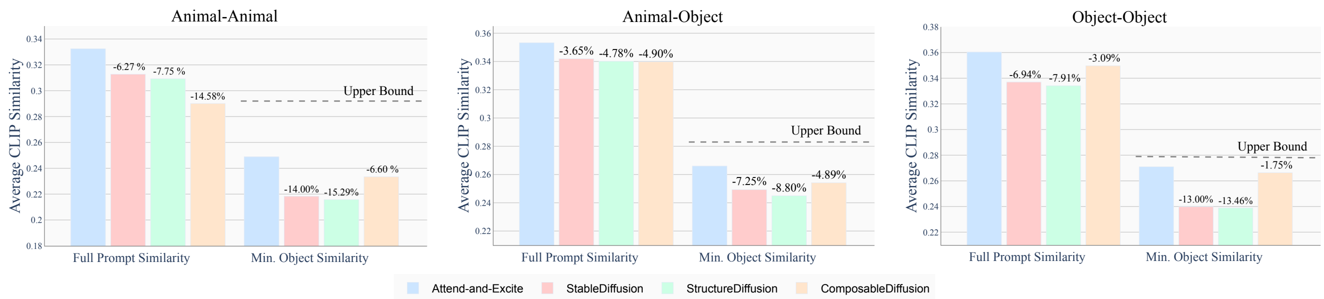


Fig. 8. Average CLIP image-text similarities between the text prompts and the images generated by each method, split by subset. The *Full Prompt Similarity* indicates the image-text similarity when considering the full text prompt while *Minimum Object Similarity* represents the average CLIP similarity for the most neglected subject. Note, the *Upper Bound* (the maximal-expected similarity) is applicable only to the *Minimum Object Similarity*.

a red clock instead of a yellow clock). In contrast, Attend-and-Excite is able to synthesize images that more faithfully contain all subjects with correctly binded colors. Although we explicitly tackle only the issue of neglect, we are able to implicitly improve attribute bindings between colors and subjects (*e.g.*, the red bench and yellow clock).

Additionally, we provide examples of complex prompts in Figure 6 and in the supplementary materials, including prompts with three or more subjects, complex attributes, and interactions between subjects. As can be seen, Attend-and-Excite is able to mitigate neglect while generating images that correspond to the input prompt, and the interactions between the subjects. For example, for the prompt "A grizzly bear catching a salmon in a crystal clear river surrounded by a forest" Attend-and-Excite mitigates the neglect over the salmon while generating images in which the bear catches the salmon, as specified in the prompt. Finally, Attend-and-Excite can also be used to correct global properties such as a background subject as shown with the "garden" in the third column.

In Figure 7 we consider prompts from the StructureDiffusion paper with more than two subjects or complex attributes (*e.g.*, "stone bench", "wooden patio"). As can be observed, StructureDiffusion fails to mitigate both semantic issues. For example, in the second row, StructureDiffusion generates a sphere or cube-like object but fails to generate both. In the third row, it fails to correctly bind attributes such as the red sauce to the rice. Conversely, Attend-and-Excite generates semantically accurate images in both cases.

In the supplementary materials, we provide additional qualitative and quantitative results on complex prompts, as well as an ablation study and additional comparisons to image editing techniques.

## 5.2 Quantitative Analysis

We quantify the performance of each method using CLIP-space distances along two fronts. First, we evaluate image-text similarities between the generated images and each text prompt. Second, several works [Liang et al. 2022; Sheynin et al. 2022] have analyzed the existence of a modality gap between CLIP's image and text embeddings. To overcome this gap, we consider an additional text-only metric.

*Text-Image Similarities.* For each prompt, we compute the average CLIP cosine similarity between the text prompt and the corresponding set of 64 generated images. We denote this as the *Full Prompt Similarity*. Yet, considering the full text may not accurately reflect the existence of neglect. It has been observed [Paiss et al. 2022] that CLIP's similarities resemble a bag-of-words behavior where a high score can be achieved even if the image does not fully correspond to the semantic meaning of the prompt. For example, an image of a cat may obtain a high similarity to "a cat and a dog" even though a dog is not present. In such cases, considering only the full-text similarity will not capture the existence of neglect.

Table 1. Average CLIP text-text similarities between the text prompts and captions generated by BLIP over the generated images.

| Method | Animal-Animal | Animal-Object | Object-Object |
|---|---|---|---|
| Stable Diffusion | 0.767 (-5.08%) | 0.793 (-4.74%) | 0.765 (-5.89%) |
| Composable Diffusion | 0.692 (-16.47%) | 0.769 (-7.94%) | 0.759 (-6.85%) |
| StructureDiffusion | 0.761 (-5.91%) | 0.781 (-6.31%) | 0.762 (-6.49%) |
| **Attend-and-Excite** | **0.806** | **0.830** | **0.811** |

As such, we evaluate the CLIP similarity for the *most* neglected subject independently of the full text. To this end, we split the prompt into two sub-prompts, each containing a single subject (*e.g.*, "a cat", "a dog"). We then compute the CLIP similarity between each sub-prompt and each generated image. Given the two scores for each image, we are interested in maximizing the smaller of the two as this would correspond to minimizing neglect. We average the smaller of the two scores across all seeds and prompts and denote this as the *Minimum Object Similarity*. To provide intuition for the scale of the best-achievable Minimum Similarity, we compute an *Upper Bound*. For each subject, we collect 50 images from classification and detection datasets [Banerjee 2022; Lin et al. 2014] and the internet. We then compute the average CLIP similarity between the collected images and the subject prompt (*e.g.*, "a cat"). To obtain the bound for each subset, we average the scores of all subjects in the set.

Figure 8 presents the results of the CLIP text-image metrics for all three subsets (Animal-Animal, Animal-Object, Object-Object). Observe that Attend-and-Excite outperforms all baselines across all subsets and for both metrics. Additionally, we provide the relative decrease in similarity (in percentage) compared to Attend-and-Excite.

Notice that StructureDiffusion obtains scores similar to those of Stable Diffusion (albeit slightly lower). Attend-and-Excite significantly improves the Minimum Object Similarity in comparison to both by a gap of at least 7% across all test cases, indicating that our method substantially improves the issue of neglect. For some

Table 2. User study conducted with 65 respondents. We randomly select 10 prompts from each subset and apply the same 4 randomly-selected seeds to all methods. Users are asked to select the set of images that best corresponds to the input prompt. Results are averaged across all prompts in the subset.

| Method | Animal-Animal | Animal-Object | Object-Object |
|---|---|---|---|
| Stable Diffusion | 2.32% | 13.92% | 5.71% |
| Composable Diffusion | 0% | 1.69% | 9.82% |
| StructureDiffusion | 6.98% | 6.75% | 7.31% |
| **Attend-and-Excite** | **90.70%** | **77.64%** | **77.16%** |


Fig. 9. Limitations. Left: Out-of-distribution results due to the limited expressive power of Stable Diffusion. Right: When the subject combination is not natural ("elephant", "sombrero"), the results may be less realistic.

subsets, Composable Diffusion achieves results closest to those obtained by Attend-and-Excite. This can be attributed to a deficiency in the image-based metric where a high score can be achieved even when only a portion of a subject is present. As mentioned, Composable Diffusion often generates an object that is a mixture of the subjects in the input text. In such cases, the similarity to both subjects could be high, even though they are not generated separately. For example, an image featuring a car shaped like a bird may obtain a high similarity for both "a bird" and "a car" since the shape corresponds to "a bird" while the object itself is a car. We refer the reader to the supplementary materials for examples of such behavior. To overcome this limitation, we explore a text-based metric below.

*Text-Text Similarities.* Given the 64 generated images for a given input prompt, we generate matching image captions using a pre-trained BLIP image-captioning model [Li et al. 2022]. We then compute the average CLIP similarity between the prompt and all captions. This process is repeated for each subset and the results are averaged across the prompts in the subset. The choice of CLIP to compute the text-text similarity arises from the strong semantic prior of CLIP. We are less concerned with the exact phrasing and order of subjects in the captions. Instead, our focus is on capturing all subjects and attributes in the original prompt.

We present the text-text similarity results in Table 1. As shown, Attend-and-Excite outperforms all alternative methods across each of our three subsets by at least 4.7%. Additionally, observe that ComposableDiffusion is the lowest-performing approach when considering the text-text similarity metrics, indicating that the text-text metric captures the subject-mixing behavior discussed above.

*User Study.* Finally, we perform a user study to analyze the fidelity of the generated images. For each of the three evaluation subsets, we randomly sample 10 prompts and generate images with each approach using the same 4 randomly-selected seeds. For each prompt, we ask the respondents to select which set of images best reflects the prompt. The final score for each approach is calculated as the number of times respondents selected the approach averaged across all the prompts in the set (*e.g.*, a score of 90% indicates that 90% of responses preferred the approach over all others).

The study results are shown in Table 2. Attend-and-Excite received the highest percentage of votes across *all* subsets, with 90.70% of responses preferring our method in the Animal-Animal subset, 77.64% for the Animal-Object category, and 77.16% for Object-Object. When evaluating each prompt individually, Attend-and-Excite is always preferred over the baselines by a majority of respondents. Even our lowest performing prompt received 59.09% of votes (with SD and StructureDiffusion tied for second, each receiving 16% of votes). This substantiates the effectiveness of Attend-and-Excite in alleviating semantic issues in text-based image generation.

## 6 LIMITATIONS

While our method offers increased fidelity with respect to the given prompt, there are several limitations to consider. First, our method is limited by the expressive power of the generative model since we do not apply additional training. In cases where the prompt resides outside the distribution of the textual descriptions the model learned, our method could lead to latents that are out of distribution, resulting in images that do not correspond to the text prompt.

Second, when synthesizing subjects that naturally do not appear together, the generated images may be less realistic (*e.g.*, paintings). We attribute this to the fact that such combinations tend to reside outside the distribution that Stable Diffusion has learned for real images. Examples for these limitations are shown in Figure 9.

Finally, while we tackle two core semantic issues, the path to achieving semantically-accurate generation is still long, and there exist additional challenges to be addressed such as complex object compositions (*e.g.*, "riding on", "in front of", "beneath"). Additionally, while we have not explored applying Attend-and-Excite over a negation (*i.e.*, "not"), this could potentially be achieved by demanding a *low* attention value for the subject.

## 7 CONCLUSIONS

Can a diffusion process be corrected once it takes a wrong turn? In this work, we introduce the concept of *Generative Semantic Nursing* (GSN), which refers to a careful manipulation of latents during the denoising process of a pre-trained text-to-image diffusion model. We then present *Attend-and-Excite*, a specific form of GSN that encourages all subject tokens in the text to be attended to by some image patch. We demonstrate that by applying this intuitive optimization, we are able to alleviate two core semantic issues on the fly, thus correcting the generator after it has taken a wrong turn.

Similar to extrapolating text-driven gradients in classifier-free guidance, our approach aims to strengthen the text conditioning along the image generation process. While we explore the notion of GSN for mitigating semantic issues of text-conditioned generation, we believe GSN can potentially be applied to any image editing and generation task by defining an appropriate loss objective. Moreover, this guidance need not be through text and does not require conditioning at all, but is defined only by the task itself.

## ACKNOWLEDGMENTS

We thank Daniel Livshen, Elad Richardson, Matan Cohen, Or Patashnik, Rinon Gal, and Yotam Nitzan for their early feedback and insightful discussions. The first author is supported by the Council for Higher Education in Israel. This work was supported in part by the Israel Science Foundation under Grant No. 2366/16 and Grant No. 2492/20.

## REFERENCES

Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. *ArXiv* abs/2005.00928 (2020).

Oron Ashual and Lior Wolf. 2019. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4561–4569.

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2022a. Blended Latent Diffusion. *arXiv preprint arXiv:2206.02779* (2022).

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2022b. SpaText: Spatio-Textual Representation for Controllable Image Generation. *arXiv preprint arXiv:2211.14305* (2022).

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *ArXiv* abs/2211.01324 (2022).

Sourav Banerjee. 2022. Animal Image Dataset. https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals.

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by word. *arXiv preprint arXiv:2103.10951* (2021).

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800* (2022).

Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. 2022a. Image-Based CLIP-Guided Essence Transfer. (2022).

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 397–406.

Hila Chefer, Idan Schwartz, and Lior Wolf. 2022b. Optimizing Relevance Maps of Vision Transformers Improves Robustness. In *Thirty-Sixth Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=upuYKQiyxa_

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*. Springer, 88–105.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. *arXiv preprint arXiv:2212.05032* (2022).

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131* (2022).

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022b. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1219–1228.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276* (2022).

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488* (2022).

Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18062–18071.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086* (2022).

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053* (2022).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional Visual Generation with Composable Diffusion Models. *arXiv preprint arXiv:2206.01714* (2022).

Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems*. 1–23.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2. *arXiv preprint arXiv:2204.13807* (2022).

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

Roni Paiss, Hila Chefer, and Lior Wolf. 2022. No token left behind: Explainability-aided image classification and generation. In *European Conference on Computer Vision*. Springer, 334–350.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv* (2022). https://arxiv.org/abs/2204.06125

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242* (2022).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv* abs/2205.11487 (2022).

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. KNN-Diffusion: Image Generation via Large-Scale Retrieval. *ArXiv* abs/2204.02849 (2022).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16515–16525.

Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. 2022. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477* (2022).

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv preprint arXiv:2210.14896* (2022).

Sam Witteveen and Martin Andrews. 2022. Investigating Prompt Engineering in Diffusion Models. *arXiv preprint arXiv:2211.15462* (2022).

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

Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. 2021. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423* (2021).

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 833–842.

Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. 2019. Image Generation from Layout. In *CVPR*.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5802–5810.