Training-Free Layout Control with Cross-Attention Guidance

Minghao Chen Iro Laina Andrea Vedaldi
Visual Geometry Group, University of Oxford
{minghao, iro, vedaldi}@robots.ox.ac.uk

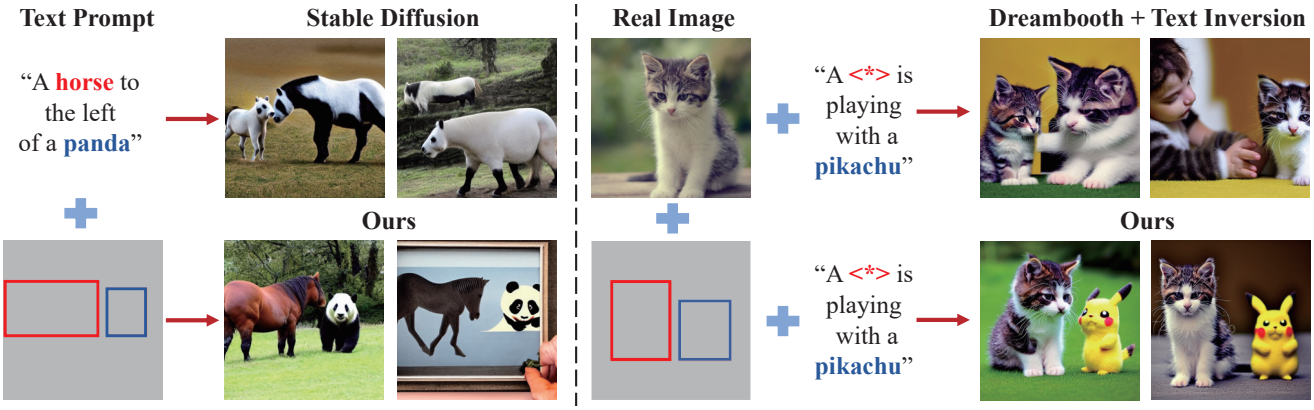

Figure 1. Left: our method controls the layout of an image generated by a pre-trained diffusion model, such as Stable Diffusion [38], without any training or finetuning. It also alleviates the problem of such generators omitting certain objects present in the prompt. Right: given a single real image, our method can be also used to edit the position and context of a subject (represented by $\langle * \rangle$).

Abstract

Recent diffusion-based generators can produce high-quality images from textual prompts. However, they often disregard textual instructions that specify the spatial layout of the composition. We propose a simple approach that achieves robust layout control without the need for training or fine-tuning of the image generator. Our technique manipulates the cross-attention layers that the model uses to interface textual and visual information and steers the generation in the desired direction given, e.g., a user-specified layout. To determine how to best guide attention, we study the role of attention maps and explore two alternative strategies, forward and backward guidance. We thoroughly evaluate our approach on three benchmarks and provide several qualitative examples and a comparative analysis of the two strategies that demonstrate the superiority of backward guidance compared to forward guidance, as well as prior work. We further demonstrate the versatility of layout guidance by extending it to applications such as editing the layout and context of real images.

1. Introduction

Generative AI is one of the most disruptive technologies that emerged in the past years. In computer vision, new text-to-image generation methods, such as DALL-E [36], Imagen [41], and Stable Diffusion [38], have demonstrated that machines are capable of generating images of a quality high enough for use in numerous applications, multiplying the productivity of professional artists as well as lay people.

Despite this success, however, many practical applications of image generation, particularly in a professional setting, require a high level of *control* that such methods lack. Specifications in language-based image generators are textual; and while text can tap into a vast library of high-level concepts, it is a poor vehicle for expressing fine-grained visual nuances in an image. Specifically, text is often inadequate for describing the exact *layout of a composition*.

In fact, as shown in previous work [16], state-of-the-art image generators struggle to correctly interpret simple layout instructions specified via text. For example, when prompting such models with a phrase such as *"a dog to the left of a cat"*, the *"left of"* relationship is not always depicted accurately in the generated images. In fact, prompts

of this nature often cause models to produce erroneous semantics, for example, an image of a cat-dog hybrid. This limitation is exacerbated by unusual compositions, *e.g.*, *"horse on top of a house"*, which fall outside the typical compositions the model observes during training.

This work provides a better understanding of this limitation and contributes a mechanism to overcome it. To this end, we introduce a method that achieves *layout control* without the need for further training of the image generator, while still maintaining the quality of the generated images.

We note that, while layout cannot be easily controlled via textual prompting, one can *intervene* directly in the cross-attention layers, steering the generation in a direction of choice with user-specified inputs, such as bounding boxes, which we refer to as *layout guidance*. We consider and compare two alternative strategies for such an intervention: "forward guidance" and "backward guidance". Forward guidance directly biases the cross-attention layers to shift activations in the desired pattern, letting the model incorporate the guidance via the iterated application of its denoising steps. Our main contribution is backward guidance, which uses backpropagation to update the image latents to match the desired layout via energy minimization.

While layout control has already received some attention, with some methods following the forward paradigm [2,43], we show that backward guidance is a more effective mechanism. Our second contribution is then an in-depth investigation of the factors that influence the layout during the image generation process, shedding light on the shortcomings of forward guidance and discussing how backward guidance addresses these. We show that, while there is an intuitive correlation between different concepts and their visual extent, this correlation is more nuanced than one might think, and, perhaps counter-intuitively, even the special tokens in the prompt (start tokens and padding tokens) contribute to shaping the layout.

Finally, we show that our backward guidance outperforms existing methods and seamlessly integrates into applications such as real-image layout editing.

2. Related Work

Text-to-Image Generation. For several years, generative adversarial networks (GANs) [17] have been the dominant approach in image generation from textual prompts [37, 46, 49, 54–56]. Alternative representations for text, such as scene graphs, have also been considered [26]. More recently, the focus has shifted onto text-conditional autoregressive [10, 14, 36, 53] and diffusion models [18, 31, 35, 38, 41], with impressive results in generating images of remarkable fidelity, while avoiding common GAN pitfalls such as training instability and mode collapse [9]. A substantial increase in both the data scale [42] and the size and capabilities of transformer models [34] has played a crucial role in enabling this shift. Typically, these models are designed to accept a textual prompt as input, which may pose a challenge for accurately conveying all details of the image. This problem is exacerbated with longer prompts or when describing atypical scenes. Recent studies have demonstrated the effectiveness of classifier-free guidance [22] in improving the faithfulness of the generations with respect to the input prompt. Others focus on improving compositionality, *e.g.*, by combining multiple diffusion models with different operators [30], and attribute binding [5, 13].

Layout Control in Image Generation. Image generation with spatial conditioning is closely related to layout control and typically done with bounding boxes or semantic maps [12, 32, 44, 45, 50, 58]. These methods do not use text prompts and rely on a closed-set vocabulary to generate images, *i.e.*, the labels of the training distribution (*e.g.*, COCO [29]). Recent image-text models such as CLIP [34] are now enabling the extension to open-vocabulary. However, the precise layout of a composition is still challenging to convey through text alone; even then, the *spatial* fidelity of image generators is extremely limited [16]. Thus, jointly conditioning on text and layout [14, 20, 25] and predicting layout from text [23] have also been considered.

Recent works [1, 2, 4, 6, 28, 43, 48, 51] propose to extend the state-of-the-art Stable Diffusion [38] with spatial conditioning. GLIGEN [28] and ReCo [51] fine-tune the diffusion model with gated self-attention layers and additional regional tokens, respectively. Other works [2, 4, 6, 43, 48] follow a training-free approach. MultiDiffusion [4] adopts the idea from [30] by combining masked noise. eDiff-I [2] and HFG [43] share a similar idea with our forward guidance, directly intervening in the cross-attention. However, they overlook the significance of special tokens in the process. Concurrently with our work, ZestGuide [6] and BoxDiff [48] propose to compute a loss on cross-attention to achieve layout control, which is closer to our backward guidance. Unlike prior work, we use an objective function that does not rely on precise segmentation masks to be provided by the user, and we provide an in-depth analysis of the factors that affect the layout, and consequently, the behavior of both forward and backward strategies. Finally, building on top of diffusion, some recent works show controllable image generation from various other conditioning signals [3, 24, 57], such as depth or edge maps.

Diffusion-Based Image Editing. Most aforementioned methods lack the ability to control or edit an already generated image, or even the ability to edit real images. For example, simply changing a word in the original prompt typically leads to a drastically different generation. This can be circumvented by providing or generating masks for the objects of interest [7, 31]. Prompt-to-prompt [19] addresses this issue with simple text-based edits by exploiting the fact that the cross-attention layers present in most state-of-the-

art architectures connect word tokens to the spatial layout of the generated images. Text-based image editing can also be achieved through single-image model fine-tuning [27, 47]. However, these approaches, while successful at semantically editing entities can only apply such edits *in-place* and do not allow editing of the spatial layout itself.

3. Method

We consider the problem of *layout-guided* text-to-image generation. Text-based image generators allow to sample images $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ from a conditional distribution $p(\boldsymbol{x} \mid y)$ where y is language description. Given one such generator off-the-shelf, we wish to steer its output to match a desired layout for the generated composition, *without further training or finetuning*. In other words, our objective is to investigate whether pre-trained text-to-image generators can adhere to a layout specified by the user during inference, without having been trained with explicit layout conditioning. In the simplest case, given the text prompt y, the index i of a word y_i in the text prompt, and a bounding box B, we would like to generate an image $\boldsymbol{x}$ that contains y_i *inside* B, essentially modifying the generator to sample from a new distribution $p(\boldsymbol{x} \mid y, B, i)$ with additional controls.

3.1. Preliminaries: Stable Diffusion

We first briefly review the technical details of Stable Diffusion (SD) [38], a publicly accessible, state-of-the-art text-to-image generator representative of an important class of image generators based on diffusion [36, 38, 41]. SD consists of an image encoder and decoder, a text encoder, and a denoising network that operates in latent space.

The text encoder $Y = \phi(y)$ maps the input prompt into a tensor of fixed dimension $Y \in \mathbb{R}^{N\times M}$. This works by prepending a start symbol `[SoT]` to y and appending $N - |y| - 1$ padding symbols `[EoT]` at the end, to obtain N symbols in total. Then, the function ϕ, implemented as a large language model (LLM), takes the padded sequence of words as input and produces a corresponding sequence of token vectors $Y_i \in \mathbb{R}^M$ with $i \in \{1, \dots, N\}$ as output.

While not crucial for our discussion, SD's encoding network h maps images $\boldsymbol{x}$ to corresponding latent codes $\boldsymbol{z} = h(\boldsymbol{x}) \in \mathbb{R}^{4\times \frac{H}{s}\times \frac{W}{s}}$, where s divides H and W. The function h is an autoencoder with a left inverse h^*, such that $\boldsymbol{x} = h^* \circ h(\boldsymbol{x})$. The main purpose of this component is to replace the problem of modeling $p(\boldsymbol{x} \mid y)$ with the problem of modeling $p(\boldsymbol{z} \mid y)$, reducing the spatial resolution s-fold.

A key component of SD is the iterative conditional denoising network D. This network is trained to output a conditional sample $\boldsymbol{z} \sim p(\boldsymbol{z} \mid y)$ of the latent code $\boldsymbol{z}$. It is designed to take a noised sample $\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t$, as input, where $\boldsymbol{\epsilon}_t$ is normally distributed noise and α_t is a decreasing sequence, from $\alpha_0 \approx 1$ to $\alpha_T \approx 0$, representing the noise schedule. Then, the network D returns an estimate

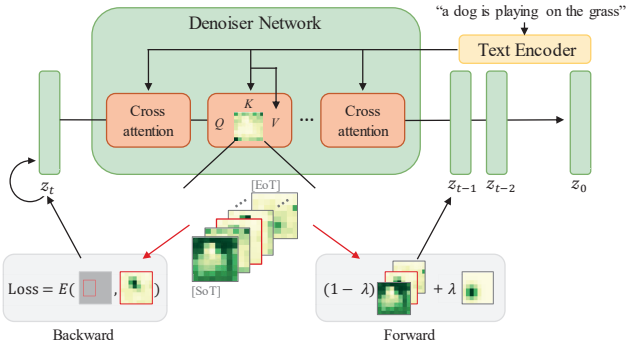

Figure 2. Overview of the two layout guidance strategies. The cross-attention map for a chosen word token is marked with a red border. In forward guidance, the cross-attention maps of the word, start and padding tokens are biased spatially. In backward guidance, we compute instead a loss function and perform backpropagation during the inference process to optimize the latent.

of the noised sample $\boldsymbol{z}_t$: $D(\boldsymbol{z}_t, y, t) \approx \boldsymbol{\epsilon}_t$. To sample an image, one first samples $\boldsymbol{z}_T$, which is normally distributed, and applies D iteratively, to obtain the intermediate codes $\boldsymbol{z}_{T-1}, \dots, \boldsymbol{z}_1, \boldsymbol{z}_0 \approx \boldsymbol{z}$. Finally, $\boldsymbol{z}$ is converted back to an image via the image decoder $\boldsymbol{x} = h^*(\boldsymbol{z})$.

There is one final aspect of the SD architecture that is relevant for our work. While there are several design choices that make the network D work well in practice, the mechanism that is of interest in our investigation is *cross-attention*, which connects visual and textual information and allows the generation process to be conditioned on text. Each cross-attention layer takes an intermediate feature tensor $\boldsymbol{z}^{(\gamma)} \in \mathbb{R}^{C\times \frac{H}{r}\times \frac{W}{r}}$ as input, where γ is the index of the relevant layer in the network, and r is a scaling factor defining the spatial resolution at that level of the representation. The cross-attention map $A^{(\gamma)}$ associates each spatial location $u \in \{1, \dots, \frac{H}{r}\} \times \{1, \dots, \frac{W}{r}\}$ to a token indexed by $i \in \{1, \dots, N\}$:

$$A^{(\gamma)}_{ui} = \frac{\exp\langle Q^{(\gamma)}_u, K^{(\gamma)}_i\rangle}{\sum_{j=1}^{N} \exp\langle Q^{(\gamma)}_u, K^{(\gamma)}_j\rangle}, \quad \boldsymbol{a}^{(\gamma)}_u = \sum_{i=1}^{N} A^{(\gamma)}_{ui} V^{(\gamma)}_i,$$

where the value $V^{(\gamma)}_i$ and the key $K^{(\gamma)}_i$ are linear transformations of the token embedding Y_i provided by the textual encoder, $Q^{(\gamma)}$ is a linear transformation of $\boldsymbol{z}^{(\gamma)}$, and $\boldsymbol{a}^{(\gamma)}_u$ is the output of the cross-attention layer.

3.2. Layout Guidance

Text-to-image generators such as SD struggle to accurately interpret layout instructions provided through text. We thus introduce a method to guide the layout during the generation process by sampling from a distribution $p(\boldsymbol{x} \mid y, B, i)$ with additional controls, *e.g.*, user-specified bounding boxes B corresponding to selected text tokens y_i. This can be achieved via manipulation of the attention response in certain cross-attention layers in the architecture.

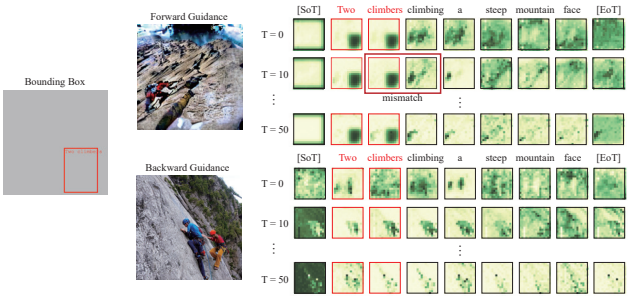

Figure 3. Cross-attention maps during forward and backward guidance. Spatial dependencies between different words negatively affect forward guidance, while backward guidance softly encourages all dependent tokens to match the desired layout.

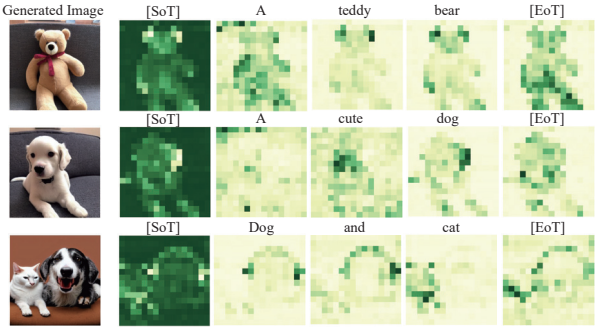

Figure 4. Cross-attention maps of different text prompts at the generation process, indicating that start `[SoT]` and padding `[EoT]` tokens carry rich semantic and layout information.

It has already been shown that cross-attention layers regulate the spatial layout of a generated image [19]. Specifically, $A_{ui}^{(\gamma)}$ determines how strongly each location u in layer γ is associated with each of the N text tokens y_i. Since the sum of association strengths $\sum_{i=1}^{N} A_{ui}^{(\gamma)} = 1$ for each spatial location u, the different tokens can be seen as "competing" for a location. To control the image layout using a bounding box B corresponding to token y_i, the attention can be biased such that locations $u \in B$ within the target box are strongly associated with y_i (while other locations are not). As we discuss below, this can be done without fine-tuning the image generator or training additional layers.

Next, we present a comprehensive investigation of two strategies to achieve training-free layout control: forward and backward guidance (Fig. 2). While instances of *forward* guidance have been discussed in recent work [2, 43], we hereby formalize this approach, identify its limitations, and propose backward guidance as a more effective alternative.

Forward Guidance. In forward guidance, the bounding box B is represented as a smooth windowing function $g_u^{(\gamma)}$ which is equal to a constant $c > 0$ inside the box and quickly falls to zero outside.[1] We rescale the windowing function such that $\|g^{(\gamma)}\|_1 = 1$. Then, we bias a cross-attention map by replacing it with:

$$A_{ui}^{(\gamma)} \leftarrow (1-\lambda) A_{ui}^{(\gamma)} + \lambda g_u^{(\gamma)} \sum_v A_{vi}^{(\gamma)}, \qquad (1)$$

where $\lambda \in [0,1]$ defines the strength of the intervention. In practice, we normalize the right side of Eq. (1) with a softmax function along the text token dimension, keeping the sum of per-pixel attention equal to 1. Note that (1) only the cross-attention map $A_{:,i}^{(\gamma)}$ of the i-th token is manipulated, and (2) the window is weighed by the mass $\sum_v A_{vi}^{(\gamma)}$ so as to leave the latter unchanged.

This intervention is applied for a number of iterations of the denoiser network D at selected layers $\gamma \in \Gamma$. This means that the activation maps computed by each selected layer are independently modified following Eq. (1).

A critical analysis reveals that forward guidance is a simplistic approach that suffers from inherent constraints hindering its ability to provide effective layout control. As we discuss in Section 3.3, this is primarily due to various factors that influence the layout during the generation process, including spatial dependencies among text tokens and spatial information "hidden" in the initial noise.

Backward Guidance. To address the shortcomings of forward guidance, we introduce an alternative mechanism, which we refer to as backward guidance. Instead of directly manipulating attention maps, in backward guidance, we bias the attention by introducing an energy function

$$E(A^{(\gamma)}, B, i) = \left(1 - \frac{\sum_{u \in B} A_{ui}^{(\gamma)}}{\sum_u A_{ui}^{(\gamma)}}\right)^2. \qquad (2)$$

Optimizing this function encourages the cross-attention map of the i-th token to obtain higher values inside the area specified by B. Specifically, at each application of the denoiser D, when layer $\gamma \in \Gamma$ is evaluated, the gradient of the loss (2) is computed via backpropagation to update the latent $\boldsymbol{z}_t (\equiv \boldsymbol{z}_t^{(0)})$:

$$\boldsymbol{z}_t \leftarrow \boldsymbol{z}_t - \sigma_t^2 \eta \nabla_{\boldsymbol{z}_t} \sum_{\gamma \in \Gamma} E(A^{(\gamma)}, B, i), \qquad (3)$$

where $\eta > 0$ is a scale factor controlling the strength of the guidance and $\sigma_t = \sqrt{(1-\alpha_t)/\alpha_t}$. By updating the latent, the cross-attention maps of all tokens are indirectly influenced by backward guidance. To generate an image, we alternate between gradient updates and denoising steps.

3.3. Analysis and Discussion

Next, we detail a comparative analysis between the forward and backward strategies. To motivate backward guidance and understand its effectiveness, we shed light on the significance of all tokens and the influence of the initial noise in shaping the layout during the generation process.

[1] For simplicity, in our implementation, we put a Gaussian blob with σ decided by the resolution, height, and width of the bounding box.

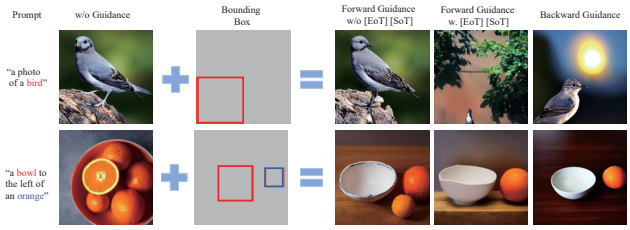

Figure 5. Comparison between forward and backward guidance, including guidance of start and padding tokens.

The Role of Word Tokens. One important consideration is that the text encoder fuses information from different words when processing a prompt due to self-attention. This results in a "semantic overlap": information from one token being encoded by another token. In other words, text embeddings capture both word-specific *and contextual* information, *e.g.*, subject-verb-object dependencies. This overlap is then transferred from the text encoder into the diffusion process via the cross-attention layers, resulting in *spatial* overlap. The example in Figure 3 illustrates this overlap in the cross-attention maps of different words. It also shows the behavior of forward and backward guidance when providing spatial conditioning for the phrase "two climbers". It becomes evident that the mismatch between the attention map of the conditioned phrase and its spatial dependencies with other words ("climbing","a") causes forward guidance to disregard the layout condition. Instead, backward guidance indirectly drives all attention maps toward the layout condition as necessary, because it acts on the latent codes.

The Role of Special Tokens. Another crucial finding is that the cross-attention maps of `[SoT]` and `[EoT]` tokens, which do not correspond to content words in the input text, still carry significant semantic and layout information. As we show in Figure 4, the cross-attention maps of `[EoT]` tokens correspond to salient regions in the generated image, *i.e.*, typically the union of individual semantic entities in the text prompt. `[SoT]` behaves complementarily to `[EoT]`, emphasizing the background. For forward guidance to be effective, it is thus necessary to intervene not only on selected content tokens but also on the special ones. We use the union of the input boxes as guidance for `[EoT]` and the reverse for `[SoT]`. However, we have empirically found that this sometimes results in overly aggressive guidance, which harms image fidelity. Backward guidance, on the other hand, does not suffer from such drawbacks, as it optimizes the latent. We discuss this further in the supplement.

The Role of Initial Noise. Finally, the initial noise of the diffusion process plays an important role in shaping the layout of the images. We have empirically observed that the noise contains an intrinsic layout; *e.g.*, when prompting the model with phrases like "an image of a dog" and "an image of a cat" using the same seed, it generates images with consistent layouts, placing the dog and the cat in the same locations. We provide examples in the supplement.

An initial noise with an intrinsic layout close to the one given by users is easier to optimize and results in higher fidelity. Therefore, selecting a noise pattern that aligns with the desired layout can further boost the effectiveness of the guidance. In backward guidance, the loss applied to the cross-attention maps can, in fact, double as a metric for initial noise selection. Specifically, we sample different noise patterns and evaluate Eq. (2) after applying backward guidance for a few steps. This allows us to pick the best-aligned initial noise. Please see the supplement for detailed results.

Forward *vs.* Backward. In summary, forward and backward guidance use different mechanisms to manipulate cross-attention. Forward guidance *directly* modifies cross-attention to conform to the prescribed pattern, which is "forced" repeatedly for a number of denoising iterations. While it does not incur any extra computational cost, it struggles to provide robust control over the layout, as non-guided tokens may cause the generation to deviate from the desired pattern. In contrast, backward guidance uses a loss function to evaluate whether the attention follows the desired pattern. While slower than forward guidance, backward guidance is more refined, as it indirectly encourages all tokens (guided and non-guided ones) to adhere to the layout through latent updates.

3.4. Real-image Layout Editing

Layout guidance can be used in combination with other techniques that build on diffusion-based image generators. We demonstrate this for the task of real-image editing. To this end, we incorporate backward guidance into two methods that are commonly used for personalization of diffusion models given real images, namely Textual Inversion (TI) [15] and Dreambooth [40]. TI extends an existing image generator with a new concept given one or several images as examples, by optimizing a learnable text token $\langle * \rangle$ for the concept. Dreambooth attempts to capture the appearance of a particular subject of which several images are available by fine-tuning a pre-trained text-to-image model. Then, new images of the learned concept can be generated.

Neither method supports *localized* spatial control over the newly generated images; their edits are usually global and semantic. To achieve this, we apply backward guidance on the Dreambooth-finetuned model and the TI-optimized token as part of a prompt. This allows us to control the layout of the generated images while preserving the identity of the original object represented by $\langle * \rangle$.

4. Experiments

In this section, we evaluate our approach for training-free layout guidance, quantitatively comparing variants of forward and backward guidance and providing comparisons

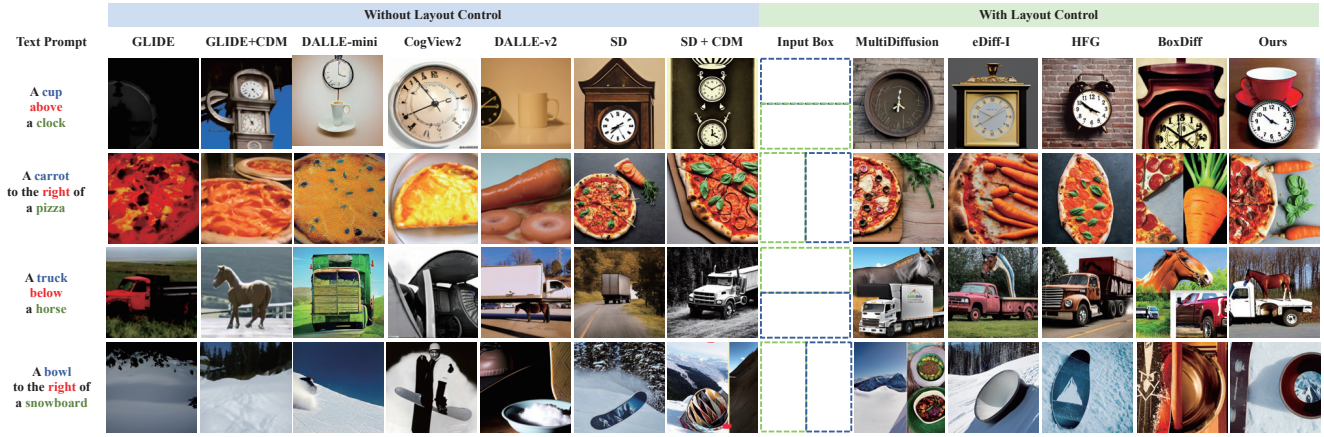

Figure 6. Qualitative comparison of different text-to-image models with text prompts defined in [16]. As stated in [16], current text-to-image models fail to understand spatial relationships without explicit layout conditioning. However, we achieved control of the generated images with the help of guidance on cross-attention maps.

Model	OA (%)	VISOR (%) uncond	VISOR (%) cond	Runtime
Stable Diffusion	27.4	16.4	59.8	∼ 4 sec/image
Ours (FG)	25.9	23.5	90.7	∼ 4 sec/image
Ours (FG*)	27.6	26.1	95.0	∼ 5 sec/image
Ours (BG)	38.8	37.6	**96.9**	∼ 8 sec/image
Ours (BG + NS)	**43.7**	**42.3**	**96.9**	∼ 9 sec/image

Table 1. Comparison of the forward (FG) and backward (BG) strategies, including noise selection (NS). FG*: forward guidance includes `[SoT]` and `[EoT]` tokens. We randomly sampled 1000 text prompts and compute metrics based on VISOR [16].

Model	OA (%)	VISOR (%) uncond	VISOR (%) cond
GLIDE [31]	3.36	1.98	59.06
GLIDE + CDM [30]	10.17	6.43	63.21
DALLE-mini [8]	27.10	16.17	59.67
CogView2 [11]	18.47	12.17	65.89
DALLE-v2 [35]	**63.93**	37.89	59.27
SD [38]	29.86	18.81	62.98
SD + CDM [30]	23.27	14.99	64.41
SD + Ours	40.01	**38.8**	**95.95**

Table 2. Comparison of backward guidance (ours) with text-to-image generation models based on the VISOR [16] protocol.

to prior and concurrent work on three benchmarks.

4.1. Experimental setup

Implementation Details. We utilize Stable-Diffusion (SD) V-1.5 [38] trained on the LAION-5B dataset [42] as the default pre-trained image generator, if not specified. For a detailed description of the architecture and noise scheduler please see the supplement.

For forward guidance, we apply Eq. (1) to every layer of the denoiser network for the first 40 steps of the diffusion process and set $\lambda = 0.8$. For backward guidance, we calculate the loss on the cross-attention maps of the mid-block and the first block of the up-sampling branch of the denoising network (U-Net [39]) as we found this to be the optimal setting to balance control and fidelity. We set $\eta = 30$ by default but found that values between 30-50 work well across most settings. Since the layout of the generated image is typically established in the early stages of inference, backward guidance is performed during the initial 10 steps of the diffusion process and repeated 5 times at each step.

Method	COCO 2014 FID (↓)	COCO 2014 mAP (↑)	Flickr30K FID (↓)	Flickr30K AP_P (↑)	Flickr30K mAP (↑)
MultiDiffusion [4]	70.7	22.3	84.1	21.6	11.9
eDiff-I [2]	72.5	21.7	85.3	21.4	9.7
HFG [43]	72.2	21.5	85.6	22.4	10.7
BoxDiff [48]	72.6	24.1	78.7	26.0	16.6
Stable Diffusion [38]	72.3	19.2	**76.4**	19.4	8.7
Stable Diffusion + Ours	73.3	35.7	78.9	35.6	17.9
GLIGEN [28]	69.1	62.8	77.3	87.2	31.4
GLIGEN + Ours	**66.7**	**65.1**	78.1	**88.9**	**32.7**

Table 3. Comparison with other layout-to-image models. Our approach improves spatial fidelity (suggested by higher AP/mAP scores). mAP is calculated with an IoU threshold of 0.3.

Evaluation Benchmarks. We quantitatively evaluate our approach on three benchmarks: VISOR [16], COCO 2014 [29], and Flickr30K Entities [33, 52]. We discuss the ethical concerns of the dataset usage in the supp. VISOR proposes metrics to quantify the spatial understanding abilities of text-to-image models. For COCO 2014, we follow the same setup adopted by prior work [4], which uses only a subset of the annotated objects per image. Finally, we introduce the Flickr30K Entities dataset as another benchmark

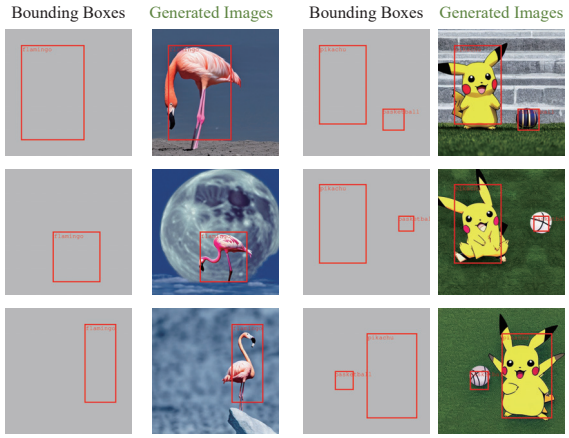

Figure 7. Our method controls the objects inside the generated images with user-specified bounding boxes. On the left, the size and position of *flamingo* changes according to the bounding box. On the right, we show the ability to control multiple objects.

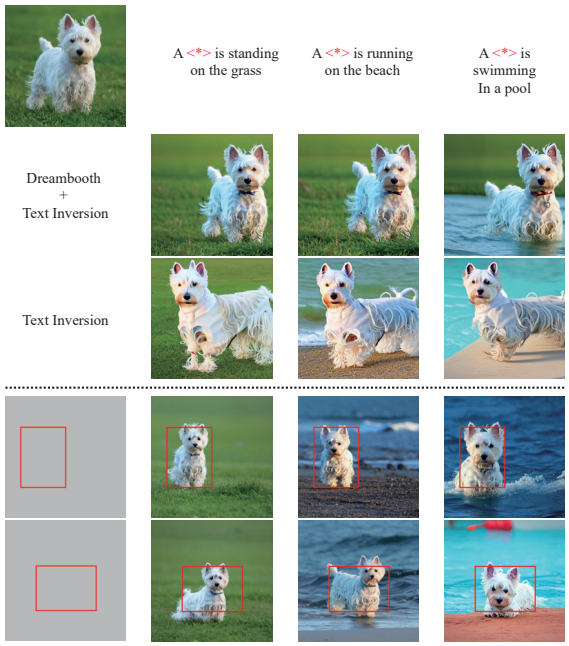

Figure 8. The top left is the real image input. The images above the dash are generations using only text inversion (TI) [15] and Dreambooth [40]. The images under the line are generated by our method on top of Dreambooth and TI.

to evaluate layout control, since it contains image-caption pairs with visual grounding. Details for all benchmarks and metrics are provided in the supplementary material.

4.2. Forward *vs.* Backward Guidance

First, we compare the two different modes of guidance (forward and backward) in Table 1 using the VISOR protocol with 1,000 randomly chosen text samples. The biggest advantage of forward guidance is that the computation overhead is negligible, thus leading to a faster inference time. However, we observe that, compared to (unguided) SD, forward guidance does not significantly increase the object accuracy (OA), while the backward mechanism yields a notably higher OA. In terms of evaluating the generated spatial relationships (VISOR conditional/unconditional metrics), both forward and backward guidance obtain significantly better results than the SD baseline. We also find that the inclusion of `[SoT]` and `[EoT]` tokens improves forward guidance, which confirms our analysis and insights in Section 3.3, yet backward guidance still achieves superior performance. Finally, noise selection using backward guidance offers a significant boost in all metrics.

We provide a qualitative comparison of the forward and backward mechanisms in Figure 5, including the impact of special tokens on forward guidance. Backward guidance achieves a better alignment between the generated objects and the input bounding boxes. It also helps to address the issue of objects occasionally being omitted from the generated images in diffusion models.

4.3. Comparisons to Prior Work

In Table 2, we compare our method with text-to-image generation methods that do not use layout control. We note that comparisons are fair since, in this setting (VISOR), manual user input is not required for guidance (see supplement). Our method exhibits remarkable performance under the $\text{VISOR}_{\text{cond}}$ metric, achieving an accuracy of 95.95%, and higher OA compared to the baseline (SD). Although OA does not directly assess layout, the improvement can be explained by the fact that unguided SD often fails to generate correct semantics in atypical compositions. We also note that, while DALLE-v2 [35] achieves the highest OA overall, it appears to struggle more with layout instructions compared to SD, as indicated by a lower $\text{VISOR}_{\text{cond}}$ score.

In Table 3, we compare our backward guidance to other mechanisms for layout conditioning. Apart from the entries in the last two rows, all methods are based on Stable Diffusion [21] V1.5. Remarkably, our backward guidance surpasses other layout conditioning methods by a significant margin, achieving over a 9-point improvement in mAP and AP_{P} on COCO and Flickr30K. Notably, in direct comparison with the concurrent BoxDiff model [48], we achieve gains of 11.6 in mAP and 9.6 in AP_{P}, all while maintaining analogous image quality. Finally, we show that our approach can be used complementarily to methods like GLIGEN [28] that train additional layers for layout conditioning, further improving their performance.

In Figure 6, we qualitatively compare different text-to-image models using prompts sampled from [16]. Methods that do not use layout control are not capable of inferring the spatial relationships between objects based purely on

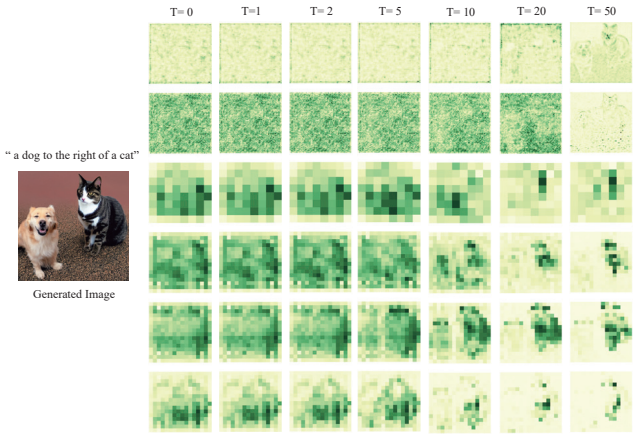

Figure 9. The cross-attention map of the word "cat" at different layers (top to bottom) across different timesteps (left to right).

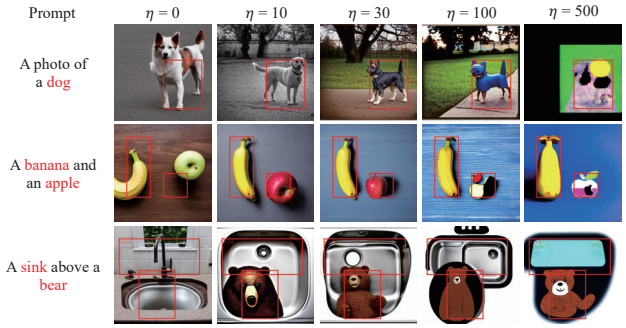

Figure 10. Qualitative comparison of different loss scales in the backward guidance. We increase the loss scale from left to right keeping the same prompt and random seed. With increasing scale, the objects are more tightly constrained inside the bounding boxes. However, for very high scales, fidelity decreases significantly.

textual input and often fail to generate one or both objects. We also observe that even methods with layout conditioning struggle in this setting, especially those that adopt a forward guidance paradigm (eDiff-I [2], HFG [43]). In the case of BoxDiff [48], the lower quality could potentially be due to overlooking the impact of special tokens and the loss function design. In contrast, our approach (backward-guided SD) can accurately position objects within a scene, even when they are rarely seen together, such as "snowboard" and "bowl", and achieves the best adherence to the prompt without loss of image fidelity. More examples of our approach are shown in Figure 7, demonstrating precise control over the *size* and *position* of one or more objects, including unconventional object categories, such as "flamingo" or "pikachu", and atypical scene compositions.

4.4. Further Analysis and Applications

Real-Image Layout Editing. We showcase the potential of backward layout guidance for editing real images in Figure 8, confirming its effectiveness at changing the position, gesture, and orientation of the "dog" (based on the aspect ratio of the bounding box) to fit the new context, without altering its identity. As shown in the same figure, the capability to precisely control object size and position cannot be attained through Dreambooth/TI alone. This highlights the potential of our method in a wide range of applications related to image editing and manipulation.

Cross-attention Layers and Guidance Steps. We also investigate the layers and the number of guidance steps that are necessary to achieve layout control. Cross-attention maps at various layers of the denoising network are presented in Figure 9. We observe that the first layers of (down-sampling) do not capture much information about the object (here, the "cat"). We found it most effective to perform backward guidance only on the mid and up-sampling blocks of the architecture. The figure also illustrates that object outlines are typically generated in the early steps of the diffusion process, before $T = 20$. Based on our experimentation, we find that 10-20 steps are generally suitable for guidance. Additional quantitative analysis and examples are presented in the supplement.

Loss Scale Factor. In Figure 10, we qualitatively analyze the impact of the loss scale factor η. We observe that increasing the loss weight leads to stronger control over the generated images, but at the cost of some fidelity, particularly with higher scales. The optimal loss scale setting depends on the difficulty of the text prompt. For example, an atypical prompt like *"a sink above a bear"* requires stronger guidance to generate both objects successfully (without guidance, *i.e.*, $\eta = 0$, the bear is not generated). This suggests that layout guidance helps the generator "recognize" multiple objects in the text prompt.

5. Conclusions

In this paper, we investigated the potential of manipulating the spatial layout of images generated by large, pre-trained text-to-image models without additional training or fine-tuning. Through our exploration, we discovered that both the cross-attention maps and the initial noise of the diffusion play a dominant role in determining the layout and that even the cross-attention maps of special tokens contain valuable semantic and spatial information. We identify and analyze the mechanism behind most prior work: forward guidance. Moreover, based on our analysis, we propose a new technique "backward guidance" that overcomes the shortcomings of forward guidance. Finally, we demonstrate the versatility of our training-free strategy by extending it to applications such as real-image layout editing.

Acknowledgements. This research is supported by ERC-CoG UNION 101001212.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 2

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 4, 6, 8

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. *arXiv preprint arXiv:2302.07121*, 2023. 2

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2, 6

[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. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[6] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. *arXiv preprint arXiv:2306.13754*, 2023. 2

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[8] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall· e mini, 2021. 6

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[11] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, pages 16890–16902, 2022. 6

[12] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 2

[13] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 2

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 5, 7

[16] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 1, 2, 6, 7

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NeurIPS*, 2014. 2

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 4

[20] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686*, 2019. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 7

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[23] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 2

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

[25] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 91–109. Springer, 2022. 2

[26] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3

[28] 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. *arXiv preprint arXiv:2301.07093*, 2023. 2, 6, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2, 6

[30] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 2, 6

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 6

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[33] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision (IJCV)*, 2017. 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, page 3, 2022. 2, 6, 7

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2, 3

[37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 6

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6

[40] 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. *arXiv preprint arXiv:2208.12242*, 2022. 5, 7

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2, 6

[43] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5997–6006, 2023. 2, 4, 6, 8

[44] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019. 2

[45] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2647–2655, 2021. 2

[46] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 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*, pages 16515–16525, 2022. 2

[47] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 3

[48] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. *arXiv preprint arXiv:2307.10816*, 2023. 2, 6, 7, 8

[49] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[50] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022. 2

[51] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael

Zeng, et al. Reco: Region-controlled text-to-image generation. *arXiv preprint arXiv:2211.15518*, 2022. 2

[52] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 6

[53] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[54] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[55] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[56] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

[57] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[58] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 2