BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion

Jinheng Xie[1] Yuexiang Li[2*] Yawen Huang[2] Haozhe Liu[2,3] Wentian Zhang[2]
Yefeng Zheng[2] Mike Zheng Shou[1*]
[1] Show Lab, National University of Singapore [2] Jarvis Lab, Tencent
[3] AI Initiative, King Abdullah University of Science and Technology
{sierkinhane,mike.zheng.shou}@gmail.com

https://github.com/showlab/BoxDiff

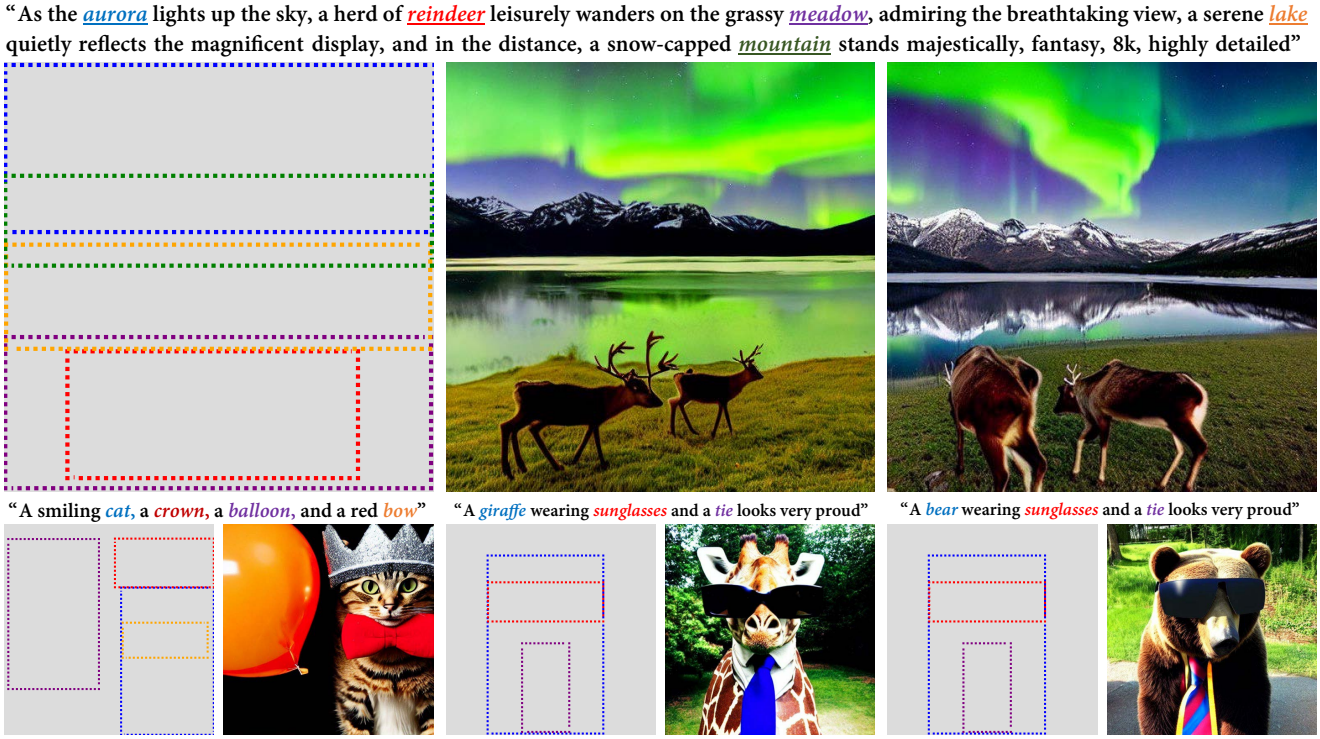

Figure 1: In a training-free manner, BoxDiff consumes the simplest form of user-provided conditions, such as box or scribble, to control the location and scale of contents in the image synthesized by the pre-trained text-to-image diffusion model.

Abstract

Recent text-to-image diffusion models have demonstrated an astonishing capacity to generate high-quality images. However, researchers mainly studied the way of synthesizing images with only text prompts. While some works have explored using other modalities as conditions, considerable paired data, e.g., box/mask-image pairs, and fine-tuning time are required for nurturing models. As such paired data is time-consuming and labor-intensive to acquire and restricted to a closed set, this potentially becomes the bottleneck for applications in an open world. This paper focuses on the simplest form of user-provided conditions, e.g., box or scribble. To mitigate the aforementioned problem, we propose a training-free method to control objects and contexts in the synthesized images adhering to the given spatial conditions. Specifically, three spatial constraints, i.e., Inner-Box, Outer-Box, and Corner Constraints, are designed and seamlessly integrated into the denoising step of diffusion models, requiring no additional training and massive annotated layout data. Extensive experimental results demonstrate that the proposed constraints can control what and where to present in the images while retaining the ability of Diffusion models to synthesize with high fidelity and diverse concept coverage.

* Corresponding Author

1. Introduction

Due to the large-scale publicly available image-text paired data from websites, recent text conditional auto-regressive and diffusion models, such as DALL-E 1 & 2 [26, 25], Imagen [30], and Stable Diffusion [27], have demonstrated as one of the panaceas in generating images with high fidelity and diverse concept coverage. The excellent capacity of image synthesis increases the potential of these models for practical applications, *e.g.,* art creation. However, most existing models can only be conditioned on class labels or text prompts. A few studies tried to use other modalities as conditions, *e.g.,* spatial conditions, to further control the object or context synthesis. More fine-grained control on the location or scale of synthesized objects or contexts would widen the applications of text conditional generative models for the realistic scenario. For example, users can interactively design objects or contexts for human-in-the-loop art creation with additional spatial conditioning input. As a more user-friendly solution, this interactive cooperation with artificial intelligence (AI) would stimulate more potential for content creation.

Layout-to-image literature [19, 32, 33, 36, 40, 12, 2, 1] has studied on the way to synthesize images adhering to the spatial conditioning input. However, the setting of these studies is restricted to the limited closed-set categories, which is infeasible to novel categories in open-world situations. Moreover, the previous studies followed the fully-supervised learning pipeline; hence, considerable paired box/skeleton/mask-image data is required for high-quality training. Since pixel-level annotation is time-consuming and labor-intensive to acquire, label efficiency gradually becomes the bottleneck of fully-supervised layout-to-image methods. Beyond text prompts as conditions, Stable Diffusion [27] and ControlNet [38] have also studied other modalities as conditioning input and provided qualitative results. In contrast to closed-set layout-to-image synthesis methods, Stable Diffusion, and ControlNet nurtured from large-scale image-text pairs have a strong perception of diverse visual concepts, *e.g.,* different kinds of objects and contexts. Nevertheless, they also follow the general pipeline of layout-to-image literature in a fully-supervised manner, in which massive paired image-layout data is indispensable for high-quality training. Besides, the training period is time-consuming for train-from-scratch or fine-tuning.

In this paper, we focus on the most efficient setting for conditional image synthesis. Specifically, the simplest spatial conditions (or termed constraints), *e.g.,* **box or scribble**, from users are adopted to seamlessly control object and context synthesis during the denoising step of Stable Diffusion models, **requiring no additional model training on the substantial paired layout-image data**. As shown in [15], conditioning mechanisms incorporated in Stable Diffusion provide explicit cross-attentions between

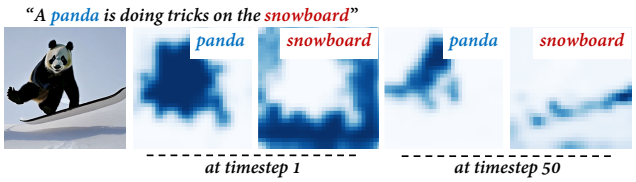

Figure 2: Cross-attentions between target text tokens, *e.g.,* panda, bamboo, snowboard, and intermediate features of the denoiser, *i.e.,* a UNet, in the Stable Diffusion model.

the given text prompt and intermediate features of the denoiser. Specific spatial attention maps for objects or contexts in the text prompt can be accordingly extracted. As the cross-attentions shown in Fig. 2, the spatial location of high-response attention, *i.e.,* panda and snowboard, is perceptually equivalent to that of objects or contexts in the synthesized images. Hence, a simple idea to control the spatial location and scale of objects/contexts to be synthesized is adding guidance or constraints on the extracted cross-attentions. To achieve this goal, we propose a **training-free** approach, namely Box-Constrained Diffusion (BoxDiff), by adding three spatial constraints, *i.e.,* Inner-Box, Outer-Box, and Corner Constraints, on the cross-attentions extracted at each denoising timestep. This plays a role in pointing out directions to update the noised latent vector, which consequently leads synthesized objects or contexts to gradually follow the given spatial conditions. Furthermore, since strong constraints applied to the cross-attentions will affect the denoising step of diffusion models, impairing the fidelity of the resulting synthesized images, we also explore a manner of representative sampling to mitigate the problem. Samples synthesized by the proposed BoxDiff can be found in Fig. 1.

The main contributions of this paper are summarized as:

- We propose a training-free approach, termed Box-Constrained Diffusion (BoxDiff), for text-to-image synthesis following the given spatial conditions, requiring no additional model training and massive paired layout-image data.
- The proposed spatial constraints can be seamlessly incorporated into the denoising step, which retains the strong perception of diverse visual concepts of Stable Diffusion. Hence, our method can synthesize various novel objects and contexts beyond the closed world.
- Extensive experiments demonstrate that the proposed training-free BoxDiff can synthesize photorealistic images following the given spatial conditions.

2. Related Work

Diffusion Models: Recently, diffusion models have ushered in a new era of image generation. It consists of a forward process, *i.e.,* adding noise, and a reverse process,

i.e., removing noise. The denoising diffusion probabilistic model (DDPM) [31, 16] learns to invert a parameterized Markovian image noising process. Given isotropic Gaussian noise samples, they can transform them into signals, *e.g.,* images, by iteratively removing the noise. Beyond pure noise to image fashion, class-conditional and image-guided synthesis have also been explored [8, 9, 22]. In contrast to denoising in the pixel space, Rombach *et al*. [27] proposed to operate on the compressed latent space by employing an autoencoder. This significantly lowers the training costs and speeds up the inference time while retaining the ability to generate high-quality images.

Text-to-Image Models: Recently, large-scale image-text pairs available on the Internet dramatically enabled generative models, *e.g.,* DALL-E [26], Imagen [30], and Stable Diffusion [27], to synthesize images in higher quality and richer diversity. [10] and recently introduced [6] operated on the cross-attention for better content consistency to subjects in text prompts. However, they are chiefly conditioned on the text prompts or class labels. As compensation, a few works [12, 1, 2] have been proposed to handle additional spatial layout conditions, but most of them require additional model training or have only a limited scope of knowledge. For example, Gafni *et al*. [12] incorporated semantic maps to control objects & contexts in the generated images. However, it is restricted to the closed-set world (only 158 categories). Concurrently, Balaji *et al*. [2] illustrated that a modification of the attention map can lead to corresponding changes in the synthesized objects. Motivated by the above, we are interested in controlling object synthesis by the simplest form of conditions, *e.g.,* box or scribble, from users, potentially motivating simpler and more efficient interactive cooperation of image synthesis.

Layout-to-Images Models: Traditional layout-to-image literature [19, 32, 33, 36, 40, 12] has been focused on how to synthesize images adhering to the given bounding boxes of object categories. Generally, they follow the pipeline, *i.e.,* training and validation, to obtain layout-to-image models, and promising results have been obtained. However, they are trapped in a dilemma of time-consuming and labor-intensive annotation like box/mask-image paired data. In addition, they are greatly restricted to a fixed number of categories, failing to synthesize novel categories in the open world. The image quality of such models is also lower than that of the recently introduced large-scale image-text-pairs-driven generative models. Recently, fine-tuning Stable Diffusion models to adhere to additional layout information has also been explored in [18] and [37]. Compared to the above methods, we propose a training-free approach by adding the simplest constraints, *e.g.,* box or scribble, from users to the denoising step of Stable Diffusion models. It requires no additional training and paired layout-image data. Besides, the proposed approach has the ability to synthesize a wide range of visual concepts rather than the limited closed set. Note that, there are many concurrent works [23, 7, 3, 21] that studied a similar area. For example, both Chen *et al*. [7] and Phung *et al*. [23] operated constraints on the cross-attention to control the synthetic contents. More recently, VisorGPT [35], LMD [20], LayoutGPT [11], and ControlGPT [39] have been proposed to plan visual layouts for image synthesis models. Along with image personalization models such as DreamBooth [29], Textual Inversion [13], Mix-of-Show [14], and Perfusion [34], our BoxDiff has become increasingly feasible to create a more complex scene with personalized contents from Diffusion Models.

3. Preliminaries: Stable Diffusion

Different from [16, 9], the Stable Diffusion model efficiently operates on the latent space. Specifically, an autoencoder consisting of an encoder $\mathcal{E}$ and decoder $\mathcal{D}$ is trained with a reconstruction objective. Given an image $\mathbf{x}$, the encoder $\mathcal{E}$ maps it to a latent $\mathbf{z}$, and the decoder $\mathcal{D}$ reconstructs the image from the latent, *i.e.,* $\widetilde{\mathbf{x}} = \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathcal{E}(\mathbf{x}))$. In this way, at each timestep t, a noisy latent $\mathbf{z}_t$ can be obtained. Beyond the routine training scheme, Stable Diffusion devises conditioning mechanisms to control the synthesized image content by an additional input, *e.g.,* a text prompt $\mathbf{y}$. The text prompt is first pre-processed to text tokens $\tau_\theta(\mathbf{y})$ by the text encoder of pre-trained CLIP [24]. The DDPM model ϵ_θ can then be trained via:

$$\mathcal{L}_{DDPM} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{y},\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{y}))||_2^2\right], \tag{1}$$

where UNet [28] enhanced with self-attention and cross-attention layers is adopted as the denoiser ϵ_θ. During training, given a noised latent $\mathbf{z}_t$ at timestep t and text tokens $\tau_\theta(\mathbf{y})$, denoiser ϵ_θ is tasked with predicting the noise ϵ added to the current latent.

In inference, a latent $\mathbf{z}_T$ is sampled from the standard normal distribution $\mathcal{N}(0, 1)$ and the DDPM is used to iteratively remove the noise in $\mathbf{z}_T$ to produce $\mathbf{z}_0$. In the end, the latent $\mathbf{z}_0$ is passed to the decoder $\mathcal{D}$ to generate an image $\widetilde{\mathbf{x}}$.

4. Methodology

In this section, we present the proposed BoxDiff approach and spatial constraints in detail.

4.1. Cross-Modal Attention

Conditioning mechanisms in the Stable Diffusion model can explicitly form the cross-attentions between text tokens and intermediate features of the denoiser ϵ_θ. In the denoising step, given the conditioning text tokens $\tau_\theta(\mathbf{y})$ and intermediate features $\varphi(\mathbf{x}_t)$, the cross-attention $\mathbf{A}$ can be accordingly acquired:

$$\mathbf{A} = \text{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}), \tag{2}$$

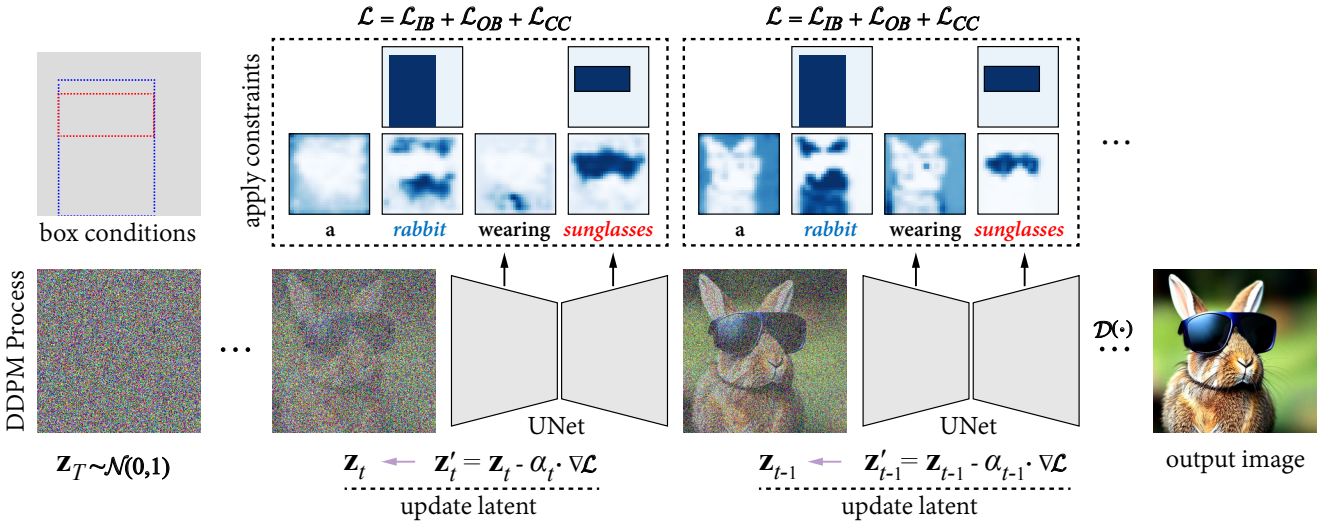

Figure 3: Overview of our BoxDiff. Given the box conditions, we transform them into a set of binary spatial masks. A latent $\mathbf{z}_T$ sampled from Normal distribution $\mathcal{N}(0,1)$ is passed to the denoiser, *i.e.,* UNet, to obtain the denoised latent. At timestep t, $\mathbf{z}_t$ is first passed to UNet to get the cross-attention maps, on which the proposed constraints, *i.e.,* $\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}$, are applied. Subsequently, the current latent $\mathbf{z}_t$ can be updated by the gradient $\nabla\mathcal{L}$ to get $\mathbf{z}'_t$ for the denoising step.

$$\mathbf{Q} = \mathbf{W}_Q \varphi(\mathbf{x}_t),\ \mathbf{K} = \mathbf{W}_K \tau_\theta(\mathbf{y}), \tag{3}$$

where $\mathbf{Q}, \mathbf{K}$ are the projection of intermediate features $\varphi(\mathbf{x}_t)$ and text tokens $\tau_\theta(\mathbf{y})$ by two learnable matrices $\mathbf{W}_Q, \mathbf{W}_K$, respectively. At each timestep t, given $\tau_\theta(\mathbf{y})$ with N text tokens $\{\mathbf{s}_1, \cdots, \mathbf{s}_N\}$, the cross-attention $\mathbf{A}^t$ containing N spatial attention maps $\{\mathbf{A}^t_1, \cdots, \mathbf{A}^t_N\}$ can be consequently obtained. Here, following [6], we remove the cross-attention between the start-of-text token (*i.e.,* [sot]) and intermediate features before applying Softmax$(\cdot)$. Besides, a Gaussian filter is applied to smooth the cross-attentions along the spatial dimension. Therefore, the aforementioned operations bring an enhancement on cross-attentions between actual subject tokens, *e.g.,* object or context, with the intermediate features. Cross-attention can be performed at different scales, *i.e.,* $64 \times 64, 32 \times 32, 16 \times 16$, and 8×8. Following [15], we operate the proposed constraints on the cross-attentions with a resolution of 16×16 as the inherent sufficient semantic information.

4.2. Box-Constrained Diffusion

Given a text prompt with a set of target tokens $\mathcal{S} = \{\mathbf{s}_i\}$ and a set of user-provided object or context locations $\mathcal{B} = \{\mathbf{b}_i\}$ as spatial conditions, a set of corresponding spatial cross-attention maps $\mathcal{A}^t = \{\mathbf{A}^t_i\}$ between target tokens and intermediate features can be accordingly obtained. For example, cross-attention over target tokens such as "rabbit" and "sunglasses" can be yielded (as shown in Fig. 3). Each location $\mathbf{b}_i$ contains the user-provided top-left and bottom-right coordinates $\{(x^i_1, y^i_1), (x^i_2, y^i_2)\}$.

It can be observed from Fig. 2 that, during the denoising step of the Stable Diffusion model, the location and scale of high response regions in the cross-attention map are perceptually equivalent to that of synthesized objects in the decoded image $\widetilde{\mathbf{x}}$. This motivates us that constraints can be added on the cross-attention to control the synthesis of target objects in the image $\widetilde{\mathbf{x}}$. As shown in Fig. 3, given user-provided location, *i.e.,* $\{\mathbf{b}_i\}$, a set of binary spatial masks $\mathcal{M} = \{\mathbf{M}_i\}$ can be transformed from the top-left and bottom-right coordinates, where each $\mathbf{M}_i \in \mathbb{R}^{16\times16}$. Our target is to synthesize target objects approaching the mask regions. To achieve this goal, we propose three spatial constraints, *i.e.,* Inner-Box, Outer-Box, and Corner Constraints, over the target cross-attention maps $\mathcal{A}^t$ to gradually update the latent $\mathbf{z}_t$ such that the location and scale of synthesized objects will be consistent with the mask region. Henceforward, the diffusion model with our three constraints is named as Box-Constrained Diffusion (BoxDiff).

Inner-Box Constraint: To ensure the synthesized objects will approach the user-provided locations, a simple solution is to ensure that high responses of cross-attention are only in the mask regions. To this end, we propose the inner-box constraint as below:

$$\mathcal{L}^1_{\mathbf{s}_i} = 1 - \frac{1}{P} \sum \mathbf{topk}\left(\mathbf{A}^t_i \cdot \mathbf{M}_i, P\right), \tag{4}$$

$$\mathcal{L}_{IB} = \sum_{s_i \in \mathcal{S}} \mathcal{L}^1_{\mathbf{s}_i}, \tag{5}$$

where $\mathbf{topk}(\cdot, P)$ means that P elements with the highest response would be selected. As observed in the experiments, constraints added on all elements in the cross-attention map potentially lead to a collapse of image fidelity. Besides, constraints on only a few elements with high responses are sufficient to affect the synthesis of objects, which can reduce the impact of constraints and prevent the failure of denoising. Hence, only P elements are

constrained to update the latent $\mathbf{z}_t$. Binary mask $\mathbf{M}_i$ in Eq. (4) aims to mask out elements of the cross-attention maps within the mask regions and $\mathcal{L}_{IB}$ plays a role to maximize the response of the mask-out elements.

Outer-Box Constraint: However, the involvement of Inner-Box Constraint can only ensure that the user-provided regions in $\widetilde{\mathbf{x}}$ will contain objects. It cannot guarantee that no object pixels are synthesized out of the user-provided boxes. To prevent the object from moving out of the target regions, we propose the outer-box constraint as follows:

$$\mathcal{L}^2_{\mathbf{s}_i} = \frac{1}{P}\sum \textbf{topk}\left(\mathbf{A}^t_i \cdot (1-\mathbf{M}_i), P\right), \tag{6}$$

$$\mathcal{L}_{OB} = \sum_{s_i \in \mathcal{S}} \mathcal{L}^2_{\mathbf{s}_i}. \tag{7}$$

In Eq. (6), we get the reversion of mask $(1-\mathbf{M}_i)$ to mask out elements of the cross-attention map beyond the target regions. Here, $\mathcal{L}_{OB}$ aims to minimize the response of cross-attentions out of the target regions. Note that two constraints $\mathcal{L}_{IB}$ and $\mathcal{L}_{OB}$ work in a complementary manner.

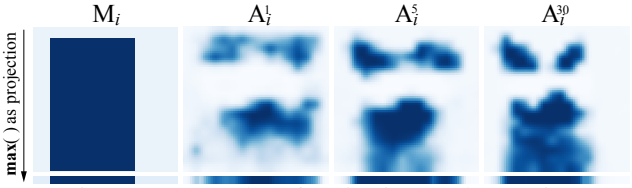

Figure 4: Examples of projection on the x-axis.

Corner Constraint: Since there are only weak spatial conditions, *i.e.,* box or scribble, from users, exact boundary pixels of objects or contexts are not available to restrict the scale. Hence, there will be some tricky solutions, *e.g.,* the target objects are synthesized on a smaller scale than the box regions, with only the above two constraints. In this regard, we propose the corner constraint at the projection of the x-axis and y-axis, respectively. First, we project each target mask $\mathbf{M}_i$ and cross-attention $\mathbf{A}^t_i$ on the x-axis via the **max** operation as below:

$$\mathbf{m}_x(k) = \textbf{max}_{j=1,\cdots,H}\{\mathbf{M}_i(j,k)\}, \tag{8}$$

$$\mathbf{a}^t_x(k) = \textbf{max}_{j=1,\cdots,H}\{\mathbf{A}^t_i(j,k)\}, \tag{9}$$

where $\mathbf{m}_x \in \mathbb{R}^W$ and $\mathbf{a}^t_x \in \mathbb{R}^W$. $\mathbf{m}_x$ is employed as the target and we aim to optimize $\mathbf{a}^t_x$ close to $\mathbf{m}_x$:

$$\mathcal{L}^3_{\mathbf{s}_i} = \frac{1}{L}\sum \textbf{sample}\left(\{|\mathbf{m}_x(k) - \mathbf{a}^t_x(k)|\}^W_{k=1}, L, x^i_1, x^i_2\right), \tag{10}$$

where $\textbf{sample}(\cdot, L, x^i_1, x^i_2)$ indicates a uniform sampling of L error terms from the set $\{|\mathbf{m}_x(k) - \mathbf{a}^t_x(k)|\}^W_{k=1}$ around the given corner coordinates x^i_1 and x^i_2 at x-axis. It plays the same role as the **topk** sampling in Eqs. (4) and (6).

For y-axis projection, same operations are performed:

$$\mathbf{m}_y(j) = \textbf{max}_{k=1,\cdots,W}\{\mathbf{M}_i(j,k)\}, \tag{11}$$

$$\mathbf{a}^t_y(j) = \textbf{max}_{k=1,\cdots,W}\{\mathbf{A}^t_i(j,k)\}, \tag{12}$$

$$\mathcal{L}^4_{\mathbf{s}_i} = \frac{1}{L}\sum \textbf{sample}\left(\{|\mathbf{m}_y(j) - \mathbf{a}^t_y(j)|\}^H_{j=1}, L, y^i_1, y^i_2\right). \tag{13}$$

The corner constraint is the summation of $\mathcal{L}^3_{\mathbf{s}_i}$ and $\mathcal{L}^4_{\mathbf{s}_i}$ as below:

$$\mathcal{L}_{CC} = \sum_{s_i \in \mathcal{S}} \mathcal{L}^3_{\mathbf{s}_i} + \mathcal{L}^4_{\mathbf{s}_i}. \tag{14}$$

At each timestep, **overall constraints** is formulated as:

$$\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}. \tag{15}$$

Having computed the loss $\mathcal{L}$, the current latent $\mathbf{z}_t$ can be updated with a step size of α_t as follow:

$$\mathbf{z}'_t \leftarrow \mathbf{z}_t - \alpha_t \cdot \nabla\mathcal{L}, \tag{16}$$

where α_t decays linearly at each timestep. With a combination of the aforementioned constraints, $\mathbf{z}_t$ at each timestep gradually moves toward the direction of generating high-response attention in the given location and with a similar scale to the box, which leads to a synthesis of target objects in the user-provided box regions.

Note: We prioritize enabling users to provide conditions in the possibly simplest way, *i.e.,* bounding boxes. Beyond that, BoxDiff can interact with other types of conditions such as scribble. More details are in the appendix. Additionally, BoxDiff can be used as a plug-and-play component in many diffusion models, including GLIGEN [18].

5. Experiments

5.1. Experimental Setup

Datasets: Current layout-to-image methods are mainly trained on paired layout-image data of COCO-Stuff [5] or VG [17]. It is unfair to directly make comparisons between our training-free BoxDiff and the fully-supervised methods. Hence, we propose to compare performance on a new dataset. Details can be found in the appendix. Specifically, we collect a set of images (no intersection with COCO and VG) and use YOLOv4 [4] to detect objects. For evaluation, we consider two types of situations: i) a single instance, *i.e.,* "a {}"; ii) multiple instances, *i.e.,* "a {}, a {}". In this way, 189 different text prompts are combined with conditional boxes for image synthesis.

Evaluation Metrics: To validate the effectiveness of our BoxDiff, YOLOv4 is employed to detect object-bounding boxes and predict classification scores on the synthesized images. YOLO score [19], including AP, AP_{50} and AP_{75}, is adopted to evaluate the precision of the conditional synthesis. Additionally, we employ a metric of Text-to-Image Similarity (T2I-Sim) to explicitly evaluate the correctness of semantics in the synthesized images. In particular, synthesized images and the text prompts, *e.g.,* "a photo of {}"

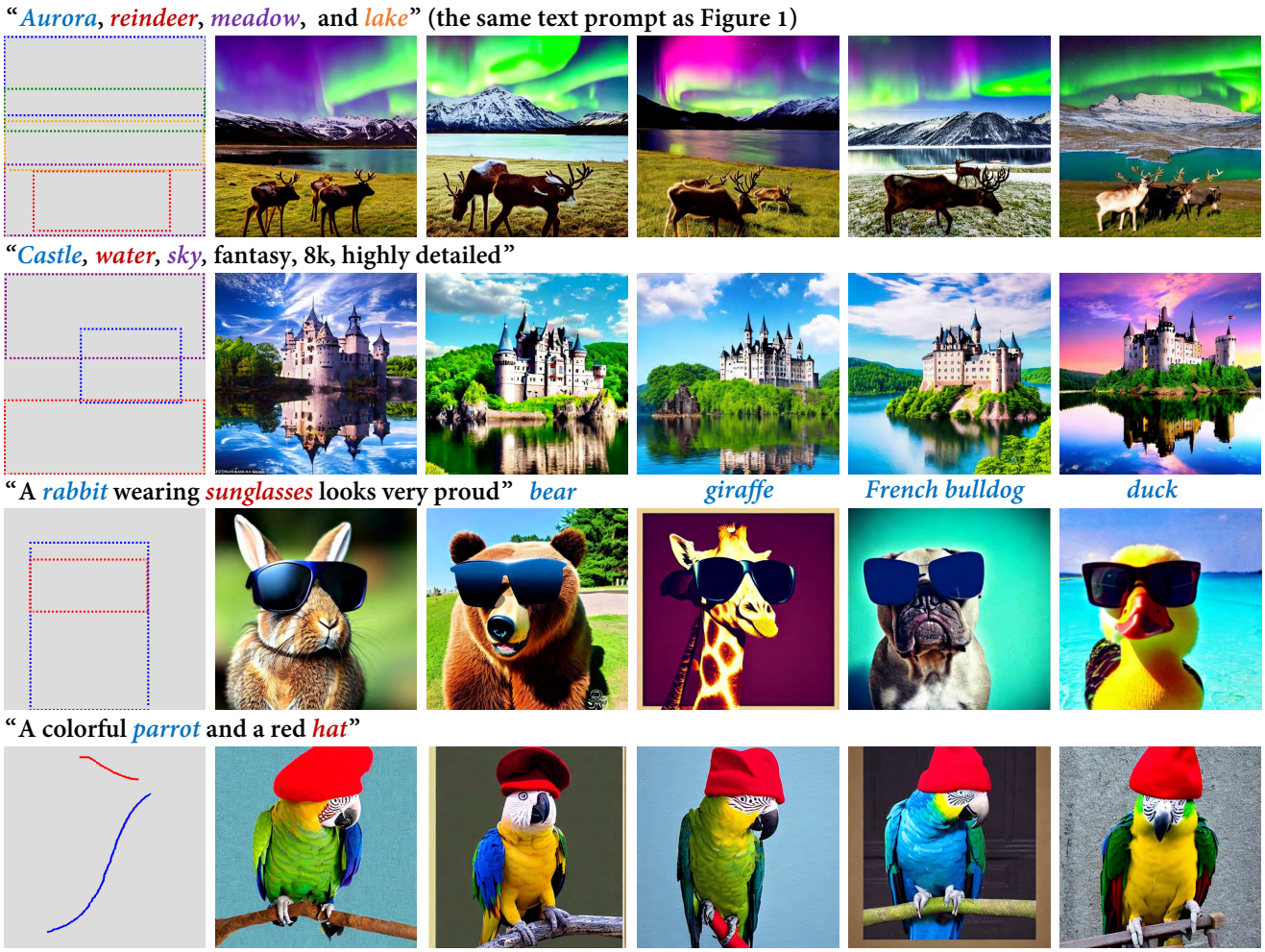

Figure 5: Multiple samples synthesized with fixed spatial conditioning inputs.

Table 1: Ablation studies on various components.

$\mathcal{L}_{IB}$	$\mathcal{L}_{OB}$	$\mathcal{L}_{CC}$	**sample**$(\cdot)$	**topk**$(\cdot)$	T2I-Sim (↑)	AP (↑)
Stable Diffusion					0.3511	2.8
✓			✓	✓	0.3516	9.8
✓	✓		✓	✓	0.3518	20.2
✓	✓	✓	✓	✓	0.3513	22.3
✓	✓	✓	✓		0.3489	24.8
✓	✓	✓		✓	0.3472	7.7

or "a photo of {} and {}", are passed to the image and text encoder of pre-trained CLIP [24], respectively, to calculate their similarity (*i.e.,* T2I-Sim). In CLIP feature space, the similarity can reflect whether the semantics of objects or contexts are correctly presented in the images.

5.2. Ablation Studies

Impact of Various Constraints: To validate the impact of $\mathcal{L}_{IB}$, $\mathcal{L}_{OB}$, and $\mathcal{L}_{CC}$, we perform ablation studies on different combinations of constraints, and the results are listed in Table 1. As shown, the model achieves a T2I-Sim of 0.3516 and an AP of 9.8 in terms of YOLO score with only the inner-box constraint $\mathcal{L}_{IB}$. Such a result reveals that the synthesized objects are mostly not consistent with the conditional spatial input. As $\mathcal{L}_{IB}$ and $\mathcal{L}_{OB}$ work complementary to restrict the cross-attention of objects inside the conditional boxes, a higher YOLO score of 20.2 AP is achieved on the synthesized images. When corner constraint $\mathcal{L}_{CC}$ is involved to limit the corner elements on the projection of cross-attention, the scales of synthesized objects are guaranteed to be consistent with the given bounding box conditions, which accordingly increases the AP from 20.2 to 22.3. Obviously, the proposed constraints are effective in controlling the location and scale of synthesized objects. Visual variations can be found on the left in Fig. 6.

Impact of Representative Sampling: As aforementioned, adding constraints to all elements in the cross-attentions may potentially affect image synthesis. The quantitative evaluation is presented in Table 1. Without **topk**$(\cdot)$ in Eq. (4) and Eq. (6), though there is an improvement of AP, T2I-Sim of the synthesized images decreases. This accordingly represents that the consistency between semantics synthesized in the images and the given text prompts is impaired, and the image quality is decreased.

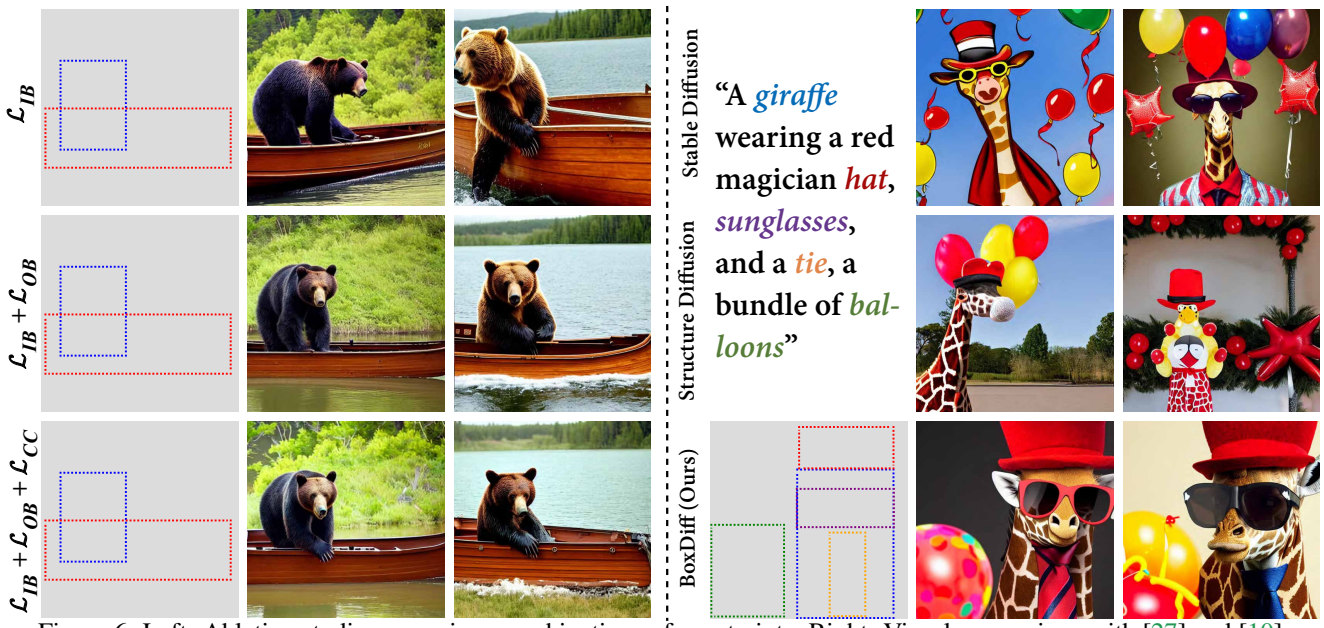

Figure 6: Left: Ablation studies on various combinations of constraints. Right: Visual comparison with [27] and [10].

Table 2: Comparison among various sampling manners.

	All Sampling	Random Sampling	**topk**$(\cdot)$ (Ours)
AP (↑)	24.8	21.4	22.3
T2I-Sim (↑)	0.3489	0.3491	0.3513

When **sample**$(\cdot)$ in Eq. (10) and Eq. (13) is removed, T2I-Sim of synthesized images significantly degrades. The removal of **sample**$(\cdot)$ also impairs the consistency of synthesized objects to the conditional input, leading to a lower AP. Hence, we adopt **topk**$(\cdot)$ and **sample**$(\cdot)$ for the better image quality and consistency with the text prompts.

Impact of sampling in $\mathcal{L}_{IB}$ and $\mathcal{L}_{OB}$. In Table. 2. One can observe that while sampling all pixels in $\mathcal{L}_{IB}$ and $\mathcal{L}_{OB}$ can lead to a more precise synthesis adhering to the conditions, the quality of synthetic contents will be correspondingly degraded (lower T2I-Sim than that of **topk**$(\cdot)$). Randomly sampling pixels cannot effectively maximize the activation of foreground pixels and may activate background regions in cross-attention, leading to significant degradation of the AP and T2I-Sim. To balance these trade-offs, we propose **topk**$(\cdot)$, which achieves the best synthetic quality while maintaining a relatively good AP.

5.3. Visualization Results

Fixing Locations and Scales: Fig. 5 presents synthesized samples using BoxDiff and samples at each row have the same spatial conditioning input. Given a text prompt "a {} wearing sunglasses", and the conditioning layout, the location and scale of the animals and sunglasses in the images are consistent with that of the conditional boxes. In addition, one can observe in other rows, the mountains, aurora, castle, and hats are also nearly consistent with the locations and scales of the given conditional boxes.

Visual Comparison: We present visual comparison between the proposed BoxDiff with the state-of-the-art text-to-image synthesis models such as Stable Diffusion [27] and Structure Diffusion [10] in Fig. 6. Beyond text prompts as conditional input for image synthesis, additional spatial layout, *e.g.,* box, is used in BoxDiff. One can observe from the figure that, in Stable Diffusion and Structure Diffusion, some subjects are occasionally missing in the synthesized images, *e.g.,* the tie in the second column. Besides, these methods may yield unexpected subjects like the soldier, in which the helmet is actually the target object. In contrast, given spatial conditions, the proposed BoxDiff can correctly synthesize target objects we want in the images. In addition, objects are relatively consistent with the conditional boxes.

Varying Locations: Fig. 7 presents synthesized samples using BoxDiff and samples at each row have one fixed and one varying box constraint. Specifically, given a text prompt of "a castle in the middle of a calm lake", the same conditioning inputs are applied for the calm lake, and the conditioning of the castles varies from right to left at the top. Clearly, the lake is always synthesized in the bottom, obeying the given fixed red dashed box. Besides, the location of the castle is changed from the leftmost to the rightmost according to the varying conditioning, *i.e.,* blue dashed box. The same visual variations can also be found in the third row, in which the location of the synthesized castle is moved according to the conditional box. Note that while the text prompts contain the words, *e.g.,* "in the middle of", indicating the positional relation between objects, the proposed constraints added on the cross-attentions have a stronger impact on the position of synthetic contents.

Varying Scales: We further probe the controllability of the object scale of the proposed BoxDiff, and visual results

Figure 7: Synthesized samples obtained with various spatial conditioning inputs, *e.g.,* location, scale, and content.

are illustrated in the second and third rows of Fig. 7. Given a text prompt of "a castle in the middle of a marsh", we constrain the location and scale of the marsh by a fixed red dashed box and vary the scale of the castle by expanding the red dashed box. Clearly, the size of castles varies from small to large following the given conditional boxes, and the location and scale of the marsh are kept unchanged.

Multi-level Variations: Beyond variation at a single aspect, we simultaneously vary at multiple aspects, *e.g.,* scale, and content, to further demonstrate the effectiveness of our method. As shown in the third row of Fig. 7, given a text prompt of "a castle in the middle of {}" and a fixed red dashed box, BoxDiff can successfully synthesize different contents, *i.e.,* "the grasslands", "a sea of sunflowers", and "a marsh", in the red dashed box while controlling the scale of the castle from small to large.

5.4. Quantitative Results

As shown in Table 3, we compare fully-supervised layout-to-image methods, *e.g.,* LostGAN [32], LAMA [19], and TwFA [36] using the newly collected spatial conditions. One can observe from the table, BoxDiff significantly outperforms those fully-supervised ones in terms of the YOLO score. Besides, BoxDiff also achieves the best T2I-Sim, which equivalently represents a better precision of semantic synthesis. Besides, when BoxDiff is integrated, the performance of GLIGEN [18] can be further improved. This validates that our BoxDiff can be used as a plug-and-play component to improve the existing models.

Table 3: Comparison to fully-supervised methods. †: model inference with FP16 due to the memory cost.

Methods	Layout Data	T2I-Sim (↑)	YOLO score		
			AP (↑)	AP_{50} (↑)	AP_{75} (↑)
$LostGAN_{TPAMI'21}$ [32]	COCO-Stuff	0.2279	5.3	8.9	5.6
$LAMA_{ICCV'21}$ [19]	COCO-Stuff	0.2396	10.2	15.3	11.7
$TwFA_{CVPR'22}$ [36]	COCO-Stuff	0.2443	10.6	14.7	12.6
Stable Diffusion [27]	None	0.3511	2.8	9.2	1.1
Stable Diffusion [27] + BoxDiff	None	0.3513	22.3	46.8	20.2
GLIGEN† [18]	COCO-Stuff	0.3489	29.7	45.8	33.9
GLIGEN† [18] + BoxDiff	COCO-Stuff	0.3511	40.2	62.0	46.2

6. Conclusion and Discussion

This paper proposed a training-free approach, *i.e.,* BoxDiff, to controlling object synthesis in spatial dimensions. In contrast to conventional layout-to-image methods, the proposed constraints are seamlessly applied to the denoising step of Diffusion models, requiring no additional training. Extensive results demonstrated that BoxDiff enabled the Diffusion models to control objects and contexts where to synthesize.

To exploit semantic information effectively, we only applied spatial constraints to the cross-attentions at the scale of 16×16. Resolution potentially restricts the precision of the control of object and context synthesis. We believe that as only the simplest form of conditions, *e.g.,* box or scribble, are required, BoxDiff can be potentially extended to data synthesis adhering to additional bounding box conditions, from which a lot of downstream tasks, such as open-vocabulary, weakly- and semi-supervised detection, would benefit. More discussions are included in the appendix.

Acknowledgment This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008, Mike Zheng Shou's Start-Up Grant from NUS, and the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2022) Award 22-5406-A0001. Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang and Yefeng Zheng are funded by Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001) and the Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (No. 2020AAA0104100). Haozhe Liu is also partially supported by the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

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. *arXiv preprint arXiv:2211.14305*, 2022. 2, 3

[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, 3

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPR*, pages 843–852, 2023. 3

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[6] 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. *arXiv preprint arXiv:2301.13826*, 2023. 3, 4

[7] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 3

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, pages 8780–8794, 2021. 3

[10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 3, 7

[11] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. 3

[12] 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 *ECCV*, pages 89–106, 2022. 2, 3

[13] 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. 3

[14] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Chen Yunpeng, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Shan Ying, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 3

[15] 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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pages 6840–6851, 2020. 3

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5

[18] 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. *CVPR*, 2023. 3, 5, 8

[19] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, pages 13819–13828, 2021. 2, 3, 5, 8

[20] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 3

[21] Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint arXiv:2302.13153*, 2023. 3

[22] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[23] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 3

[24] 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 *ICML*, pages 8748–8763, 2021. 3, 6

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

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 2, 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 7, 8

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[29] 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. In *CVPR*, pages 22500–22510, 2023. 3

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 3

[32] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *TPAMI*, 44(9):5070–5087, 2021. 2, 3, 8

[33] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *AAAI*, number 3, pages 2647–2655, 2021. 2, 3

[34] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *arXiv preprint arXiv:2305.01644*, 2023. 3

[35] Jinheng Xie, Kai Ye, Yudong Li, Yuexiang Li, Kevin Qinghong Lin, Yefeng Zheng, Linlin Shen, and Mike Zheng Shou. Visorgpt: Learning visual prior via generative pre-training. *arXiv preprint arXiv:2305.13777*, 2023. 3

[36] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *CVPR*, pages 7764–7773, 2022. 2, 3, 8

[37] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 3

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

[39] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023. 3

[40] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, pages 8584–8593, 2019. 2, 3