Grounded Text-to-Image Synthesis with Attention Refocusing

Quynh Phung   Songwei Ge   Jia-Bin Huang

University of Maryland, College Park

{quynhpt, songweig,jbhuang}@umd.edu

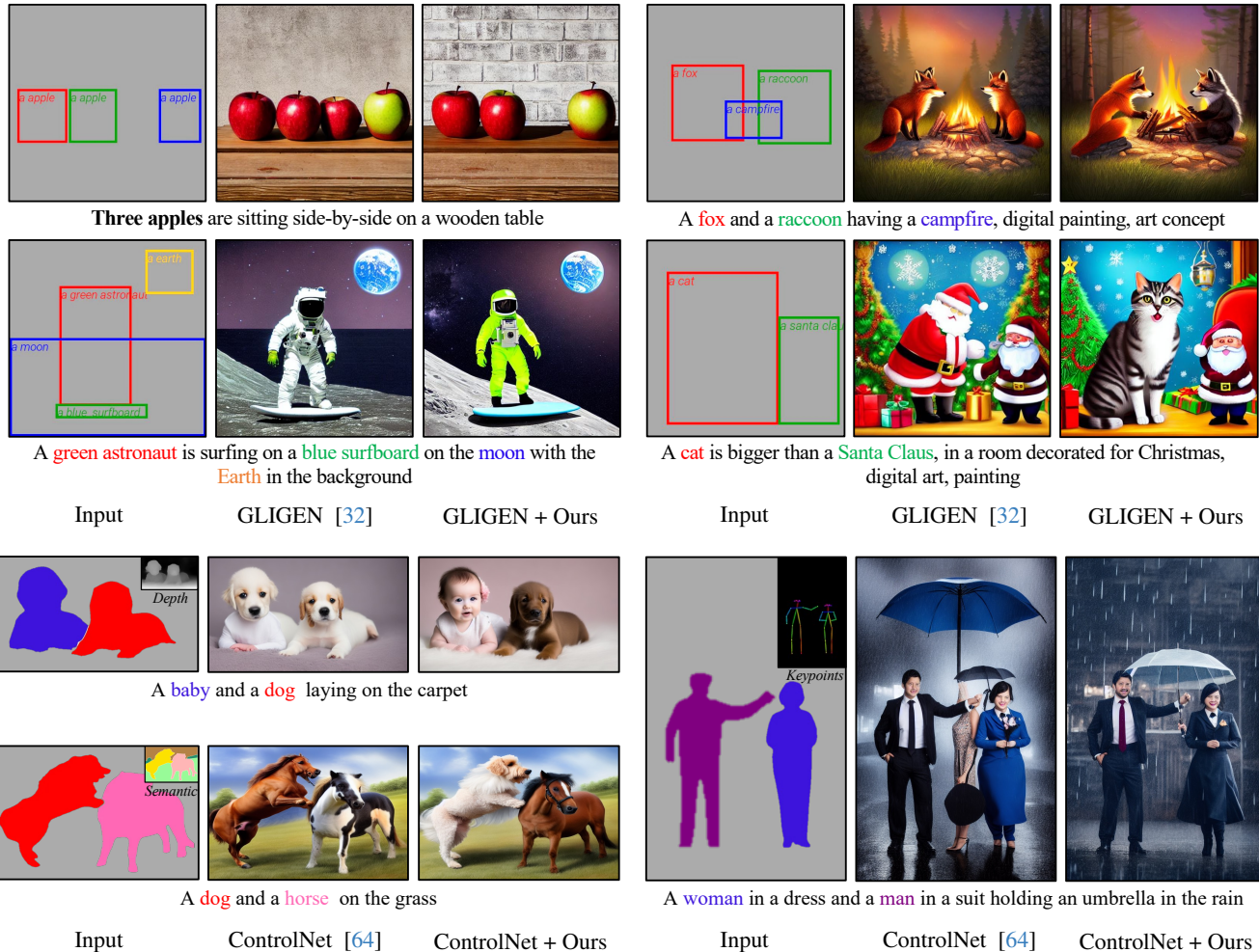https://attention-refocusing.github.io/

Figure 1. **Controllable text-to-image synthesis with attention refocusing.** We introduce a new framework to improve the controllability of text-to-image synthesis. Given a text prompt, we first leverage GPT-4 to generate spatial layouts and then use grounded text-to-image methods to generate the images given the layouts and prompts. However, the detailed information, like the quantity, identity, and attributes, often remains incorrect or mixed using existing models. We propose a training-free approach — attention-refocusing — to substantially improve the controllability. Our method is model-agnostic and can be applied to enhance the control capacity of methods like GLIGEN [32] and ControlNet [64]

Abstract

Driven by the scalable diffusion models trained on large-scale datasets, text-to-image synthesis methods have shown compelling results. However, these models still fail to precisely follow the text prompt involving multiple objects, attributes, or spatial compositions. In this paper, we reveal the potential causes in the diffusion model's cross-attention

and self-attention layers. We propose two novel losses to refocus attention maps according to a given spatial layout during sampling. Creating the layouts manually requires additional effort and can be tedious. Therefore, we explore using large language models (LLM) to produce these layouts for our method. We conduct extensive experiments on the DrawBench, HRS, and TIFA benchmarks to evaluate our proposed method. We show that our proposed attention refocusing effectively improves the controllability of existing approaches.

1. Introduction

Despite the unprecedented zero-shot capacity and photorealism achieved by the recent progress in text-to-image synthesis [4, 28, 45–47, 49, 63], existing models still struggle with text prompts containing multiple objects and attributes with complex spatial relationships among them [3, 10, 11, 14, 42, 59]. Some objects, attributes, and spatial compositions specified in the text prompts are often mixed, swapped, or even completely missing in the synthesized image. Our work aims to mitigate this problem by grounding the text-to-image synthesis using explicit spatial layouts *without extra training*.

The deep level of language understanding exhibited by the text-to-image models can be attributed to using pretrained language models [43] as the text encoder [49]. The computed text embeddings are processed using the *cross-attention layers*, key for various problems [25, 26], in the denoising models [38, 39]. Upon careful analysis of the failure examples generated by Stable Diffusion [47], we identify a potential cause of the failure above in the attention layers [55], where the pixels with similar features produce similar attention queries and consequently attend to a similar set of regions or tokens. The information of these pixels is thus *mixed* through these attention layers. Note that such pixels can come from two different objects with similar features. For example, given the prompt "A dog on the right of a cat", a pixel associated with the token "dog" could have similar features to the pixels in the "cat" region. As a result, the model could incorrectly attend to the "cat" token through the cross-attention layers or the "cat" region through self-attention layers, causing the missing object or blended attribute issues.

Previous studies propose to mitigate this issue by manipulating the cross-attention maps during the sampling process [10, 11, 14]. However, they overlook a similar issue in self-attention layers, where distinguishing between pixels of the same object and those of different objects with similar features becomes a challenge. To this end, we leverage *explicit layout representations* for grounded synthesis following the previous works [11, 32].

In this paper, we propose two novel losses based on the input layout during the sampling process to *refocus* the at-

tention in both self- and cross-attention layers. Our analysis shows that with our losses the attention can be effectively *refocused* to the desired region instead of similar but irrelevant regions. We also explore using LLMs to generate explicit layout representations. We demonstrate that these models have strong spatial reasoning capabilities and can predict the plausible layout of the objects when using our designed prompts with in-context learning. We will release code and data in the future.

We show that when combining the bounding boxes generated by GPT4 [40] and our attention-refocusing losses, our method significantly and consistently improves over several strong baselines on the DrawBench [49], HRS benchmarks [3], and TIFA benchmark [24]. Our main contributions are summarized below:

- We propose attention-refocusing losses to regularize both cross- and self-attention layers during the sampling to improve the controllability given the layout and text prompt;
- We explore using LLMs to generate layouts given text prompts, allowing the exploitation of the up-to-date LLMs with trained text-to-image models;
- We conduct a comprehensive experiment on existing methods of grounded text-to-image generation and show that our method compares favorably against the state-of-the-art models.

2. Related work

Large-scale text-to-image models High-resolution text-to-image synthesis has been dramatically advanced by the development of large-scale text-to-image models [4, 16, 28, 45, 47, 49, 63]. Such rapid progress can be attributed to several critical techniques. First, the availability of large-scale text-image datasets [8, 51] makes it possible to train data-hungry models on a massive volume of samples from diverse resources. The development of the scalable model architectures, including GANs [28, 50], autoregressive models [9, 13, 45, 63], and diffusion models [4, 22, 38, 46, 49], together with various training and inference techniques [21–23, 52]. Our work focuses on the problem of improving the *controllability* of the generated images with respect to the input text with large-scale diffusion models.

Improving the controllability of text-to-image models Enhancing the user control of large-scale text-to-image models has drawn great attention recently. Previous work proposes to boost the controllability through various input formats such as rich text [17], personal images [30, 48], edge maps, segmentation masks, depth maps [64], and bounding boxes [2, 6, 32]. There are also works focusing on strengthing the controllability of the original input text, motivated by the observation that existing models often fail to fulfill the description from the input text accurately [3, 10, 14, 41]. For example, when multiple objects and attributes occur in the text prompt, some are of-
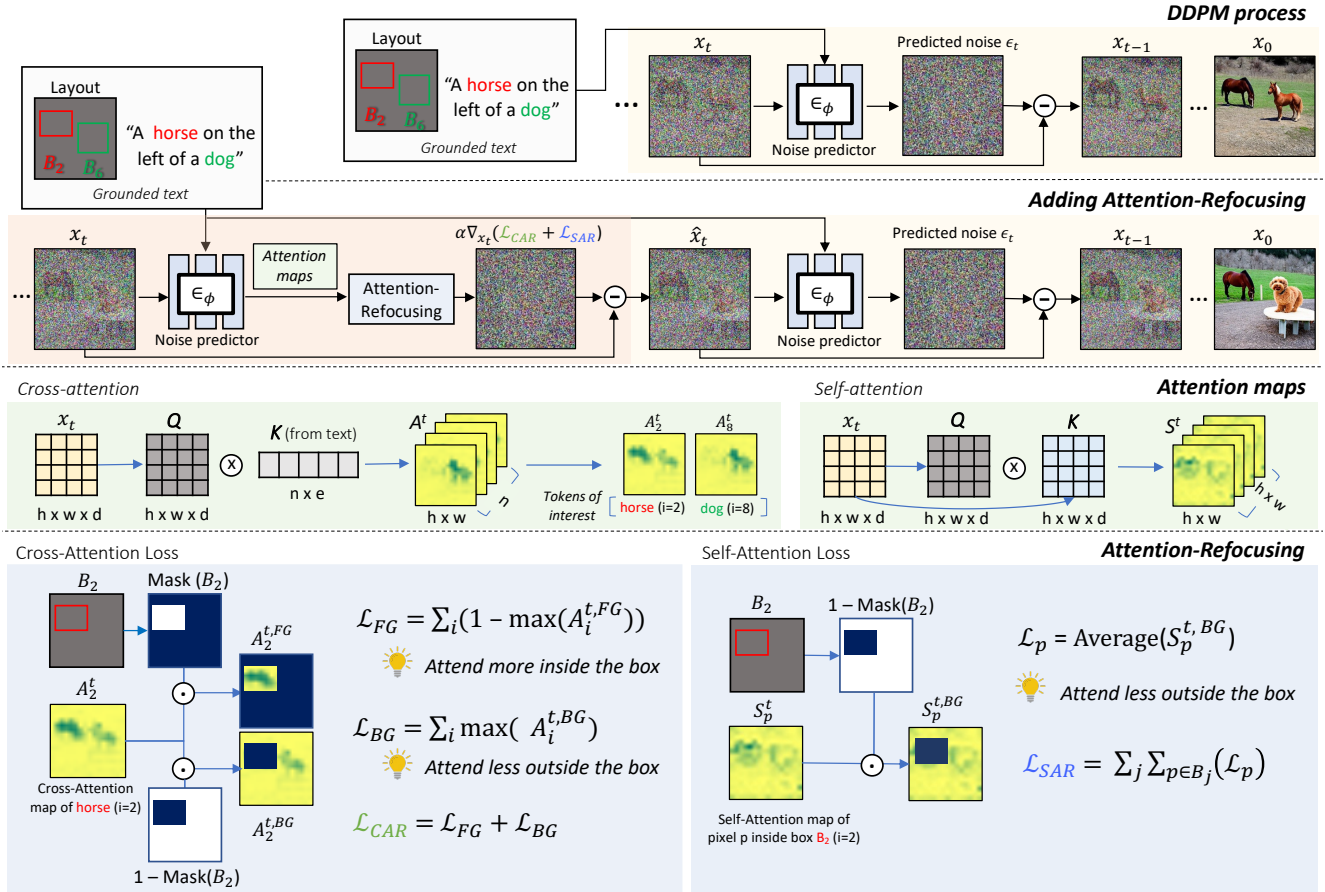
Figure 2. **The proposed Attention-Refocusing framework.** At each denoising step, we update the noised sample by optimizing our $\mathcal{L}_{CAR}$ and $\mathcal{L}_{SAR}$ losses (red block) before denoising with the predicted noise (yellow block). For each cross-attention map, $\mathcal{L}_{CAR}$ is designed to encourage a region to attend more to the corresponding token while discouraging the remaining region from attending to that token. For each self-attention map, $\mathcal{L}_{SAR}$ prevents the pixels in a region from attending to irrelevant regions ($\mathcal{L}_{CAR}$ and $\mathcal{L}_{SAR}$ in blue blocks).

ten missing or mixed in the synthesized images [10, 14]. Attend-and-Excite [10] proposes optimizing cross-attention maps during sampling to ensure all the tokens are attended. Similarly, A-STAR [1] tries to separate attention region of different objects, avoiding objects mixing. Several studies finetune the existing models with human feedback [31, 58] or use improved language models [36, 41, 53, 62, 65] to enhance the text-image alignment. Similar to these recent efforts, our work also focuses on improving the alignment between the generated images and input texts. However, we leverage an intermediate spatial layout generated by LLMs [40, 43, 44, 53] and ground the image synthesis on the layout.

Layout-conditioned text-to-image synthesis. Several approaches have been developed to extend the Stable Diffusion [47] to condition its generation on the layouts through finetuning on layout-image pairs [2, 32, 38, 64] or modifying the sampling process [4–6, 11]. For example, GLIGEN [32] finetunes a gated self-attention layer to incorporate

the box information from the input to the Stable Diffusion model. Mixture-of-Diffusion [27] and MultiDiffusion [30] perform a denoising process on each region and fuse the predicted scores. Others, including SD-Pww [4], layout-predictor [57], direct-diffusion [37] Layout-guidance [11], BoxDiff [61], and R&b [60] directly optimize the cross-attention layers during the sampling process. However, our approach not only optimizes cross-attention maps but also self-attention maps, which is not commonly addressed by these methods. Unlike the optimization of multiple values in these methods, which can lead to image quality degradation, our method iteratively optimize peak values in the attention maps, preserving image quality. Universal guidance [5] leverages a trained object detector and constructs a loss to force the generated images to match location guidance. DenseDiffusion [29] directly modifies all attention maps based on mask guidance without any optimization steps. Differing from DenseDiffusion, our method optimizes the latent space, indirectly influencing attention maps
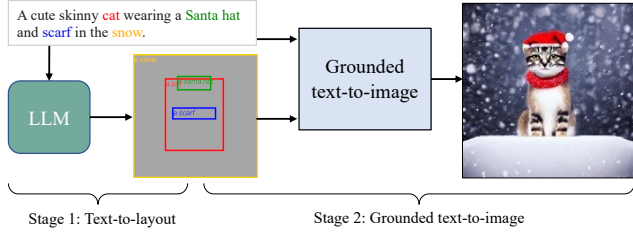
Figure 3. Our pipeline. Our approach includes 1) text-to-layout using GPT-4 model and 2) grounded text-to-image using a pretrained diffusion model with our attention-refocusing.

under mask guidance. Our approach yields improvements demonstrated in our experiments. Our proposed method to ground the text-to-image generation on the layout uses both *cross-attention* and *self-attention* layers without needing extra training or additional models. We demonstrate that adding the proposed attention-based guidance to various base models improves their performance consistently.

Layout predictions. Several concurrent works leverage the potential of large language models for enhancing text-to-image models. Similar to ours, the concurrent work LLM-grounded Diffusion [33] uses GPT-4 as a layout generator. Their extended work, LDM [34], produces dynamic scene layouts rather than single layouts, guiding a diffusion model for video generation. Another concurrent work, LayoutGPT [15], leverages GPT to create layouts from text conditions, then uses GLIGEN [32] to generate images from the created layouts. Cho et al. [12] finetune an open-source language model for the specific text-to-layout task and use standard layout-to-image models for image generation. However, as demonstrated in the experimental results, existing grounded text-to-image models still fail to fulfill the details in the text prompts, like quantity, identity, and attributes. Our work further improves the controllability using attention-based guidance.

3. Method

In this section, we discuss our method for grounded text-to-image generation. Our approach includes two main phases, as shown in Fig. 3: 1) text-to-layout and 2) grounded text-to-image. In both phases, we use off-the-shelf pretrained models *without any extra training*. We exploit the spatial understanding ability in the latest large language models (LLMs) to produce visual representations such as bounding boxes as the layout given a text prompt.

3.1. Preliminaries

Text-to-image diffusion models. The key to the text-to-image diffusion model is the iterative denoising process. A UNet model is trained to progressively denoise the random Gaussian noise by computing the score $\epsilon_t = U(\mathbf{x}_t; \mathbf{c})$, where t is the time step and $\mathbf{c}$ is the embedding for conditional information. Next, we briefly describe the two types of attention layers used in our method.

Cross-attention layer. Text-to-image diffusion models condition its generation on the text prompt via cross-attention layers. Specifically, a pretrained CLIP encoder [43] is often used to encode the text prompt $\mathbf{w} = (w_1, w_2, \cdots w_n)$ and obtain the text embeddings $\mathbf{c} = f_{\text{CLIP}}(\mathbf{w}) \in \mathbb{R}^{n \times e}$, where e is the embedding dimension. The *key* $\mathbf{K} \in \mathbb{R}^{n \times d}$ and *value* $\mathbf{V} \in \mathbb{R}^{n \times d}$ are obtained from text embedding $\mathbf{c}$ with a linear mapping (d is the feature dimension). As shown in the third row of the Fig. 2, given a set of queries $\mathbf{Q} \in \mathbb{R}^{hw \times d}$ computed from the features map of size $h \times w$, the cross-attention map A^t at the step t is computed as:

$$A^t = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \in [0,1]^{hw \times n}, \qquad (1)$$

which is formed by n attention maps $\{A_1^t, ..., A_n^t\}$, where $A_i^t \in [0,1]^{hw}$ denotes the strength of association between a word token w_i and each spatial location in the feature map.

Self-attention layer. It is used to facilitate the use of global information. It propagates the feature at each spatial location to a similar region in the feature map of resolution $h \times w$. With all key-value-query obtained from the same feature map through linear mappings and Eq. (1), the self-attention map is denoted as $S^t \in [0,1]^{hw \times hw}$. Similarly, we use $S_p^t \in [0,1]^{h \times w}$ to denote the self-attention map of all pixels attending to pixel p.

3.2. Grounded text-to-image generation

We now introduce two losses on the attention layers to improve the controllability of the layout-conditioned image synthesis. We consider spatial layouts defined by k bounding boxes $B \in (\mathbb{Z}^+)^{k \times 4}$. Each box is associated with a box caption describing the content inside the box. Given the captions associated with every box, we denote their indices in the input text prompt $\mathbf{w}$ as $I = \{i_1 \cdots i_q\}$ (q is the number of those tokens of interest). Note that each token index i can relate to one or more bounding boxes B_i. Let $\text{Mask}(B_i)$ be the binary mask generated from the boxes B_i, where the regions inside the boxes are one, and the rest are zero, as shown in the fourth row of Fig. 2.

3.2.1 Cross-Attention Refocusing (CAR)

When generating images using GLIGEN, given the text prompt "three parrots", we notice in the cross-attention layer that the token "parrot" incorrectly attends to the unrelated regions, as visualized in Fig. 4a, leading to four parrots generated in the result. To this end, we propose a loss to *refocus* the cross-attention of these tokens according to the layout.

Post-processing cross-attention maps. First, we skip the attention maps of $< sot >$ token, then use $Softmax$ for the remaining cross-attention maps. After that, we apply Gaussian Smoothing to the attention maps following [10].
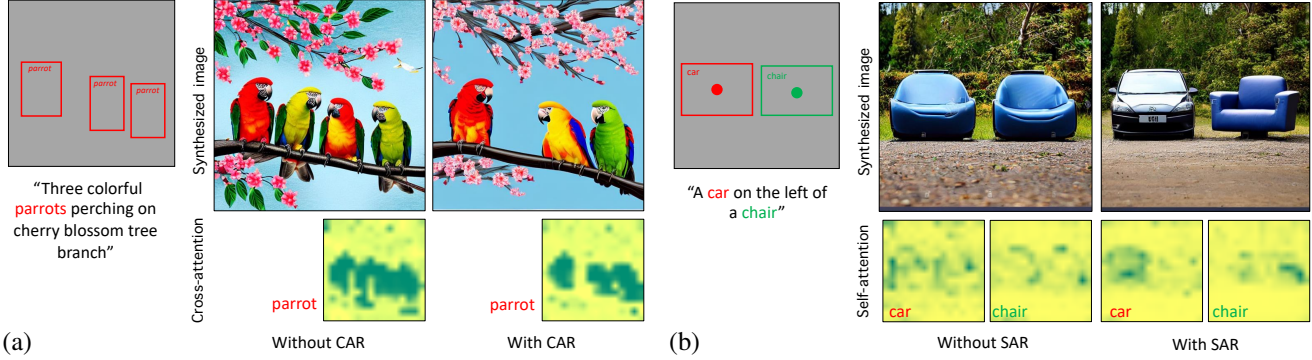
Figure 4. (a) **Cross-Attention-Refocusing (CAR) visualization.** Without CAR, the token "parrot" attends to background regions. Using CAR calibrates the cross-attention map to attend to the correct regions. (b) **Self-Attention-Refocusing (SAR) visualization.** The dots in each box represent the pixel query of the self-attention map. Applying SAR loss helps refocus the self-attention layer to attend *less* to the incorrect regions.

Designing cross-attention refocusing loss. To encourage the generated objects to present in corresponding boxes, we aim to boost the scores of masked attention map $A_i^t \cdot \text{Mask}(B_i)$:

$$\mathcal{L}_{FG} = \frac{1}{q}\sum_{i\in I}(1 - \max(A_i^t \cdot \text{Mask}(B_i))) \qquad (2)$$

We also propose a background loss to discourage the tokens from being attended by the irrelevant regions:

$$\mathcal{L}_{BG} = \frac{1}{q}\sum_{i\in I}\max(A_i^t \cdot (1 - \text{Mask}(B_i))) \qquad (3)$$

The overall CAR loss is then defined as $\mathcal{L}_{CAR} = \mathcal{L}_{FG} + \mathcal{L}_{BG}$. As shown in Fig. 4a, when applying with our loss, the model effectively mitigates the incorrect attention to the grounded tokens and synthesizes three parrots as desired.

3.2.2 Self-Attention Refocusing (SAR)

Similar to the observation in the cross-attention layers, as shown in Fig. 4b, the pixels of one region (e.g., "car") may attend *outside* of the region to similar regions (e.g., "chair") in self-attention layers. As a result, the attributes of the two regions get mixed in the generation. To this end, we develop a loss to help self-attention *refocus* to the correct regions.

Recall that S_p^t is the self-attention map of pixel p. For each pixel $p \in B_i$, we denote $S_p^{t,BG}$ as the background region of the self-attention map:

$$S_p^{t,BG} = S_p^t \cdot (1 - \text{Mask}(B_i)) \qquad (4)$$

We aim to ensure each pixel $p \in B_i$ attends less to the regions outside the boxes B_i. To achieve this, we define self-attention loss for each pixel p as follows:

$$\mathcal{L}_p = \frac{\sum(S_p^{t,BG})}{\sum(1 - \text{Mask}(B_i))} \qquad (5)$$

The overall self-attention loss is defined as follows:

$$\mathcal{L}_{SAR} = \frac{1}{q}\sum_{i\in I}\sum_{p\in B_i}\mathcal{L}_p \qquad (6)$$

As shown in Fig. 4b, using self-attention loss helps each box to focus less on the irrelevant regions, and the model consequently generates distinct attributes for each region.

3.2.3 Sampling with the attention-refocusing losses

With the CAR and SAR losses, we modify the noised sample x_t at each denoising step to minimize the loss using gradient descent. We show the update process in Fig. 2:

$$\hat{x}_t \leftarrow x_t - \alpha\nabla_{x_t}(\mathcal{L}_{CAR} + \mathcal{L}_{SAR}), \qquad (7)$$

where α is the step size that controls the influence of the optimization in the denoising process. However, a single update step is often insufficient to refine the cross-attention and self-attention maps. We thus update τ times every early denoising step. After finishing the τ updates, we feed the output to the diffusion UNet to resume the denoising process and compute x_{t-1}. Intuitively, we use the gradient derived from attention-refocusing losses to guide the denoising process. More details about τ setting and algorithm can be found in the supplementary material.

3.3. Text-to-layout prediction

Generating an image from text requires strong text comprehension and reasoning capacity. The limited power of text encoders could be another reason existing methods fail. However, once a text-to-image model is trained with a specific language model, upgrading the text encoder without additional (costly) training becomes nontrivial. Such schema could hinder the existing text-to-image models from benefiting from recent large language models (LLMs) breakthroughs. Given this challenge, we explore the direct use of LLMs to generate intermediate visual presentations, such as box layouts. We exploit GPT-4 [40], the

Method	CAR & SAR	Counting F1 ↑	Compositions Spatial ↑	Size ↑	Color ↑
Stable Diffusion [47]	×	58.31	8.48	9.18	12.61
	✓	60.62 (+2.3)	24.45 (+16.0)	16.97 (+7.7)	23.54 (+10.9)
Attend-and-excite [10]	×	60.47	9.98	10.58	19.56
	✓	62.71 (+2.2)	20.76 (+10.8)	14.17 (+3.6)	20.83 (+1.3)
Layout-guidance [11]	×	56.22	16.47	12.38	14.39
	✓	63.01 (+6.8)	25.84 (+9.4)	15.56 (+3.2)	21.50 (+7.1)
MultiDiffusion [6]	×	55.18	14.27	10.58	17.15
	✓	57.37 (+2.19)	22.65 (+8.2)	10.78 (+0.2)	24.59 (+7.3)
GLIGEN [32]	×	66.58	30.74	26.75	18.78
	✓	67.54 (+0.7)	40.22 (+9.5)	27.74 (+1.0)	26.32 (+7.5)
Boxdiff [61]	-	67.02	33.93	28.54	22.50

Table 1. The CAR and SAR losses increase the F1 score in counting and accuracy(%) in all spatial, size, and color categories of the HRS benchmark. The small, green numbers represent value improvement.

Method	CAR & SAR	Counting	Spatial	Average
Stable Diffusion 1.4 [47]	×	68 .15	72.01	78.38
	✓	69.37 (+1.22)	73.33 (+1.32)	78.87 (+0.49)
Stable Diffusion 2.1 [47]	×	73.63	76.11	81.84
	✓	74.44 (+0.81)	76.29 (+0.18)	81.89 (+0.05)

Table 2. TIFA score(↑) in two baselines: Stable Diffusion 1.4 and 2.1.The small, green numbers represent value improvement.

Method	IoU ↑	Clip Score ↑	SOA-I ↑
SD-Pww [4]	23.76 ± 0.50	0.2800 ± 0.0005	73.92 ± 1.84
DenseDiffusion [29]	34.99 ± 1.13	0.2814 ± 0.0005	77.61 ± 1.75
Stable Diffusion + Our	**38.97** ± **0.56**	**0.3177** ± **0.0011**	**78.80** ± **1.27**

Table 3. Evaluation of image generation based on mask guidance, highlighting the performance of our approach (± indicates standard deviation). The results of other methods are directly taken from DenseDiffusion [29].

state-of-the-art large language model that can understand the number and spatial compositions of objects in our experiments. Specifically, given the input text for image generation, we create a new prompt to request GPT-4 to generate box coordinates and the label of objects in each box. We outline the details of this in-context learning in the supplementary.

4. Experiments

We evaluate our methods on several benchmarks, conduct ablation experiments on each component.

4.1. Experiment setup

Dataset. For text-to-image tasks, we utilize the benchmark **HRS** [3] and **Drawbench** [49] to evaluate the text-to-image generation performance on various categories, including counting, spatial, color, and size compositions. To

	Stable Diffusion [47]		GLIGEN [32]	
CAR&SAR	×	✓	×	✓
	20.82	21.03	20.63	20.37

Table 4. FID (↓) in Stable Diffusion and GLIGEN with and without CAR & SAR in COCO 2014 [35]
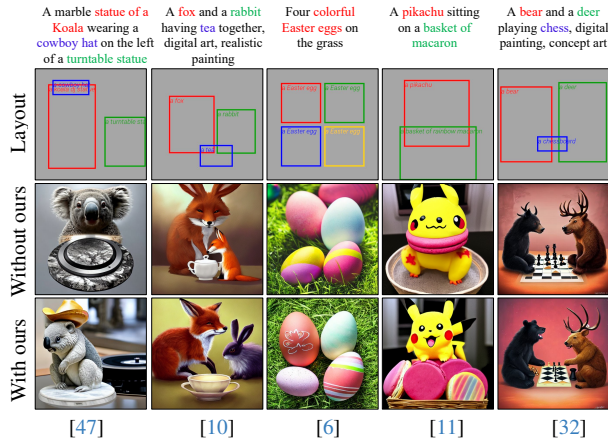

Figure 5. **Plug & play use of our attention-based guidance.** Our method applies to various base models. Here we show improved controllability across multiple text-to-image methods: Stable Diffusion [47], Attend-and-excite [10], MultiDiffusion [6], Layout-guidance [11], GLIGEN [32]

further assess the alignment of the generated image and input text, we use **TIFA** [24] benchmark. We include more details about these datasets in the supplementary material.

For quality evaluation, we utilize the **COCO2014** [35] validation dataset to assess images generated from textual descriptions and corresponding bounding boxes.

Our mask-and-text-to-image evaluation utilizes the dataset provided by DenseDiffusion [29].

Evaluation metrics. Regarding text-to-image evaluation, we follow the protocol in HRS [3] to compute the metrics on individual categories: counting, spatial, size, and color. For TIFA benchmark, we assess the alignment between the generated images and input texts using the TIFA score [24].

In box-and-text evaluations, we report FID [19] in COCO2014 [35], using available bounding boxes from this benchmark as input layouts for grounded text-to-image models.

For mask guidance evaluation, we measure the alignment with IoU used in DenseDiffusion [29], assess the textual similarity with CLIP Score [18], and ensure the object presence using the SOA-I score [20] using YOLOv7 [56] for object detection. More details about these metrics are presented in the supplementary.

Implementation details. In terms of bounding boxes guidance, we evaluate our method by plugging it into various open-source text-to-image models and methods, including Stable Diffusion (SD) V-1.4 [47], Attend-and-

CAR	SAR	Counting			Spatial	Size	Color
		Precision ↑	Recall ↑	F1 ↑	Acc. ↑	Acc. ↑	Acc. ↑
✗	✗	78.81	59.44	66.58	30.74	26.75	18.78
✗	✓	79.76	59.34	67.03	36.43	**30.34**	18.39
✓	✗	**82.11**	59.35	**67.59**	36.92	28.94	23.88
✓	✓	81.25	**59.39**	67.54	**40.22**	27.74	**26.32**

Table 5. **Abaltion study** of the CAR and SAR losses using the GLIGEN model on the HRS benchmark.

excite [10], Layout-guidance [11], MultiDiffusion [30], and GLIGEN [32]. For mask guidance, we integrate our losses to Stable diffusion [47] and compare with other free-training methods: DenseDiffusion [29] and Pww [4]. All the mentioned methods are configured as default settings.

4.2. Quantitative results

We show the results on the HRS benchmark [3] in Table 1. Using our losses consistently enhances F1 scores in counting by an average of 2%, with Layout-guidance [11] models showing a notable 7% improvement. Spatial accuracy gains an average of 10% with our losses. Our method boosts accuracy by up to 10.9% in the size and color categories. It can be seen that Attention-refocusing losses outperforms Boxdiff [61] in counting, spatial and color categories, especially 7% improvement in spatial composition. Specifically, GLIGEN [32] sees an increase in spatial and color accuracy by around 10% and 8%, respectively. Stable Diffusion [47] , which relies solely on text input, lags behind other grounded text-to-image models. However, with our attention-guided enhancements, it outperforms several methods like Layout-guidance and MultiDiffusion. A similar trend also can be observed in the DrawBench [49] results in our supplementary material. The **TIFA** evaluation in Table 2 further demonstrates the efficacy of our CAR and SAR losses, particularly enhancing counting and spatial accuracy across all baseline versions without detriment to other categories. This is reflected in the overall TIFA score improvements. Moreover, our attention-refocusing losses improve textual alignment in models like Stable Diffusion and GLIGEN without affecting image quality, ensuring that **FID scores** remain compatible with the originals. The evaluation of FID can be seen in Table 4.

For mask guidance evaluation, we report the quantitative evaluation in Table 3. Our method outperforms the current state-of-the-art approaches with an IoU of 38.97 ± 0.56 and a leading SOA-I score and Clip score, indicating improved layout fidelity and object detection in generated images.

4.3. Qualitative results

Note that we generate each pair of images with the same initial noise and the layouts in all figures are generated from GPT-4 api. Fig. 5 illustrates the qualitative comparison of various methods with and without our losses. In all the
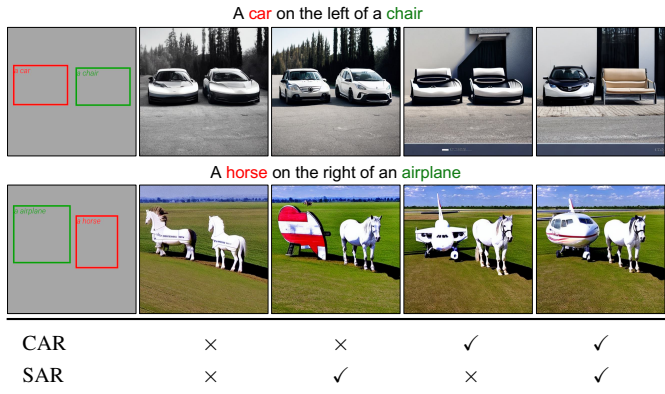

CAR	✗	✗	✓	✓
SAR	✗	✓	✗	✓

Figure 6. **Ablation study.** We show sample grounded text-to-image generation demonstrating the effects of the two proposed attention guidance.
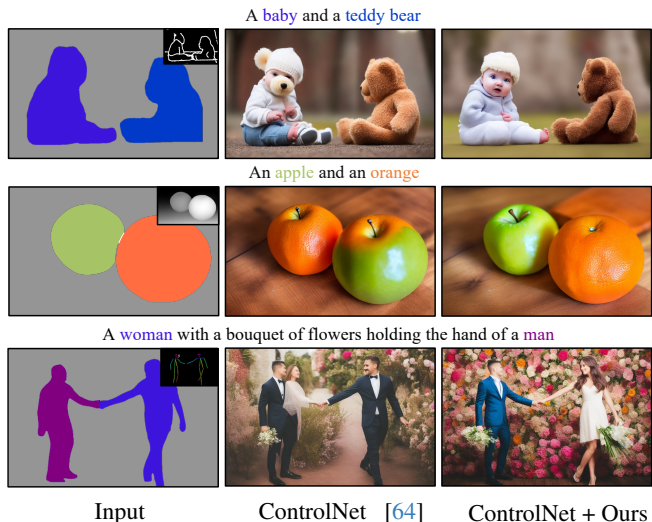

Input ControlNet [64] ControlNet + Ours

Figure 7. **ControlNet with attention-based guidance.** The input of ControlNet is a small image in the top right of the first column. With an extra segmentation map (the bigger images in the first column), our losses can refine the attribute blending of ControlNet.

cases, our losses help generate images with more precise spatial locations, colors, and numbers of objects. For example, attention-refocusing loss helps to mitigate the attribute mixing problem of Layout-guidance [11] (in the fourth column).

In Fig. 8, we show the results using prompts from the HRS benchmark. While MultiDiffusion [30] and Layout-guidance [11] often do not respect the input layouts, particularly in smaller boxes, GLIGEN incorrectly aligns objects with grounded input or mismatch colors. In contrast, integrating our attention-refocusing losses with GLIGEN enhances alignment and color accuracy. The results of our approach underscore our method's effectiveness in creating novel spatial configurations and attributes.

Additionally, as shown in Fig. 7, our CAR and SAR losses can adapt to segmentation mask guidance, refining ControlNet's output by reducing attribute mixing and avoid-

A dog on the right of a horse

A car on the left of a bus

An orange dog and a green cat

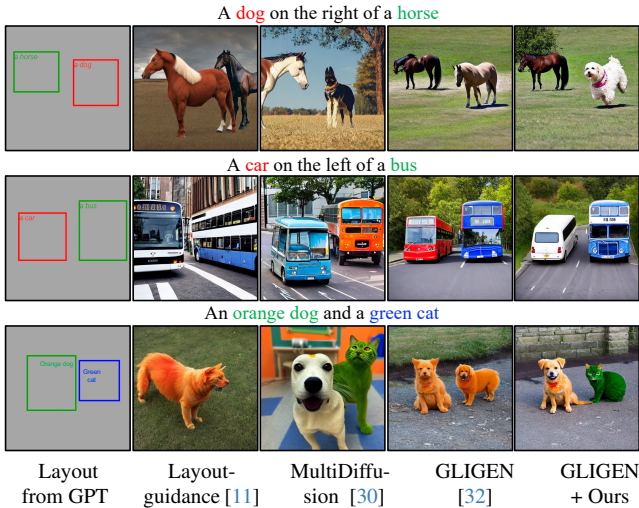| Layout from GPT | Layout-guidance [11] | MultiDiffu-sion [30] | GLIGEN [32] | GLIGEN + Ours |

Figure 8. **Visual comparisons on HRS benchmark.** Here, we apply our attention-based guidance to grouned text-to-image models. All methods take the same grounded texts as inputs. The results show the capability of our method in synthesizing novel spatial compositions and attributes.

ing generating additional and irrelevant objects.

4.4. Ablation studies

We ablate the two losses using GLIGEN [32] as the baseline method in Fig. 6. GLIGEN sometimes struggles with prompts including multiple objects, especially objects in the same category or size. CAR loss can mitigate this problem, but the generated objects still have attributes blended from others. For instance, with only CAR loss, a generated car might have a mixed chair feature and vice versa. Incorporating CAR and SAR losses further mitigates the attribute blending problem.

We further perform quantitative evaluation in Table 5 for all categories in the HRS benchmark. Adding CAR or SAR loss to the GLIGEN model improves the baseline in all four categories. Particularly in spatial relationships, using SAR or SAR can improve GLIGEN by approximately 6%. When using both losses, we can achieve an around 10% accuracy improvement.

4.5. Large language model evaluation

We examine the latest large language models, GPT-4 [40] GPT-3 [7], Llama 1 [53] and Llama 2 [54] (version 13b-chat) by evaluating their ability to comprehend the visual concept. We randomly chose 200 prompts from four categories in the HRS benchmark and report three metrics: Format, Validness, and Correctness. (The details of three metrics can be found in the supplementary.)

In Table 6, GPT-4 outperforms Llama 1, Llama 2, and GPT-3 in three metrics. Leveraging the leading large language model, our two-stage text-to-image model surpasses the single-stage Stable Diffusion [47] in understanding object relationships and textual alignment, as shown in Fig. 9.

Model	Format ↑	Valid ↑	Correct ↑
Llama 1 [53]	67.5	46.0	38.5
Llama 2 [54]	98.5	84.0	63.5
GPT-3 [7]	**98.5**	97.5	83.5
GPT-4 [40]	**98.5**	**98.5**	**88.5**

Table 6. Performance evaluation of LLMs using 200 random prompts in the HRS benchmark (%)

An Owl riding a bicycle with the moon in the background, studio lighting, quality photography by Wes Anderson, trending on art station

Anthro furry humanoid racoon, military uniform, holding a machinegun and sitting on the tank, character concept design, painting, ruined city background

| Layout from GPT | GLIGEN + Ours | Stable Diffusion [47] |

Figure 9. **Comparisons of Stable Diffusion and our two-stage pipeline.** Our two-stage pipeline excels over Stable Diffusion [47] in prompt understanding.

Four elephants and four people playing frisbee in a large open field.
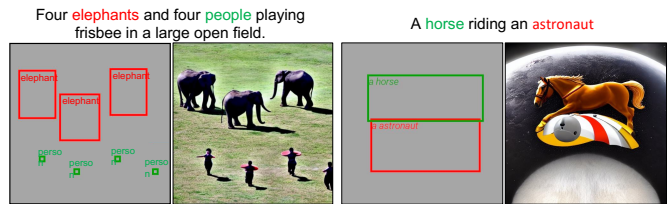
A horse riding an astronaut

Figure 10. **The failures cases of our framework.** GPT-4 sometimes misinterprets object quantity or size and instances of the text-to-image model not aligning with GPT-4's layout

5. Limitation

The Fig. 10 illustrates failure cases where our framework struggles. When dealing with prompts describing a large number of objects, GPT-4 occasionally produces an incorrect count or generates small boxes. Additionally, there are instances where GPT-4 accurately generates the layout, yet the grounded text-to-image model fails to adhere to these out-of-distribution layouts (the second example).

6. Conclusion

In this paper, we propose a novel attention-refocusing approach to improve the alignment of cross- and self-attention layers given layouts during the sampling process. Furthermore, we explore the usage of Large Language Models for generating visual layouts from text prompts. Our proposed losses can be easily incorporated into existing text-to-image diffusion models. The comprehensive experiments show favorable performance against state-of-the-art grounded text-to-image models.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *ICCV*, 2023. 3

[2] 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. *CVPR*, 2023. 2, 3

[3] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *ICCV*, 2023. 2, 6, 7

[4] 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, 6, 7

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

[6] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 2, 3, 6

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 8

[8] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, Li Yuanzhen, and Krishnan Dilip. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[10] 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, 3, 4, 6, 7

[11] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 2, 3, 6, 7, 8

[12] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 4

[13] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022. 2

[14] 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 diffu-

[15] 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. In *ICLR*, 2023. 4

[16] 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*, 2022. 2

[17] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *ICCV*, 2023. 2

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[20] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[24] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 2, 6

[25] Chuong Huynh, Yuqian Zhou, Zhe Lin, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Abhinav Shrivastava. Simpson: Simplifying photo cleanup with single-click distracting object segmentation network. In *CVPR*, 2023. 2

[26] Chuong Huynh, Seoung Wug Oh, , Abhinav Shrivastava, and Joon-Young Lee. Maggie: Masked guided gradual human instance matting. In *CVPR*, 2024. 2

[27] Álvaro Barbero Jiménez. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*, 2023. 3

[28] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 2

[29] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 3, 6, 7

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 3, 7, 8

sion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 2, 3

[31] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[32] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 1, 2, 3, 4, 6, 7, 8

[33] 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. 4

[34] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 4

[35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 6

[36] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*, 2022. 3

[37] Wan-Duo Kurt Ma, Avisek Lahiri, JP Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *AAAI*, 2024. 3

[38] 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. In *ICML*, 2022. 2, 3

[39] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2022. 2

[40] OpenAI. Gpt-4 technical report, 2023. 2, 3, 5, 8

[41] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV*, 2023. 2, 3

[42] Khoi Pham, Chuong Huynh, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *CVPR*, 2024. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Krueger Gretchen, and Sutskever Ilya. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[45] 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*, 2021. 2

[46] 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

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 6, 7, 8

[48] 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. *CVPR*, 2023. 2

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Ho Jonathan, J Fleet David, and Norouzi Mohammad. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2, 6, 7

[50] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *ICML*, 2023. 2

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Schramowski Patrick, Kundurthy Srivatsa, Crowson Katherine, Schmidt Ludwig, Kaczmarczyk Robert, and Jitsev Jenia. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 2

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 8

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 8

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[56] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, 2023. 6

[57] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *ICCV*, 2023. 3

[58] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023. 3

[59] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[60] Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qing-ming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. *arXiv preprint arXiv:2310.08872*, 2023. 3

[61] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wen-tian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 3, 6, 7

[62] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raf-fel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Compu-tational Linguistics*, 2022. 3

[63] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gun-jan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yin-fei Yang, Burcu Karagol Ayan, Hutchinson Ben, Han Wei, Parekh Zarana, Li Xin, Zhang Han, Baldridge Jason, and Wu Yonghui. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learn-ing Research*, 2022. 2

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 2, 3, 7

[65] Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *ACM MM*, 2023. 3