Grounded Text-to-Image Synthesis with Attention Refocusing

Quynh Phung Songwei Ge Jia-Bin Huang
University of Maryland College Park
https://attention-refocusing.github.io/

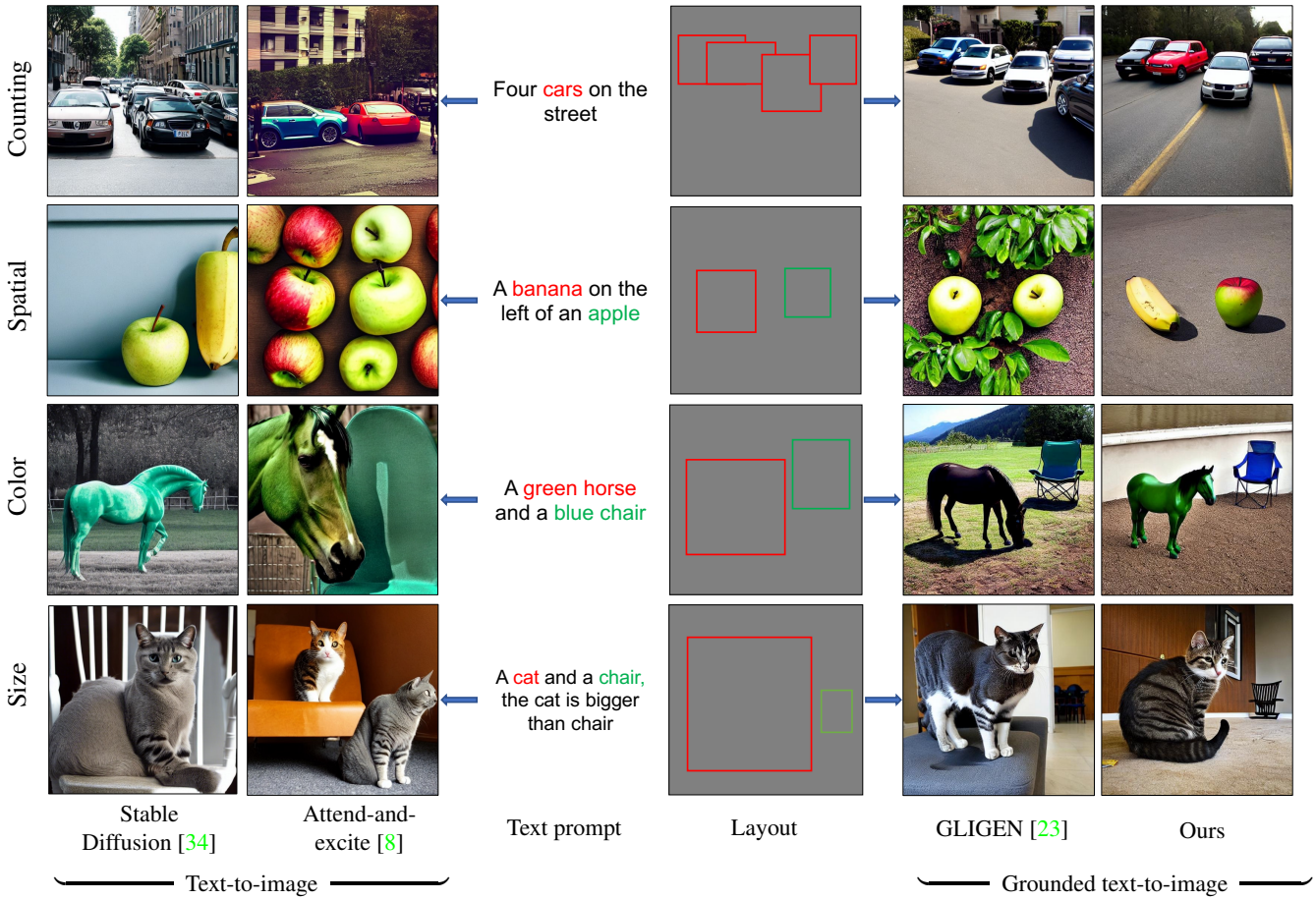


Figure 1. **Controllable text-to-image synthesis.** Existing text-to-image synthesis approaches [8, 34] struggle to generate images that respect the text prompts (with the correct object counts, spatial relationships, and color and size attributes). Using grounded texts (texts and layout guidance), GLIGEN [23] shows improvement but still does not offer full control. Our proposed attention-based guidance substantially improves the controllability and fidelity in text-to-image synthesis.

Abstract

Driven by scalable diffusion models trained on large-scale paired text-image datasets, text-to-image synthesis methods have shown compelling results. However, these models still fail to precisely follow the text prompt when multiple objects, attributes, and spatial compositions are involved in the prompt. In this paper, we identify the potential reasons in both the cross-attention and self-attention layers of the diffusion model. We propose two novel losses to refocus the attention maps according to a given layout during the sampling process. We perform comprehensive experiments on the DrawBench and HRS benchmarks using layouts synthesized by Large Language Models, showing that our proposed losses can be integrated easily and effectively into existing text-to-image methods and consistently improve their alignment between the generated images and the text prompts.

1. Introduction

Despite the unprecedented zero-shot capacity and photorealism achieved by the recent progress in text-to-image synthesis [3, 20, 32–34, 36, 46], the current state-of-the-art models still struggle with text prompts containing multiple objects and attributes with complex spatial relationships among them [2, 8, 9, 12, 44]. Some objects, attributes, and spatial compositions specified in the text prompts are often swapped or completely missing in the synthesized image. Our work aims to mitigate this problem by grounding the text-to-image synthesis using explicit layouts *without extra training and auxiliary models*.

The deep level of language understanding exhibited by these models can be attributed to using pretrained language models [30] as the text encoder [36]. The computed text embeddings are processed using the *cross-attention layers* in the denoising models [26, 27]. Upon careful analysis of the failure example generated by Stable Diffusion [34], we identify a potential cause of the failure above in the attention layers [41], 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 could thus be *mixed* through the attention layers. Note that such pixels could 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 [8, 9, 12]. However, existing work neglects that the same feature mixing issue also occurs in the self-attention layers. One immediate question when addressing this issue in self-attention layers is how to discriminate the pixels that are truly from the same object and those pixels that have similar features. To this end, we leverage *explicit layout representations* following the previous works [9, 23]. In this paper, we propose two novel losses based on the input layout during the sampling process to *refocus* the attention in both self- and cross-attention layers. Our attention-refocusing losses show that the attention can be effectively *refocused* to the desired region instead of a similar but irrelevant region.

We use up-to-date LLMs to generate explicit layout representations in our attention-refocusing losses. We demonstrate that these models have strong spatial reasoning capabilities. We design prompts with in-context learning to query LLMs about the spatial relationship of the objects given a challenging text prompt. The LLMs are asked to either predict the location of the objects or draw visual representations like vector graphics directly.

We show that when combining the bounding boxes generated by GPT4 [28] and our attention-refocusing losses, our method significantly and consistently improves over several strong baselines on the DrawBench [36] and HRS benchmarks [2]. 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;
- 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 perform 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.

look

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 [3, 14, 20, 33, 34, 36, 46]. Such rapid progress can be attributed to several critical techniques. First, the availability of large-scale text-image datasets [6, 38] 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 [20, 37], autoregressive models [7, 11, 33, 46], and diffusion models [3, 16, 26, 32, 36], together with various training and inference techniques [16–18, 39] and using pretrained large language models (LLMs) [28, 30, 31, 40] as text encoders. Our work studies 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 [15], personal images [21, 35], edge maps [47], and bounding boxes [1, 5, 23]. There are also works focusing on strengthing the controllability with the original input text, motivated by the observation that existing models often fail to fulfill the description from the input text accurately [2, 8, 12, 29]. For example, when multiple objects and attributes occur in the text prompt, some are often missing or mixed in the synthesized images [8, 12]. Attend-and-Excite [8] proposes optimizing cross-attention maps during sampling to ensure all the tokens are attended to. Several studies finetune the existing

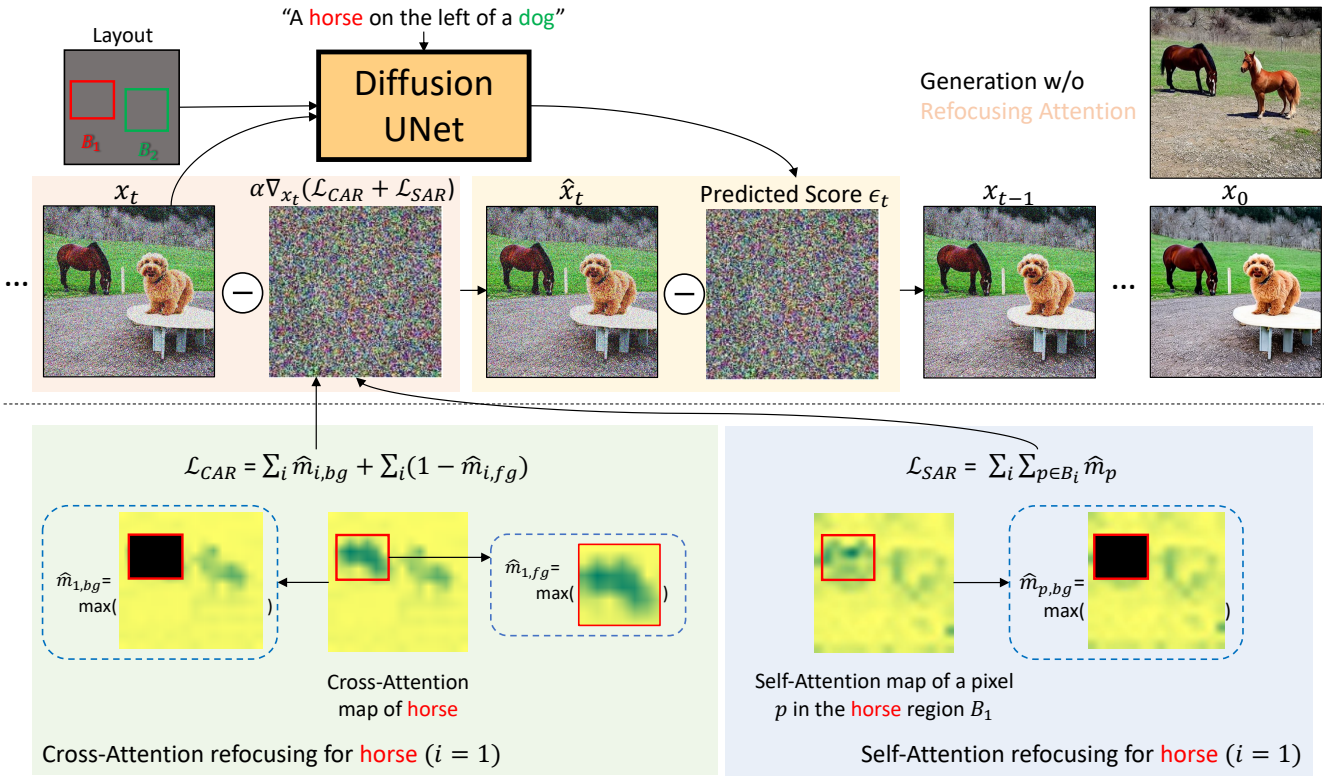


Figure 2. **The proposed Attention-Refocusing framework.** At each denoising step, we update the noise sample by optimizing our $\mathcal{L}_{CAR}$ and $\mathcal{L}_{SAR}$ losses (red block) before denoising with the predicted score (yellow block). For each cross-attention map, $\mathcal{L}_{CAR}$ is designed to encourage a region to attend more to the corresponding token while discouraging the remaining region from attending to that token (green block). For each self-attention map, $\mathcal{L}_{SAR}$ prevents the pixels in a region from attending to irrelevant regions (blue block).

models with human feedback [22, 43] or use improved language models [24, 29, 40, 45, 48] 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 [28, 30, 31, 40] and ground the image synthesis on the layout.

Layout-conditioned text-to-image synthesis. Several approaches have been developed to extend the Stable Diffusion [34] to condition its generation on the layouts through finetuning on layout-image pairs [1, 23, 26, 47] or manipulating the sampling process [3–5, 9]. For example, GLIGEN [23] finetunes a gated self-attention layer to incorporate the box information from the input to the Stable Diffusion model. Mixture-of-Diffusion [19] and MultiDiffusion [21] perform a denoising process on each region and fuse the predicted scores. Paint-by-word [3], layout-predictor [42], direct-diffusion [25] and Layout-guidance [9] directly optimize the cross-attention layers during the sampling process. Universal guidance [4] leverages a trained object detector and constructs a loss to force the generated images to match location guidance. Our proposed method to grounded the text-to-image generation on the layout uses both *cross-attention* and *self-attention* layers without the need for 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. LayoutGPT [13] leverages GPT to create layouts from text conditions, then generate images from created layouts. Cho et al. [10] 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 may still fail to generate the correct objects and their attributes. Our work focuses on improving 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

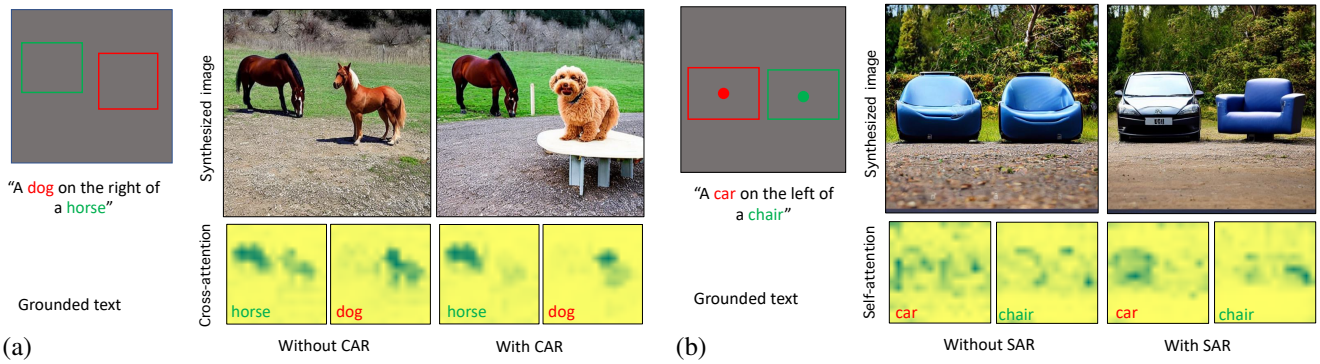


Figure 3. (a) **Cross-Attention-Refocusing (CAR) visualization.** Without CAR, the token "horse" attend to both the regions of "horse" and "dog". Using CAR calibrates the cross-attention map to attend to the correct regions. (b) **Self-Attention-Refocusing (SAR) visualization.** The dots in each box represent the pixel query of the self-attention map. Applying SAR loss helps refocus the self-attention to attend less on the incorrect regions.

phases: 1) text-to-layout prediction and 2) grounded text-to-image. In both phases, we use pretrained models **without any extra training**. We exploit the spatial understanding ability in the latest large language models (LLMs) to produce visual representations like bounding boxes to obtain the layout given a text prompt. [1] Given the generated layout, we propose novel attention-refocusing losses to guide the pre-trained text-to-image models. We first present the background in Sec. 3.1. We then derive our proposed loss aiming to improve the alignment with respect to the layouts in Sec. 3.2 In Sec. 3.3, we describe how we prompt LLMs to generate layouts.

3.1. Preliminaries:

Text-to-image diffusion models. We use the Diffusion model [16, 34, 39] as the image generation model in this paper. The key to text-to-image model diffusion is the iterative denoising process. A UNet model is trained to progressively denoise the random Gaussian noise by computing the score $\epsilon_t = U(x_t; c)$, where t is the time step and c is the embedding for conditional information. Next, we briefly describe the two types of attention layers used in our method.

Cross-attention layer. The large-scale text-to-image diffusion model condition its generation on the text prompt via cross-attention layers. Specifically, a pretrained CLIP encoder [30] is often used to process the text prompt $\mathbf{w} = (w_1, w_2, \cdots w_n)$ to obtain the text embedding features $c = f_{\text{CLIP}}(\mathbf{w}) \in \mathbb{R}^{n \times e}$, where $e = 77$ 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 c with a linear mapping. Given a set of queries $\mathbf{Q} \in \mathbb{R}^{n_q \times d}$ computed from the features map of resolution $h' \times w'$, where $n_q = h' \times w'$, the attention map is derived as

$$A^c = \text{softmax}\left(\frac{\mathbf{QK}}{\sqrt{d}}\right) \in [0,1]^{n_q \times n}. \quad (1)$$

At denoising step t, by reshaping and indexing A^c, we have $A^c_i \in [0,1]^{h' \times w'}$ denoting the attention map between a word token w_i and each spatial location in the feature map.

Self-attention layer To facilitate the use of global information, self-attention layers are used in Stable Diffusion. 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 the same formula in Eq. (1), the self-attention map is computed as $A^s \in [0,1]^{n_q \times n_q}$. Similarly, we use $A^s_p \in [0,1]^{h' \times w'}$ to denote the self-attention map of all pixels attending to pixel p.

3.2. Grounded text-to-image generation

In this section, as shown in Figure 2, we introduce two losses regarding the attention layers to facilitate the controllability of the layout-conditioned image synthesis. First, we describe the **Cross-Attention Refocusing (CAR)** loss, improving the spatial alignment of the text prompt with the layout. Then, we discuss the **Self-Attention Refocusing (SAR)** loss that encourages the regions to attend less to the irrelevant areas. Both losses can serve as a plugin and readily integrate with existing models, such as GLIGEN [23], MultiDiffusion [5], and Stable Diffusion [34] to further improve the model controllability.

We consider layouts defined by k bounding boxes $B \in (\mathbb{Z}^+)^{k \times 4}$ where $B_j = (x^1_j, y^1_j, x^2_j, y^2_j)$ denotes (top, bottom, left, right) coordinates of j-th box. Each box is also associated with a textual phrase describing the content inside the box.

[1] Other spatial representations, segmentation maps, edges, vector graphics, and ASCII art, can also be extracted. We focus on using bounding boxes as the spatial layout in this work.

3.2.1 Cross-Attention Refocusing (CAR)

As shown in Fig. 3a, a token "horse" may incorrectly attend to the region of a dog in the Stable Diffusion cross-attention layers given the prompt "A dog on the right of the horse". This can be explained by Eq. (1) that the query $\mathbf{Q}$ has similar values in both regions and yield similar attention scores given the same $\mathbf{K}$. To this end, we propose a loss to *refocus* the cross-attention of these tokens according to the layout.

We denote the set of coordinates in the box B_j as $Fg(B_j) = \{(x, y) \,|\, x_j^1 \leq x \leq x_j^2, y_j^1 \leq y \leq y_j^2\}$. We first apply Gaussian Smoothing to the attention map following [8]. Given the attention map A_i^c of the token $w_i \in \mathbf{w}$, the peak value of region $Fg(B_j)$ is defined as:

$$M(Fg(B_j), A_i^c) = \max\{A_i^c(x, y) \,|\, (x, y) \in Fg(B_j)\}, \tag{2}$$

where $A_i^c(x, y)$ denotes the attention score at y-th row and x-th column in A_i^c. To encourage the model to attend more inside the boxes that correspond to the token i, we propose to minimize the following objective:

$$\mathcal{L}_{fg} = \sum_{i=1}^{k} \hat{m}_{i,fg} = \sum_{i=1}^{k} M(Fg(B_i), A_i^c), \tag{3}$$

Minimizing $\mathcal{L}_{fg}$ boosts the scores of regions associated with w_i. We further propose a loss to discourage the tokens associated with boxes from attending the remaining region. Let $P = \{(x, y) \,|\, 0 \leq x < w, 0 \leq y < h\}$ denote the set of all coordinates in A^c. Then the background coordinate set $Bg(B_j) = P \setminus Fg(B_j)$. The background loss is defined as the sum of all maximum scores of background regions in each attention map:

$$\mathcal{L}_{bg} = \sum_{i=1}^{k} \hat{m}_{i,bg} = \sum_{i=1}^{m} M(Bg(B_i), A_i^c) \tag{4}$$

The overall CAR loss is then defined as $\mathcal{L}_{CAR} = \mathcal{L}_{fg} + \mathcal{L}_{bg}$. As shown in Fig. 3a, our loss effectively reduces the incorrect attention from grounded tokens.

3.2.2 Self-Attention Refocusing (SAR)

Similar to the observation in the cross-attention layers, as shown in Fig. 3b, the pixels of one region (e.g., "car") may attend *outside* of the region with similar colors (e.g., "chair") in self-attention layers. To this end, we develop a loss to help self-attention *refocus* to the correct regions. We denote the set of coordinates that are not overlapped with any box that contains the coordinate p as:

$$B^p = P \setminus \{(x, y) \,|\, x_j^1 \leq x \leq x_j^2, y_j^1 \leq y \leq y_j^2, \\ \forall B_j \in B \; s.t. \; p \in fg(B_j)\} \tag{5}$$

We then propose to minimize the background loss for B^p so that the pixels within each box attend less outside of the box:

$$\mathcal{L}_{SAR} = \sum_{i=1}^{k} \sum_{p \in Fg(B_i)} \hat{m}_p = \sum_{j=1}^{k} \sum_{p \in Fg(B_i)} M(B^p, A_p^s) \tag{6}$$

As shown in Figure 3b, using our loss helps each box to focus less on the irrelevant regions.

3.2.3 Sampling with the attention refocusing losses

After having the CAR and SAR losses, we modify the noised sample x_t at each step to minimize the loss similar to [8]. The update process is shown in Fig. 2 with the formula:

$$x_t' \leftarrow x_t - \alpha \nabla_{x_t} \left(\mathcal{L}_{CAR} + \mathcal{L}_{SAR}\right), \tag{7}$$

with α as the step size that controls the influence of the optimization in the denoising process. However, we find that a single update step is often not strong enough to refine the cross-attention and self-attention maps. We thus make τ times of update at each denoising step. We also observe that applying optimization in later steps can lead to quality degradation. Therefore, we only update in the first t' steps such that $t\prime < t$, with $t\prime$ as the number of steps using CAR and SAR losses. After finishing τ times of update, we feed the output to the diffusion UNet to resume the denoising process and compute x_{t-1}. We discuss the implementation details in the experiment section.

3.3. Text-to-layout prediction

Generating an image from text requires solid 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, it becomes non-trivial to upgrade the text encoder without additional (costly) training. Such schema could hinder the existing text-to-image models from benefiting from recent breakthroughs in large language models (LLMs). In light of this challenge, we explore directly using LLMs to generate intermediate visual presentations such as box layouts, segmentation maps, edges, etc.

We leverage GPT-4 [28], the state-of-the-art large language model which exhibits the capability to understand the number and spatial compositions of objects in our experiments. Specifically, given the input text for image generation, we create a prompt to request GPT-4 to generate a corresponding coordinate box layout and predict the label of objects in each box. To further enhance the model's capacity for the task, we supplement our prompt with multiple examples, which also help the model output the desired form of bounding boxes and their corresponding labels. The

Table 1. Our proposed losses improve the baselines in the HRS Counting benchmark.

Method	CAR & SAR	Precision ↑	Recall ↑	F1 ↑
Stable Diffusion	×	71.86	52.19	58.31
	✓	81.54 (+10.1)	51.18 (-0.9)	60.61 (+2.3)
Attend-and-excite	×	73.10	54.79	60.47
	✓	75.94 (+2.8)	56.31 (+1.5)	62.71 (+2.2)
Layout-guidance	×	80.60	45.83	56.22
	✓	78.15 (-2.5)	55.65 (+9.8)	63.01 (+6.8)
MultiDiffusion	×	79.00	45.35	55.45
	✓	83.21 (+4.2)	45.65 (+0.3)	57.29 (+1.8)
GLIGEN	×	80.04	60.67	68.28
	✓	87.20 (+7.2)	62.03 (+1.4)	71.83 (+3.6)

Table 2. The CAR and SAR losses increase the accuracy(%) in all spatial, size, and color categories of the HRS benchmark.

Method	CAR & SAR	Spatial	Size	Color
Stable Diffusion [34]	×	8.48	9.18	12.61
	✓	24.45 (+16.0)	16.97 (+7.7)	23.54 (+10.9)
Attend-and-excite [8]	×	9.98	10.58	19.56
	✓	20.76 (+10.8)	14.17 (+3.6)	20.83 (+1.3)
Layout-guidance [9]	×	16.47	12.38	14.39
	✓	25.84 (+9.4)	15.56 (+3.2)	21.50 (+7.1)
MultiDiffusion [5]	×	14.47	10.58	17.15
	✓	22.65 (+8.2)	10.78 (+0.2)	24.59 (+7.3)
GLIGEN [23]	×	45.71	36.13	17.84
	✓	54.19 (+8.5)	39.72 (+3.6)	29.46 (+11.6)

process of generating the box layout given an textual prompt is shown in the 8. The textual input is sent to GPT-4, which then returns box coordinates along with a corresponding label for each box. These box coordinates is visualized into layout.

4. Experiments

In this section, we describe the experiments to validate the effectiveness of the proposed methods. We present the details of benchmark datasets, evaluation metrics, and implementation in Sec. 4.1. We show quantitative results on the two benchmarks with various tasks in Sec. 4.2. The Sec. 4.3 illustrates the generated images with and without our proposed pipeline between different text-to-image generation frameworks. We validate the effectiveness of each of our proposed losses via the ablation in Sec. 4.4.

Table 3. Quantitavie evaluation on the DrawBench benchmark.

Method	CAR & SAR	Counting			Spatial
		Precision ↑	Recall ↑	F1 ↑	Accuracy ↑
Stable Diffusion [34]	×	73.32	70.00	71.55	12.50
	✓	78.53 (+5.2)	73.63 (+0.6)	75.81 (+4.6)	43.50 (+31.0)
Attend-and-excite [8]	×	77.64	74.85	76.20	20.50
	✓	74.06 (-3.6)	77.58 (+2.7)	75.66 (-0.5)	38.00 (+18.0)
Layout-guidance [9]	×	79.15	70.61	74.48	36.50
	✓	78.45 (-0.7)	75.45 (+4.8)	76.82 (+2.3)	52.50 (+16.0)
MultiDiffusion [5]	×	75.37	65.61	69.90	38.00
	✓	84.30 (+8.9)	68.03 (+2.4)	75.20 (+5.3)	54.50 (+16.5)
GLIGEN [23]	×	81.66	80.89	81.18	48.00
	✓	90.28 (+8.6)	86.21 (+5.3)	88.16 (+7.0)	64.00 (+16.0)

4.1. Experiment setup

Dataset We use two standard benchmarks HRS [2] and DrawBench [36], to evaluate the performance on various tasks. **The HRS dataset** contains various prompts divided into three main topics: accuracy, robustness, and generalization. Our method focuses on accuracy improvement, which has four main categories: *spatial relationship, color, size, and counting*. Each prompt in the dataset is tagged with the object's name and corresponding labels intended for evaluation. For example, in spatial relationships, the labels include objects and their relative positions, such as "on the left" or "on the right". The prompts for each category counting/spatial/size/color are 3000/1002/501/501, respectively. Depending on the number of objects and their relationship, the difficulty level of each prompt is labeled as easy, medium, and hard with roughly the same amount. **The DrawBench dataset** consists of 39 prompts about *Counting* and *Positional (or spatial relationship)*. Since there are no labels for this benchmark, we manually create the label for each prompt based on the number of objects mentioned and their relationships.

Evaluation metrics We follow HRS [2] to compute the evaluation metrics, where each category has specific metrics to assess the accuracy. In counting, the number of objects detected in generated images is compared to the ground truths in text prompts to measure the precision, recall, and F1 score. False positive samples happen when the number of generated objects is smaller than the ground truths. In contrast, the false negative objects are counted for the missing objects in the synthesized images. For other categories, we use accuracy as the evaluation metric. Depending on the category, the image is counted as a correct prediction when all detected objects are correct, either for spatial relationships, color, or size.

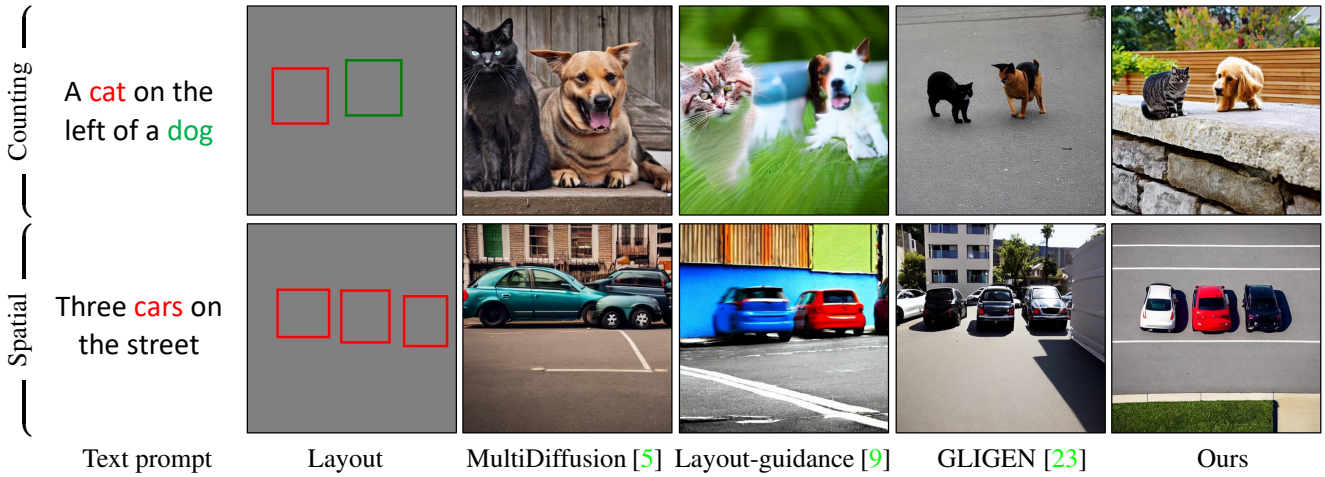


Figure 4. **Visual comparisons on DrawBench.** Here we apply our attention-based guidance on GLIGEN. All methods take the same grounded texts as inputs. Ours outperforms others in counting and managing spatial relationships.

Implementation details We evaluate the framework by plugging it into various open-source text-to-image models and methods, including Stable Diffusion (SD) V-1.4 [34], Attend-and-excite [8], Layout-guidance [9], MultiDiffusion [21], and GLIGEN [23]. We apply both Cross-Attention Refocusing and Self-Attention Refocusing losses on the attention maps of resolution 16×16. All images are generated with 50 steps of denoising, but the number of steps applying our losses t' is configured as 30. The initial step sizer α is set to 3 and decreases by 1 after each 10 denoising steps to prevent the deterioration of image quality in later steps of diffusion models. We initiate the number of refinement iterations τ for each step as 5 and decrease to 3 and then 1 at 10-th and 20-th step. We use the same set of hyperparameters for both benchmarks and all the backbone methods in our experiment.

For the text-to-layout prediction, we use GPT-4 model via API to generate bounding boxes given input prompts. To generate the correct prompt, we provide guidance and example, including the maximum size of bounding boxes as well as the output format, for the best context understanding. The table 5 is an extra prompt including the instruction of the task in the first row, the unwanted cases in the next four rows, and the most format-compatible example in the two final rows.

4.2. Quantitative results

We compare existing baselines, including text-to-image and grounded text-to-image models, before and after applying our proposed losses. In the HRS benchmark, the results are shown in Table 1 for counting and Table 2 for spatial, color, and size compositions.

In the counting category, all considered methods with our losses improve the base models by a large margin of an average of 3% in the F1 score. Remarkably, in the context of Layout-guidance, adding our losses improves the baseline by approximately 7% in the F1 score. Table 2 demonstrates the effectiveness of our losses in managing spatial relationships. Using CAR and SAR improves the baseline methods' performance by around 10% in accuracy. As for the size and color categories, the usage of our losses provides a boost of up to 12% in accuracy.

In DrawBench, our proposed losses show a comparable performance boost to HRS. By integrating our losses, we compare favorably or comparatively against baselines in the counting procedure. Moreover, our losses substantially improve the accuracy of the spatial category.

For both benchmarks, the standard Stable Diffusion, which uses text only as input, performs worse than other grounded text-to-image models. However, with the proposed attention guidance, it achieves better performance than other non-retrained methods like Attend-and-excite, Layout-guidance, or MultiDiffusion.

4.3. Qualitative results

Fig. 5 illustrates the qualitative comparison of various methods with and without our losses. In all the cases, our losses help generate images with more precise spatial locations, colors, and numbers of objects. For example, although the chair and banana in the fourth example are located correctly, their size is not correctly reflected in the generation. GLIGEN works well in generating objects considering layout locations and sizes; however, the number of objects remains a problem. Including the proposed CAR and SAR losses mitigates this issue.

In Fig. 4 and Fig. 7, we show the results using prompts from DrawBench and HRS, comparing four ground text-to-image models and our losses on the GLIGEN base model. MultiDiffusion and Layout-guidance occasionally do not respect the layout. For instance, the generated objects are

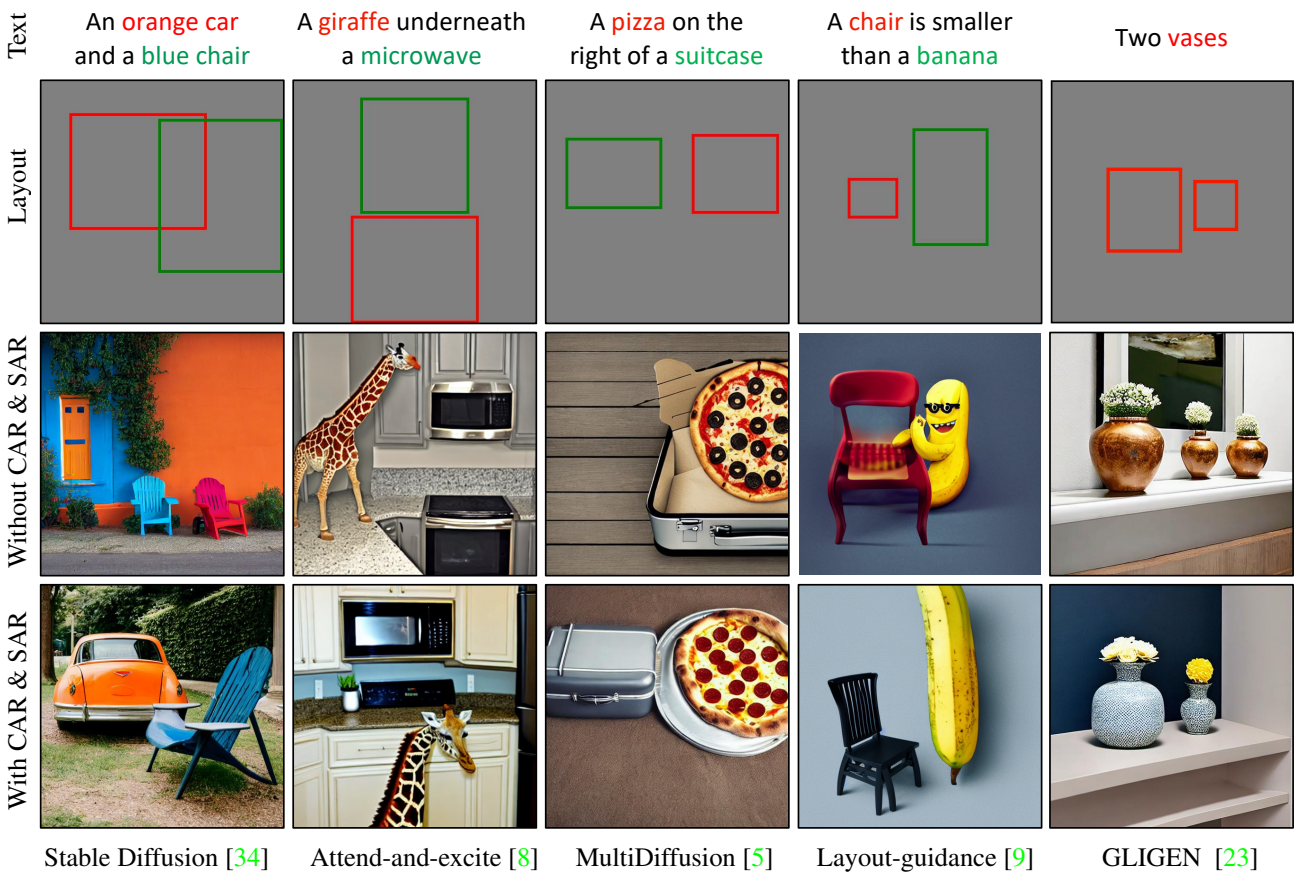


Figure 5. **Plug & play use of our attention-based guidance.** Our method is applicable to various base models. Here we show improved controllability across multiple text-to-image methods.

not tightly fit inside the corresponding boxes, especially the small boxes. Moreover, the image quality in those methods sometimes exhibits artifacts. The GLIGEN method produces objects that do not align with a grounded text prompt. For example, for "a car on the left of a chair", the generated images tend to generate two cars or two chairs. Besides, this method fails to correctly bind colors to corresponding objects (eg. the model generates an orange car instead of a blue one given the prompt "an orange horse and a blue car"). The outcomes of our approach in the final column underscore our method's effectiveness in creating novel spatial configurations and attributes.

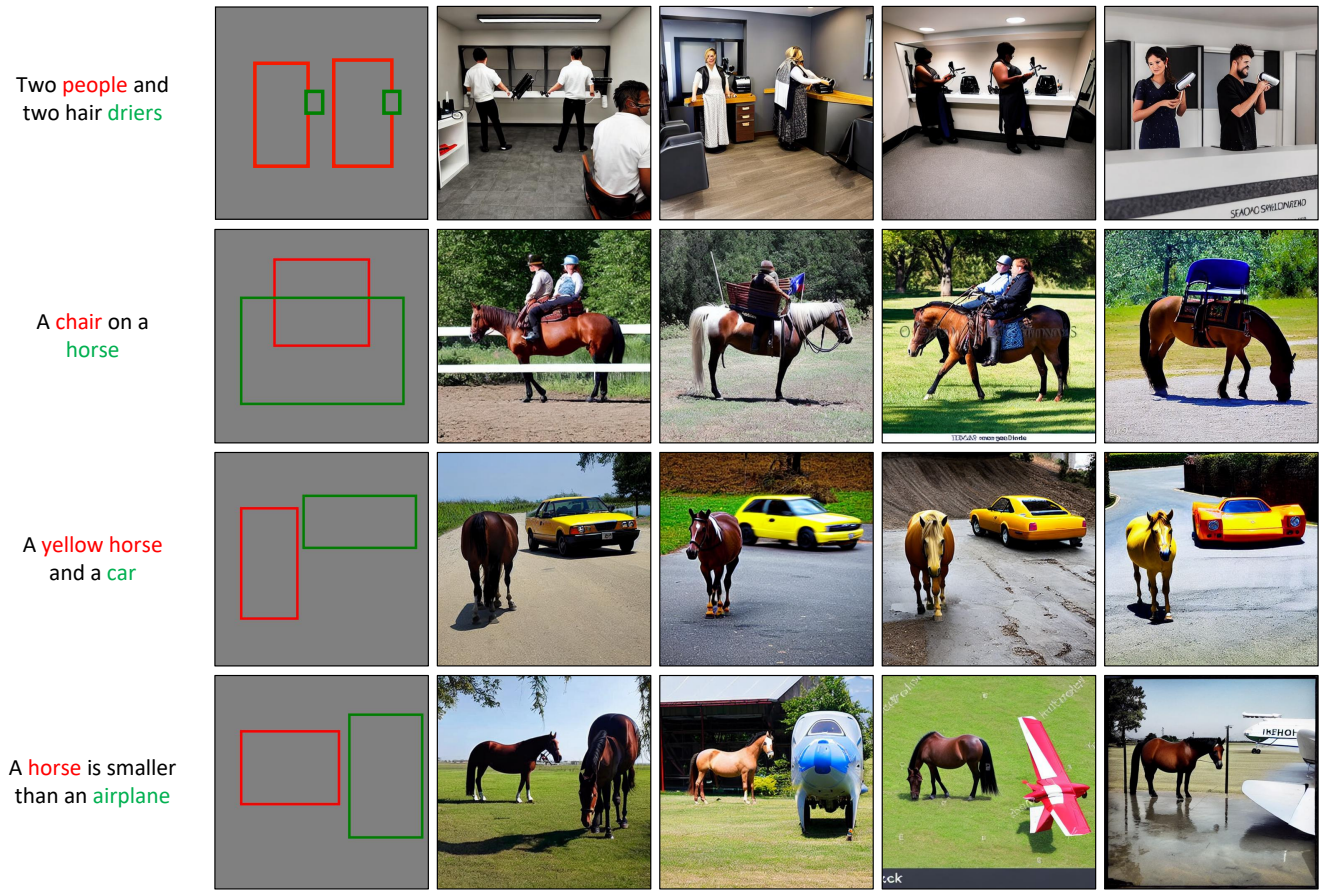
4.4. Ablation studies

We ablate the two losses using GLIGEN [23] as the baseline method in Table 4 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, while SAR can improve GLIGEN by approximately 5%, CAR loss outperforms it with around a 9% improvement. When using both losses, we can achieve a 14% accuracy improvement.

Table 4. **Abaltion study** of the CAR and SAR losses using the GLIGEN model on the HRS benchmark.

CAR	SAR	Counting			Spatial	Size	Color
		Precision ↑	Recall ↑	F1 ↑	Acc. ↑	Acc. ↑	Acc. ↑
×	×	81.04	60.67	68.28	40.32	31.94	18.11
×	✓	84.25	61.89	70.43	45.81	34.73	19.46
✓	×	84.76	61.79	70.62	49.60	**39.92**	22.92
✓	✓	**87.20**	**62.03**	**71.83**	**54.19**	39.72	**29.46**

Besides the quantitative results, we show the visual comparison in Fig. 6. These guidances can mitigate the problem of generating extra objects in the background as well as the missing object issue. For example, in the first-row images, GLIGEN generates three people without any drier. The number of people is fixed after adding the SAR loss while the driers appeared with the help of the CAR loss. The combination of two losses corrects the generated images in object locations and numbers. Other samples also



Cross Attention Refocusing	×	×	✓	✓
Self Attention Refocusing	×	✓	×	✓

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

show the favorable results of CAR in object color binding (e.g., the yellow horse in the third row) and SAR in counting and spatial location (e.g., the airplane in the last row).

5. 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.

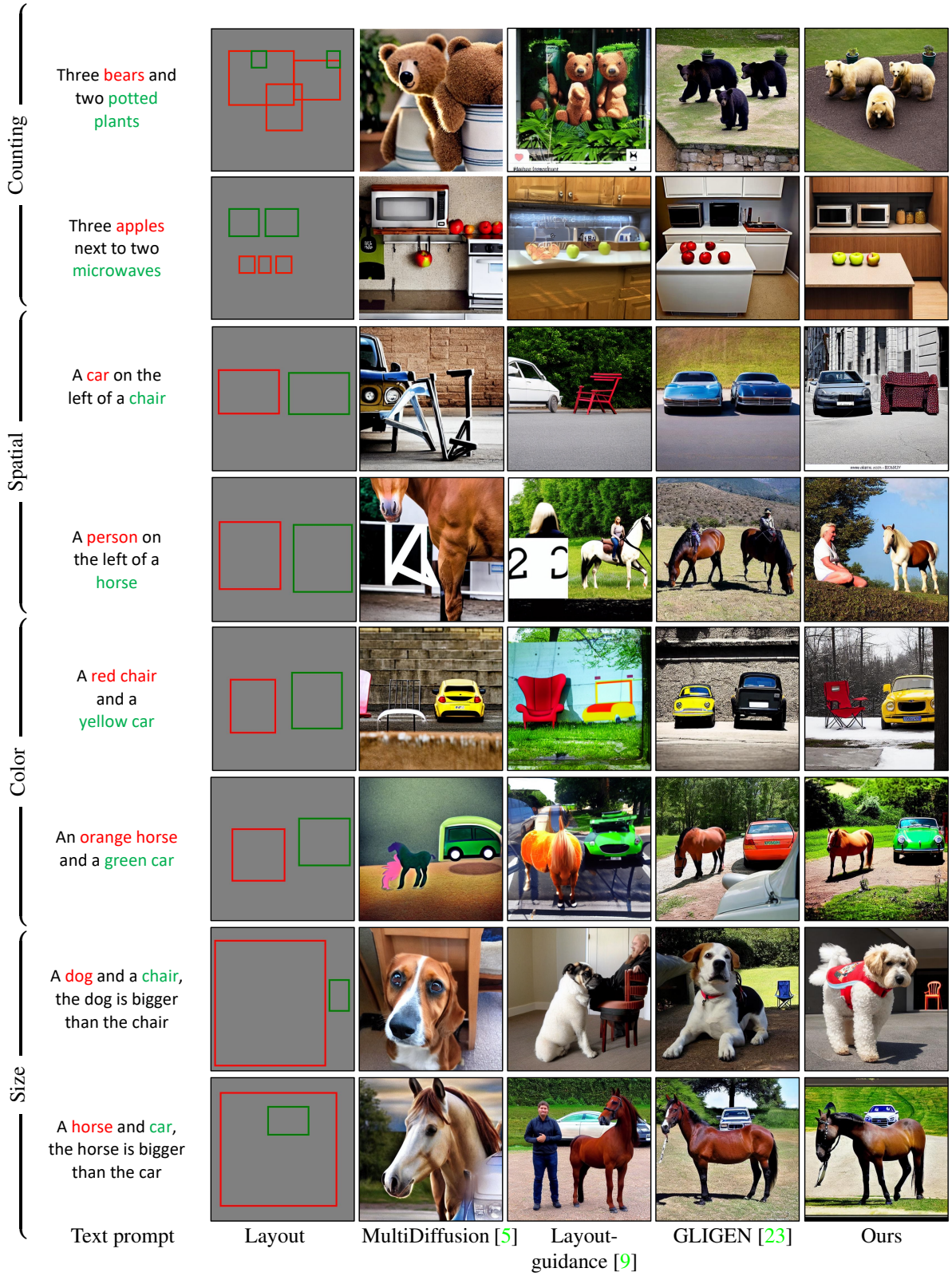


Figure 7. **Visual comparisons on HRS benchmark.** Here we apply our attention-based guidance on GLIGEN. All methods take the same grounded texts as inputs. The results show the capability of our method in synthesizing novel spatial compositions and attributes.

Table 5. The example prompt for ChatGPT-4 API to generate box layout given a text prompt

Role	Content
System	”You are ChatGPT-4, a large language model trained by OpenAI. Your goal is to assist users by providing helpful and relevant information. In this context, you are expected to generate specific coordinate box locations for objects in a description, considering their relative sizes and positions and the number of objects. The size of the image is 512*512.”
User	”Provide box coordinates for an image with a cat in the middle of a car and a chair. Make the size of the boxes as big as possible.”
Assistant	”cat: (230, 196, 297, 301); car: (80, 270, 202, 352); chair: (341, 231, 447, 308)”
User	”The sizes of objects do not reflect the object’s relative sizes. The car’s box size is similar chair’s box size, and the cat’s box size is quite similar to the car and chair. the car should be larger than the chair, and the chair should be larger than the cat.”
Assistant	”Apologies for the mistake. Here are the corrected coordinates: cat: (245, 176, 345, 336); car: (10, 128, 230, 384); chair: (353, 224, 498, 350)”
User	”Provide box coordinates for an image with three cats on the field.”
Assistant	”cat: (51, 82, 399, 279);cat: (288, 128, 472, 299); cat: (27, 355, 418, 494)”

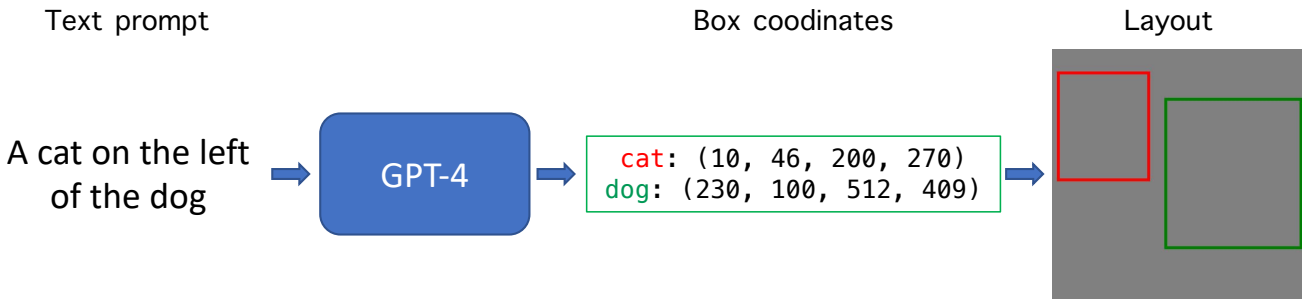


Figure 8. Layout generation process using gpt4 api.

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. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *arXiv preprint arXiv:2304.05390*, 2023. 2, 6

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Karras Tero, and Liu Ming-Yu. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 3

[4] 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. 3

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning (ICML)*, 2023. 2, 3, 4, 6, 7, 8, 10

[6] 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

[7] 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

[8] 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. In *ACM SIGGRAPH*, 2023. 1, 2, 5, 6, 7, 8

[9] 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, 10

[10] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 3

[11] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2

[12] 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. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[13] 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 *International Conference on Learning Representations (ICLR)*, 2023. 3

[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 *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022. 2

[15] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. *arXiv preprint arXiv:2304.06720*, 2023. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4

[17] 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

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

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

[20] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 2, 3, 7

[22] 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

[23] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4, 6, 7, 8, 10

[24] 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

[25] 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

[26] 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 *International Conference on Machine Learning (ICML)*, 2022. 2, 3

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2022. 2

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

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

[30] 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 *International Conference on Machine Learning (ICML)*, 2021. 2, 3, 4

[31] 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. 2, 3

[32] 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

[33] 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 (ICML)*, 2021. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 4, 6, 7, 8

[35] 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. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[36] 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 *Neural Information Processing Systems (NeurIPS)*, 2022. 2, 6

[37] 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. *arXiv preprint arXiv:2301.09515*, 2023. 2

[38] 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. *Neural Information Processing Systems (NeurIPS)*, 2022. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 4

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2, 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[42] 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. *arXiv preprint arXiv:2304.03869*, 2023. 3

[43] 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

[44] 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

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

[46] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei 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 Learning Research*, 2022. 2

[47] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3

[48] Shanshan Zhong, Zhongzhan Huang, Wushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. *arXiv preprint arXiv:2305.05189*, 2023. 3