ReGround: Improving Textual and Spatial Grounding at No Cost

Phillip Y. Lee Minhyuk Sung

KAIST
{phillip0701,mhsung}@kaist.ac.kr

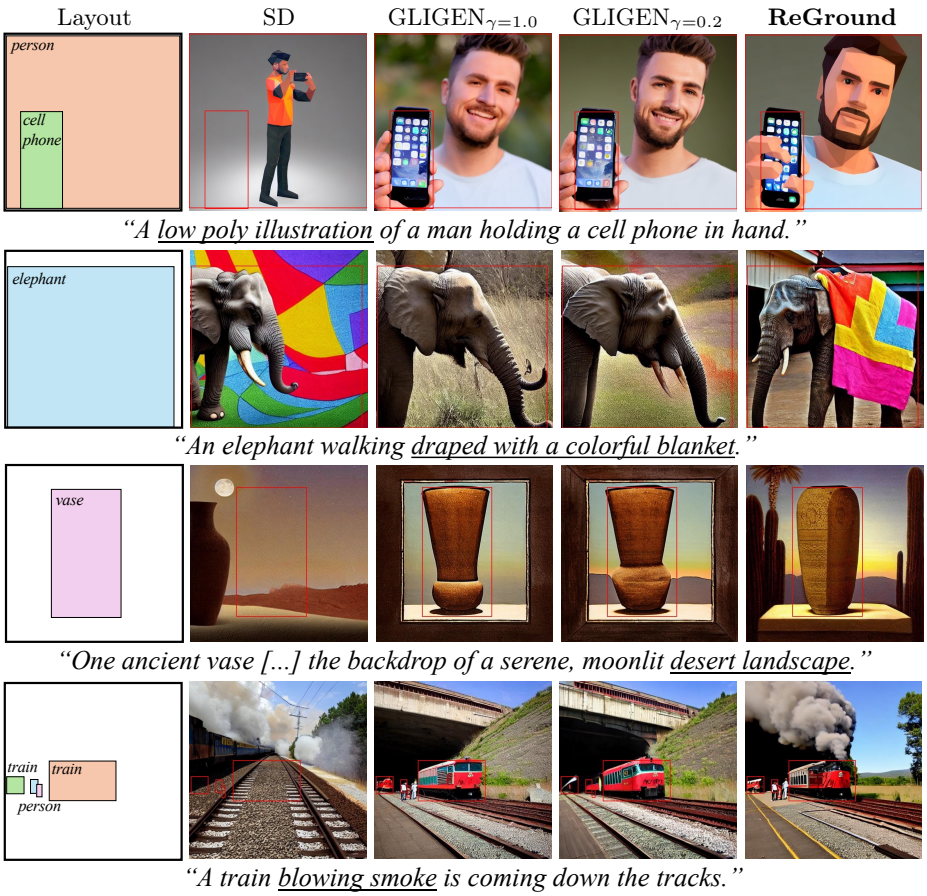


Fig. 1: Comparison across Stable Diffusion (SD) [41], GLIGEN [28], and our ReGround. SD (2nd column) can generate an image aligned with the input prompt (shown below each row), while it does not allow taking spatial constraints such as bounding boxes and labels. GLIGEN (3rd column) enables spatial grounding using gated self-attention, although it often disregards some descriptions in the input prompt due to a bias towards bounding box conditions. Such trends also occur when only activating gated self-attention for 0.2 fraction of the initial denoising steps (4th column). Our ReGround (last column) resolves the issue of description omission while accurately reflecting the bounding box information.

Abstract. When an image generation process is guided by both a *text* prompt and *spatial* cues, such as a set of bounding boxes, do these elements work in harmony, or does one dominate the other? Our analysis of a pretrained image diffusion model that integrates gated self-attention into the U-Net reveals that spatial grounding often outweighs textual grounding due to the *sequential* flow from gated self-attention to cross-attention. We demonstrate that such bias can be significantly mitigated without sacrificing accuracy in either grounding by simply rewiring the network architecture, changing from sequential to *parallel* for gated self-attention and cross-attention. This surprisingly simple yet effective solution does not require any fine-tuning of the network but substantially reduces the trade-off between the two groundings. Our experiments demonstrate significant improvements from the original GLIGEN to the rewired version in the trade-off between textual grounding and spatial grounding. The project webpage is at `https://re-ground.github.io`.

Keywords: Textual Grounding · Spatial Grounding · Network Rewiring

1 Introduction

The emergence of diffusion models [17,44,45] has markedly propelled the field of text-to-image (T2I) generation forward, allowing users to generate high-quality images from text prompts. In a bid to further augment the creativity and controllability, recent efforts [2–5, 8, 11, 25, 28, 32, 36, 54, 55, 57] have focused on enabling these models to understand and interpret *spatial instructions*, such as layouts [4, 8, 28, 32, 36, 54, 55], segmentation masks [2–5, 11, 25] and sketches [51, 57].

Among them, *bounding boxes* are extensively employed in downstream image generation tasks [4, 8, 28, 32, 36, 55]. GLIGEN [28] is a pioneering work in terms of enhancing existing T2I models with the capability to incorporate additional spatial cues in the form of bounding boxes. Its core component, *gated self-attention*, is a simple attention module [50] that is plugged into each U-Net [42] layer of a pretrained T2I model such as Stable Diffusion [41], and is trained to accurately position various entities in their designated areas. A notable advantage of GLIGEN is that the original parameters of the underlying model remain unchanged, inheriting the generative capability of the T2I model while introducing the novel functionality of spatial grounding using bounding boxes. This capability has been leveraged by numerous studies to facilitate high-quality, layout-guided image generation [9, 12, 31, 54].

However, our analysis reveals that GLIGEN's integration of the gated self-attention into an existing T2I model is not optimal for blending new spatial guidance from bounding boxes with the original textual guidance. It often leads to the *omission of specific details* from the text prompts. For instance, in the first row and third column of Fig. 1, GLIGEN fails to reflect the description *"low poly illustration"* from the input text prompt. Also in the second row, a crucial detail in the text prompt, *"draped with a colorful blanket"*, is neglected in the output image. We refer to this issue as **description omission**. Such outcomes imply

that the current architectural design of GLIGEN does not effectively harmonize the new spatial guidance and the text conditioning in the given T2I model. Considering the widespread applications of GLIGEN in various layout-based generation tasks [9, 12, 31, 36, 53, 54, 59], these limitations represent a significant bottleneck.

To address the observed neglect of textual grounding in GLIGEN, we first analyze the root causes. Our investigation reveals that the issue arises from the *sequential* arrangement of the spatial grounding and textual grounding modules. Specifically, the output of the gated self-attention is directed to a cross-attention module in each layer of the U-Net architecture (Fig. 2-(b)).

Building on this insight, we propose a straightforward yet impactful solution: *network rewiring*. This approach alters the relationship between the two grounding modules from sequential to *parallel* (Fig. 2-(c)). Remarkably, this network modification significantly reduces the grounding trade-off between textual and spatial groundings without necessitating any adjustments to the network parameters. Importantly, **this rewiring does *not* require additional network training, extra parameters, or changes in computational load and time.** Simply reconfiguring the attention modules of the pretrained GLIGEN, originally trained with the sequential architecture, during inference dramatically enhances performance.

In our experiments on MS-COCO [30] and our newly introduced NSR-1K-GPT datasets, we demonstrate that rewiring the pretrained GLIGEN substantially reduces the trade-off between textual and spatial groundings. This is evidenced by the evaluation of text prompt alignment (measured using CLIP score [38], PickScore [26] and user study) and bounding box alignment (assessed by YOLO score [52]). Furthermore, we show that our rewiring also leads to better outcomes in other frameworks using GLIGEN as a backbone, including BoxDiff [54].

2 Related Work

2.1 Zero-Shot Guidance in Diffusion Models

The progress in diffusion models [17, 44, 45] has significantly elevated the capabilities of text-to-image (T2I) generation, resulting in foundation models [6, 37, 39–41] that exhibit remarkable generative performance. Leveraging the robust performance of these models, recent studies [3–5, 8, 11, 25, 28, 32, 36, 54, 55, 57] have introduced efficient guidance techniques designed to further improve the image generation process. Notably, numerous works [8, 28, 32, 36, 43, 54, 55] focus on the internal architecture (Fig. 2-(a)) of the denoising U-Net of Latent Diffusion Models (LDMs) [41], where self-attention and cross-attention modules are intertwined to facilitate inter-pixel communication and text conditioning. The self-attention of U-Net can be utilized to improve image quality [19] or facilitate image translation [49] and image editing tasks [7]. Since text conditions are integrated via cross-attention, the intermediate attention maps have been leveraged

to improve text faithfulness [13] or enable spatial manipulation of the generation process [35]. Recently, FreeU [43] analyzed the contributions of the backbone and residuals of the U-Net and proposed a *free-lunch* strategy to enhance image quality: reweighting the backbone and residual features maps. In contrast to previous works that only deal with self- and cross-attention in standard LDMs, we introduce a method to enhance GLIGEN [28] by reconnecting its gated self-attention with the other attention modules, thereby achieving performance improvement in zero-shot without any tuning of the network parameters.

2.2 Layout-Guided Image Generation

The use of layouts, particularly in the form of bounding boxes, has become a popular intermediary to bridge the gap between textual inputs and the images generated [14, 18, 21, 27, 29, 46, 47, 56, 58]. Layout2Im [58] samples object latent codes from a normal distribution, eliminating the need to predict instance masks as done in prior works [18, 21]. LostGAN [46] controls the style of each object by devising an extension of the feature normalization layer used in StyleGAN [22–24], while OC-GAN [47] incorporates the spatial relationships between objects using a scene-graph representation. LAMA [29] introduces a mask adaptation module that mitigates the semantic ambiguity arising from overlaps in the input layout. While these developments have greatly improved user control over image generation, their applicability is confined to the categories found in the training data, such as those of the MS-COCO [30] dataset.

In contrast, recent studies [5, 8, 10, 28, 36, 54, 55, 60] have extended layout-guided image generation towards *open-vocabulary*, building on the advancements of foundational text-to-image (T2I) models [41]. Training-free approaches [3, 5, 8, 36, 54] aim to improve the spatial grounding of T2I models through straightforward guidance mechanisms. GLIGEN [28], on the other hand, introduces gated self-attention, which is injected into the U-Net architecture of the Latent Diffusion Model [41], and is trained to equip the underlying model with spatial grounding abilities. Given the simple architecture of GLIGEN and its robust grounding accuracy with the input bounding boxes, numerous studies [36, 53, 54, 59] build upon its framework and propose further refinements to increase performance. In this work, we identify and address a significant performance bottleneck in GLIGEN related to description omission and propose a simple yet effective solution.

3 Background — Latent Diffusion Models [41]

Rombach *et al.* [41] proposed Latent Diffusion Model (LDM), a text-to-image (T2I) diffusion model with a U-Net as the noise prediction network. It is trained to generate an image from an input text prompt by predicting the noise $\epsilon(\mathbf{x}_t, t, c)$ conditioned both on the timestep t and the text embedding c. Each layer of LDM's U-Net consists of three core components: a convolutional residual block, followed by a self-attention (SA), and a cross-attention (CA) module (Fig. 2-(a)). In each l-th layer of the U-Net, its residual block first extracts intermediate visual features $F = (f_1, ..., f_{N_l})^T$ from the output of the previous layer.

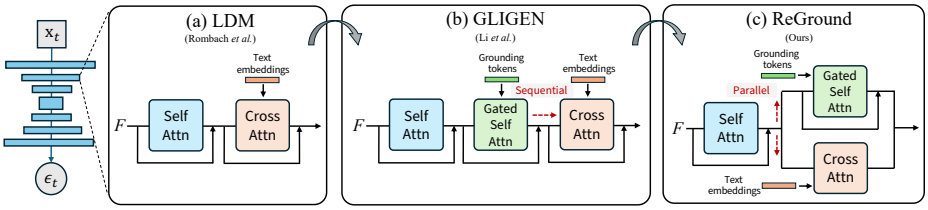


Fig. 2: Comparison between the U-Net architectures of (a) Latent Diffusion Model (LDM) [41], (b) GLIGEN [28] and (c) our ReGround. From LDM, GLIGEN enables spatial grounding by injecting Gated Self-Attention before cross-attention, forming a *sequential* flow of them. Based on GLIGEN, our ReGround changes the relationship of the two attention modules to become *parallel*, resulting in noticeable improvement in textual grounding while preserving the spatial grounding capability. (The residual block before self-attention is omitted.)

The self-attention module then facilitates interaction between the features in F. Subsequently, the cross-attention module enables the interaction between each visual feature f_i and the text embedding c. Throughout this process, the output feature of the previous module is also forwarded through a residual connection, as illustrated in lines 4-5 and 7 of Alg. 1 and also in Fig. 2-(a).

4 GLIGEN [28] and Description Omission

In this section, we review GLIGEN [28] and its key idea of employing gated self-attention for spatial grounding. Then, we present our key observations on the description omission issue that occurs due to the addition of gated self-attention.

4.1 Gated Self-Attention

Li *et al.* [28] propose a plug-in spatial grounding module, named *gated self-attention*, which adopts the gated attention mechanism [1] to equip a pretrained T2I diffusion model [41] with spatial grounding capabilities (Fig. 2-(b)). Given a set of bounding boxes and text labels for each of them, let b_i be the xy-coordinates of the i-th bounding box's top-left and bottom-right corners, and p_i be the corresponding text label. Then, the i-th grounding token is defined as $g_i := \mathcal{G}\left(\mathcal{T}\left(p_i\right), \mathcal{F}\left(b_i\right)\right)$, where $\mathcal{T}(\cdot)$ is a pretrained text encoder [20, 38], $\mathcal{F}(\cdot)$ is the Fourier embedding [33,48] and $\mathcal{G}(\cdot,\cdot)$ is a shallow MLP network that concatenates the two given embeddings, respectively. Given a set of grounding tokens $\{g_i\}$, gated self-attention learns the self-attention among the unified feature set $(f_1, ..., f_{N_l}, g_1, ..., g_M)$ where $\{f_i\}$ is the set of intermediate visual features in the l-th layer of U-Net, and M is the number of bounding boxes.

As shown in Fig. 2-(b), gated self-attention receives the output of the self-attention along with the residual features as its input and forwards the output features to the cross-attention module. By incorporating gated self-attention into

each layer of the U-Net, the model enables the placement of the entity specified in the text label p_i at the location indicated by the bounding box b_i. Note that the integration of gated self-attention does not require training the network from scratch or fine-tuning it, but can be accomplished simply by training the gated self-attention parameters while keeping all other parameters in the backbone model frozen.

Alg. 1 shows the pseudocode of the U-Net forward-pass including the plug-in of gated self-attention in line 6. Note that β_t is set to 1 for GLIGEN. If $\beta_t = 0$, the algorithm is identical to that of LDM [41].

Algorithm 1: Noise Prediction U-Net with Gated Self-Attention.

Parameters : β_t; `// Weight for GSA.`
Inputs: $\mathbf{x}_t, c, \{g_i\}_{i=0\cdots N-1}$; `// Noisy data at timestep` t`, text condition, and grounding tokens`
Outputs: ϵ_t; `// Noise at timestep` $t-1$`.`
1 **Function** `U-Net`$(\mathbf{x}_t, c, \{g_i\})$**:**
2 $F \leftarrow \mathbf{x}_t$
3 **for** $i = 0, \ldots, L-1$ **do**
4 $F_{\text{RS}} \leftarrow \text{Conv}(F) + F$; `// Residual block.`
5 $F_{\text{SA}} \leftarrow \text{SA}(F_{\text{RS}}) + F_{\text{RS}}$; `// Self-Attention module.`
6 $F_{\text{GSA}} \leftarrow \beta_t \cdot \text{GSA}(F_{\text{SA}}, \{g_i\}) + F_{\text{SA}}$; `// Gated Self-Attention module.`
7 $F \leftarrow \text{CA}(F_{\text{GSA}}, c) + F_{\text{GSA}}$; `// Cross-Attention module.`
8 $\epsilon_t \leftarrow F$;
9 **return** ϵ_t;

4.2 Description Omission

Despite its high accuracy in spatial grounding, GLIGEN [28] frequently struggles to capture essential attributes specified in the input text prompt. As illustrated in Fig. 3, the leftmost image shows *"a person"* and *"a skateboard"* accurately placed in their designated regions. However, a critical detail from the input text prompt, *"black and white photography"*, is absent in the output image. This discrepancy often emerges when the input comprises distinct but equally important descriptions regarding the image, presented through text prompts and bounding boxes. Such omissions not only fail to convey the stylistic intent of the image but also tend to overlook significant objects mentioned within the text prompt. Additional examples of this problem are showcased in Fig. 1, where the second row demonstrates the absence of a *"blanket"* in the generated image, a key element from the text prompt. This limitation significantly hampers GLIGEN's fidelity to user-provided text prompts, a challenge we term as **description omission**.

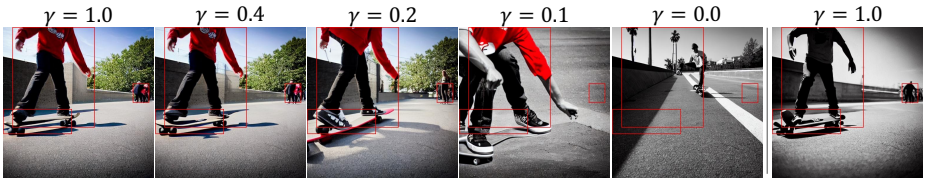


Fig. 3: (a) Images generated by GLIGEN [28] with varying activation duration of gated self-attention γ in scheduled sampling (Sec. 5.1). The **red** words in the text prompt denote the words used as labels of the input bounding boxes. Note that for GLIGEN to reflect the underlined description in the text prompt in the final image, γ must be decreased to 0.1, which compromises spatial grounding accuracy. (b) In contrast, our ReGround reflects the underlined phrase even when $\gamma = 1.0$, therefore achieving high accuracy in both textual and spatial grounding.

5 ReGround: Rewiring Attention Modules

Gated self-attention and cross-attention each play a crucial role in enabling spatial and textual groundings, by taking bounding boxes and text prompts as inputs, respectively. To tackle the issue of description omission, we first examine the impact of attention modules on the groundings they do not address: the effect of gated self-attention on textual grounding (Sec. 5.1), and the influence of cross-attention on spatial grounding (Sec. 5.2). Building on this analysis, we propose an approach for network reconfiguration, modifying the connections among self-attention, gated self-attention, and cross-attention modules (Sec. 5.3).

5.1 Impact of Gated Self-Attention on Textual Grounding

As the issue of description omission arises due to the newly added gated self-attention in GLIGEN [28], we first attempt to mitigate the impact of gated self-attention by using *scheduled sampling* [28], activating gated self-attention only in a few initial steps of the denoising process. This approach is inspired by the observation that the coarse structure of the final image is established within the first few denoising steps. The scheduling is applied by setting the weight of gated self-attention β_t (line 6 of Alg. 1) as

$$\beta_t = \begin{cases} 1 & (t \leq \gamma \cdot T) \\ 0 & (t > \gamma \cdot T), \end{cases} \tag{1}$$

where $\gamma \in [0, 1]$ represents the fraction of the initial denoising steps to activate gated self-attention.

Fig. 3-(a) shows an example of generated images while incrementally adjusting γ from 1.0 to 0.0. As γ is reduced from 1.0 to 0.0, the details specified in the text prompt, *"a black and white photograph"*, begin to be reflected starting

at $\gamma = 0.1$, demonstrating that longer activation of gated self-attention may interfere with the alignment of the output image with the text prompt. However, as gated self-attention is activated for shorter durations, the spatial grounding diminishes, as shown in the objects' reduced alignment with the input bounding boxes. This phenomenon illustrates the inherent trade-off between spatial and textual grounding, which cannot be resolved by controlling the duration of gated self-attention activation.

5.2 Impact of Cross-Attention on Spatial Grounding

We also investigate whether cross-attention has influence on spatial grounding. For this, we conduct a toy experiment by removing cross-attention modules in GLIGEN [28], allowing the output of the gated self-attention to be directly passed to the next layer of the U-Net. This modification is equivalent to changing line 7 of Alg. 1 to $F \leftarrow F_{GSA}$.

The results are displayed in Fig. 4. Note that, while the appearance of the background and objects changes, the silhouettes of the cat (left) and the individuals (right) remain precisely positioned within their respective bounding boxes *without* cross-attention. This observation indicates that while gated self-attention that is performed before cross-attention may compromise textual grounding, cross-attention that processes the output of gated self-attention does not affect spatial grounding.

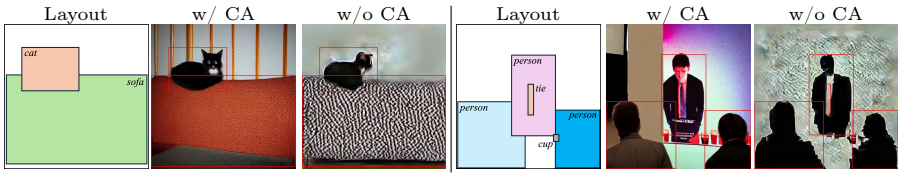


Fig. 4: Comparison of the output of GLIGEN [28] with and without cross-attention. While the absence of cross-attention reduces realism and quality of the image, the silhouette of objects remains grounded within the given bounding boxes, as shown in the third column of each case.

5.3 Network Rewiring: From Sequential to Parallel

Building on the analyses above, we propose a simple yet effective modification to the grounding mechanism, changing the relationship between gated self-attention and cross-attention from sequential to *parallel*. This change eliminates the placement of gated self-attention before cross-attention, thus preventing the reduction of text grounding caused by gated self-attention. Moreover, in this parallel arrangement, the preservation of spatial grounding is assured, as gated self-attention for spatial grounding does not require subsequent cross-attention.

Specifically, recall that in GLIGEN [28], the output of gated self-attention is added to the residual from self-attention, which is then passed to the cross-attention module as follows:

$$\begin{aligned} F_{GSA} &\leftarrow \underbrace{\mathrm{GSA}(F_{SA}, \{g_i\})}_{\textit{spatial grounding}} + F_{SA}; \\ F &\leftarrow \underbrace{\mathrm{CA}(F_{GSA}, c)}_{\textit{textual grounding}} + F_{GSA}; \end{aligned} \tag{2}$$

We propose to transform this sequence grounding pipeline into two parallel processes as follows:

$$F \leftarrow \underbrace{\mathrm{GSA}(F_{SA}, \{g_i\})}_{\textit{spatial grounding}} + \underbrace{\mathrm{CA}(F_{SA}, c)}_{\textit{textual grounding}} + \underbrace{F_{SA}}_{\textit{residual}}; \tag{3}$$

Refer to Fig. 2 for the visualization of network architecture changes ((b) $\rightarrow$ (c)). This *network rewiring* is feasible because the input to gated self-attention remains unchanged, while the input to cross-attention shifts to F_{SA}, for which it was originally designed in the context of Latent Diffusion Models [41].

It is important to note that the modification is effective even when applied to the pretrained GLIGEN, which was trained with the sequential structure of the attention modules. Therefore, **our rewiring does not require any additional training or fine-tuning, introduces no extra parameters, and does not affect computation time or memory usage during the generation process.** The only requirement is the simple reconfiguration of the attention modules at inference time.

6 Experiments

In this section, we show the effectiveness of our ReGround by evaluating the spatial grounding on existing layout-caption datasets [12, 30] and the textual grounding on text-image alignment metrics [26, 38]. We use the official GLIGEN [28] checkpoint which is trained based on Stable Diffusion v1.4 [41].

6.1 Datasets

MS-COCO. We use the validations sets of both MS-COCO-2014 and MS-COCO-2017 datasets [30]. Each dataset provides image-captions pairs and the xy-coordinates of bounding boxes along with their corresponding object categories.

NSR-1K-GPT. We also use the NSR-1K benchmark [12] for evaluation. Based on each subset of NSR-1K—*Counting* and *Spatial*—we develop a new benchmark, *NSR-1K-GPT*, augmenting each original caption in NSR-1K using GPT-4 [34]. The instructions for augmentation are to (i) elaborate on the descriptions of each mentioned entity and (ii) provide additional details about the background of the image. More details on the evaluation datasets are provided in the **Supplementary (Sec. S1)**.

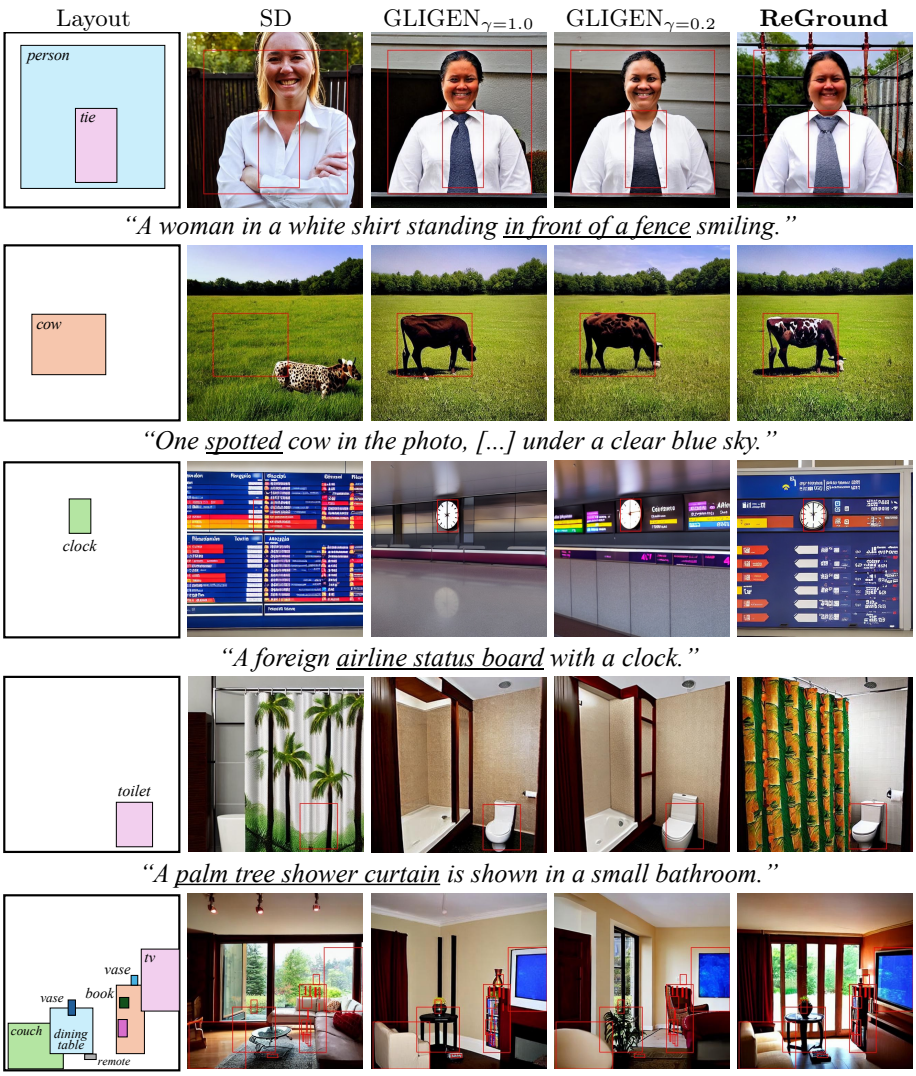
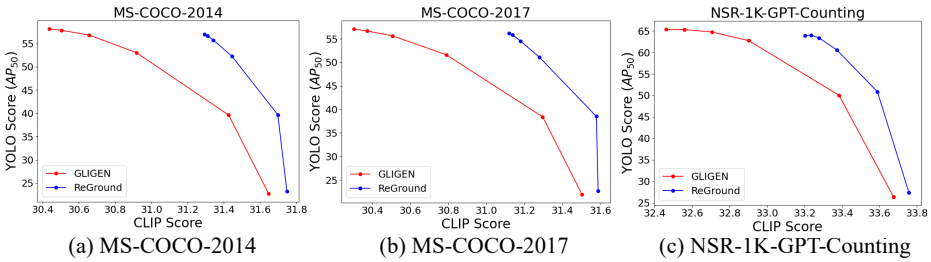


Fig. 5: Qualitative comparisons. Stable Diffusion (SD, 2nd column) generates images that align with the given text descriptions, including the underlined phrase in each row, but cannot take bounding boxes as input. GLIGEN (3rd column) creates images that match the input layouts but suffers from *description omission*, failing to reflect the underlined descriptions. Scheduled sampling strategy (4th column) can partially address this issue (for instance, in the 5th row, where *"window"* appears in the room), but it results in a noticeable decline in spatial accuracy (as seen in the 1st row, where the tie is not generated). In contrast, our method (last column) accurately incorporates the underlined text descriptions while maintaining precise spatial representation.



(a) MS-COCO-2014 (b) MS-COCO-2017 (c) NSR-1K-GPT-Counting

Fig. 6: Comparisons on MS-COCO [30] and NSR-1K-GPT. Each plot shows the relationship between textual grounding (*i.e.* CLIP score [15]) and spatial grounding (*i.e.* YOLO score [52]) accuracy of GLIGEN [28] and our method. Note that the plot of our ReGround is positioned in the top-right quadrant relative to GLIGEN, signifying that it alleviates the inherent trade-off between textual and spatial grounding.

6.2 Evaluation Metrics

- **YOLO score:** Spatial grounding accuracy is assessed using YOLO score [52]. We employ YOLOv7 [52] to detect objects in each generated image and compute the average precision (AP) based on the ground truth bounding box annotations from MS-COCO [30].
- **CLIP score:** Textual grounding accuracy is assessed using CLIP score [15].
- **FID:** Image quality and diversity are evaluated using FID [16].
- **User Study and PickScore**: We conduct a user study to assess human preferences for the generated images based on each input text prompt. Additionally, we use PickScore [26], a human preference predictor, to further analyze the results.

6.3 Comparison with GLIGEN

Textual-Spatial Grounding Trade-off. We first examine the trade-off between textual and spatial groundings for both GLIGEN [28] and our ReGround, the rewired version of GLIGEN, while varying the scheduled sampling parameter γ from 1.0 to 0.1.

Fig. 6-(a), (b) present the graphs of CLIP score [15] and YOLO score [52] measured on the MS-COCO datasets [30]. In MS-COCO-2014, when reducing γ from 1.0 to 0.1, the CLIP score of GLIGEN varies from 30.44 to 31.65, while the YOLO score significantly drops from 58.13 to 22.75 (red in Fig. 6). In contrast, our ReGround, (blue in Fig. 6), demonstrates a notably superior trade-off between textual and spatial grounding. Specifically, with γ set to 1.0, ReGround already achieves a CLIP score of 31.29, accounting for 70.25% of GLIGEN's total improvement in CLIP score when γ is reduced from 1.0 to 0.1. Despite this significant increase in CLIP score, the YOLO score remains largely unchanged, marking 56.96 which represents only a 3.31% decrease in the range of YOLO score variation for GLIGEN when γ is adjusted from 1.0 to 0.1. Moreover, when varing the γ, the plot for ReGround (blue) is constantly on the upper right

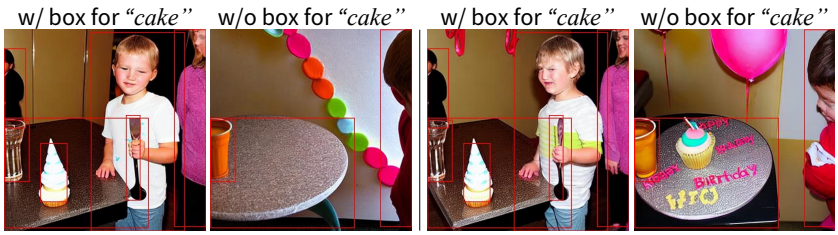


Fig. 7: Generated images from the text prompt and bounding boxes from the MS-COCO-2017 (left of each column) and our COCO-Drop (right of each column). While GLIGEN [28] fails to generate *"a birthday cupcake"* when the corresponding bounding box is removed, our ReGround successfully generates a cupcake on the table.

side of GLIGEN (red), signifying a more advantageous trade-off across varying γ. The same pattern is observed in MS-COCO-2017, where our ReGround achieves 68.33% of the increase in CLIP score of GLIGEN while only compromising YOLO score by 2.62% compared to the decrease for GLIGEN.

Fig. 6-(c) further shows a quantitative comparison on the *Counting* subset of the newly generated NSR-1K-GPT benchmark. The plot reveals a consistent trend with the MS-COCO datasets. By reducing γ from 1.0 to 0.1, GLIGEN's CLIP score is increased from 32.46 to 33.67, while the YOLO score is decreased from 65.36 to 26.38. In contrast, when $\gamma = 1.0$, ReGround achieves a CLIP score of 33.20, which is equal to 61.16% of GLIGEN's total improvement in CLIP score, while the compromise in YOLO score is equal to only 3.69% of the total decrease in the YOLO score of GLIGEN from $\gamma = 1.0$ to $\gamma = 0.1$. Moreover, a comparison on the *Spatial* subset of NSR-1K-GPT is provided in the **Supplementary (Sec. S2)**. These results highlight that the advantage of our ReGround holds robustly for the realistic image captions provided in the MS-COCO [30], as well as for diverse text prompts generated by GPT-4 [34].

Random Box Dropping. To further assess the extent of description omission in each method, we modify the MS-COCO-2017 dataset [30] to make the *COCO-Drop* dataset. In this version, the bounding boxes for 50% of the categories are randomly removed from each image, thereby preventing every entity described in the text prompt from being included within the bounding boxes.

Fig. 8 shows the quantitative comparison of ReGround and GLIGEN on COCO-Drop. In this case, ReGround shows a larger advantage over GLIGEN in CLIP score, obtaining a gap in CLIP score which is 1.57 times that of the original MS-COCO-2017 dataset before box dropping for $\gamma = 1.0$. Such a larger gap in CLIP score demonstrates that compared to GLIGEN, our ReGround better reflects the text prompts even when some entities in the text prompt are not provided as a bounding box. Fig. 7 displays a representative example, where GLIGEN fails to generate a *"cupcake"* when its corresponding bounding

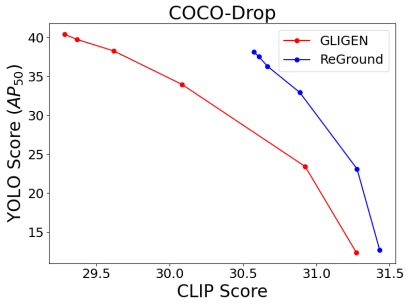


Fig. 8: Comparison on the COCO-Drop dataset.

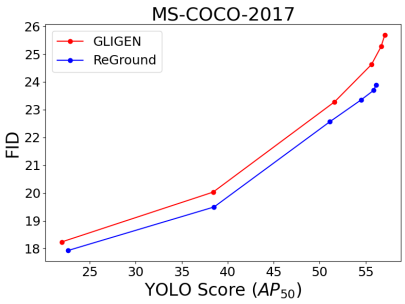


Fig. 9: Comparison of FID [16] on MS-COCO-2017 [30] dataset.

box is removed in COCO-Drop, whereas our ReGround robustly generates the cupcake even when it it not provided as a bounding box.

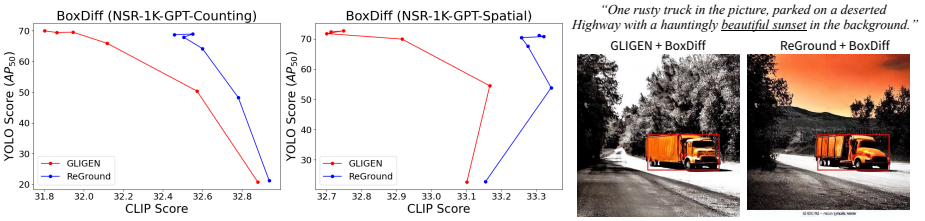
Image Quality. Fig. 9 displays the relationship between YOLO score [52] and FID [16] for each method on MS-COCO-2017. Note that the FID of ReGround is constantly lower than that of GLIGEN [28], meaning that our network rewiring also results in higher image quality and diversity.

User Study. We conducted a user study to compare GLIGEN and our ReGround in terms of faithfulness to input text prompts. We used GPT-4 [34] to generate 100 prompts each containing two different objects, along with a bounding box for each object. Participants were given the text prompt along with two images—one from each method—and asked to choose the image that "`better includes all the objects from the prompt.`" Among the 92 out of 100 participants who passed the vigilance tests, our ReGround surpassed GLIGEN, with a preference rate of 70.05% compared to 29.95%. Further details on the user study are provided in the **Supplementary (Sec. S1)**.

PickScore. We further compare the PickScore [26] of GLIGEN [28] and our ReGround given each input text prompt. On MS-COCO-2017, ReGround is preferred over GLIGEN by 55.66% to 44.34%, and on COCO-Drop, ReGround is preferred by 57.57% to 42.43%.

6.4 Impact of ReGround as a Backbone

We demonstrate that applying our rewiring of attention modules can also improve text-image alignment in other layout-guided generation methods that use GLIGEN as a backbone. For instance, BoxDiff [54] is a notable example that uses GLIGEN as its foundation and improves spatial grounding with respect to the bounding boxes by leveraging cross-attention maps as additional spatial cues in a zero-shot manner. Our network rewiring can also be combined with



(a) NSR-1K-GPT-Counting (b) NSR-1K-GPT-Spatial (c) BoxDiff Outputs

Fig. 10: Comparison of applying BoxDiff [54] on GLIGEN [28] and our ReGround, respectively. (a) and (b) show that our ReGround further improves the grounding quality of BoxDiff on NSR-1K-GPT datasets. (c) While BoxDiff with GLIGEN (left) also shows description omission—omitting *"beautiful sunset"* from the text prompt—BoxDiff with our ReGround contains the sunset in the final image (right).

the zero-shot guidance of BoxDiff. Fig. 10 illustrates the results on the NSR-1K-GPT datasets (a) when BoxDiff uses GLIGEN as the base, and (b) when it uses our ReGround, the rewired GLIGEN, as the base. It depicts that for the same range of spatial grounding accuracies, ReGround obtains noticeably higher textual grounding (*i.e.* CLIP score [15]). Also, as shown in Fig. 10-(c), our network rewiring allows for a more detailed description to accurately appear in the final image, both for the entities in the bounding boxes (*"truck"*) and the entities that are given as a text prompt (*"sunset"*).

7 Conclusion

We have demonstrated that a simple network rewiring of attention modules, making the gated self-attention and cross-attention parallel, surprisingly improves the trade-off between textual and spatial grounding at no additional cost — without introducing any new parameters, any fine-tuning of the network, or any changes in generation time and memory. Using the pretrained GLIGEN [28], which was trained with the original sequential architecture of the two attention modules, the reconfiguration at inference time has led to achieving higher CLIP scores, indicating the noticeable improvement in textual grounding accuracy. Moreover, our ReGround improves the textual grounding while preserving the spatial grounding accuracy – achieving 70.25% and 68.33% of GLIGEN's total improvement with the scheduled sampling in CLIP score while compromising YOLO score only 3.31% and 2.62% for the MS-COCO-2014 and MS-COCO-2017 datasets, respectively. We also showcased that this simple yet effective solution for the textual-spatial grounding trade-off can lead to improvements in diverse frameworks using GLIGEN as a base.

Supplementary. Due to limited space, we provide the following contents in the Supplementary: details on the evaluation setup (Sec. S1), additional quantitative (Sec. S2) and qualitative (Sec. S4) comparisons, and more results of ReGround as a backbone for other layout-guided generation methods (Sec. S3).

Acknowledgments

This work was supported by NRF grant (RS-2023-00209723), IITP grants (RS-2019-II190075, RS-2022-II220594, RS-2023-00227592, RS-2024-00399817), and Alchymist Project Program (RS-2024-00423625) funded by the Korean government (MSIT and MOTIE), and grants from the DRB-KAIST SketchTheFuture Research Center, NAVER-intel, Adobe Research, Hyundai NGV, KT, and Samsung Electronics.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
2. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: CVPR (2023)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: ICLR (2024)
5. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023)
6. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions (2023), `https://cdn.openai.com/papers/dall-e-3.pdf`
7. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: ICCV (2023)
8. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: WACV (2024)
9. Chen, W.G., Spiridonova, I., Yang, J., Gao, J., Li, C.: Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. arXiv preprint arXiv:2311.00571 (2023)
10. Cheng, J., Liang, X., Shi, X., He, T., Xiao, T., Li, M.: Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. arXiv (2023)
11. Couairon, G., Careil, M., Cord, M., Lathuilière, S., Verbeek, J.: Zero-shot spatial layout conditioning for text-to-image diffusion models. In: ICCV (2023)
12. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: LayoutGPT: Compositional visual planning and generation with large language models. In: NeurIPS (2023)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: ICLR (2022)
14. Herzig, R., Bar, A., Xu, H., Chechik, G., Darrell, T., Globerson, A.: Learning canonical representations for scene graph to image generation. In: ECCV (2020)
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)

16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2018)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
18. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: CVPR (2018)
19. Hong, S., Lee, G., Jang, W., Kim, S.: Improving sample quality of diffusion models using self-attention guidance. In: ICCV (2023)
20. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773
21. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR (2018)
22. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020)
25. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: ICCV (2023)
26. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. NeurIPS (2024)
27. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: CVPR (2019)
28. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: GLIGEN: Open-set grounded text-to-image generation. In: CVPR (2023)
29. Li, Z., Wu, J., Koh, I., Tang, Y., Sun, L.: Image synthesis from layout with locality-aware mask adaption. In: CVPR (2021)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
31. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
32. Ma, W.D.K., Lewis, J.P., Lahiri, A., Leung, T., Kleijn, W.B.: Directed diffusion: Direct control of object placement through attention guidance. arXiv preprint arXiv:2302.13153 (2023)
33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
34. OpenAI: Chatgpt, https://chat.openai.com/
35. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM TOG (2023)
36. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023)
37. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
39. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
40. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
42. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
43. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)
44. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2021)
45. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
46. Sun, W., Wu, T.: Image synthesis from reconfigurable layout and style. In: ICCV (2019)
47. Sylvain, T., Zhang, P., Bengio, Y., Hjelm, R.D., Sharma, S.: Object-centric image generation from layouts. In: AAAI (2021)
48. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: NeurIPS (2020)
49. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
51. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. ACM TOG (2023)
52. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: CVPR (2023)
53. Xiao, J., Li, L., Lv, H., Wang, S., Huang, Q.: R&b: Region and boundary aware zero-shot grounded text-to-image generation. In: ICLR (2024)
54. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: ICCV (2023)
55. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: CVPR (2023)
56. Yang, Z., Liu, D., Wang, C., Yang, J., Tao, D.: Modeling image composition for complex scene generation. In: CVPR (2022)
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
58. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: CVPR (2019)
59. Zhao, P., Li, H., Jin, R., Zhou, S.K.: Loco: Locally constrained training-free layout-to-image synthesis. arXiv preprint arXiv:2311.12342 (2023)
60. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: CVPR (2023)