# GOAL: Grounded text-to-image Synthesis with Joint Layout Alignment Tuning

Anonymous Authors

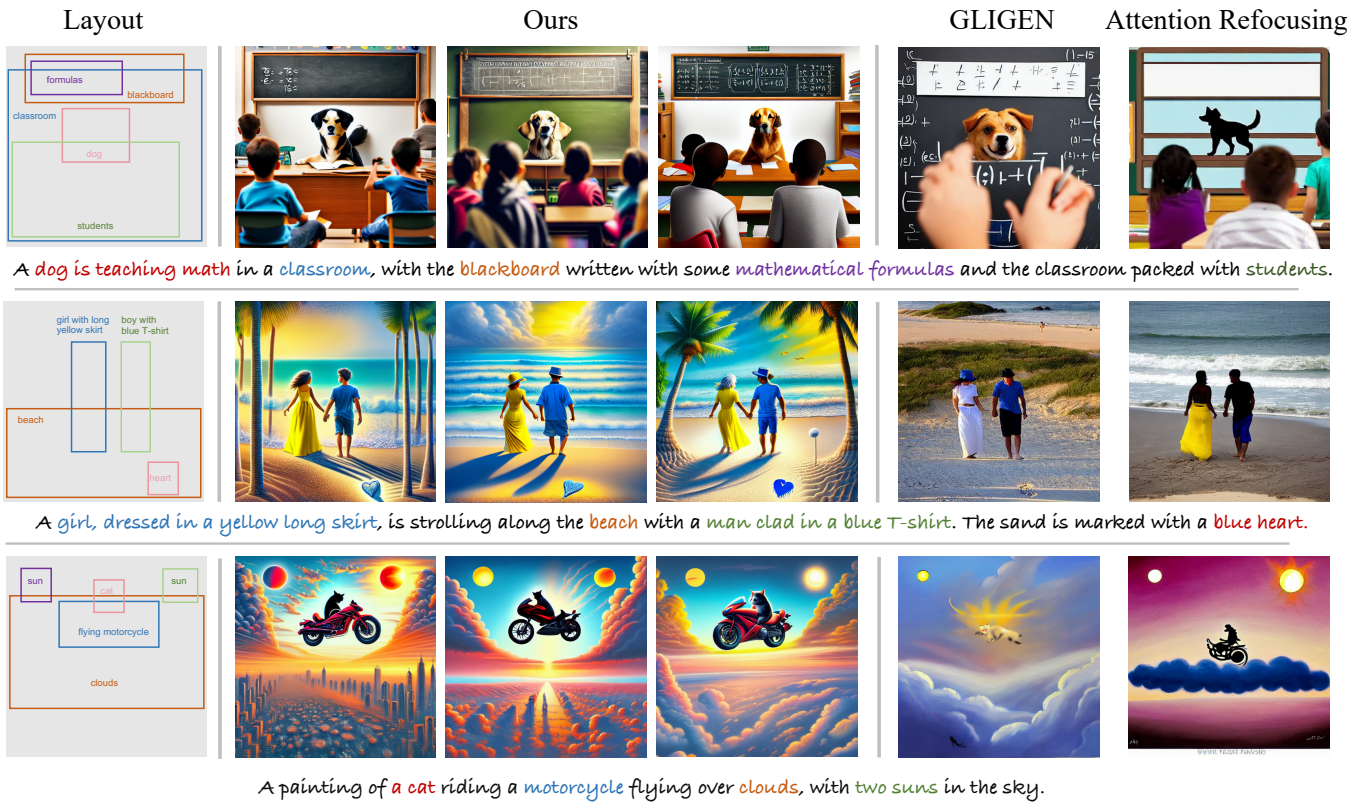


**Figure 1: Given a user-specified text prompt and layout condition of complex scenes, baseline models (*e.g.*, GLIGEN [26], Attention Refocusing [34]) fail to solve text-image misalignment problems such as attribute errors and relation mistakes. Our method exhibits outstanding generation performance for text-to-image synthesis task, which generates images faithfully capturing the details of text prompts.**

## ABSTRACT

Recent text-to-image (T2I) synthesis models have demonstrated intriguing abilities to produce high-quality images based on text prompts. However, current models still face Text-Image Misalignment problem (*e.g.*, attribute errors and relation mistakes) for compositional generation. Existing models attempted to condition T2I models on grounding inputs to improve controllability while ignoring the explicit supervision from the layout conditions. To tackle this issue, we propose Grounded jOint lAyout aLignment (GOAL), an effective framework for T2I synthesis. Two novel modules, discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign), are proposed and incorporated in this framework to improve the text-image alignment. DSAlign leverages discriminative tasks at the region-wise level to ensure low-level semantic alignment. MAAlign provides high-level attention alignment by guiding the model to focus on the target object. We also build a dataset GOAL2K for model fine-tuning, which composes 2000 semantically accurate image-text pairs and their layout annotations. Comprehensive evaluations on T2I-Compbench, NSR-1K, and Drawbench demonstrate the superior generation performance of our method. Especially, there are improvements of 19%, 13%, and 12% in color, shape, and texture metrics for T2I-Compbench. Additionally, Q-Align metrics demonstrate that our method can generate images of higher quality.

**Unpublished working draft. Not for distribution.**


*ACM MM, 2024, Melbourne, Australia*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

## KEYWORDS

Text-to-Image Generation; Diffusion Model; Large Language Model

## 1 INTRODUCTION

Recently, text-to-image (T2I) diffusion models have made tremendous progress in generating high-fidelity and diverse images in response to textual prompts [8, 13, 19, 40–42]. However, models still struggle with the Text-Image Misalignment problem [4, 9, 15, 29] for compositional T2I generation, which means they often fail to compose multiple objects with various attributes (*e.g.*, color, shape, texture) and complex spatial relations, as shown in the failure cases of GLIGEN [26] in Figure 1.

Previous works exploring compositional T2I generation can be classified into two main groups: training-free [9, 21, 34] and training-based approaches [26, 42, 52]. Specifically, training-free methods [9, 21] focus on directly altering the latent and cross-attention maps or adding objects by continuous editing [27, 53]. For instance, Attention Refocusing [9] employs inference-time optimization on the cross-attention map to align intermediate outputs with the layout conditions. However, it adds considerable cost during inference and may lead to image quality degradation, as shown in the last column of Figure 1.

Meanwhile, other training-based methods [42, 52] incorporate additional modules for controllable image generation. For example, as shown in Figure 2(b), GLIGEN [26] integrates grounding information into designed gated self-attention layers, supporting existing T2I models on layout inputs. Nevertheless, these layout-aware models ignore inherently explicit supervision in layout conditions, leaving them only as guidance for estimating the added global noise during training. We hypothesize that this objective is insufficient for complex generation tasks. Therefore, a natural question arises: Can we find fine-grained and explicit supervision with layout conditions as auxiliary training objectives to enhance text-image alignment? To achieve this, we adopt multiple phrase-region pairs in layout conditions as 'ground truth' labels, allowing the diffusion model to learn region-wise information within the image's context in an atomistic manner. Given that annotation of the "ground-truth" labels (*e.g.*, layout) from human is costly, we borrow large language model's (*e.g.*, GPT-4 [2]) strong scene understanding ability for layout planning.

In this work, we propose Grounded jOint lAyout aLignment (GOAL), an effective framework that utilizes layout conditions to provide explicit region-wise supervision for text-image alignment. It is achieved by incorporating discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign) as auxiliary training objectives. As shown in Figure 2(c), GOAL obtains a denoised version of the clean image via a single denoising step during training, DSAlign then utilizes discriminative tasks at the region-wise level to refocus on refining the generated context, ensuring low-level semantic alignment.

Considering that a more detailed denoised image can provide richer information for semantic alignment, we refine image details by applying region-wise MAAlign in high-level feature space. This process involves shifting attention towards the target region while suppressing the attention of unrelated areas. As a result, sharper and more detailed images are generated, making semantic alignment more effective and further enhancing text-image alignment.

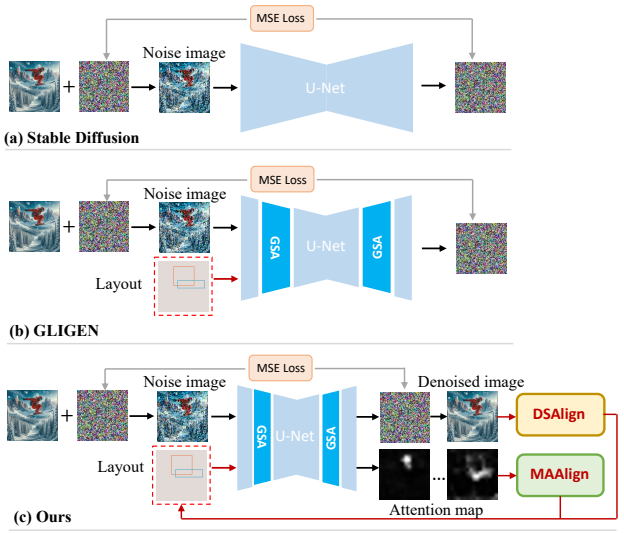


**Figure 2: (a) Stable Diffusion [40] is optimized to estimate the added noise with MSE loss. (b) GLIGEN [26] integrates grounding information into designed gated self-attention (GSA) layers, supporting existing T2I models on layout inputs. (c) GOAL provides explicit region-wise supervision by incorporating discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign).**

By conducting extensive experiments, the model exhibits significant improvement on several benchmarks, including T2I-Compbench [20], NSR-1K [16] and Drawbench [42]. Importantly, these enhancements are achieved without incurring additional inference costs. Along with our proposed training framework, we also publish a curated dataset GOAL2k for training, which consists of 2000 multi-modal samples with layout annotation for improving text-image alignment especially in complex scenes. Moreover, images included in GOAL2k are generated by outstanding T2I model DALLE-3 [39], demonstrating both high-detail and semantically accurate characteristics.

The main contributions of this paper are summarized as follows:

- We propose a Grounded jOint lAyout aLignment (GOAL), which is a novel layout-aware training framework. Discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign) are incorporated in this framework to improve the text-image alignment.
- We build a dataset GOAL2K to study the effectiveness of our alignment-based objectives, which composes 2000 semantically accurate image-text pairs and their layout annotations for model fine-tuning.
- We conduct comprehensive experiments on existing methods of T2I generation on T2I-Compbench, NSR-1K and Drawbench, and show that our method compares favorably against the state-of-the-art models.

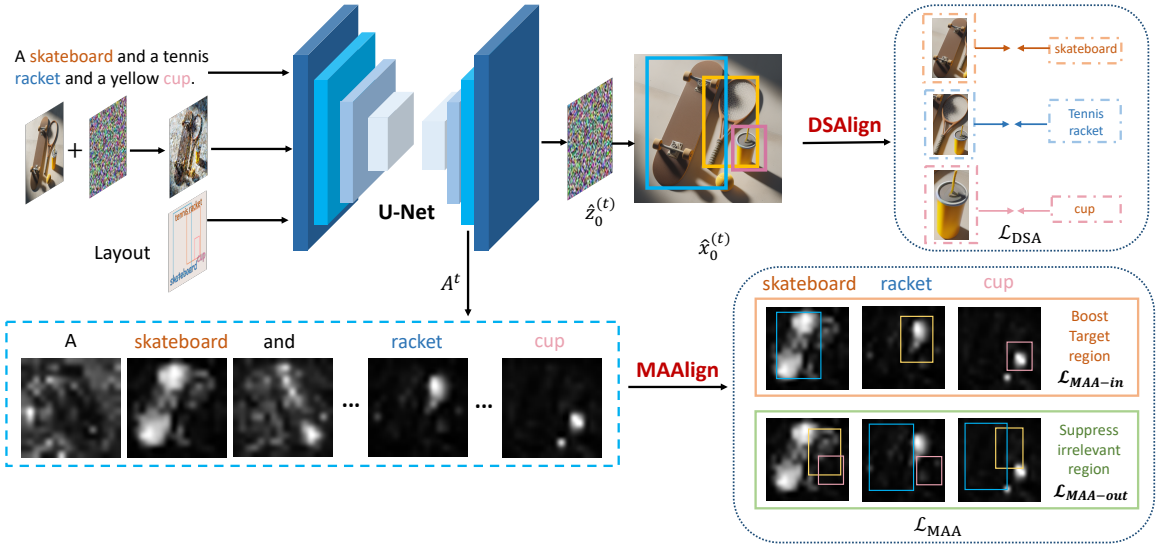


**Figure 3: An overview of the grounded joint layout alignment (GOAL) framework, which provides explicit region-wise alignment by discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign). GOAL obtains a denoised version of the clean image through a single denoising step and performs DSAlign by directly optimizing low-level semantic alignment. Furthermore, MAAlign is employed for high-level attention alignment, jointly optimizing the U-Net with DSAlign and its original denoising objective.**

## 2 RELATED WORKS

### 2.1 Text-to-Image Models

Early methods to text-to-image (T2I) generation primarily relied on Generative Adversarial Networks (GANs) [1, 12, 38, 44, 51]. However, recent advancements have shifted the focus towards diffusion models [8, 13, 19, 32, 45], which have gained prominence due to their exceptional capabilities in generating high-quality images. The Denoising Diffusion Probabilistic Model (DDPM) [19, 33, 46] introduces standard noise in the forward process and reconstructs the image from noise in the reverse process. Unlike denoising in the pixel space, the Latent Stable Diffusion Model (LDM) [40] conducts the denoising process in latent space, significantly reducing computational costs. Due to this remarkable progress, LDM has found wide application across various tasks, including image editing [7, 11, 18], image super-resolution [24, 43], inpainting [31, 50] and semantic segmentation [3, 6]. Building upon these state-of-the-art models, our work aims to enhance the capabilities of text-image alignment according to given layouts.

### 2.2 Training-Free Layout-Aware Generation

Despite the remarkable image generation capabilities demonstrated by Diffusion models, they encounter challenges in compositional generation, particularly within complex scenes. Recent methods [9, 34] primarily tackle this issue during the inference stage by incorporating layout-aware attention supervision at specific steps. BoxDiff [49] controls the noise map by adjusting cross-attention and self-attention layers. Attend-and-Excite [9] improves the generation of missing objects by maximizing the attention score for each object. Attention Refocusing [9] employs guidance functions to align intermediate outputs with layout conditions. Paint-with-words [5] enhances the cross-attention scores between image and text tokens corresponding to the same object based on segmentation masks. Continuous Layout Editing [53] disentangles various object concepts and facilitates continuous editing to align images with layouts. However, these methods incur computational costs during inference. Moreover, due to the direct optimization in the attention map without training, they may cause image distortion.

### 2.3 Training-Based Text-image Alignment

Several works aim to improve text-image alignment by fine-tuning diffusion models [10, 14, 26, 52] to integrate layout conditions. GLIGEN [26] integrates grounding information by the gated self-attention mechanism, enabling existing pre-trained T2I diffusion models to be conditioned on grounding inputs. LayoutLLM-T2I [36] introduces a relation-aware attention module, integrating semantic relation to generate high-fidelity images. Frido [14] performs multi-scale coarse-to-fine denoising to generate images of complex scenes. ReCo [52] incorporates spatial coordinates to achieve precise region control for arbitrary objects. Inspired by these works, we propose layout alignment at the semantic and attention level simultaneously to effectively fine-tune pre-trained T2I models without non-negligible costs.

## 3 METHOD

### 3.1 Preliminaries

Latent Diffusion Models (LDM) [40] are widely used in conditional image generative tasks. Given an image $x_0 \in \mathbb{R}^{H\times W\times 3}$, VAE $\mathcal{E}$ is adopted to encode image into latent space as $z_0 = \mathcal{E}(x_0)$. Then

Gaussian noise $\epsilon$ is added to the latent $z_0$ with a randomly sampled timestep $t$, yielding $z_t$ as:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \tag{1}$$

where $\alpha_t$ defines the level of noise. To achieve conditional image generation, the caption describing the image is encoded by text encoder $\varphi$ as the text embedding $\varphi(y)$ and then injected into the U-Net via cross-attention. Then U-Net $\epsilon_\theta$ is trained to predict the added noise $\epsilon$, following the objective as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \varphi(y)\right)\|_2^2\right]. \tag{2}$$

To ground the generation process via the additional condition, GLIGEN [26] incorporates the semantic information of grounding entity and spatial configurations through gated self-attention as:

$$v = v + tanh(\gamma)\cdot(SelfAttn[v, h]), \tag{3}$$

where $v$ is the visual feature, $\gamma$ is a learnable scalar which is initialized as 0 and $h$ is the added condition such as layout. Following GLIGEN [26], we optimize gated self-attention for stable training.

## 3.2 Discriminative Semantic Alignment

In the current layout-aware LDM, which incorporates additional layout conditions and optimizes the loss function $\mathcal{L}_{LDM}$ for noise prediction, the relationships within the bounding box, such as attributes and objects, are not explicitly optimized. As depicted in Figure 1, this results in poor semantic-level alignment in specific regions. Particularly, when different layouts overlap, the fusion of multiple conditions may result in a dissonant visual composition. Therefore, we propose discriminative semantic alignment (**DSAlign**), leveraging discriminative tasks at the region-wise level to refocus on refining the generated context. Specifically, with the noise predictor $\epsilon_\theta$ and $z_t$, we reconstruct the latent noise $z_0$ via a single denoising step using the reverse of Equation 1, and then obtain the denoised version of the clean image $x_0$ by VAE decoder $\mathcal{D}$. The formulation is:

$$\hat{x}_0^{(t)} = \mathcal{D}\left(\hat{z}_0^{(t)}\right) = \mathcal{D}\left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta\left(z_t, y, t\right)}{\sqrt{\alpha_t}}\right). \tag{4}$$

Given the spatial layout defined by $B$ bounding boxes $b_i \in \left(\mathbb{Z}^+\right)^{1\times 4}$, where $i \in [0, B)$, we perform DSAlign by minimizing the distance between the image embeddings of the target region and the corresponding phrase $p_i$ describing the region of $b_i$. To crop the region from the image, we construct the $M_i$ from the box $b_i$ and obtained the masked image as $\hat{x}_0^{(t)} \cdot M_i$, which is then encoded by the image encoder of CLIP, denoted by $CLIP_{img}$. We adopt the text encoder from CLIP, denoted by $CLIP_{text}$, to encode the corresponding phrase captions $p_i$. Then a normalization operation is performed to align the two embeddings within a unified semantic space. This process is formulated as follows:

$$\bar{x}_{0,m_i}^{(t)} = norm(CLIP_{img}(\hat{x}_0^{(t)} \cdot M_i)), \tag{5}$$

$$\bar{p}_i = norm(CLIP_{text}(p_i)). \tag{6}$$

Subsequently, the distance between the two normalized embeddings is computed using spherical distance as follows:

$$D_{sp}\left(\bar{p}_i, \bar{x}_{0,m_i}^{(t)}\right) = \arcsin\left(\frac{\|(\bar{p}_i) - (\bar{x}_{0,m_i}^{(t)})\|_2}{2}\right)^2. \tag{7}$$

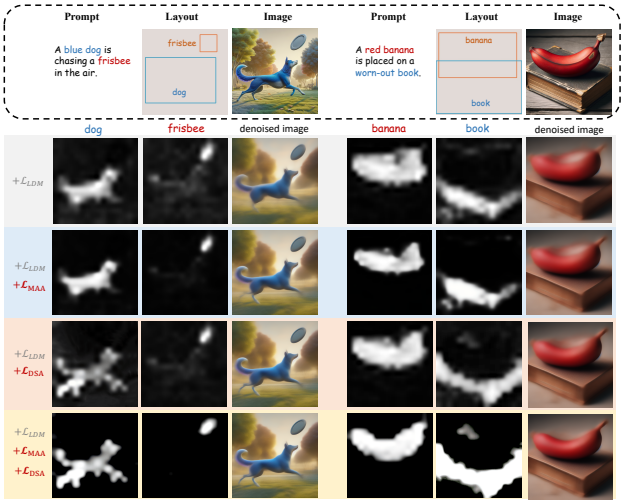


**Figure 4: Visualization of the attention maps and denoised images obtained from a single denoising step during training with different loss compositions.**

The average distance between all the $B$ target regions and the corresponding phrases is:

$$D_{local} = \frac{1}{B}\sum_0^B D_{sp}\left(\bar{p}_i, \bar{x}_{0,m_i}^{(t)}\right), \tag{8}$$

where $D_{local}$ represents the local-level semantic alignment guidance between target regions and local phrases. Moreover, we evaluate the global-level semantic alignment between the entire image and the text as follows:

$$D_{global} = D_{sp}\left(\bar{y}, \bar{x}_0^{(t)}\right) = \arcsin\left(\frac{\|(\bar{y}) - (\bar{x}_0^{(t)})\|_2}{2}\right)^2, \tag{9}$$

where $\bar{y} = norm(CLIP_{text}(y))$ and $\bar{x}_0^{(t)} = norm(CLIP_{img}(\hat{x}_0^{(t)}))$. To ensure the text-image alignment, we directly optimize local-level and global-level semantic alignment by $\mathcal{L}_{DSA}$:

$$\mathcal{L}_{DSA} = D_{local} + D_{global}. \tag{10}$$

## 3.3 Masked Attention Alignment

Given that a more detailed denoised image can provide richer information for semantic alignment performed by DSAlign, we refine image details through region-wise masked attention alignment (MAAlign) in high-level feature space. This process involves shifting attention towards the target region while suppressing the attention of unrelated areas. Specifically, for each cross-attention layer, let $Q \in \mathbb{R}^{hw\times d}$ be the input intermediate feature to the cross attention layer, which is obtained from the feature map of size $h \times w$ with feature dimension $d$, and let $K \in \mathbb{R}^{n\times d}$ be the transformed text embedding with $n$ tokens via a linear map, we can obtain cross-attention map $A^t$ at the step $t$ as follows:

$$A^t = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right). \tag{11}$$

To ensure high responses within the masked regions, we shift attention towards the target regions by boosting the values of the masked attention map:

$$\mathcal{L}_{MAA-in} = \frac{1}{B}\sum_{i\in B}\left(1-\max\left(A_j^t \cdot M_i\right)\right), \tag{12}$$

where $j \in [0, n)$ is the token index describing the content of $M_i$, and $A_j^t$ represents the corresponding attention map. Moreover, we suppress attention on irrelevant areas by preventing attention from extending beyond the target regions:

$$\mathcal{L}_{MAA-out} = \frac{1}{B}\sum_{i\in B}\max\left(A_j^t \cdot (1-M_i)\right). \tag{13}$$

The overall masked attention alignment loss is defined as:

$$\mathcal{L}_{MAA} = \mathcal{L}_{MAA-in} + \mathcal{L}_{MAA-out}. \tag{14}$$

In this way, MAAlign provides attention alignment by guiding the model to focus on the target object. Finally, the training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{LDM} + \alpha\mathcal{L}_{DSA} + \beta\mathcal{L}_{MAA}, \tag{15}$$

where $\alpha$ and $\beta$ are trade-off parameters to balance the contributions of the $\mathcal{L}_{DSA}$ and $\mathcal{L}_{MAA}$.

Figure 4 illustrates the attention map $A^t$ and denoised image $\hat{x}_0^{(t)}$ with various combinations of loss terms. It's clear that when employing $\mathcal{L}_{DSA}$ (3rd row), the cross-attention maps exhibit a closer alignment with the real image. Furthermore, when adding $\mathcal{L}_{MAA}$ alongside $\mathcal{L}_{DSA}$ (4th row), the maximum value of the attention map is boosted, indicating an enhanced focus on the target object, thus resulting in a clearer denoised image. This enhances the effectiveness of DSAlign and further improves text-image alignment.

# 4 EXPERIMENT

## 4.1 Dataset Construction

To study the effectiveness of our alignment-based objectives, we construct a compact yet robust dataset called **GOAL2k** to fine-tune our model. Initially, we first select common objects from COCO dataset [28]. Subsequently, leveraging in-context learning, we devise various templates encompassing different categories (*e.g.*, color, shape and spatial). Then we employ GPT-4 [2] to generate relational captions according to the templates and objects, resulting in a set of 1,000 template-based prompts. Additionally, we directly choose 1,000 captions from COCO dataset [28] that contain multiple objects or spatial relations as natural prompts. To guarantee the high quality and semantic accuracy of images incorporated in GOAL2K, we employ DALLE-3 [39] for generating images in the training set. Subsequently, we utilize GroundingDINO [30] to produce layout annotations. Ultimately, we manually verify the generated image-text pairs alongside the layout annotations to ensure their fidelity to accurate semantic information. **More details about dataset construction can be found in supplemental materials.**

## 4.2 Evaluation Metrics

Building upon existing works [16, 20, 35], we evaluate the precision of layout-to-image generation across various widely-adopted benchmarks, such as T2I-CompBench [20], NSR-1K [16] and Drawbench [42]. For the evaluation of text-image alignment, we adopt the recommended protocols in T2I-CompBench [20]. For NSR-1K [16], we utilize cross-modal similarities, indicated by CLIP score [37], and detection-based similarities, denoted by GLIP score [25]. As Drawbench [42] doesn't provide labels or metrics for automatic assessment, we conduct evaluator-based evaluation via user study. To evaluate the image generation quality, we calculate Q-Align [48] as image quality score and aesthetic score. Additionally, PickScore [22] is employed to evaluate human preferences regarding the quality of generated images. Furthermore, NSR-1K [16] is also utilized to evaluate the accuracy of the layout generated by different text-to-layout models [2, 17, 47] across spatial and counting scenarios.

## 4.3 Implementation Details

Our experiments are performed using GLIGEN [26], a popular layout-to-image diffusion model built upon Stable Diffusion v1.4 [40]. GPT-4 [2] is employed for layout planning during inference. We use $\mathcal{L}_{DSA}$ and $\mathcal{L}_{MSA}$ as alignment-based objectives alongside the original denoising objective $\mathcal{L}_{LDM}$ to train our model. The weights ($\alpha$ and $\beta$) for these objectives are set as 1e-2 and 1e-3, respectively. CLIP ViT-L/14 [37] is employed in discriminative semantic alignment (DSAlign) as the encoder for the image and text. We select the attention map with size $16\times16$ obtained from the decoder of the U-Net to perform masked attention alignment (MAAlign). We employ the AdamW optimizer with a constant learning rate of 5e-5 over 14,000 steps, with a batch size of 1. During training, we fine-tune the gated self-attention layers following GLIGEN [26]. Additionally, to fully exploit DSAlign and MAAlign, we fine-tune both self-attention and cross-attention layers.

## 4.4 Quantitative results

To evaluate the text-image alignment, we compare the proposed method with other state-of-the-art methods in two aspects: with text-to-image (T2I) methods such as Stable Diffusion v1.4 [40], DPT [35], Attn-Exct v1 [9], and layout-to-image methods such as GLIGEN [26], LayoutLLM-T2I [36], LayoutGPT [16] on T2I-Compbench [20]. Table 1 illustrates that our proposed method surpasses all other methods in all metrics. Particularly, GOAL achieves improvements of 19%, 13%, and 12% in color, shape, and texture metrics respectively, compared to GLIGEN. This verifies that discriminative tasks utilized by DSAlign could ensure region-wise semantic alignment between objects and corresponding attributes, and attention alignment performed by MAAlign allowing the model focusing on the target object, which is a prerequisite for attribute binding.

Additionally, we present the results of counting and spatial evaluation on the NSR-1K [16] benchmark in Table 2. The improvement, particularly in the GLIP [30] score (4% in counting and 3% in spatial metrics), further demonstrates the effectiveness of our method. From Table 1 and Table 2, we observe that layout-to-image methods outperform T2I methods in spatial and counting relationships by a large margin. This illustrates that employing layouts as an intermediate representation can significantly improve the controllable generation, particularly in scenarios involving counting and spatial relationships.

Moreover, we demonstrate the results of the proposed method compared with training-free methods such as Structure v2 [15], BoxDiff [49], Attend-and-Excite [9], and Attention Refocusing [34]

**Table 1: Comparison to the state-of-the-art text-to-image methods and layout-to-image methods on T2I-Compbench.**

| Method | Color | Shape | Texture | Spatial | Non-Spatial | Complex |
|---|---|---|---|---|---|---|
| *Text-to-Image Methods* | | | | | | |
| Stable v1.4 [40] | 37.65 | 35.76 | 41.56 | 12.46 | 30.79 | 28.18 |
| HN-DiffusionITM [23] | 36.71 | 35.48 | 39.84 | 11.22 | 30.91 | 28.05 |
| Composable v2 [29] | 40.63 | 32.99 | 36.45 | 8.01 | 29.80 | 28.98 |
| DPT [35] | 48.84 | 38.93 | 50.1 | 14.63 | 30.83 | 30.05 |
| DPT+SC [35] | 51.51 | 39.61 | 49.38 | 15.45 | 30.84 | 30.29 |
| *Layout-to-Image Methods* | | | | | | |
| GLIGEN [26] | 34.41 | 38.61 | 46.34 | 35.42 | 30.42 | 28.96 |
| LayoutGPT [16] | 33.86 | 36.35 | 44.07 | 35.06 | 30.31 | 26.36 |
| LayoutLLM-T2I [36] | 37.98 | 39.78 | 47.62 | 31.97 | 29.24 | 27.66 |
| Ours | **53.55** | **51.19** | **58.37** | **37.28** | **30.94** | **32.48** |

**Table 2: Comparison to the state-of-the-art text-to-image methods and layout-to-image methods on NSR-1K.**

| | Numerical Reasoning | | SpatialReasoning | |
|---|---|---|---|---|
| Method | Acc. (GLIP) | CLIP Sim. | Acc. (GLIP) | CLIP Sim. |
| *Text-to-Image Methods* | | | | |
| Stable v1.4 [40] | 32.22 | 0.256 | 16.89 | 0.252 |
| Stable v2.1 [40] | 42.44 | 0.256 | 17.81 | 0.256 |
| Attn-Exct v1 [9] | 38.96 | 0.258 | 24.38 | 0.263 |
| Attn-Exct v2 [9] | 45.74 | 0.254 | 26.86 | 0.264 |
| *Layout-to-Image Methods* | | | | |
| LayoutLLM-T2I [36] | 57.89 | 0.261 | 49.25 | 0.267 |
| LayoutGPT [16] | 55.64 | 0.261 | 60.64 | 0.268 |
| Attention Refocusing [34] | 57.26 | 0.244 | 61.23 | 0.251 |
| GLIGEN [26] | 56.02 | 0.258 | 60.03 | 0.265 |
| Ours | **60.25** | **0.263** | **63.12** | **0.271** |

**Table 3: Comparison to the state-of-the-art training-free methods on T2I-Compbench.**

| Component | Color | Shape | Texture | Spatial | Time (s) |
|---|---|---|---|---|---|
| Structured v2 [15] | 49.90 | 42.18 | 49.00 | 13.86 | 4.23 |
| Box Diff [49] | 50.26 | 45.11 | 53.18 | 33.01 | 4.11 |
| Attn-Exct v1 [9] | 53.31 | 38.51 | 56.13 | 10.06 | 5.13 |
| Attention Refocusing [34] | 45.38 | 43.04 | 50.62 | 36.01 | 5.29 |
| Ours | **53.55** | **51.19** | **58.37** | **37.28** | **2.03** |

on T2I-Compbench. Additionally, the inference time for generating an image is presented in Table 3. While these methods do not need additional training, they incur considerable computational costs during inference. For instance, Attention Refocusing [34] and Attend-Excite [9] take ×2.61 and ×2.52 times longer, respectively, to generate a single image compared to the proposed method. Our proposed method enables efficient inference and consistently outperforms baseline methods across all evaluation metrics.

Beyond the above evaluation, we also assessed the quality of images generated by different models using the PickScore [22] test-unique set. Results suggest that training-free methods, which directly intervene with cross-modal attention and latent noise maps during sampling, may degrade image quality. For instance, the PickScore of Attend-Excite [9] decreased by 3% compared to its baseline model Stable Diffusion v1.4 [40]. In contrast, our proposed method demonstrates exceptional performance for both text-image alignment and image quality.

**Table 4: Comparison to the state-of-the-art text-to-layout methods on PickScore test-unique set.**

| Method | PickScore | Quality | Aesthetic |
|---|---|---|---|
| Stable v1.4 [40] | 0.2481 | 3.79 | 3.08 |
| Attn-Exct v1 [9] | 0.2152 | 3.71 | 2.89 |
| Attention Refoucsing [25] | 0.1518 | 3.83 | 2.83 |
| GLIGEN [26] | 0.1686 | 3.96 | 3.05 |
| Ours | 0.2561 | 4.42 | 4.14 |

## 4.5 Qualitative results

Figure 5 demonstrates some cases from the T2I-Compbench [20] and NSR-1K [16] benchmarks across attributes such as color, spatial layout, shape, counting, and texture. We observe that 1) Layout conditions, serving as an intermediate representation, can enhance the generation controllability in scenes involving counting and spatial relationships. As shown in the last row, the T2I models Stable Diffusion [40] and Attend-and-Excite [9] fail to generate the correct number of animals, while other layout-to-image models achieve better control over the counting through guidance from the layout. 2) Our proposed method demonstrates outstanding performance across all attributes, generating images that faithfully capture semantic details. This can be attributed to the incorporation of DSAlign and MAAlign, which ensure region-wise semantic and attention alignment, respectively. 3) Compared with other methods, our proposed method can generate high-quality and aesthetic images, which verifies the improvement in text-image alignment does not result in a loss of image quality.

## 4.6 Ablation studies

**Effect of loss terms.** To explore the contribution of the loss terms $\mathcal{L}_{DSA}$ and $\mathcal{L}_{MAA}$, we conduct experiments in T2I-Compbench [20] including color, shape, texture, spatial metrics. From Table 5, we could see that both $\mathcal{L}_{DSA}$ and $\mathcal{L}_{MAA}$ could consistently promote alignment performance for T2I gengeration. Specifically, $\mathcal{L}_{DSA}$ shows considerable improvement in shape and texture metrics, while $\mathcal{L}_{MAA}$ exhibits significant enhancement in color metrics. It can be attributed to the fact that discriminative loss provided by semantic alignment improves shape and texture metrics, whereas attention-level alignment guides the model to focus on the target

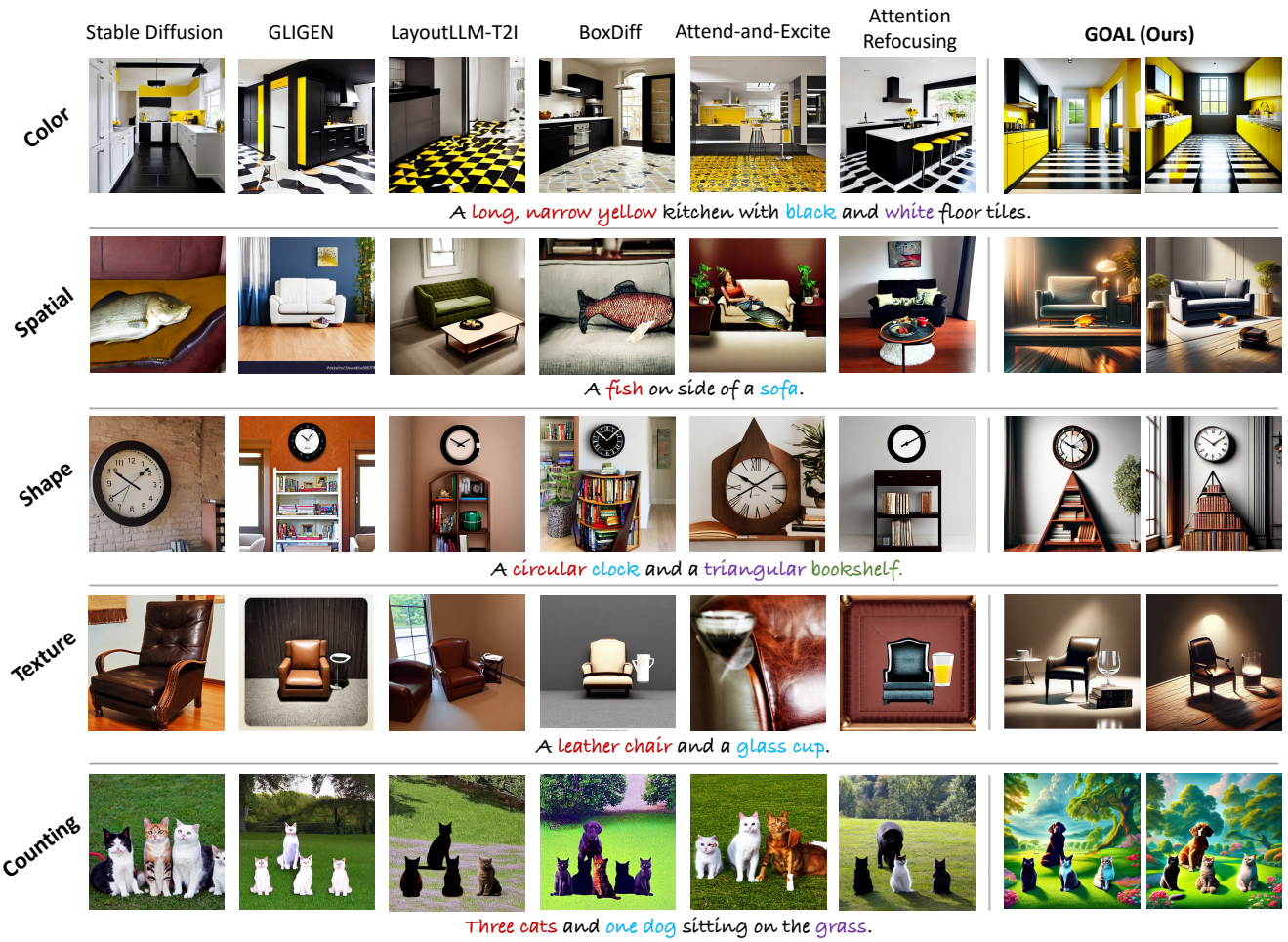


**Figure 5: Qualitative results from T2I-Compbench and NSR-1K for various attributes such as color, spatial, shape, counting and texture. We demonstrate the effectiveness of the proposed method in text-image alignment compared with Stable Diffusion v1.4 [40], GLIGEN [26], LayoutLLM-T2I [36], BoxDiff [49], Attend-and-Excite [9] and Attention Refocusing [34].**

**Table 5: Effect of loss terms.**

| Component | Color | Shape | Texture | Spatial |
|---|---|---|---|---|
| frozen | 34.41 | 38.61 | 46.34 | 35.42 |
| $\mathcal{L}_{LDM}$ | 44.17 | 44.27 | 53.86 | 36.42 |
| $\mathcal{L}_{LDM} + \mathcal{L}_{feature}$ | 49.83 | 46.89 | 55.21 | 36.77 |
| $\mathcal{L}_{LDM} + \mathcal{L}_{pixel}$ | 47.69 | 50.01 | 56.28 | 37.02 |
| $\mathcal{L}_{LDM} + \mathcal{L}_{pixel} + \mathcal{L}_{feature}$ | **53.55** | **51.19** | **58.37** | **37.28** |

**Table 6: Effect of different level of semantic alignment.**

| $D_{local}$ | $D_{global}$ | Color | Shape | Texture | Spatial |
|---|---|---|---|---|---|
| | | 49.83 | 46.89 | 55.21 | 36.77 |
| ✓ | | 52.08 | 51.12 | 57.03 | 37.10 |
| | ✓ | 50.02 | 48.49 | 56.51 | 36.99 |
| ✓ | ✓ | **53.55** | **51.19** | **58.37** | **37.28** |

object and binds accurate attributes such as color. By incorporating both $\mathcal{L}_{DSA}$ and $\mathcal{L}_{DSA}$, text-image alignment can be further enhanced.

**Effect of different level of semantic alignment.** DSAlign utilizes both local-level and global-level semantic alignment. To assess the effectiveness of semantic alignment at different levels, we conducted a detailed analysis using T2I-Compbench. Comparing the 2nd and 3rd rows of Table 6, it's evident that local-level semantic alignment plays a more significant role in enhancing text-image alignment. This can be attributed to the fact that local-level semantic alignment provides region-wise alignment, enabling fine-grained supervision for the denoising process. As a result, the model can learn region information in a more detailed manner. Furthermore, by incorporating both local-level and global-level semantic alignment, we consistently achieve improvements in text-image alignment, which verifies the effectiveness of global-level semantic alignment.

**Table 7: Effect of different cross-attention layers.**

| Layer | Resolution | Color | Shape | Texture | Spatial |
|---|---|---|---|---|---|
| middle block | $8 \times 8$ | 49.50 | 50.02 | 55.48 | 36.24 |
| decoder layer | $64 \times 64$ | 51.60 | 49.85 | 56.58 | 35.21 |
| decoder layer | $32 \times 32$ | 49.76 | 49.10 | 55.95 | 35.79 |
| decoder layer | $16 \times 16$ | **53.55** | **51.19** | **58.37** | **37.28** |

**Table 8: Comparison to the state-of-the-art text-to-layout methods on NSR-1K.**

| | Numerical Reasoning | | | SpatialReasoning |
|---|---|---|---|---|
| Method | Precision | Recall | Accuracy | Accuracy |
| LayoutTransformer [17] | 75.70 | 61.69 | 22.26 | 6.36 |
| llama2-7B [47] | 75.42 | 90.47 | 71.00 | 30.51 |
| llama2-13B [47] | 76.90 | 92.80 | 77.56 | 29.15 |
| GPT-3.5 [2] | 94.81 | 96.49 | 86.33 | 86.33 |
| GPT-4 [2] | 88.23 | 97.60 | 94.48 | 90.11 |

**Effect of different cross-attention layers.** We analyze the effectiveness of employing different cross-attention layers with MAAlign as shown in Table 7. It's clear that MAAlign is most effective when applied to the $16 \times 16$ attention map obtained from the U-Net decoder, which is consistent with findings from previous works [18, 25].

## 4.7 Text-to-Layout Model Evaluation

To evaluate the performance of text-to-layout models, we examine the transformer-based model LayoutTransformer [17], as well as the latest large language models (LLMs) including GPT-4 [2], GPT-3.5 [2], Llama 2-7B [47], and Llama 2-13B [47]. We assess their ability to comprehend visual concepts through spatial and counting scenes in the NSR-1K benchmark. From table 8, we know that the GPT-4 outperforms all other models in both spatial and counting scenes, thus it is employed for layout planning in this work. Moreover, we observe that LLMs exhibit significantly superior performance compared to LayoutTransformer, highlighting its robust cross-modal spatial reasoning abilities. Comparing Table 2 and Table 8, we observe that the text-to-layout process is proficient, and the bottleneck of image generation in complex scenes primarily lies in layout-guided image control.

## 4.8 User Study

While quantitative metrics have limitations in providing a comprehensive assessment, we complement our analysis with a user study. We selected attributes including color, description, counting, position, and conflicts from Drawbench [42], resulting in 94 prompts. For each prompt, two images were generated by different models, which were then assigned to 20 individuals for evaluation. Evaluation of the images by the participants focused on two aspects: semantic alignment and aesthetic quality. Semantic alignment is utilized to assess whether the model can generate images that faithfully capture the semantic details from the input text prompts. Aesthetic quality is employed to determine if the images generated by the model exhibit any incoherent parts or unnatural poses.

Our method is compared to several baseline models including Stable Diffusion v1.4 [40], GLIGEN [26], Attend-and-Excite [9], and

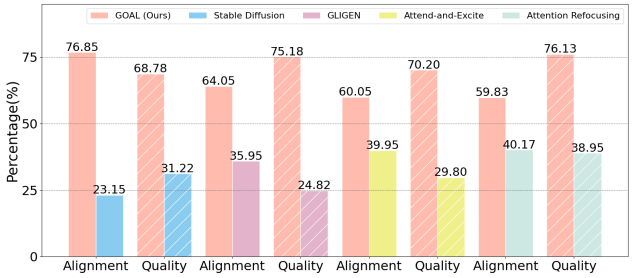


**Figure 6: User study on 94 prompts from Drawbench [42]. The ratios illustrate the participant preferences for the corresponding model. GOAL demonstrated superior performance in both image alignment and quality**


**Figure 7: Comparison with the training-free layout-aware method Attention Refocusing [34].**

Attention Refocusing [25]. Figure 6 shows that human preferences are consistent with our evaluation outcomes both in text-image alignment (*e.g.*, Table 1) and aesthetic quality (*e.g.*, Table 4), with 28.1% and 50.36% improvements compared to GLIGEN in alignment and quality respectively. Additionally, Figure 7 demonstrates that some images generated by Attention Refocusing [25] exhibit incoherent parts and unnatural poses, possibly attributed to the direct intervention of latent noise maps during sampling. In contrast, the proposed method enhances alignment without compromising image quality.

## 5 CONCLUSION

In this work, we propose Grounded jOint lAyout aLignment (GOAL) framework to handle Text-Image Misalignment issues in complex scenes. Discriminative semantic alignment (DSAlign) and masked attention alignment (MAAlign) are performed to provide explicit supervision with the layout conditions. In addition, We build a dataset GOAL2K to study the effectiveness of our alignment-based objectives, which composes 2000 semantically accurate image-text pairs and their layout annotations for model fine-tuning. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance on different benchmarks with improved image quality.

# REFERENCES


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4432–4441.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390* (2021).

[4] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20041–20053.

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).

[6] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126* (2021).

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

[9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

[10] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908* (2023).

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

[12] Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems* 28 (2015).

[13] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[14] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. 2023. Frido: Feature pyramid diffusion for complex scene image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 579–587.

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

[16] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[17] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1004–1014.

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

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[20] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 78723–78747.

[21] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7701–7711.

[22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[23] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. 2023. Are Diffusion Models Vision-And-Language Reasoners?. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[24] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479 (2022), 47–59.

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

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

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[29] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*. Springer, 423–439.

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.

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

[33] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.

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

[35] Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. 2024. Discriminative Probing and Tuning for Text-to-Image Generation. *arXiv preprint arXiv:2403.04321* (2024).

[36] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 643–654.

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

[38] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[43] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 4713–4726.

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

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

[48] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023).

[49] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7452–7461.

[50] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22428–22437.

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

[52] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14246–14255.

[53] Zhiyuan Zhang, Zhitong Huang, and Jing Liao. 2023. Continuous layout editing of single images with diffusion models. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, e14966.