# MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis

Dewei Zhou[1] You Li[1] Fan Ma[1] Xiaoting Zhang[2] Yi Yang[1†]
[1]ReLER, CCAI, Zhejiang University, Zhejiang, China [2]Huawei Technologies Ltd., China

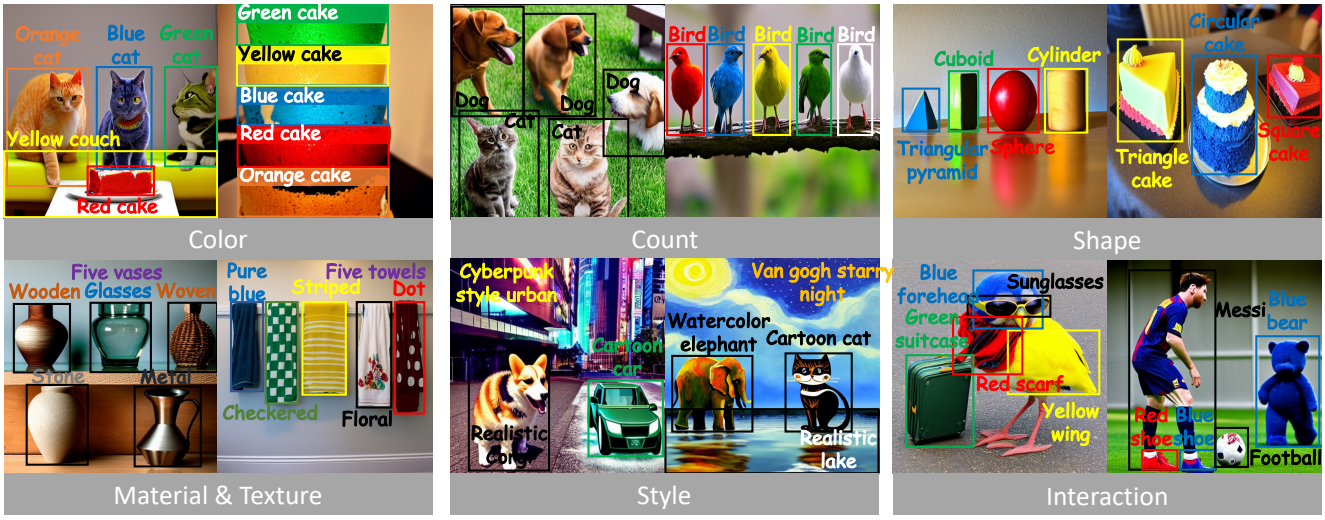


Figure 1. **Multi-Instance Generation (MIG) with our MIGC**. MIGC enables precise position control while ensuring the correctness of various attributes like color, shape, material, texture, and style in Multi-Instance Generation tasks. It can also control the number of instances and improve interaction between instances. We label the instance with a **corresponding color box**, while **the black box** represents that the instance **does not have a specified color**.

## Abstract

*We present a Multi-Instance Generation (MIG) task, simultaneously generating multiple instances with diverse controls in one image. Given a set of predefined coordinates and their corresponding descriptions, the task is to ensure that generated instances are accurately at the designated locations and that all instances' attributes adhere to their corresponding description. This broadens the scope of current research on Single-instance generation, elevating it to a more versatile and practical dimension. Inspired by the idea of divide and conquer, we introduce an innovative approach named Multi-Instance Generation Controller (MIGC) to address the challenges of the MIG task. Initially, we break down the MIG task into several subtasks, each involving the shading of a single instance. To ensure precise shading for each instance, we introduce an instance enhancement attention mechanism. Lastly, we aggregate all the shaded instances to provide the necessary information for accurately generating multiple instances in stable diffusion (SD). To evaluate how well generation models perform on the MIG task, we provide a COCO-MIG benchmark along with an evaluation pipeline. Extensive experiments were conducted on the proposed COCO-MIG benchmark, as well as on various commonly used benchmarks. The evaluation results illustrate the exceptional control capabilities of our model in terms of quantity, position, attribute, and interaction. Code and demos will be released at* https://migcproject.github.io/.

## 1. Introduction

Stable diffusion [40] has exhibited extraordinary capabilities in wild scenarios, including photography, painting, and others [11, 28, 37, 54, 57, 58]. The current research mainly focuses on Single-Instance Generation, where the generated content is only required to align with the single description, including image editing, personalized image generation, 3D generation [8, 12, 19–21, 23, 28, 41, 45, 46, 51], etc. However, more practical cases where multiple instances are simultaneously generated in one image with diverse controls

†Yi Yang is the corresponding author.

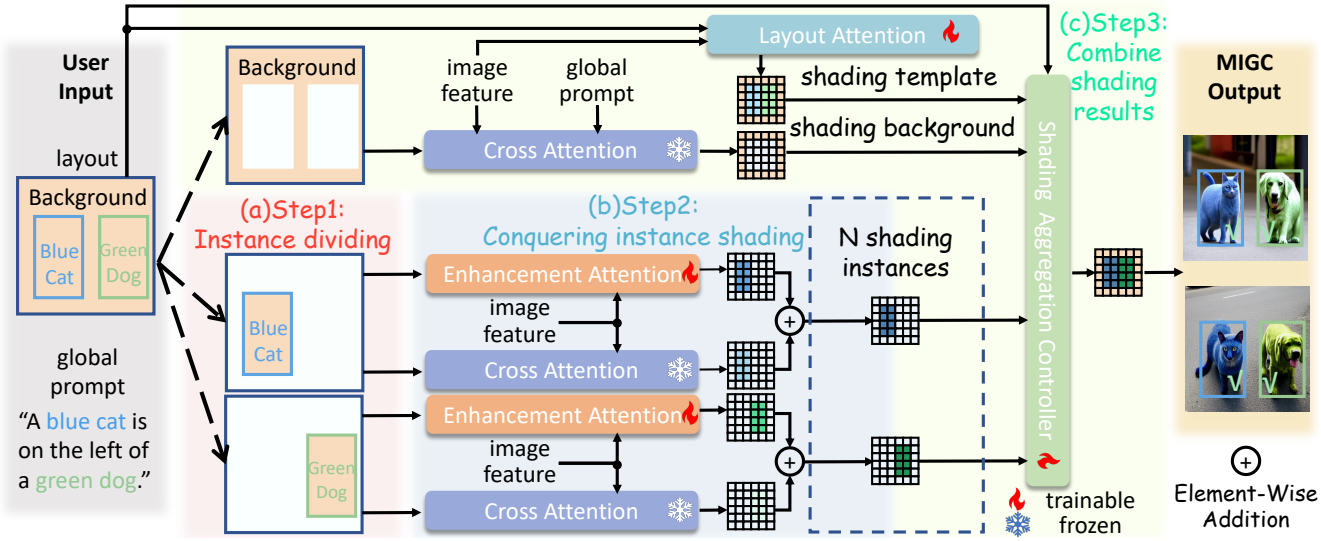


Figure 2. **Overview of our MIGC**. Stable diffusion's UNet inputs text description and image features into the Cross-Attention layer to obtain the residual feature and then adds it to the image features to determine generated content, which is like a shading process (i.e., coloring with parallel pencil lines or a block of color). In this view, MIG can be considered multi-instance shading on image features, and MIGC comprises three steps: (a) Divide MIG into single-instance shading subtasks. (b) Conquer single-instance shading with Enhancement Attention. (c) Combine shading results through Layout Attention and Shading Aggregation Controller.

have been rarely explored. This paper delves into a more general task, *i.e.*, Multi-Instance Generation (MIG), incorporating all factors such as quantity, position, attribute, and interaction control into one-time generation.

**Challenges in MIG.** MIG not only requires the instance to comply with the user-given description and layout but also ensures global alignment among all instances. Incorporating this information directly into the stable diffusion [40] often leads to failure. On the one hand, the current text encoder, like CLIP [38], struggles to differentiate each singular attribute from prompts containing multiple attributes [14]. On the other hand, Cross-Attention [47] layers in stable diffusion lack the ability to control position [6, 26, 32], resulting in difficulties when generating multiple instances within a specified region.

Motivated by the divide and conquer strategy, we propose the **M**ulti-**I**nstance **G**eneration **C**ontroller (MIGC) approach. This approach aims to decompose MIG into multiple subtasks and then combines the results of those subtasks. Although the direct application of stable diffusion in MIG is still a challenge, the outstanding capacity of stable diffusion in Single-Instance Generation could facilitate this task. Illustrated in Fig. 2, MIGC comprises three steps: 1) Divide: MIGC decomposes MIG into multiple instance-shading subtasks **only in the Cross-Attention layers of SD** to speed up the resolution of each subtask and make the generated images more harmonious. 2) Conquer: MIGC employs an Enhancement Attention Layer to enhance the shading results obtained through the frozen Cross-Attention, ensuring successful shading for each instance. 3) Combine: MIGC obtains the shading template through a Layout Attention layer and then inputs it, together with the shading background and shading instances, into a Shading Aggregation Controller to obtain the final shading result.

**Benchmark for MIG.** To evaluate how well generation models perform on the MIG task, we propose a COCO-MIG benchmark based on the COCO dataset [27], and this benchmark requires generation models to achieve strong control on position, attribute, and quantity **simultaneously**.

We conducted comprehensive experiments on the proposed COCO-MIG and the widely recognized COCO [27] and DrawBench [43] benchmarks. When applied to the COCO-MIG benchmark, our method substantially enhanced the Instance Success Rate, increasing it from **32.39**% to **58.43**%. Transitioning to the COCO benchmark, our approach exhibited noteworthy improvements in Average Precision (AP), elevating it from **40.68/68.26/42.85** to **54.69/84.17/61.71**. Similarly, on DrawBench, our method demonstrated advancements across position, attribute, and count, particularly elevating the attribute success rate from **48.20**% to **97.50**%. Moreover, MIGC maintains an inference speed close to the original stable diffusion.

Our contributions are summarized as follows:

1) To advance the development of vision generation, we present the MIG task to address prevailing challenges in both academic and industrial domains. Meanwhile, we propose the COCO-MIG benchmark to evaluate the inherent MIG capabilities of generative models.

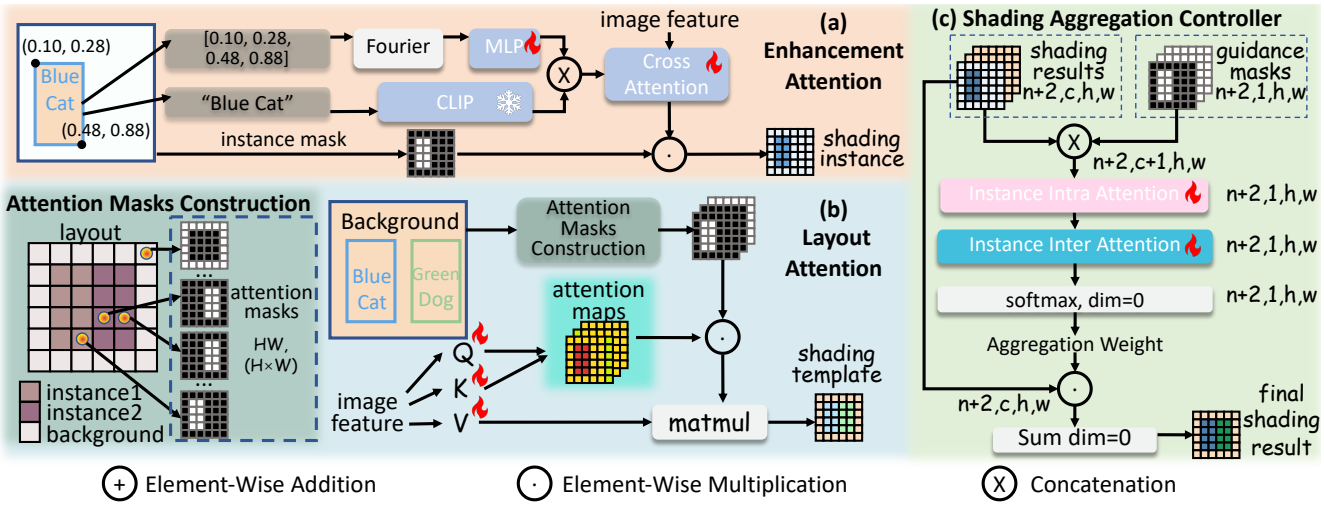


Figure 3. **Three main modules in MIGC**. (a) Architecture of Enhancement Attention Layer. (b) Architecture of Layout Attention Layer. (c) Architecture of Shading Aggregation Controller.

2) Inspired by the principle of divide and conquer, we introduce a novel MIGC approach that enhances pre-trained stable diffusion with improved MIG capabilities.
3) We conducted extensive experiments on three benchmarks, indicating that our MIGC significantly surpassed the previous SOTA method while ensuring the inference speed was close to the original stable diffusion.

## 2. Related work

### 2.1. Text-to-Image Generation

Text-to-image (T2I) Generation aims to generate high-quality images based on text descriptions. Conditional GANs [39, 50, 56] were initially used for T2I Generation, while diffusion models [2, 7, 13, 17, 30, 31, 34, 40, 43, 59, 60, 62] and autoregressive models [4, 10, 55] gradually replaced GANs as the foundational generator due to their more stable training and higher image quality.

### 2.2. Layout-to-Image Generation

As text cannot precisely control the position of generated instances. Some Layout-to-Image methods [6, 32, 34, 49, 61] extend the pre-trained T2I model [40] to integrate layout information into the generation and achieve control of instances' position. However, they struggle to isolate the attributes of multiple instances, thus generating images with mixed attributes. This paper proposes a novel MIGC approach to achieve precise position-and-attribute control.

## 3. Method

### 3.1. Preliminaries

**Stable diffusion** [40] is one of the most popular T2I models, and it uses the CLIP [38] text encoder to project texts into sequence embedding and integrate textual conditions into the generation process via Cross-Attention [47] layers.

**Attention layers**. Attention mechanisms [47, 53] play key roles in the interaction of multi-modal features. Omitting the reshape operation, our attention layers are expressed as:

$$\mathbf{R} = \text{Softmax}(\frac{\mathbf{QK}^T}{\sqrt{d}})\mathbf{V}, \mathbf{R} \in \mathbb{R}^{(H,W,C)} \quad (1)$$

where $\mathbf{R}$ represents the output residual, and $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ separately represents the Query, Key, and Value in attention layers, which are projected by linear layers.

### 3.2. Overview

**Problem Definition.** In the Multi-Instance Generation (MIG), users will give generation models the global prompt $\mathcal{P}$, instance layout bounding boxes $\mathbb{B} = \{\mathbf{b}^1, ..., \mathbf{b}^N\}$, where $\mathbf{b}^i = [x_1^i, y_1^i, x_2^i, y_2^i]$, and corresponding descriptions $\mathbb{D} = \{\mathbf{d}^1, ..., \mathbf{d}^N\}$. According to user-provided inputs, the model needs to generate an image $\mathcal{I}$, in which the instance within the box $\mathbf{b}^i$ should adhere to the instance description $\mathbf{d}^i$, and global alignment is ensured in all instances.

**Difficulties in MIG.** When dealing with Multi-Instance prompts, stable diffusion struggles with attribute leakage, i.e., 1) *Textual Leakage*. Due to the causal attention masks used in the CLIP encoder, the latter instance tokens may exhibit semantic confusion [14]. 2) *Spatial Leakage*. The Cross-Attention lacks precise position control [6], and instances will affect the generation of each others' region.

**Motivation.** Divide and conquer is an ancient but wise idea. It first *divides* a complex task into several simpler subtasks, then *conquers* these subtasks respectively, and finally obtains the solution to the original task by *combining* the solutions of the subtasks. This idea is highly applicable to MIG. For example, MIG is a complex task for most T2I models,

while Single-Instace Generation is a simpler subtask that T2I models can solve well [8, 33, 41, 46, 52]. Based on this idea, we proposed our MIGC, which extends the stable diffusion with stronger MIG ability. We will introduce the technology details by telling "how to divide," "how to conquer," and "how to combine."

### 3.3. Divide MIG into Instance Shading Subtasks

**Instance shading subtasks in Cross-Attention space.** Cross-Attention is the only way for text and image features to interact in stable diffusion, and the output determines the generated content, which looks like a shading operation on image features. In this view, the MIG task can be defined as performing correct Multi-Instance shading on image features, and $subtask_i$ can be defined as finding a single instance shading result $\mathbf{R}^i$ to satisfy the following:

$$\mathbf{R}^i = \arg\min_{\mathbf{R}^i}(\left\|\mathbf{R}^i - \mathbf{R}^{correct}\right\|_2 \cdot \mathbf{M}^i), \tag{2}$$

where $\mathbf{R}^{correct}$ represents an objectively existing correct feature, and $\mathbf{M}^i$ is an instance mask generated according to the box $\mathbf{b}^i$, with the values inside the box region are set to 1, and the rest of the positions are set to 0. That is to say, each shading instance should have the correct textual semantic in its corresponding area.

**Two benefits of division in the Cross-Attention space.** i.e., 1) Conquer more efficiently: For N-instance generation, MIGC conquers N subtasks solely on Cross-Attention layers instead of the entire Unet network, which will be more efficient; 2) Combine more harmoniously: Combining subtasks in the middle layer enhances the overall cohesiveness of the generated image compared to combining at the final output of the network.

### 3.4. Conquer Instance Shading

**Shading stage 1: shading results of Cross-Attention.** The pre-trained Cross-Attention will notice regions with high attention weight and perform shading according to the textual semantics [9, 48]. As shown in Fig. 2, MIGC uses the masked Cross-Attention output as the first shading results:

$$\mathbf{R}^i_f = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{i^T}}{\sqrt{d}})\mathbf{V}^i \cdot \mathbf{M}^i, \tag{3}$$

where $\mathbf{K}^i$ and $\mathbf{V}^i$ are obtained from text embedding of $\mathbf{d}^i$, and $\mathbf{Q}$ is obtained from the image feature map.

**Two issues of Cross-Attention shading results.** 1) *Instance Merge.* According to Eq. (3), for two instances with the same description, they will get the same $\mathbf{K}$ and $\mathbf{V}$ in the Cross-Attention layer. If their boxes are close or even overlap, the network will easily merge the two instances; 2) *Instance Missing.* The initial edit method [32] shows that the initial noise of SD largely determines the layout of the generated image, i.e., specific regions prefer to generate specific instances or nothing. If the initial noise does not tend to generate an instance according to description $\mathbf{d}^i$ in box $\mathbf{b}^i$, the $\mathbf{R}^i_f$ will be weak, leading to the instance missing.

**Grounded phrase token for solving instance merge.** To identify instances with the same description but different boxes, MIGC extends the text tokens of each instance to a combination of text and position tokens. As shown in Fig. 3(a), MIGC first projects the bounding box information to the Fourier embedding, then uses a MLP layer to get position tokens. MIGC concatenates the text tokens with position tokens to obtain the grounded phrase tokens:

$$\mathbf{G}^i = [\text{CLIP}(\mathbf{d}^i), \text{MLP}(\text{Fourier}(\mathbf{b}^i))], \tag{4}$$

where $[\cdot]$ represents the concatenation.

**Shading stage 2: Enhancement Attention for solving instance missing.** Illustrated in Fig. 2, MIGC uses a trainable Enhancement-Attention (EA) Layer to enhance the shading result. Specifically, as shown in Fig. 3(a), after obtaining the grounded phrase token, EA uses a new trainable Cross-Attention layer to obtain an enhanced shading result and adds it to the first shading result $\mathbf{R}^i_f$:

$$\mathbf{R}^i_s = \mathbf{R}^i_f + \text{softmax}(\frac{\mathbf{Q}_{ea}\mathbf{K}^{i\ T}_{ea}}{\sqrt{d}})\mathbf{V}^i_{ea} \cdot \mathbf{M}^i, \tag{5}$$

where $\mathbf{K}^i_{ea}$ and $\mathbf{V}^i_{ea}$ are obtained from the grounded phrase token $\mathbf{G}^i$, and $\mathbf{Q}_{ea}$ is obtained from the image feature map. During the training period, since $\mathbf{M}^i$ ensures precise spatial positioning, the instance shading result output by EA exclusively impacts the correct region, so the EA can easily learn: no matter what the image feature is, the EA should perform enhanced shading to satisfy the textual semantic of $\mathbf{d}^i$ and solve the issue of instance missing. *Finally, MIGC treats the enhanced result $\mathbf{R}^i_s$ as the solution of the $subtask_i$.*

### 3.5. Combine Shading Results

**Global prompt residual as shading background.** Obtaining N-instance shading results as shading foreground, the next step of MIGC is to get the shading background. Illustrated in Fig. 2(c), MIGC utilizes global prompt $\mathcal{P}$ to obtain the shading background result $\mathbf{R}^{bg}$ in a manner similar to Eq.(3), with the background mask $\mathbf{M}^{bg}$, in which positions containing the instance are assigned a value of 0, while all other positions are marked as 1.

**Layout Attention residuals as shading template.** A certain gap exists between shading instances $\{\mathbf{R}^1_s, \ldots, \mathbf{R}^N_s\}$ and the shading background $\mathbf{R}^{bg}$, as their shading process is independent. To bridge these shading results and minimize the gap, MIGC needs to learn a shading template according to the image feature maps' information. As shown in Fig. 2(c), a Layout Attention layer is used in MIGC to achieve the above goal. Illustrated in Fig. 3(b), Layout Attention performs similarly to the Self-Attention [42, 44] while instance masks $\mathbb{M}_{inst} = \{\mathbf{M}^{bg}, \mathbf{M}^1, \ldots, \mathbf{M}^N\}$ are used to construct attention masks:

| Method | Instance Success Rate(%)↑ | | | | | | mIoU↑ | | | | | | Time(s)↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Level | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | Avg | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | Avg | |
| Stable Diffusion | 6.87 | 5.01 | 3.45 | 3.27 | 2.21 | 3.61 | 18.92 | 17.44 | 15.85 | 15.17 | 14.42 | 15.80 | **9.18** |
| TFLCG | 20.47 | 12.71 | 8.36 | 6.72 | 4.36 | 8.62 | 29.34 | 25.06 | 20.82 | 18.81 | 17.86 | 20.92 | 19.92 |
| BOX-Diffusion | 24.61 | 19.22 | 14.20 | 11.92 | 9.31 | 13.96 | 32.64 | 29.88 | 25.39 | 23.81 | 21.19 | 25.14 | 44.17 |
| Multi Diffusion | 24.88 | 22.14 | 19.88 | 18.97 | 18.60 | 20.12 | 29.41 | 28.06 | 25.59 | 24.83 | 24.71 | 25.89 | 25.15 |
| GLIGEN | 42.30 | 35.55 | 32.66 | 28.18 | 30.84 | 32.39 | 37.58 | 32.34 | 29.95 | 26.60 | 27.70 | 32.25 | 22.00 |
| Ours | **67.70** | **59.61** | **58.09** | **56.16** | **56.88** | **58.43** | **59.39** | **52.73** | **51.45** | **49.52** | **49.89** | **51.48** | 15.61 |

Table 1. Quantitative results in our proposed COCO-MIG benchmark. According to the count of generated instances, COCO-MIG is divided into five levels: $L_2$, $L_3$, $L_4$, $L_5$, and $L_6$. $L_i$ means that the count of instances needed to generate in the image is i.

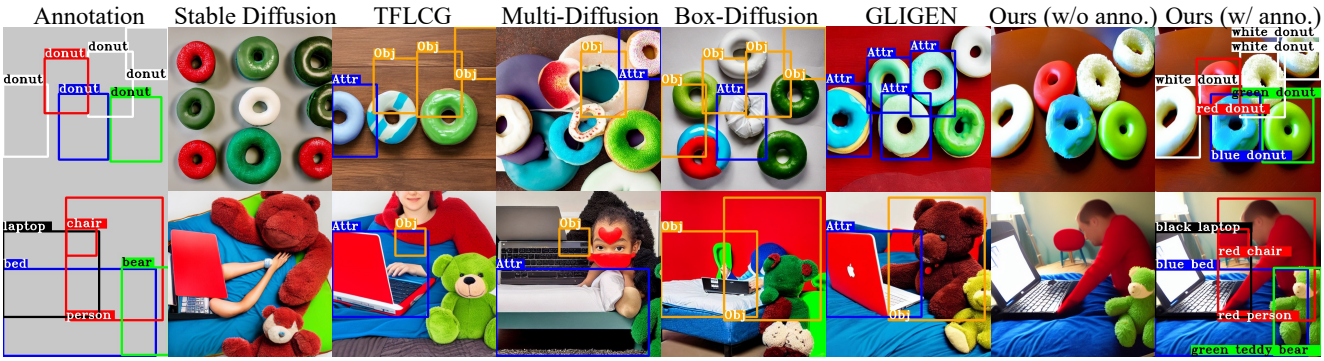


Figure 4. Qualitative comparison of our MIGC and other baselines on COCO-MIG. We use a yellow bounding box labeled "Obj" to indicate a position-wrongly generated instance and a blue bounding box labeled "Attr" to indicate an attribute-wrongly generated instance. Experimental results show that MIGC can achieve better attribute (i.e., color) control while precisely controlling the positions of instances.

$$\mathbf{A}_{(a,b),(c,d)} = \begin{cases} 1, \text{if} \exists\, \mathbf{m} \in \mathbb{M}_{inst}, \mathbf{m}_{a,b} = \mathbf{m}_{c,d} = 1 \\ -inf, \quad \text{otherwise} \end{cases} \tag{6}$$

$$\mathbf{R}_{LA} = \text{Softmax}(\frac{\mathbf{Q}_{LA}{\mathbf{K}_{LA}}^T}{\sqrt{d}} \odot \mathbf{A})\mathbf{V}_{LA}, \tag{7}$$

where $\odot$ represents the Hadamard product, and $\mathbf{A} \in \mathbb{R}^{((H,W),(H,W))}$ represents attention masks, in which $\mathbf{A}_{(a,b)(c,d)}$ determines whether pixel (a, b) should attend to pixel (c, d). The constructed attention mask $\mathbf{A}$ ensures one pixel can only attend to other pixels in the same instance region, which avoids attribute leakage between instances.

**Shading Aggregation Controller for the final fusion.** To summarize, in all the above operations, MIGC can get $\mathbb{R}_s = \{\mathbf{R}_s^1, \ldots, \mathbf{R}_s^N, \mathbf{R}^{bg}, \mathbf{R}_{LA}\} \in \mathbb{R}^{(N+2,C,H,W)}$ and $\mathbb{M} = \{\mathbf{M}^1, \ldots, \mathbf{M}^N, \mathbf{M}^{bg}, \mathbf{M}_{LA}\} \in \mathbb{R}^{(N+2,1,H,W)}$, where $\mathbf{M}_{LA}$ is the all-1 guidance mask corresponding to $\mathbf{R}_{LA}$. In order to dynamically aggregate shading results at different timesteps of the generation process, we propose the Shading Aggregation Controller (SAC). As shown in Fig.3(c), SAC sequentially performs instance intra-attention and inter-attention, and aggregation weights summing to 1 are assigned to shading results on each spacial pixel through the softmax function, resulting in the final shading.

$$\mathbf{R}_{final} = SAC(\mathbb{R}_s, \mathbb{M}), \mathbf{R}_{final} \in \mathbb{R}^{H,W,C} \tag{8}$$

### 3.6. Summary

**Training Loss.** We use the original denoising loss [18, 40]:

$$\min_{\theta'} \mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}[||\epsilon - f_{\theta,\theta'}(\mathbf{z}_t, t, \mathcal{P}, \mathbb{B}, \mathbb{D})||_2^2], \tag{9}$$

where $\theta$ represents the frozen parameters of the pre-trained stable diffusion, and $\theta'$ means the parameter of our MIGC.

Besides, to constrain generated instances within their regions and prevent the generation of additional objects in the background, we design an inhibition loss to avoid high attention weight in the background region:

$$\min_{\theta'} \mathcal{L}_{\text{ihbt}} = \sum_{i=1}^{i=N} \left|\mathbf{A}_c^i - \text{DNR}(\mathbf{A}_c^i)\right| \odot \mathbf{M}^{bg}, \tag{10}$$

where $\mathbf{A}_c^i$ denotes the attention maps for the ith instance in the frozen $16 \times 16$ Cross-Attention layer of the Unet decoder [32], and DNR$(\cdot)$ means the denoising (e.g., we use the average operation) of the background region. The final training loss is designed as follows:

$$\min_{\theta'} \mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda \mathcal{L}_{\text{ihbt}}, \tag{11}$$

we set the loss weight $\lambda$ as 0.1.

| Method | Spatial Accuracy(%) | | | | | Image Text Consistency | | Image Quality |
|---|---|---|---|---|---|---|---|---|
| | Success Ratio↑ | mIoU↑ | AP↑ | AP50↑ | AP75↑ | CLIP↑ | Local CLIP↑ | FID-6K↓ |
| Real Image | 83.75 | 85.49 | 65.97 | 79.11 | 71.22 | 24.22 | 19.74 | - |
| Stable Diffusion | 5.95 | 21.60 | 0.8 | 2.71 | 0.42 | **25.69** | 17.34 | **23.56** |
| TFLCG | 13.54 | 28.01 | 1.75 | 6.77 | 0.56 | 25.07 | 17.97 | 24.65 |
| BOX-Diffusion | 17.84 | 33.38 | 3.29 | 12.27 | 1.08 | 23.79 | 18.70 | 25.15 |
| Multi Diffusion | 23.86 | 38,82 | 6.72 | 18.65 | 3.63 | 22.10 | 19.13 | 33.20 |
| Layout Diffusion | 50.53 | 57.49 | 23.45 | 48.10 | 20.70 | 18.28 | 19.08 | 25.94 |
| GLIGEN | 70.52 | 71.61 | 40.68 | 68.26 | 42.85 | 24.61 | 19.69 | 26.80 |
| Ours | **80.29** | **77.38** | **54.69** | **84.17** | **61.71** | 24.66 | **20.25** | 24.52 |

Table 2. Quantitative results on the COCO-Position.

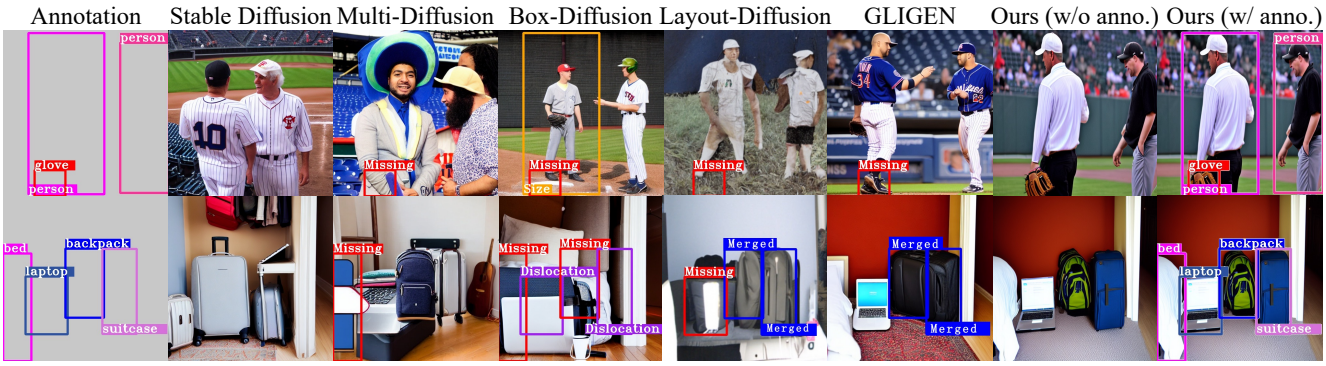


Figure 5. Qualitative comparison of MIGC and other baselines on COCO-Position. Experimental results show that our method can reduce the problem of instance missing, improve positional control, and alleviate the phenomenon of instance merging.

**Implementation Details.** We only deploy MIGC on the mid-layers (i.e., $8 \times 8$) and the lowest-resolution decoder layers (i.e., $16 \times 16$) of UNet, which greatly determine the generated image's layout and semantic information [6, 33]. In other Cross-Attention layers, we use the global prompt for global shading. We use COCO 2014 [27] to train MIGC. To get the instance descriptions and their bounding boxes, we use stanza [36] to split the global prompt and detect the instances with the Grounding-DINO[29] model. We train our MIGC based on the pre-trained stable diffusion v1.4. We use AdamW [24] optimizer with a constant learning rate of $1e^{-4}$, and train the model for 300 epochs with batch size 320, which requires 15 hours on 40 V100 GPUs with 16GB VRAM each. For inference, we use EulerDiscreteScheduler [22] with 50 sample steps and use our MIGC in the first 25 steps. We select the CFG scale[17] as 7.5. For more details, please refer to supplementary materials.

# 4. Experiments

## 4.1. Benchmarks

We evaluate models' performance on three benchmarks: COCO-MIG, COCO-Position [27], and DrawBench [43]. We use 8 seeds to generate images for each prompt.

**In COCO-MIG**, we pay attention to position, color, and quantity. To construct it, we randomly sampled 800 COCO images and assigned a color to each instance while keeping the original layout. Furthermore, We reconstruct the global prompts in the format of ' a <attr1> <obj1> and a <attr2> <obj2> and a ... ', and we divide this benchmark into five levels based on the number of instances in the generated image. Each method will generate 6400 images.

**In COCO-Position**, we sampled 800 images, using the captions as the global prompts, labels as instance descriptions, and bounding boxes as layouts to generate 6400 images.

**Drawbench** is a challenging T2I benchmark. We use GPT4 [15, 35] to extract all instance descriptions and generate the layouts for each prompt. We use a total of 64 prompts, of which 25 are related to color, 19 are related to counting, and 20 are related to position, ultimately generating 512 images.

## 4.2. Evaluation Metrics

**Position Evaluation.** We use Grounding-DINO [29] to detect each instance and calculate the maximum IoU between the detection boxes and the Ground Truth box. If the above IoU is higher than the threshold $t$=0.5, we mark it as *Position Correctly Generated*.

**Attribute Evaluation.** For a *Position Correctly Generated*

| Method | Spatial(%) ↑ | | Attribute(%) ↑ | | Count(%) ↑ | |
|---|---|---|---|---|---|---|
| | R | Human | R | Human | R | Human |
| SD1.4 | - | 13.30 | - | 57.52 | - | 23.70 |
| AAE | - | 23.13 | - | 51.50 | - | 30.92 |
| Struc-D | - | 13.12 | - | 56.5 | - | 30.26 |
| Box-D | 11.88 | 50.00 | 28.50 | 57.50 | 9.21 | 39.47 |
| TFLCG | 9.38 | 53.13 | 35.00 | 60.00 | 15.79 | 31.58 |
| Multi-D | 10.63 | 55.63 | 18.5 | 65.50 | 17.76 | 36.18 |
| GLIGEN | 61.25 | 78.80 | 51.00 | 48.20 | 44.08 | 55.90 |
| Ours | **69.38** | **93.13** | **79.00** | **97.50** | **67.76** | **67.50** |

Table 3. Evaluation on the drawbench.

instance, we use the Grounded-SAM model [25, 29] to segment it and calculate the percentage of the target color in the HSV color space. If the above percentage exceeds the threshold $S$=0.2, we denote it as *Fully Correctly Generated*.

**Metrics on COCO-MIG.** We primarily measure the *Instance Success Rate* and mIoU. The *Instance Success Rate* calculates the probability that **each instance** is *Fully Correctly Generated*, and mIoU calculates the mean of the maximum IoU for all instances. Note that if the color attribute is incorrect, we set **the IoU value as 0**.

**Metrics on COCO-Position.** We use *Success Rate*, mIoU and Grounding-DINO AP score to measure the Spatial Accuracy. The *Success Rate* represents whether **all instances** in one image are *Position Correctly Generated*. Besides, we use the Fréchet Inception Distance (FID) [16] to evaluate Image Quality. To measure Image-Text Consistency, we use CLIP score and Local CLIP score[1].

**Metrics on DrawBench.** We evaluate the *Success Rate* for images related to the position and count by checking whether all instances in each image are *Position Correctly Generated*. For color-related images, we check whether all instances are *Fully Correctly Generated*. In addition to automated evaluations, a manual evaluation is conducted.

### 4.3. Baselines

We compare our method with some SOTA layout-to-image methods: Multi-Diffusion[3], Layout Diffusion[61], GLIGEN[26], TFLCG[6], and Box-Diffusion[49]. Since Layout Diffusion cannot control color, we only run it on COCO-Position. In Drawbench, we also compare our method with some SOTA T2I methods: stable diffusion v1.4[40], AAE[5], Structure Diffusion[14]. All methods are executed using the official code and default configuration.

### 4.4. Quantitative Results

**COCO-MIG.** Tab.1 shows results in COCO-MIG. MIGC improves the *Instance Success Rate* from 32.39% to 58.43% and mIoU from 32.25 to 51.48. Improvements are consistently observed across all count-division levels, underscoring the robust control capabilities of MIGC on position, quantity, and attributes. Furthermore, MIGC runs at almost the same speed as the original stable diffusion, thanks to MIGC dividing MIG in the Cross-Attention Space, accelerating the conquering and combing of subtasks.

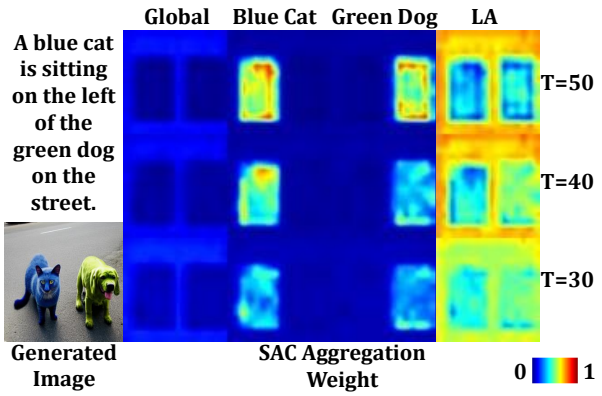


Figure 6. SAC aggregation weight at T=50, 40, and 30. T=50 means the first step, as we generated each image with 50 steps.

**COCO-Position.** Tab.2 shows quantitative results in COCO-Position, indicating that MIGC brings significant improvement in Spatial Accuracy: increased the *Success Rate* from 70.52% to 80.29%, mIoU from 71.61 to 77.38, and AP score from 40.68/68.26/42.85 to 54.69/84.17/61.71. MIGC also achieves similar FID scores compared to the stable diffusion, highlighting that MIGC can enhance position control capabilities without destroying image quality.

**DrawBench.** Tab.3 shows the results in drawbench. MIGC achieves the best performance in both mechanical metrics and human evaluation. Human evaluation doesn't rely on IoU to determine position correctness.

| SAC | EA | LA | R(%)↑ | mIoU↑ | AP↑ | AP50↑ | AP75↑ | numb. |
|---|---|---|---|---|---|---|---|---|
| | | | 7.66 | 22.71 | 0.91 | 3.18 | 0.35 | ① |
| ✔ | | | 12.10 | 29.55 | 1.89 | 7.64 | 0.49 | ② |
| ✔ | | ✔ | 34.70 | 44.08 | 11.02 | 28.64 | 6.83 | ③ |
| ✔ | ✔ | | 80.16 | 76.63 | 53.03 | 84.05 | 58.67 | ④ |
| | ✔ | ✔ | 78.12 | 75.47 | 52.05 | 83.48 | 57.16 | ⑤ |
| ✔ | ✔ | ✔ | **80.29** | **77.38** | **54.69** | **84.17** | **61.71** | ⑥ |

Table 4. Ablation on COCO-Position of **Shading Aggregation Controller(SAC), Enhancement Attention (EA), Layout Attention (LA)**.

### 4.5. Qualitative Results

Fig.4 shows qualitative results in COCO-MIG. MIGC demonstrates effective position-and-attributes control over all instances, even in complex scenarios. Fig.5 shows qualitative results in COCO-Position. MIGC achieves more precise control, ensuring all instances are generated strictly within their designated boxes without instances missing or

| Config | R(%)↑ | mIoU↑ | AP↑ | AP50↑ | AP75↑ | FID↓ |
|---|---|---|---|---|---|---|
| w/o loss | 80.20 | 77.03 | 52.46 | 82.65 | 58.05 | 24.73 |
| w/ loss 1.0 | **80.61** | **77.79** | **55.62** | **84.48** | **62.85** | 26.94 |
| w/ loss 0.1 | 80.29 | 77.38 | 54.69 | 84.17 | 61.71 | **24.52** |

Table 5. Ablation on COCO-Position of **Inhibition loss**. We conducted an ablation study on three configurations: w/o loss, loss weight 1.0, and loss weight 0.1.

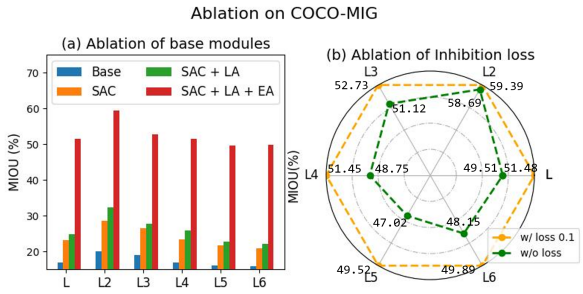


Figure 7. The results of ablation studies on COCO-MIG. (a) shows the ablation results of three components, and (b) shows the ablation results of inhibition loss.

merging. The qualitative results for DrawBench will be presented in the supplementary materials.

### 4.6. Analysis of Shading Aggregation Controller

We generate each image with 50 steps while using MIGC in the first 25 steps. Fig.6 shows SAC aggregation weights at T=50, 40, and 30 (i.e., T=50 means the first step). In the early time steps, the SAC assigns more weight to the EA layer's shading instances in the foreground while giving more weight to the LA layer's shading template in the background. In the later time steps, the SAC gradually increases the attention to the global context in the background.

### 4.7. Ablation Study

The ablation focuses on four components: (1) Enhancement Attention Layer. (2) Layout Attention Layer. (3) Shading Aggregation Controller. (4) The inhibition loss. Experiments are performed on COCO-Position and COCO-MIG.

**Shading Aggregation Controller.** From Tab.4, we find that using SAC improves the performance metrics(compare ⑤ with ⑥ and ① with ②), which is also reflected in the ablation experiments on COCO-MIG in Fig.7(a).

**Enhancement Attention Layer.** In Tab.4, the EA Layer significantly improves the *Success Rate* from 12.10% to 80.16%, mIoU from 29.55 to 76.63, and AP from 1.89 / 7.64 / 0.49 to 53.03 / 84.05 / 58.67 (Compare ② with ④). We also observe significant improvement in Fig.7(a).

**Layout Attention Layer.** The results of ④ and ⑥ in Tab.4 show that LA Layer can improve the AP. We find that SAC+LA, compared to SAC alone, has improved mIoU to some extent in Fig.7(a).

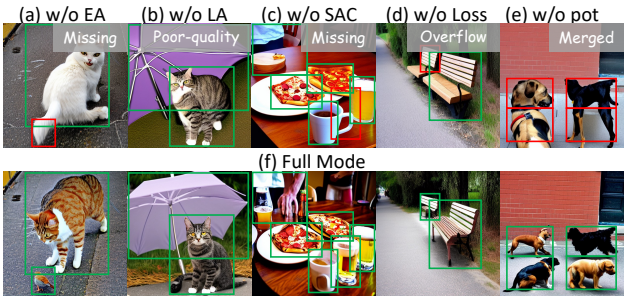


Figure 8. The qualitative results of ablation studies. Pot means the Position Token used in EA Layer. We mark incorrectly generated instances with red boxes and correctly with green boxes.

**Inhibition Loss.** We also conducted the ablation study on the Inhibition loss, with 0.1 and 1.0 loss weight. We show the results in Tab.5 and Fig.7(b). Tab.5 indicates that inhibition loss can significantly improve the AP metric in COCO-Position. We find that setting the loss function weight to 1.0 can further improve the AP metric, but it comes at the cost of a slight decrease in image quality (i.e., FID). So we finally choose loss weight as 0.1. Fig.7 (b) shows the comparison between *w/ loss 0.1* and *w/o loss* on COCO-MIG, and we observe that the inhibition loss can improve the mIoU, especially when generating images with large instance quantity.

**Qualitative Results.** We show qualitative results in Fig.8. The first column indicates that the EA Layer can effectively alleviate instance missing. The second column illustrates that the LA Layer can significantly improve generated image quality. The third column suggests that the SAC also aids in better aggregation of shading instances. The fourth column demonstrates that inhibition loss enhances the model's control capabilities. The fifth column demonstrates that position tokens effectively alleviate instance merging.

## 5. Conclusion

In this work, we define a practical and challenging MIG task and propose a MIGC approach to improve the stable diffusion's MIG ability. We divide the complex MIG task into simpler Single-Instance shading subtasks, conquer each instance shading with an Enhancement Attention layer, and combine the final shading result through a Layout Attention layer and Shading Aggregation Controller. Comprehensive experiments are conducted on our proposed COCO-MIG and popular COCO-Position and Drawbench benchmarks. Experiment results verify the efficiency and effectiveness of our MIGC. In the future, we will further explore the control of interactive relationships between instances.

**Acknowledgements.** This work was supported by the National Natural Science Foundation of China (62293554, U2336212).

## References


[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. SpaText: Spatio-textual representation for controllable image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 7

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

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 7

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. 2023. 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 7

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

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator, 2022. 3

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 1, 4

[9] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023. 4

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021. 3

[11] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 1

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation, 2023. 1

[13] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents, 2022. 3

[14] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 7

[15] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. 6

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 7

[17] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 3, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[19] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5734–5745, 2023. 1

[20] Wenjing Huang, Shikui Tu, and Lei Xu. Pfb-diff: Progressive feature blending diffusion for text-driven image editing, 2023.

[21] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18423–18433, 2023. 1

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 6

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023. 1

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 7

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

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 6

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 1

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6, 7

[30] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 3

[31] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. *arXiv preprint arXiv:2403.06135*, 2024. 3

[32] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023. 2, 3, 4, 5

[33] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 4, 6

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 3

[35] OpenAI. Gpt-4 technical report, 2023. 6

[36] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. 6

[37] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *CVPR*, 2024. 1

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

[39] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3, 5, 7

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 4

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 4

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. 2022. 2, 3, 6

[44] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 4

[45] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023. 1

[46] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 1, 4

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[48] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 4

[49] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. *arXiv preprint arXiv:2307.10816*, 2023. 3, 7

[50] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks, 2017. 3

[51] Yuanyou Xu, Zongxin Yang, and Yi Yang. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889*, 2023. 1

[52] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 4

[53] Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11794–11803, 2020. 3

[54] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024. 1

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 3

[56] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2022. 3

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[58] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 1

[59] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. *arXiv preprint arXiv:2311.16845*, 2023. 3

[60] Chen Zhao, Chenyu Dong, and Weiling Cai. Learning a physical-aware diffusion model based on transformer for underwater image enhancement. *arXiv preprint arXiv:2403.01497*, 2024. 3

[61] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22490–22499, 2023. 3, 7

[62] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023. 3