VODiff: Controlling Object Visibility Order in Text-to-Image Generation

Dong Liang[1,2] Jinyuan Jia[1,3,*] Yuhao Liu[2,*] Zhanghan Ke[2] Hongbo Fu[4] Rynson W.H. Lau[2,*]
[1]Tongji University [2]City University of Hong Kong [3]HKUST(GZ) [4]HKUST
sse_liangdong@tongji.edu.cn, jinyuanjia@hkust-gz.edu.cn, yuhaoliu7456@gmail.com,
kezhanghan@gmail.com, hongbofu@ust.hk, Rynson.Lau@cityu.edu.hk

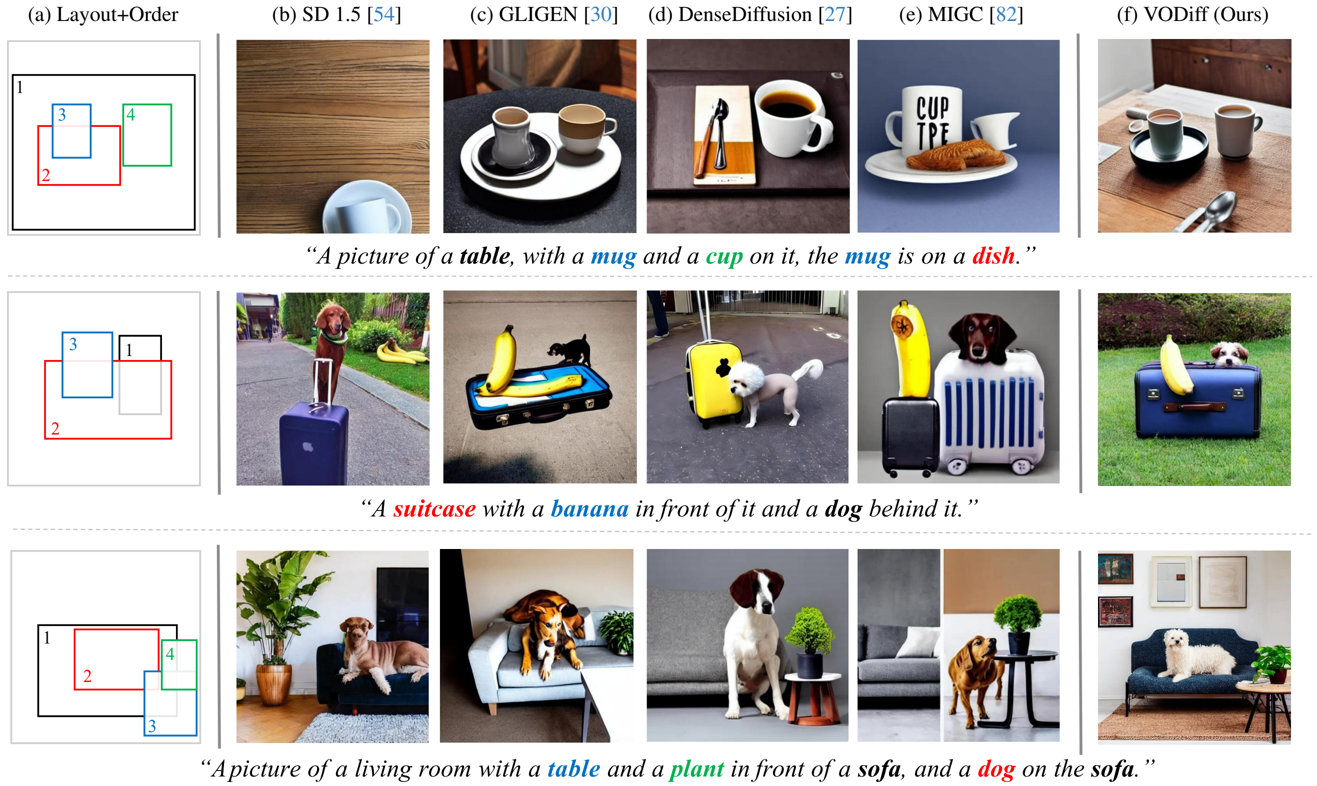

(a) Layout+Order
(b) SD 1.5 [54]
(c) GLIGEN [30]
(d) DenseDiffusion [27]
(e) MIGC [82]
(f) VODiff (Ours)
1
2
3
4
CUP
TPE
"A picture of a ***table****, with a* ***mug*** *and a* ***cup*** *on it, the* ***mug*** *is on a* ***dish****."*
1
2
3
"A ***suitcase*** *with a* ***banana*** *in front of it and a* ***dog*** *behind it."*
1
2
3
4
"A picture of a living room with a ***table*** *and a* ***plant*** *in front of a* ***sofa****, and a* ***dog*** *on the* ***sofa****."*

Figure 1. Existing T2I methods that rely on text prompts (b) and those that combine text and layout conditioning (c-e) often struggle to produce images with accurate occlusion relationships. In this work, we propose a new framework called *VODiff*, which enhances control over object arrangement by introducing their visibility orders (indicated by numbers above their bounding box in the layout map in (a)) as auxiliary constraints. *VODiff* enables the generation of images with correct spatial arrangements and occlusion relationships.

Abstract

Recent advancements in diffusion models have significantly enhanced the performance of text-to-image models in image synthesis. To enable control over the the spatial locations of the generated objects, diffusion-based methods typically utilize object layout as an auxiliary input. However, we observe that this approach treats all objects as being on the same layer and neglect their visibility order, leading to the synthesis of overlapping objects with incorrect occlusions. To address this limitation, we introduce in this paper a new training-free framework that considers object visibility order explicitly and allows users to place overlapping objects in a stack of layers. Our framework consists of two visibility-based designs. First, we propose a novel Sequential Denoising Process (SDP) to divide the whole image generation into multiple stages for different objects, each stage primarily focuses on an object. Second, we propose a novel Visibility-Order-Aware (VOA) Loss to transform the layout and occlusion constraints into an attention

[*] Joint Corresponding Authors

map optimization process to improve the accuracy of synthesizing object occlusions in complex scenes. By merging these two novel components, our framework, dubbed VODiff, enables the generation of photorealistic images that satisfy user-specified spatial constraints and object occlusion relationships. In addition, we introduce VOBench, a diverse benchmark dataset containing 200 curated samples, each with a reference image, text prompts, object visibility orders and layout maps. We conduct extensive evaluations on this dataset to demonstrate the superiority of our approach.

1. Introduction

With the advancements of diffusion models [20, 58], text-to-image (T2I) models [53, 54, 56] have revolutionized the field of image generation. These large-scale T2I models, *e.g.*, Stable Diffusion [54], can generate diverse, high-quality images conforming to given prompts, facilitating more efficient image editing [4] and artistic creation [43].

Despite the success, existing T2I models often struggle to produce satisfactory results in complex scenes involving multiple objects with occlusions. On the one hand, text descriptions alone lack spatial controllability, causing text-based methods to likely generate incorrect object positions and relationships, as shown in Fig. 1(b). On the other hand, although some condition-based methods [27, 30, 82] introduce layout conditions (*e.g.*, bbox) into pre-trained T2I models to improve positional accuracy, they do not consider the occlusion relationship of objects and may produce unsatisfactory results, as shown in Fig. 1(c)-(e).

In this work, we observe that when generating a scene with multiple objects, existing methods typically employ a one-layer-for-all strategy to synthesizes all objects simultaneously on a single canvas. Such a design requires the model to associate the objects with the input text description and accurately obtain their spatial/occlusion relationships from it, if available. This is inherently challenging and often results in mixed or even incorrect object arrangements. Even when layout constraints [3, 27, 46, 67, 69] are provided, the model may still struggle with synthesizing the correct visibility order of overlapping objects, resulting in inaccurate occlusions. To address this problem, we propose a new training-free Visibility Order Diffusion (*VODiff*) framework to explicitly model the visibility order of objects by assigning each user-specified object to a separate layer and executing a novel multi-layer denoising process to produce the output image.

We first propose a novel Sequential Denoising Process (SDP) to render the input objects one by one. SDP stratifies the objects into distinct layers, from bottom to top, according to the visibility order. For example, in the second example of Fig. 1, the layers are ordered with the *"dog"* at the bottom, the *"suitcase"* in the middle, and the *"banana"* at the top. Second, SDP divides the complete denoising process into several sub-stages, with each sub-stage primarily focusing on generating one object. Concurrently, during the denoising process, SDP assigns dynamically varying visual guidance to different objects/layers at different stages using a newly proposed smooth guidance mechanism. To prevent the misalignment of inter-object relations caused by deviations in visual guidance during denoising, we further propose a novel Visibility-Order-Aware (VOA) loss to transform layout and occlusion constraints into attention map optimization. When synthesizing an object, we first divide the corresponding layer into three regions, *overlapping*, *visible*, and *background* regions, based on its visibility *w.r.t.* those of the other objects, and then apply different constraints on these regions.

To address the lack of a benchmark in visibility-order-aware T2I generation, we also curate a new benchmark called *VOBench*, which comprises tailored visibility orders, layouts, referenced images, and prompts. To validate the effectiveness of our *VODiff*, we conduct extensive evaluations on *VOBench*. Results show that *VODiff* generates prompt-consistent objects within the user-specified layout regions and precisely delineates the occlusion relationship, as shown in Fig. 1(f).

Our key contributions can be summarized as:

- To the best of our knowledge, we are the first to consider visibility order when generating images with T2I models. We present a new training-free framework named *VODiff*, which explicitly models object visibility order to control object occlusion relationships.
- We propose two novel designs to support the explicit modeling of object visibility order: a Sequential Denoising Process (SDP) to generate different objects sequentially at various sub-stages of the denoising process, and a Visibility-Order-Aware (VOA) Loss to incorporate occlusion relationships into the attention optimization process.
- We curate a visibility-order-aware benchmark called *VOBench*, and conduct extensive evaluations on it. Results show that compared to existing SOTA methods, *VODiff* can synthesize more photorealistic images, which satisfy user-specified spatial constraints as well as object visibility order.

2. Related Work

2.1. Text-to-Image (T2I) models

Substantial progress has been made on T2I models through the use of diffusion models [20, 58, 59]. These models conceptualize image synthesis as an iterative denoising process guided by textual prompts. The denoising is conditioned on textual embeddings generated by language encoders [51, 52] and is performed either in pixel space

[2, 42, 53, 56] or latent space [12, 16, 47, 54, 73], followed by cascaded super-resolution [21] or latent-to-image decoding [11] for high-resolution image generation. By training on extensive datasets of image-text pairs [57], these models learn a unified generative space in which visual and linguistic modalities are closely associated.

Nonetheless, current T2I diffusion models still struggle to accurately interpret spatial layout instructions, making it challenging to control the occlusion relationships among multiple objects using textual descriptions alone.

2.2. Layout-guided Text-to-Image (T2I) Models

Layout-guided T2I models aim to generate images that contain objects as specified by the input text description, with their locations aligned with the input layout.

Training-based Models can be categorized into those that fine-tune pre-trained models [1, 7, 10, 23, 31, 50, 71, 75] using the specific input conditions, and those that train additional plug-ins [25, 30, 35–37, 39, 49, 66, 68, 78] to support various conditions. Some works achieve instance-level control [63, 64, 81, 82] by assigning individual prompts to each object to precisely control their attributes. However, these methods require additional training on large-scale datasets, resulting in significant computational and labor costs.

Several methods [22, 26, 70, 80] employ multiple control conditions within a single generation. Sun et al. [62] address the complex relationships among different control conditions by training a multi-control encoder to maintain harmonious results when combining multiple controls. Although these approaches can generate images with harmonious occlusion relationships, they still cannot directly control the occlusion relationship of individual objects.

Sampling-based Models often integrate multiple denoising processes, with each process targeting a specific component or region of the image. Liu *et al.* [33] interpret diffusion models as energy-based models and propose two compositional operators to combine the score functions of different conditions in a single sampling step. MultiDiff. [3] and LRDiff. [48] then propose to utilize shared parameters or constraints to ensure consistency/accuracy of each object in the generated image.

Although several methods [38, 40] have proposed training-free multimodal control for the T2I generation, they primarily focus on enhancing control techniques without specifically addressing the object occlusion issues.

Attention-based Models can be categorized into forward and backward approaches based on how they employ the attention maps. Forward methods [2, 13, 15, 27, 72] manipulate attention values directly on the feature map. The accuracy of these methods heavily depends on the precision of the layout condition (*e.g.*, segmentation masks). Backward methods [5, 8, 46, 67, 69, 72] combine the layout condition with the attention map, and utilize an energy function to translate the spatial constraints into the noisy input for iterative updates. These methods also primarily emphasize on positional accuracy while neglecting the occlusion relationships among objects, leading to the unsatisfactory generation of overlapping regions.

In addition, while several customized T2I models [9, 29, 32, 60, 61, 71, 79] attempt to integrate specific input objects into designated regions of the generated image by using additional layout input, they overlook the visibility order among the integrated objects and other objects within or near those regions. This limitation hinders their effectiveness in addressing occlusion issues.

3. Method

In this work, our goal is to enhance the ability of pre-trained text-to-image (T2I) diffusion models to correctly handle occlusion relationships in multi-object scenarios without training. We note that the overlapping regions of the objects are the main challenge, and they affect the occlusion relationship of the objects in the synthesized image. However, it is difficult for T2I models to understand occlusion relations among objects through text prompts only. In addition, current spatial layouts can only help constrain the spatial location of the objects. They do not indicate the order of objects in the overlapping regions.

To achieve our goal, we propose a training-free method, *VODiff*, which introduces object visibility order into the layout-guided T2I generation to effectively address the incorrect object arrangement problem. Fig. 2 shows our framework. It consists of two novel visibility-aware designs: a Sequential Denoising Process (SDP) and a Visibility-Order-Aware (VOA) loss. When generating an image, our SDP strategy divides the whole denoising process into several sub-stages according to the number of user-specified objects and generates the specified objects in each stage sequentially in a bottom-up manner. Meanwhile, at each denoising step, our VOA loss applies dynamic losses to regions with different visibility states, ensuring accurate occlusion relationships among the generated objects.

For the rest of this section, we first briefly introduce the preliminary information of diffusion models and attention layers in Sec. 3.1. We then present how we design the SDP in Sec. 3.2 and detail the VOA loss function in Sec. 3.3.

3.1. Preliminaries

Diffusion Models [20, 58, 59] learns to approximate a data distribution by progressively denoising a random variable sampled from a unit Gaussian distribution $\mathcal{N}(0, I)$. At a denoising step $t \in [1, \cdots, T]$, the forward-time Markov chain is formulated as $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_{t-1}$, where $\epsilon \sim \mathcal{N}(0, I)$, and β_t is a pre-defined variance schedule. The reverse-time denoising process can then be defined as:

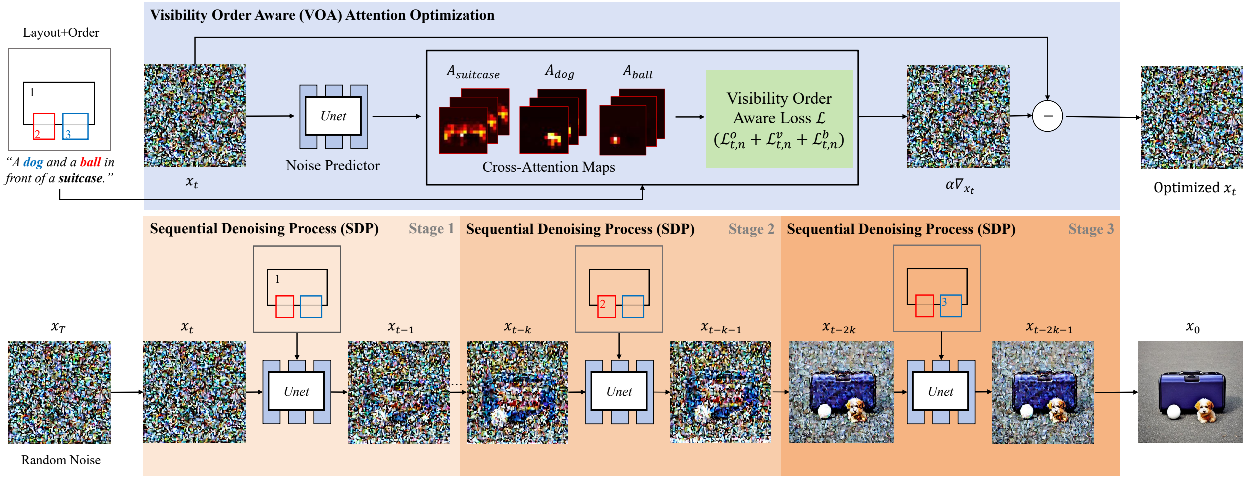

Layout+Order
1
2
3
"A dog and a ball in front of a suitcase."
Visibility Order Aware (VOA) Attention Optimization
x_t
Unet
Noise Predictor
$A_{suitcase}$
A_{dog}
A_{ball}
Cross-Attention Maps
Visibility Order Aware Loss $\mathcal{L}$ $(\mathcal{L}^o_{t,n} + \mathcal{L}^v_{t,n} + \mathcal{L}^b_{t,n})$
$\alpha\nabla_{x_t}$
−
Optimized x_t
Sequential Denoising Process (SDP)
Stage 1
Stage 2
Stage 3
x_T
Random Noise
x_t
x_{t-1}
x_{t-k}
x_{t-k-1}
x_{t-2k}
x_{t-2k-1}
x_0

Figure 2. Our *VODiff* incorporates two visibility-aware designs: a sequential denoising process (SDP) and a visibility-order-aware (VOA) loss function. SDP divides the denoising process into several stages according to the visibility order, and performs denoising on different parts sequentially. Meanwhile, the VOA loss is utilized to eliminate the local spatial shifts in the attention layers, further preventing position deviations and incorrect occlusion relationships. Given a text prompt, the spatial layout, and the visibility order of the objects to be synthesized as inputs, our *VODiff* can generate images that simultaneously adhere to all the constraints.

$x_{t-1} = \tilde{\alpha}_t x_t + \tilde{\beta}_t \hat{s}_t + \sigma_t \epsilon_t$, where $\tilde{\alpha}_t$, $\tilde{\beta}_t$, and σ_t are calculated coefficients. To generate an image, given a noisy input x_T sampled from $\mathcal{N}(0, I)$, we run denoiser ϵ_θ iteratively to obtain intermediate representations $x_{T-1}, \cdots, x_1, x_0$, where $x_0 \approx x$ is the final result.

Attention Layers are the cornerstone of Stable Diffusion [54]. They used to condition the generation based on the text prompt. A pretrained CLIP encoder [51] is usually utilized to encode the text prompt $c = (c_1, c_2, \cdots, c_n)$, where n is the number of tokens, into text embeddings. These tokens are then used as keys $K \in \mathbb{R}^{n\times d}$ and values $V \in \mathbb{R}^{n\times d}$ in the Cross-Attention layer. The symbol d denotes the feature dimension. For a set of queries $Q \in \mathbb{R}^{hw\times d}$ computed from the feature map of size $h \times w$, the cross-attention map A_t at time step t is calculated as: $A_t = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \in [0, 1]^{hw\times n}$, where A_t contains n attention maps $\{A^1_t, \cdots, A^n_t\}$, and $A^i_t \in [0, 1]^{h\times w}$ indicates the strength of association between each word token c_i and each spatial location on the feature map. A higher attention value in A^i_t indicates a higher probability of generating the object c_i at that spatial location.

3.2. Sequential Denoising Process (SDP)

Instead of directly generating all objects on a single layer at once, like what existing works [27, 38, 46] do, we are inspired by the painter's algorithm [41], which sorts all polygons according to their z values and then render them starting from the farthest polygon, and the layer concept in image editing tools such as Photoshop. We propose to divide the generation of the target scene into layers, where each object is assigned a separate layer, and synthesize the objects sequentially, layer by layer from bottom to top, with the following steps.

First, we stratify the target objects into distinct layers according to their visibility order. Specifically, given an input text prompt C containing N target objects; a layout condition represented by bounding boxes $B = \{b_1, b_2, \ldots, b_N\}$, where b_n is the bounding box of the n-th target object; and visibility orders $V = \{v_1, v_2, \ldots, v_N\}$, where v_n is the visibility order of the n-th target object, we convert the layout condition B into N masks, $M = \{m_1, m_2, \ldots, m_N\}$, where m_n is the mask corresponding to the bounding box b_n with a value of 1 inside the bounding box region and 0 elsewhere. Meanwhile, we extract N object entities from the input text prompt C and denote the set as $\{c_1, c_2, \cdots, c_N\}$. Each bounding box b_n is associated with an object prompt c_n, which describes the content of the target object inside b_n. We order the N target objects in an increasing object visibility order specified by the user, even though there may be non-overlapping objects. The layers from bottom to top can thus be represented as $L = \{(m_1, c_1, v_1), (m_2, c_2, v_2), \cdots, (m_N, c_N, v_N)\}$.

Second, we divide the reverse-time denoising process (with a total of T steps) of our SDP into two main parts, $[T, T_g]$ and $[T_g, 0]$, where T_g denotes the start time of global denoising. The global denoising is to delineate the overall scene based on the input text prompt C. The first part of our SDP focuses on per-layer generation based on layers L. We divide it into N sub-stages, $S = \{s_1, s_2, \cdots, s_N\}$, starting from the bottom layer (*i.e.*, the object with the lowest visibility order), Each sub-stage s_n contains k denoising steps, ranging from $[s_n, s_n + k]$, *i.e.*, $T - T_g = kN$. Since the diffusion denoising process is reversed, $s_n + k$ represents the start time for substage s_n. At each sub-stage, to force c_n to be generated within the object region, we explicitly incorporate visual guidance into this region. However, we observe that focusing exclusively on a single object n at each stage can lead to the failure of generating other objects, particularly those whose start time is near T_g, due to the missing of structure information [2] at later denoising steps. To address this problem, at each stage n, we apply visual guidance not only for the target object n but also for other objects j ($j \neq n$), to ensure the generated layout adheres to the overall requirements. Our scheme for adding visual guidance can be formulated as:

$$x_{t,n} = x_t + \lambda_{t,n} \odot m_n, \tag{1}$$

where $\lambda_{t,n}$ represents the visual guidance scale of the target object n at time step t. Note that we allocate a dynamically varying smooth $\lambda_{t,n}$ to each object to ensure correct occlusion relationships. Specifically, when processing object n at sub-stage n, the SDP assigns a guidance scale a to object n and $a/2$ to all other objects. As the process transitions from sub-stage n to sub-stage $n+1$, the guidance scale for object n decreases linearly to $a/2$, while the guidance scale for object $n+1$ increases linearly to a, and guidance scale for the other objects remains unchanged, as:

$$\lambda_{t,n} = \begin{cases} a - \frac{a}{2}\frac{t-s_{n-1}}{k} & \text{if } s_{n-1} < t <= s_{n-1} + k, \\ \frac{a}{2} + \frac{a}{2}\frac{t-s_n}{k} & \text{if } s_n < t <= s_n + k, \\ \frac{a}{2} & \text{otherwise}, \end{cases} \tag{2}$$

where a denotes the initial visual guidance scale (refer to Sec. D.4 of the Supp. for more discussion) and $s_0 = T$. By assigning a smoothly transitioning visual guidance, we ensure that each sub-stage focuses more on a specific object while maintaining correct occlusion relationships between it and other objects, such that in overlapping regions, objects with a higher visibility order occlude those with a lower one. Finally, we forward the updated noisy input $x_{t,n}$ to the denoising model ϵ_θ to estimate noise for each object separately and then combine them together, as:

$$\epsilon_\theta(x_t, t, C) = \sum_{n=1}^{N} \frac{m_n}{\sum_{n=1}^{N} m_n} \otimes \epsilon_\theta(x_{t,n}, t, c_n), \tag{3}$$

where $\epsilon_\theta(x_t, t, C)$[1] is the estimated noise used to update x_{t-1} via the standard diffusion scheduler.

3.3. Visibility-Order-Aware (VOA) Loss

Although SDP can help constrain target generation within the specified region using visual guidance, the consecutive conv. layers of the denoising model introduce local shifts, which are then amplified by the attention layers. As a result, adding visual guidance at the input level (*i.e.*, x_t) as a constraint is insufficient and tends to cause object position deviation and false occlusion relationships (see Fig. 5 column 2 for a visual illustration).

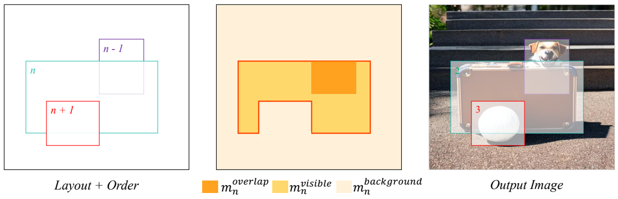


Figure 3. Region separation illustration in the VOA loss. m_n^{overlap} represents the regions where the current object n occludes other objects, m_n^{visible} and $m_n^{\text{background}}$ denote where the current object should and should not be generated in the final image.

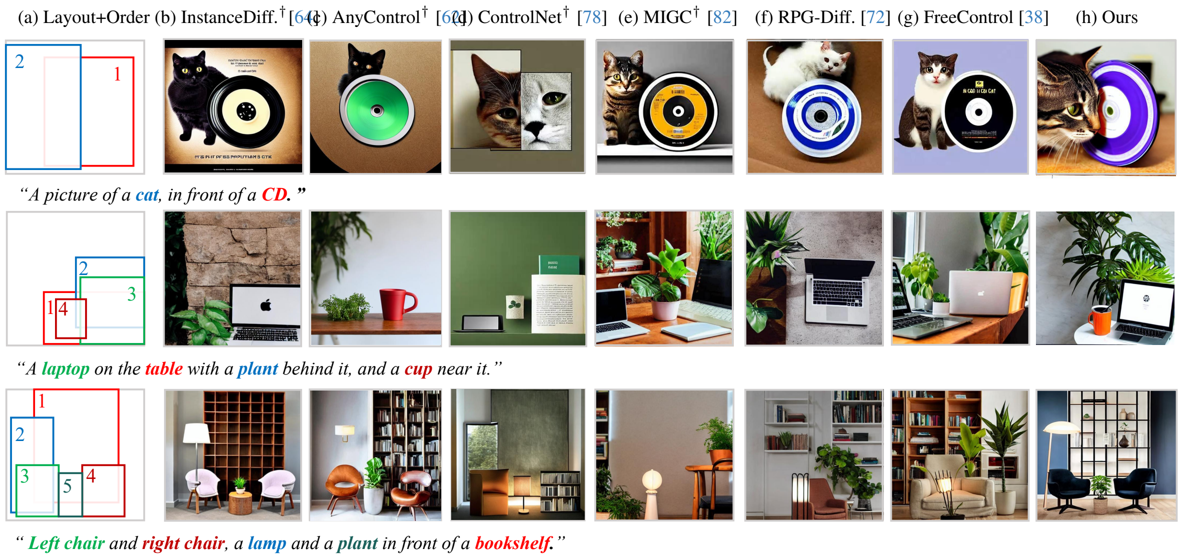
To address the above problem, we propose constraining the spatial location and occlusion relationship at feature-level. Specifically, when generating an image with multiple target objects, different object prompts usually have different activations in different regions of the cross-attention map. At a given spatial location, the higher the attention value of a prompt, the more likely the corresponding object will be generated in that region [17]. In addition, when two objects overlap each other, their occlusion relationship is directly influenced by their visibility order. Once the order changes, the content of the overlapping region changes accordingly. Based on this, we transform the spatial layout and occlusion relationship among the objects to the cross-attention map and optimize it for correct image generation.

As shown in Fig. 3, when processing target object n, we first divide the mask m_n into three parts, overlapping regions m_n^{overlap}, visible regions m_n^{visible}, and background regions $m_n^{\text{background}}$, based on the visibility order and masks of all objects in the layer set L. Specifically, for m_n^{overlap}, we compute the overlapping regions (*i.e.*, dark orange region) of the current object mask m_n and the masks of all objects with lower visibility order than the current object (*i.e.*, $\{m_j\}$, with $j \in [1, n-1]$) as $m_n^{\text{overlap}} = \bigcup_{j=1}^{n-1} m_n \odot m_j$. This is because an object located at an upper layer should obtain a higher attention score in the overlapping regions. For m_n^{visible}, we first compute the overlapping regions of the current mask m_n with the masks of all objects that have a higher visibility order (*i.e.*, $\{m_j\}$, with $j \in [n+1, N]$). We then subtract these overlapping regions from the current object mask, which can be expressed as $m_n^{\text{visible}} = m_n - \bigcup_{j=n+1}^{N} m_n \odot m_j$. $m_n^{\text{background}}$ is then obtained by inverting m_n^{visible} as $m_n^{\text{background}} = 1 - m_n^{\text{visible}}$.

During the denoising phase for the n-th target object at time step t, as indicated in Eq. 3, we extract the cross-attention maps referred to by the object prompt c_n from each layer and average them to produce $\mathbf{A}_{t,n}$. We then compute the loss between the attention map $\mathbf{A}_{t,n}$ and each of the regions. For the overlapping regions, which are affected by the visibility order, we ensure that objects in the upper layers obtain higher attention scores, as:

$$\mathcal{L}_{t,n}^{\text{o}} = \alpha_o \cdot \left[1 - \min(\mathbf{A}_{t,n} \odot m_n^{\text{overlap}})\right]. \tag{4}$$

[1] For simplicity, we omit the classifier-free guidance [19] here.


(a) Layout+Order (b) InstanceDiff.† [64] (c) AnyControl† [62] (d) ControlNet† [78] (e) MIGC† [82] (f) RPG-Diff. [72] (g) FreeControl [38] (h) Ours

"A picture of a ***cat****, in front of a* ***CD****. "*

"A ***laptop*** *on the* ***table*** *with a* ***plant*** *behind it, and a* ***cup*** *near it."*

" ***Left chair*** *and* ***right chair****, a* ***lamp*** *and a* ***plant*** *in front of a* ***bookshelf****."*

Figure 4. Qualitative comparison with six SOTA methods in Tab. 1. Training-based methods are labeled with †.

We then encourage target object n to be generated within the visible regions by increasing the attention score through $\mathcal{L}^{\mathrm{v}}_{t,n}$, and prevent the target deviation by weakening the attention score of the background regions via $\mathcal{L}^{\mathrm{b}}_{t,n}$:

$$\mathcal{L}^{\mathrm{v}}_{t,n} = \alpha_{vis} \cdot \left[1 - \max(\mathbf{A}_{t,n} \odot m_n^{\text{visible}})\right], \tag{5}$$

$$\mathcal{L}^{\mathrm{b}}_{t,n} = \alpha_{bac} \cdot \max(\mathbf{A}_{t,n} \odot m_n^{\text{background}}), \tag{6}$$

where α_o, α_{vis} and α_{bac} are coefficients for the three loss functions. Finally, the total loss is:

$$\mathcal{L}_t = \sum_{n}^{N} (\mathcal{L}^{\mathrm{o}}_{t,n} + \mathcal{L}^{\mathrm{v}}_{t,n} + \mathcal{L}^{\mathrm{b}}_{t,n}). \tag{7}$$

With the visibility-order-aware loss, we modify the noise sample x_t at each denoising step to minimize the loss using gradient descent $\hat{x}_t \leftarrow x_t - \alpha \nabla_{x_t} \mathcal{L}_t$, where α controls the influence of the optimization in the denoising process. In each step t, we iterate the gradient descent process τ times. By applying the VOA loss on all objects at each sub-stage, *VODiff* alleviates the guidance deviation and achieves more accurate object spatial position and occlusions.

4. Experiment

4.1. Metrics

We adopt a few metrics to evaluate our model:

- **FID**: We use Fréchet Inception Distance [18] to evaluate the image quality. It compares the distribution of generated images to that of real images. A lower FID score indicates a higher image quality, reflecting closer alignment with real images on visual features and diversity.
- **CLIP-Score**: We use the CLIP-Score to compute the distance between input textual features and generated image features, evaluating the fidelity of the generated image to the textual input.
- **Layout Alignment (LA)**: We use Grounding-DINO [34] to assess Layout alignment accuracy. It detects bounding boxes for each instance and then compute the mean IoU between the detected and Ground Truth boxes. The box threshold is set to 0.5 to avoid incorrect box detection.
- **Occlusion Alignment (OA)**: We employ GPT-4V to determine the occlusion relationships between objects in the generated images by prompting it with the image and the object names contained in the image. We then compare the visual order returned by GPT-4V with the ground truth visibility order. If the two match with each other, we count it as accurate. Finally, we compute the percentage of accurate occlusion relationships from all the samples.

4.2. VOBench

Since there are no visibility-order-aware datasets available for evaluation, we curate a VOBench for this purpose. All the raw image data in VOBench are manually collected from the Internet. The data annotation process is as follows. We first employ GroundingDINO [34] and GPT-4V to generate the bounding boxes and text prompts for each object in the image, and manually assign the visibility order to each object box. Based on the number of boxes detected by GroundingDINO, we then divide the images into four categories containing 2-5 objects. For each category, we manually select 50 images based on the degree of overlap between boxes to make sure that various occlusion relations are in our VOBench.

4.3. Comparison with State-of-the-Art Methods

We compare *VODiff* with 12 layout-guided text-to-image generation methods, including seven training-free methods: BoxDiff [69], R&B [67], Attention-Refocusing [46], MultiDiffusion [3], DenseDiffusion [27], FreeControl [38] and RPG-Diffusion [72], and five training-based methods: SmartControl [35], ControlNet [78], AnyControl [62], MIGC [82] and InstanceDiffusion [64]. Since these methods do not support visibility order as an input, we include

Table 1. Quantitative comparison of our *VODiff* with training-based and training-free methods. LA and OA represent the layout and occlusion alignment accuracy, respectively. AR denotes the average ranking from the user study, while Q, L, and O indicate that the ranking is based on Quality, Layout, and Occlusion. The best and second-best performances are marked in **bold** and underlined, respectively.

Metrics	**Training-based**					**Training-free**							
	SmartCtrl [35]	ControlNet [78]	MIGC[82]	InstanceDiff. [64]	AnyControl [62]	BoxDiff [69]	R&B [67]	AR [46]	MultiDiff. [3]	DenseDiff. [27]	FreeCtrl [38]	RPG [72]	**Ours**
FID ↓	10.58	18.91	**9.82**	10.21	10.35	21.99	12.37	14.87	16.34	21.68	17.35	11.66	10.03
CLIP-Score ↑	29.45	29.34	29.61	29.53	29.63	29.46	28.72	29.59	27.42	29.33	29.12	29.57	**29.73**
LA ↑	54.81	52.73	54.62	53.58	54.77	13.50	30.88	29.79	16.60	24.11	36.14	51.15	**55.11**
OA ↑	55.00	51.50	65.00	63.50	69.50	18.50	34.50	32.00	21.50	24.00	28.50	51.50	**82.50**
AR-Q ↓	5.33	12.13	2.47	6.20	3.05	8.15	10.05	4.13	12.68	7.10	9.00	5.12	**1.50**
AR-L ↓	2.16	6.21	4.04	5.11	3.12	12.98	9.02	10.07	12.03	11.14	8.08	7.17	**1.73**
AR-O ↓	6.25	5.12	3.18	4.02	2.05	12.74	9.01	10.11	12.04	11.09	8.03	7.10	**1.26**

Table 2. Ablation study on components. *Constant* and *Dynamic* represent the state of change of guidance scale in the SDP.

SDP		VOA	FID ↓	Clip-Score ↑	LA ↑	OA ↑
Constant	*Dynamic*					
	✓		**9.98**	27.85	32.33	54.50
		✓	13.34	29.01	43.25	73.50
✓		✓	11.68	29.33	43.36	69.50
	✓	✓	10.03	**29.73**	**55.11**	**82.50**

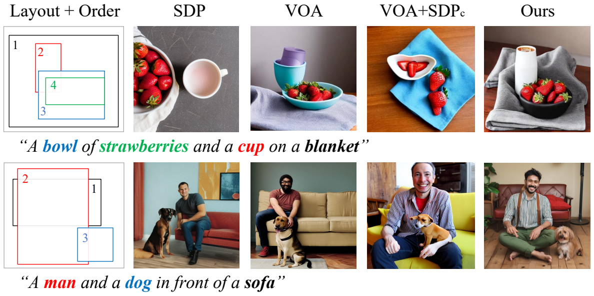

Layout + Order
SDP
VOA
VOA+SDPc
Ours
1
2
4
3
"A bowl of strawberries and a cup on a blanket"
2
1
3
"A man and a dog in front of a sofa"

Figure 5. Visual comparison of the ablation study in Tab. 2. SDP$_c$ denotes that the *constant* guidance scale is used. By default, the dynamic guidance scale is used in SDP.

descriptions of object occlusion relationships in the text prompts to ensure fairness.

Qualitative Comparison. As shown in Fig. 4, we visually compare *VODiff* with six top-performing methods. The images generated by *VODiff* exhibit superior spatial consistency and occlusion handling. Specifically, our method handles varying numbers of objects and degrees of occlusion remarkably well. In contrast, other methods exhibit issues such as incorrect spatial positioning (*e.g.*, the 2nd row, (b-f)) or missing objects (*e.g.*, the 3rd row, (d-f)). In addition, although some methods manage to generate correct spatial relationships, they fail in occlusion handling (*e.g.*, the 1st row, (b,e,f,g)).

Quantitative Comparison. As shown in Tab. 1, we evaluate our method against both training-free and training-based approaches. Our method significantly outperforms the training-free methods across all metrics. We achieve higher CLIP-Score (29.73), LA (55.11), and OA (82.50) scores, indicating superior image quality, layout alignment, and occlusion handling. Despite being training-free, our method outperforms training-based methods in six out of seven metrics. We achieve higher CLIP-Score (29.73 *vs* 29.63 by AnyControl), LA (55.11 *vs* 54.81 by SmartControl), and OA (82.50 *vs* 69.50 by AnyControl) scores. This demonstrates the effectiveness of our method in generating text-/layout-aligned and occlusion handled images without additional training. Although the training-based method, MIGC, achieves the best FID score of 9.82, our method achieves a similarly FID of 10.03. The slight reduction can be attributed to the optimization on attention maps [46].

User Study. We also conduct a user study with 63 participants to assess various methods through subjective evaluation. We present users with ten sets of 8 images along the same input conditions (*i.e.*, prompt and layout map), and ask them to rank the given images based on image quality (AR-Q), layout (AR-L), and occlusion accuracy (AR-O). Each pair of images are shown randomly, and we collect ranking scores from different participants. We then compute the average ranking (AR) for three aspects: quality, layout, and occlusion as shown in Tab. 1. The results show a clear preference for our method among the users.

4.4. Ablation Study

Component Analysis. We ablate our *VODiff* by removing SDP and VOA separately, and replacing the smooth guidance with constant guidance (denoted as VOA+SDP$_c$). Tab. 2 and Fig. 5 show the quantitative and qualitative results. We have three conclusions. ❶ Using only SDP (*Dynamic*) results in the worst LA, OA accuracy and Clip-Score but produces better FID. This is because SDP only adds visual guidance to guide the denoising direction, which does not impose stronger spatial constraints, and thus does not significantly affect the denoising process. ❷ VOA significantly improves the layout and occlusion accuracy by optimizing the cross-attention map, but at the same time, causes the model to produce less diversified results, thus impacting the FID metrics. ❸ Although constant visual guidance can improve FID, CLIP-Score, and LA, when combined with VOA, it confuses the order of objects in overlapping regions during the denoising process, interfering with attention optimization and worsening the OA. Finally, combining all the

Table 3. Ablation study on the VOA loss. $\mathcal{L}_\text{v}$, $\mathcal{L}_\text{o}$ and $\mathcal{L}_\text{b}$ represent the loss on visible, occluded and background regions, respectively.

$\mathcal{L}_\text{v}$	$\mathcal{L}_\text{o}$	$\mathcal{L}_\text{b}$	FID ↓	Clip-Score ↑	LA ↑	OA ↑
✓			11.41	28.05	42.13	60.50
✓		✓	10.23	29.25	46.42	61.50
✓	✓		14.65	28.42	44.47	77.50
	✓	✓	11.68	28.10	29.14	58.00
✓	✓	✓	10.03	29.73	55.11	82.50

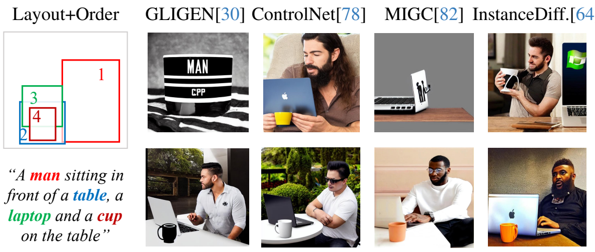

Layout+Order
GLIGEN[30]
ControlNet[78]
MIGC[82]
InstanceDiff.[64]
1
3
4
2
"A man sitting in front of a table, a laptop and a cup on the table"

Figure 6. Enhancement of existing methods using *VODiff*. Row 1 and 2 show the vanilla and the enhanced results, respectively.

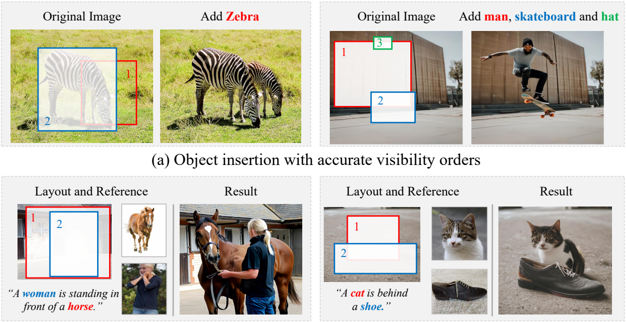
proposed components helps achieve the best results.

Visibility-Order-Aware (VOA) loss. We verify the efficacy of different loss items in the VOA by removing them separately. Tab. 3 shows the quantitative results, and Supp. Fig. 3 shows the visual results. We can see that ❶ $\mathcal{L}_v$ is indispensable for layout and occlusion constraints (4th row), but with $\mathcal{L}_v$ alone is also insufficient (1st row); ❷ $\mathcal{L}_v + \mathcal{L}_b$ produces inferior OA results as the overlapping regions lack constraints; ❸ $\mathcal{L}_v + \mathcal{L}_o$ has the remarkable OA results, but the worst FID because this combination only considers each object by itself but ignores background constraints (*e.g.*, misaligned "man" and "dog" in the 1st row, 5-th column of Fig. 3 in Supp.). Finally, combining all items helps achieve the best performance on all metrics.

4.5. Diverse Applications

Our *VODiff* is training-free and can be easily integrated into many existing layout-to-image methods [30, 64, 78, 82], significantly enhancing the accuracy of occlusion relationship generation in current pretrained models. As shown in Fig. 6, the result from each method without (or with) using our approach is presented in the first (or second) row. Our *VODiff* effectively assists these training-based methods in addressing issues such as object disappearance (columns 2–4), as well as incorrect positional (column 5) and occlusion (columns 2 and 5) relationships.

Our *VODiff* supports customized generation by integrating it with existing techniques, *e.g.*, DreamBooth [55]. In Fig. 7(a), we first demonstrate the insertion of newly generated objects into an existing image via Inversion [58], where the inserted objects adhere to the preset visibility order relative to the original objects in the image, while preserving the background unchanged. Fig. 7(b) shows the integration of our method with DreamBooth [55], enabling the generated objects to not only maintain the identity of the input objects but also ensure that the positions and visibility order among multiple objects adhere to the preset arrangement.



(a) Object insertion with accurate visibility orders

(b) Exemplar based image editing with visibility orders

Figure 7. Customized generation with *VODiff* in object insertion (row 1) and exemplar-based editing (row 2).

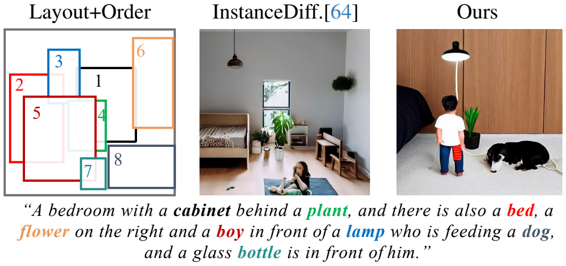

Layout+Order
InstanceDiff.[64]
Ours
1
2
3
4
5
6
7
8
"A bedroom with a cabinet behind a plant, and there is also a bed, a flower on the right and a boy in front of a lamp who is feeding a dog, and a glass bottle is in front of him."

Figure 8. A failure case of *VODiff*. From left to right are: inputs, results of InstanceDiff. [64] and Our method, respectively.

5. Conclusion

In this paper, we have introduced *VODiff* to address the the challenges of generating complex scenes with object occlusions. We first propose a Sequential Denoising Process that performs object generation sequentially from the bottom to the top layer based on the visibility order of objects. We then propose a Visibility-Order-Aware loss to constrain object positions and occlusion relationships by transforming the layout and visibility order into cross-attention map optimization. The results on our curated *VOBench* show superior performances of our method in controlling object spatial position and occlusion relationships.

VODiff does has limitations. For example, as shown in Fig. 8, like existing method [64], our method may struggle to handle scenarios where the number of objects is exceptionally high due to the challenge of balancing the image quality with precise spatial and occlusion relationships.

Acknowledgements

This work is supported in part by Ant Group and General Program of National Natural Science Foundation of China

(NSFC, No. 6207071897) and a GRF from the Research Grants Council of Hong Kong (Ref.: 11205620).

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 *CVPR*, pages 18370–18380, 2023. 3

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022. 3, 5

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *ICML*, pages 1737–1752, 2023. 2, 3, 6, 7, 5

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2

[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. *ACM TOG*, 42(4):1–10, 2023. 3, 1

[6] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv:2305.03374*, 3, 2023. 1

[7] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, HONG Lanqing, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. In *ICLR*, 2024. 3

[8] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, pages 5343–5353, 2024. 3, 2

[9] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, pages 6593–6602, 2024. 3, 1, 9

[10] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023. 3

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 3

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 3

[13] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv:2212.05032*, 2022. 3

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv:2208.01618*, 2022. 1

[15] Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, and Yu Liu. Check locate rectify: A training-free layout calibration system for text-to-image generation. In *CVPR*, pages 6624–6634, 2024. 3

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 3

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 5

[18] 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. In *NIPS*, 2017. 6

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NIPS Workshop*, 2021. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NIPS*, 2020. 2, 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47):1–33, 2022. 3

[22] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *CVPR*, pages 4754–4763, 2024. 3

[23] Chengyou Jia, Minnan Luo, Zhuohang Dang, Guang Dai, Xiaojun Chang, Mengmeng Wang, and Jingdong Wang. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *AAAI*, pages 2480–2488, 2024. 3

[24] Yuming Jiang, Nanxuan Zhao, Qing Liu, Krishna Kumar Singh, Shuai Yang, Chen Change Loy, and Ziwei Liu. Groupdiff: Diffusion-based group portrait editing. *arXiv:2409.14379*, 2024. 1

[25] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. Scedit: Efficient and controllable image diffusion generation via skip connection editing. In *CVPR*, pages 8995–9004, 2024. 3

[26] Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Scalable and composable multimodal text-to-image diffusion models. *arXiv:2305.15194*, 2023. 3

[27] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, pages 7701–7711, 2023. 1, 2, 3, 4, 6, 7, 5

[28] Leheng Li, Weichao Qiu, Xu Yan, Jing He, Kaiqiang Zhou, Yingjie Cai, Qing Lian, Bingbing Liu, and Ying-Cong Chen. Omnibooth: Learning latent control for image synthesis with multi-modal instruction. *arXiv:2410.04932*, 2024. 1

[29] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv:2306.12624*, 2023. 3, 1

[30] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 1, 2, 3, 8, 5, 6, 10

[31] Yaqi Li, Han Fang, Zerun Feng, Kaijing Ma, Chao Ban, Xianghao Zang, LanXiang Zhou, Zhongjiang He, Jingyan Chen, Jiani Hu, et al. Goal: Grounded text-to-image synthesis with joint layout alignment tuning. In *ACM MM*, pages 7055–7064, 2024. 3

[32] Chang Liu, Xiangtai Li, and Henghui Ding. Referring image editing: Object-level image editing via referring expressions. In *CVPR*, pages 13128–13138, 2024. 3, 1

[33] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, pages 423–439, 2022. 3

[34] 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:2303.05499*, 2023. 6, 1, 4

[35] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. *arXiv:2404.06451*, 2024. 3, 6, 7, 5

[36] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *CVPR*, pages 8217–8227, 2024.

[37] Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. In *CVPR*, pages 9048–9058, 2024. 3

[38] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, pages 7465–7475, 2024. 3, 4, 6, 7, 5

[39] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 3

[40] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. In *ECCV*, pages 93–110, 2025. 3

[41] Martin E Newell, RG Newell, and Tom L Sancha. A solution to the hidden surface problem. In *Seminal graphics: pioneering efforts that shaped the field*, pages 27–32. 1998. 4

[42] 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. *arXiv:2112.10741*, 2021. 3

[43] NovelAI. Novelai: Ai generated stories and art. 2021. 2

[44] Lingzhi Pan, Tong Zhang, Bingyuan Chen, Qi Zhou, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Coherent and multi-modality image inpainting via latent space optimization. *arXiv:2407.08019*, 2024. 1

[45] Yulin Pan, Chaojie Mao, Zeyinzi Jiang, Zhen Han, and Jingfeng Zhang. Locate, assign, refine: Taming customized image inpainting with text-subject guidance. *arXiv:2403.19534*, 2024. 1

[46] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *CVPR*, pages 7932–7942, 2024. 2, 3, 4, 6, 7, 5

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 3, 5

[48] Zipeng Qi, Guoxi Huang, Chenyang Liu, and Fei Ye. Layered rendering diffusion model for zero-shot guided image synthesis. In *ECCV*, 2024. 3, 1

[49] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NIPS*, 2023. 3

[50] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *ACM MM*, pages 643–654, 2023. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 4

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. 2

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 2, 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[55] 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 *CVPR*, pages 22500–22510, 2023. 8, 1, 9

[56] 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. Photorealistic text-to-image diffusion models with deep language understanding. In *NIPS*, pages 36479–36494, 2022. 2, 3

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo

Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NIPS*, pages 25278–25294, 2022. 3, 1

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 8, 1, 6, 9

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2, 3

[60] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, pages 18310–18319, 2023. 3, 1

[61] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *CVPR*, pages 8048–8058, 2024. 3, 1, 9

[62] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: Create your artwork with versatile control on text-to-image generation, 2024. 3, 6, 7, 5

[63] Zhenhong Sun, Junyan Wang, Zhiyu Tan, Daoyi Dong, Hailan Ma, Hao Li, and Dong Gong. Eggen: Image generation with multi-entity prior learning through entity guidance. In *ACM MM*, pages 6637–6645, 2024. 3

[64] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, pages 6232–6242, 2024. 3, 6, 7, 8, 5, 10

[65] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, pages 15943–15953, 2023. 1

[66] Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv:2409.08240*, 2024. 3

[67] Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. In *ICLR*, 2023. 2, 3, 6, 7, 5

[68] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv:2409.11340*, 2024. 3

[69] 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. In *ICCV*, pages 7452–7461, 2023. 2, 3, 6, 7, 5

[70] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *CVPR*, pages 8682–8692, 2024. 3

[71] 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 *CVPR*, pages 18381–18391, 2023. 3, 1, 9

[72] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *ICML*, 2024. 3, 6, 7, 5

[73] Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, and Bin Cui. Cross-modal contextualized diffusion models for text-guided visual generation and editing. *arXiv:2402.16627*, 2024. 3

[74] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv:2403.11627*, 2024. 1

[75] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, pages 14246–14255, 2023. 3

[76] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 1

[77] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv:2310.19784*, 2023. 1

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3, 6, 7, 8, 5, 10

[79] Lirui Zhao, Tianshuo Yang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Rongrong Ji. Diffree: Text-guided shape free object inpainting with diffusion model. *arXiv:2407.16982*, 2024. 3, 1

[80] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NIPS*, 2023. 3

[81] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *arXiv:2407.02329*, 2024. 3

[82] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, pages 6818–6828, 2024. 1, 2, 3, 6, 7, 8, 5, 10