Layered Rendering Diffusion Model for Controllable Zero-Shot Image Synthesis

Zipeng Qi*[1], Guoxi Huang*†[2], Chenyang Liu[1], and Fei Ye[3]

[1] Beihang University
[2] University of Bristol
[3] Mohamed bin Zayed University of Artificial Intelligence

Abstract. This paper introduces innovative solutions to enhance spatial controllability in diffusion models reliant on text queries. We first introduce vision guidance as a foundational spatial cue within the perturbed distribution. This significantly refines the search space in a zero-shot paradigm to focus on the image sampling process adhering to the spatial layout conditions. To precisely control the spatial layouts of multiple visual concepts with the employment of vision guidance, we propose a universal framework, **L**ayered **R**endering **Diff**usion (LRDiff), which constructs an image-rendering process with multiple layers, each of which applies the vision guidance to instructively estimate the denoising direction for a single object. Such a layered rendering strategy effectively prevents issues like unintended conceptual blending or mismatches while allowing for more coherent and contextually accurate image synthesis. The proposed method offers a more efficient and accurate means of synthesising images that align with specific layout and contextual requirements. Through experiments, we demonstrate that our method outperforms existing techniques, both quantitatively and qualitatively, in two specific layout-to-image tasks: bounding box-to-image and instance mask-to-image. Furthermore, we extend the proposed framework to enable spatially controllable editing. The project page is available here.

Keywords: Diffusion Models · Controlled image generation · Image Editing

1 Introduction

Large-scale Text-to-Image (T2I) diffusion models trained at scale (e.g., 250 million captioned images for DALL·E [46]) have recently shown remarkable capabilities in generating high-fidelity images, covering diverse concepts. Meanwhile, the excellent data synthesis capabilities of diffusion models have been extensively leveraged in diverse fields, encompassing 3D modelling [16,56], training data creation [55], video generation [23], among others.

Despite the versatility of text, diffusion models relying solely on text input encounter challenges in achieving spatial controllability. This hinders fine control over the layout of generated results. Existing methods to tackle this issue mainly fall into two categories: (1) Inputting additional spatial layout entities (e.g. semantic maps [64] or serialised

* Equal Contribution: Z. Qi and G. Huang
† Corresponding author: G. Huang (guoxi.huang@bristol.ac.uk)

bounding boxes [31]) and extra parameterised components through fine-tuning models; (2) Manipulating the attention map through gradient computation aims to enhance the attention score of noise features and text within specific areas [2,9]. Although the former can achieve competitive results with precise spatial alignment, it is noteworthy that they incur substantial computational costs for fine-tuning models and labour costs for data curation. The latter modifies all features simultaneously, presenting a challenge in distinguishing adjacent objects with the same category (see the giraffe example in Fig. 2). Besides, the latter, which directly updates the attention map through gradients, will increase the latency due to frequent backpropagation.

In this paper, our focus is on achieving controllable image synthesis without model re-training or fine-tuning. We propose a universal framework, a two-stage Layered Rendering Diffusion (LRDiff), specifically designed for the above task in a zero-shot paradigm. LRDiff aims to process multiple visual concepts without blending their representations and aligns the results with the input spatial conditions, such as bounding boxes or instant masks. The denoising process in LRDiff is divided into two separate sections. In the first section, we estimate the denoising direction of each object in layers to ensure the accuracy of layouts, employing an innovative concept termed 'vision guidance'. The second denoising section focuses on enriching the texture details and aligning the high-level concepts, guided by the global context of the original caption. Vision guidance, constituting one of the cores of LRDiff, provides a spatial cue for explicitly estimating the denoising direction of each object in layers to ensure the accuracy of its location, shapes, or contour without gradient computation, as illustrated in Fig. 1(a). The implementation of vision guidance empowers LRDiff with zero-shot capabilities for each object and allows adaptation to two common controllable image synthesis tasks, including box-to-image and mask-to-image.

Our experimental results demonstrate that the proposed LRDiff provides excellent spatial controllability for T2I diffusion models while generating photorealistic scene images. Compared to previous methods, including BoxDiff [57], DenseDiffusion [29], and Paint-with-Words (eDiffi-Pww) [2], among others, our results show improved performance, both quantitatively and qualitatively, as demonstrated in Fig. 2 and Fig. 3. The main contributions of this paper are summarised as follows:

- We introduce a universal framework for controllable image synthesis, which is a two-stage layered rendering diffusion model to process multiple visual concepts in layers while aligning the results with the global text.
- We propose visual guidance, independent of the network structure, and incorporate it into each layer to achieve spatial controllability for each object. Vision guidance provides a spatial cue in a zero-shot paradigm without the need for backpropagation.
- Three applications are enabled by the proposed framework: bounding box-to-image, instance mask-to-image, and controllable image editing.

2 Related Works

Text-to-Image (T2I) Models. To adhere to some specifications described by free-form text in image generation, T2I typically models the image distribution along with the encoded latent embeddings of the text prompts as the condition entities via pre-trained lan-

guage models, such as CLIP [44]. Large-scale text-to-image models can be categorised as auto-regressive models [46,11,14,61] and diffusion-based models [38,47,45,49]. Inspired by non-equilibrium statistical physics, Dickstein *et al*. [51] pioneeringly introduced the diffusion model, the concept of which is revisited in Sec. 3. To accelerate training and sampling speed, the latent diffusion model [47], a.k.a. Stable Diffusion (SD) is developed to operate the diffusion process in the latent space [12] instead of the pixel space. However, when it comes to the intricate spatial semantic arrangement of multiple objects within a scene, T2I diffusion models fall short, exhibiting object leakage and a lack of awareness regarding spatial dependencies.

Layout-to-Image (L2I) Diffusion Models. Current layout-guided generation methods can be broadly classified into two main categories based on the necessity of a training process: (1) methods that involve fine-tuning diffusion models [31,59,64,7,8,36,58,42,66,28,25], and (2) training-free approaches [9,43,57,17,4,29,63,40,65,33,67,62,3,30]. The former achieves locality-awareness by incorporating layout information as an additional condition to the pre-trained T2I diffusion model. Methods necessitating fine-tuning, like ControlNet[64], GLIGEN [31] and T2I-Adaptor [36] integrate extra modules into the backbone network. These modules work in concert with spatial control entities to ensure the generated images match the specified spatial conditions. ReCo [59] and GeoDiffusion [7] augment the textual tokens by incorporating new positional tokens arranged in sequences akin to short natural language sentences. However, it's worth noting that these approaches require further training on curated datasets collected with paired annotations, which imposes significant computational and labour costs, bottlenecking applications in an open world. On the other hand, the second group of methods, such as Paint-with-words (eDiffi-Pww) [2] and ZestGuide [9], endows T2I diffusion models with localisation abilities through manipulating the cross-attention maps in the estimator network, amplifying the attention score for the text tokens that specify an object. However, when the initial noise does not tend to generate the target objects, it is difficult to change the direction of denoising by only manipulating the attention map. Besides, these methods are likely to cause the blending of appearances of adjacent objects sharing the same visual concept. This issue is exemplified in the first and second columns of Fig. 2, where the phenomenon is clearly observable.

Image Editing with Diffusion Models. Image editing, as a fundamental task in computer graphics, can be achieved by modifying a real image by inputting auxiliary entities, including scribble [34], mask [1], or reference image [6]. Recent models for text-conditioned image editing [19,26,34,10] harness CLIP [44] embedding guidance combined with pre-trained T2I diffusion models, achieving excellent results across a range of editing tasks. Research in this field primarily advances along three directions: (1) zero-shot algorithms that steer the denoising process towards a desired CLIP embedding direction by manipulating the attention maps of the cross-attention mechanisms [18,39,34,35,27]; (2) textual token vector optimisation [48,15]; (3) fine-tuning T2I diffusion models on curated datasets with matched annotations [6,37]. In contrast to prior works that rely on text prompts to guide the editing, we aim to leverage the proposed vision guidance information to better assist the generation process.

3 Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs). DDPM [21] involves a forward-time diffusion process and a reverse-time denoising process from a prior distribution. Let $\mathbf{x}_0 \in \mathbb{R}^{h\times w\times D}$ be a sample from the data distribution $p_0(\mathbf{x})$. When using a total of T noise scales in the forward-time diffusion process, the discrete Markov chain is

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_{t-1}, \quad t = 1, \cdots, T \quad , \tag{1}$$

where ϵ_t denotes the noise sampled from $\mathcal{N}\big(\mathbf{0}, \mathbf{I}\big)$ at timestep t, $\{\beta_t\}_{t=1}^{T}$ is a pre-defined variance schedule. When recursively applying the noise perturbations Eq. (1) to a real sample $\mathbf{x}_0 \sim p_0(\mathbf{x})$, it ends up with $\mathbf{x}_T \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}\big)$. The reverse-time denoising process can be defined as:

$$\begin{aligned} \mathbf{x}_{t-1} &= \tilde{\alpha}_t\mathbf{x}_t + \underbrace{\tilde{\beta}_t\nabla_{\mathbf{x}}\log p_t(\mathbf{x})}_{\text{direction pointing to } \mathbf{x}_t} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}} \quad , \\ &\approx \tilde{\alpha}_t\mathbf{x}_t + \underbrace{\tilde{\beta}_t\hat{\mathbf{s}}_t}_{\text{estimated direction pointing to } \mathbf{x}_t} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}} \quad . \end{aligned} \tag{2}$$

where $\tilde{\alpha}_t$, $\tilde{\beta}_t$, and σ_t denote the coefficients, the values of which can be derived from β. Practically, the score of the perturbed data distribution, $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$ for all t, can be estimated with a score network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ optimised by using score matching [24,53]. After training to get the optimal solution $\hat{\mathbf{s}}_t = \mathbf{s}^{*}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}}\log p_t(\mathbf{x})$, new samples can be generated by starting from $\mathbf{x}_T \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}\big)$ by recursively applying the estimated reverse-time process. Following *classifier-free guidance* [22], the implicit update direction $\hat{\mathbf{s}}_t$ can be considered in the following form

$$\hat{\mathbf{s}}_t = \gamma\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, c) + (1-\gamma)\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \varnothing), \tag{3}$$

where $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \varnothing)$ is referred to as as an unconditional model, $\gamma \geqslant 1$ controls the guidance strength of a condition $c \in \mathcal{C}$. Trivially increasing γ will amplify the effect of conditional input. Condition space $\mathcal{C}$ can be further defined to be text words, called *text prompt* [38], fundamentalising current T2I models.

4 Method

Given a global text caption and the layout condition (bounding boxes or instant masks), our framework can generate images with accurate spatial alignments in a zero-shot paradigm. The remainder of this section is organised as follows: First, we introduce vision guidance, constituting one of the cores of our framework, acting as a foundational spatial cue for the score estimate network to guide the denoising direction of a single visual concept within a specified region. Subsequently, we detail the image-rendering process of LRDiff, where vision guidance is employed in layers for multiple visual concepts while aligning high-level concepts of the images with the global caption. The overall pipeline is shown in Fig. 1.

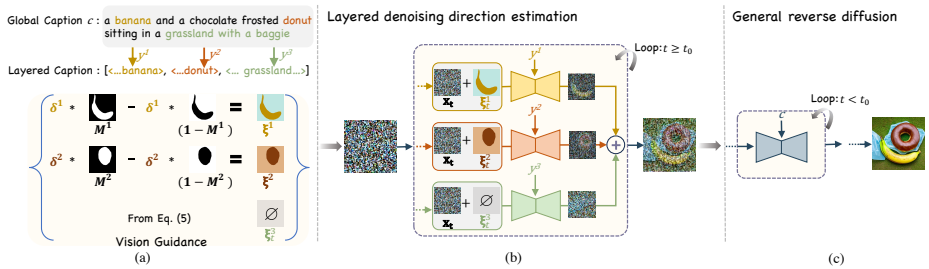

Fig. 1: Overview of our framework. (a) For synthesising a sense, the user provides the global caption, the layered caption, as well as the spatial layout entities which are used to construct the vision guidance. LRDiff divides the reverse-time diffusion process into two sections: (b) When $t \geqslant t_0$, each vision guidance is employed into separate layers to alter the denoising direction, ensuring each object contour generates within specific regions. (c) When $t < t_0$, we perform the general reverse diffusion process to generate texture details that are consistent with the global caption.

4.1 Vision Guidance

We introduce vision guidance, denoted as $\boldsymbol{\xi} \in \mathbb{R}^{h \times w \times D}$, as an additional feature map to the score estimate network, forming $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_{(t)},\ t,\ c,\ \boldsymbol{\xi})$. Furthermore, the vision guidance entities are input into the network in a zero-shot form. A significant advantage of this zero-shot paradigm is that the introduction of the additional condition has no re-training requirement for off-the-shelf conditional diffusion models, thereby substantially reducing computational costs. More analysis of vision guidance can be found in the supplementary material.

The Definition. We factorise the vision guidance into two components: a vector $\boldsymbol{\delta} \in \mathbb{R}^D$ and a binary mask $\boldsymbol{\mathcal{M}} \in \{0,1\}^{h \times w}$. Each element $\xi_{j,k,l}$ of $\boldsymbol{\xi}$ is defined as follows:

$$\begin{aligned} \xi_{j,k,l} &= \delta_l \cdot \mathcal{M}_{j,k} - \delta_l \cdot (1 - \mathcal{M}_{j,k}), \\ &= \delta_l \cdot (2\mathcal{M}_{j,k} - 1), \end{aligned} \tag{4}$$

where $\mathcal{M}_{j,k}$ is assigned the value 1 if the spatial position (j,k) falls within the expected object region. For the region containing an object, we add $\boldsymbol{\delta}$ to enhance the generation tendency of that object. Conversely, for areas outside the target region, we subtract $\boldsymbol{\delta}$ to suppress the generation tendency of the object. The binary mask $\boldsymbol{\mathcal{M}}$ can be derived from user input, such as converting a bounding box or instance mask provided by the user into the binary mask. Next, we introduce two distinct approaches to compute the vector $\boldsymbol{\delta}$.

`Constant vector:` A naïve approach for the configuration is to set the vector $\boldsymbol{\delta}$ to some constant values. When the diffusion model operates in the RGB space, we can set $\boldsymbol{\delta}$ to constant values corresponding to some colour described by the text prompt (*e.g.*, $[0.3, 0.3, 0.3]$ corresponding to a white colour with transparency). When operating in the latent space of VAE, $\boldsymbol{\delta}$ can be set to the latent representation of the constant values when operations such as dimension expansion and tensor repeat are required. Although the manual adjustment of $\boldsymbol{\delta}$ to some constant values is versatile for generating objects with various visual concepts, it necessitates human intervention, such as defining the colour of the object.

`Dynamic vector`**:** Beyond simply assigning constant values to $\boldsymbol{\delta}$, we propose to dynamically adapt the values of $\boldsymbol{\delta}$ based on the input text conditions in order to reduce human intervention during generation. In this context, we consider the implementation of Stable Diffusion [47] wherein text tokens are interconnected with the visual features via cross-attention modules. At the initial denoising step, *i.e.*, $t = T$, we extract the cross-attention map $\mathbf{A} \in \mathbb{R}^{|c| \times hw}$ from an intermediate layer in the U-Net. For a more straightforward illustration, we will consider the synthesis of an image containing a single object, corresponding to the i-th text token from the text prompt c. Subsequently, to derive the vector $\boldsymbol{\delta}$ in Eq. Eq. (4), we perform the following operations:

$$\begin{aligned} \boldsymbol{S} &= \{(j,k) | \mathbf{A}^i_{j,k} > \text{Threshold}_K(\mathbf{A}^i)\}, \\ \boldsymbol{\delta} &= \frac{\lambda}{|\boldsymbol{S}|} \sum \{\mathbf{x}_t(j,k) | (j,k) \in \boldsymbol{S}\}, \end{aligned} \tag{5}$$

where $\mathbf{x}_t(j,k)$ denotes the element at spatial location (j,k) in $\mathbf{x}_t$. The $\sum$ operation sums up all items within the $\boldsymbol{S}$ set. Additionally, the operation $\text{Threshold}_K(\cdot)$ selects the K-th largest value from the top K values in $\mathbf{A}^i$. The strength of vision guidance is modulated by the coefficient λ, alongside the classifier-free guidance coefficient γ. Given the presence of multiple cross-attention blocks within the score network, we opt to select the block following the down-sampling in each stage and subsequently average their outputs.

4.2 Layered Rendering

Considering an upcoming image drawing n objects, we encapsulate all the condition entries of the diffusion model into the following

$$\begin{aligned} \text{Global Caption}: &\quad c, \\ \text{Layered Captions}: &\quad [y^1(c), \cdots, y^n(c)], \\ \text{Vision Guidance}: &\quad [\boldsymbol{\xi}^1, \cdots, \boldsymbol{\xi}^n], \\ \text{Layered Masks}: &\quad [\boldsymbol{\mathcal{M}}^1, \cdots, \boldsymbol{\mathcal{M}}^n], \end{aligned} \tag{6}$$

where we define a set of mappings $y^i : \mathcal{C} \to \mathcal{Y}$, and $y^i(c)$ [4] represents the text condition for layer i. $\boldsymbol{\xi}^i$ is the vision guidance for the i_{th} object constructed by the $\boldsymbol{\mathcal{M}}^i$ and δ^i. Appendix A from our supplementary material details the implementation of layered caption construction.

As mentioned in Sec. 1, our layered rendering algorithm divides the full reverse-time denoising process into two denoising sections, *i.e.* $[T, \cdots, t_0]$, $[t_0{-}1, \cdots, 1]$. At each timestep $t \in [T, \cdots, t_0]$, the denoising process is given by

$$\mathbf{x}_{t-1} = \tilde{\alpha}_t \mathbf{x}_t + \tilde{\beta}_t [\gamma \, \boldsymbol{\Phi}_t + (1-\gamma) \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \varnothing)] + \sigma_t \epsilon_t, \tag{7a}$$

$$\boldsymbol{\Phi}_t = \sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t + \boldsymbol{\xi}^i, t, y^i). \tag{7b}$$

[4] We denote $y^i(c)$ by y^i for simplifying notations.

As shown in Fig. 1, we construct $\boldsymbol{\Phi}_t$ by fusing estimated noises in each layer and sending it into the next denoising loop. Equation 7 assures that the prediction $\mathbf{x}_{t-1}$ for any step from $[T, \cdots, t_0]$ is within the data distribution $p_t(\mathbf{x})$, so that the score network $\mathbf{s}_{\boldsymbol{\theta}}$ needs no fine-tuning. The derivation process of Equation 7 is provided in the supplementary material. We carry out the layered generative process along with vision guidance as expressed in Equation 7 for $[T, \cdots, t_0]$. For timesteps in the second denoising section, *e.g.*, $[t_0{-}1, \cdots, 1]$, we perform the standard denoising process but with the global caption c as the solid condition information, illustrated in the third column in Fig. 1. The overall generative process is described by Algorithm 1. The framework of our method is illustrated in Fig. 1. We assign a null value $\varnothing$ to the vision guidance $\boldsymbol{\xi}^n$ in the final layer (*e.g.*, the 3$^{\text{rd}}$ layer in Fig. 1), encompassing scene descriptions, such as 'a beautiful grassland' or other objects without defined layouts.

Algorithm 1 Layered Rendering Diffusion

Input: c, $[y^1, \cdots, y^n]$, $[\boldsymbol{\mathcal{M}}^1, \cdots, \boldsymbol{\mathcal{M}}^n]$,
  Pre-trained diffusion model $\mathbf{s}_{\boldsymbol{\theta}}$;
  Initialise $\mathbf{x}_T$ ▷ Noise initialisation
  $[\boldsymbol{\xi}^1, ..., \boldsymbol{\xi}^n] \leftarrow$ Calculate Eq. (5)
for $t = T, \cdots, t_0$ **do**
  ▷ Estimate layered denoising direction
    $\boldsymbol{\Phi}_t \leftarrow$ Calculate Eq. (7b)
    $\mathbf{x}_{t-1} \leftarrow$ Calculate Eq. (7a)
    $\mathbf{x}_t = \mathbf{x}_{t-1}$
for $t = t_0{-}1, \cdots, 1$ **do**
  ▷ Estimate general denoising direction
    $\hat{\mathbf{s}}_t = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \gamma\big(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathcal{C}) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big)$
    $\mathbf{x}_{t-1} = \tilde{\alpha}_t \mathbf{x}_t + \tilde{\beta}_t \hat{\mathbf{s}}_t + \sigma_t \epsilon_t$
    $\mathbf{x}_t = \mathbf{x}_{t-1}$
Output: $\mathbf{x}_0$

5 Experiments

Dataset and Implementation Details. All the experiments are run on a single NVIDIA Tesla V100. Unless specified otherwise, we use the DDIM sampler [52] with 50 sampling steps for the reverse diffusion process with a fixed guidance scale of 7.5; t_0 is set to 15 by default. We construct our dataset by selecting 1134 captions with one or more objects and corresponding bounding boxes or instance masks from the MS-COCO validation set [32]. For a fair comparison, we implement LRDiff based on a diffusion model with a version similar to that used by other methods, aiming to eliminate differences in results caused by variations in the capabilities of the foundational diffusion model.

Evaluation Metrics. For evaluating synthesised images with both bounding box and instance mask inputs, we employ two distinct metrics: image-score and align-score. The image-score specifically measures the fidelity of the synthesised image to the text prompt, incorporating sub-indicators such as T2I-Sim [57] and the CLIP score [20] for a nuanced assessment. On the other hand, the align-score evaluates the image's spatial alignment with the given layout condition, using the AP results predicted by YOLOv4

as a benchmark for alignment accuracy. Additionally, regarding instance mask inputs, we assess the precision of object contours using the IoU scores produced by employing YOLOv7 [54] and the ground truth.

Bounding Box	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	mAP↑	AP_{50}↑	AP_{75}↑
SD [47]	0.292	0.316	-	-	-
TwFA [60]	0.210	0.179	9.9	16.3	9.0
eDiffi-Pww [2]	0.279	0.299	4.2	8.6	4.0
BoxDiff [57]	0.295	0.319	5.8	17.2	3.0
LRDiff(Ours)	0.281	0.292	**17.4**	**35.6**	**15.5**

Table 1: The quantitative results of bounding box input.

Instance Mask	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	AP_{50}↑	AP_{75}↑	IOU↑
SD [47]	0.292	0.316	-	-	-
eDiffi-Pww [2]	0.287	0.304	16.5	6.4	33.11
MultiDiff [4]	0.277	0.281	30.9	8.2	47.59
DenseDiff. [29]	0.289	0.310	11.3	1.3	27.65
LRDiff(Ours)	0.280	0.295	**35.0**	**15.4**	**49.06**

Table 2: The quantitative results of instance mask input.

5.1 Bounding boxes as layout condition

Quantitative comparison. In Table 1, we present a comparative analysis of our method against BoxDiff [57], eDiffi-Pww [2], and TwFA [60]. Our evaluation revealed a trade-off relationship between the align-score and image-score among the compared methods. For instance, while TwFA [60] achieved superior alignment scores compared to other diffusion methods, its limited image generation capabilities led to lower-quality generated backgrounds. Consequently, this facilitated easier foreground differentiation by YOLO [5], resulting in higher detection metrics. To ensure fairness in comparisons, we established SD [47] as the baseline for the image-score. Notably, our results exhibited higher AP values in contrast to eDiffi-Pww and TwFA, while closely aligning with the baseline in terms of image-score. Furthermore, our image-score values are close to those of BoxDiff, but our three alignment score sub-metrics exceed BoxDiff by 11.6%, 17.6%, and 12.5%, respectively.

Qualitative comparison. Fig. 2 presents qualitative comparisons among various methods in a multi-object layout. According to the results, LRDiff can effectively mitigate the issue of visual blending between adjacent objects within the same category, which is a challenge for other methods such as BoxDiff [57] and eDiffi-Pww [2]. This effectiveness is exemplified in the synthesis of giraffes, as shown in the second column of Fig. 2. Furthermore, our method surpasses eDiffi-Pww when synthesising images within intricate layouts. Notably, in the third column, where a cat, bed and laptop are closely arranged, our LRDiff proficiently handles mutual occlusion, underscoring its capability to manage complex scene compositions. Additionally, our approach shows high fidelity in generating small-scale objects, which remains challenging for the listed methods that rely on manipulating cross-attention maps to control layout.

5.2 Instance masks as layout condition

Quantitative comparison. The instance masks, serving as layout guides, provide both positional and contour details. As highlighted in Table 2, our approach outperforms other diffusion-based zero-shot techniques [29,4,2] in the IoU metric. Similar to the analysis in

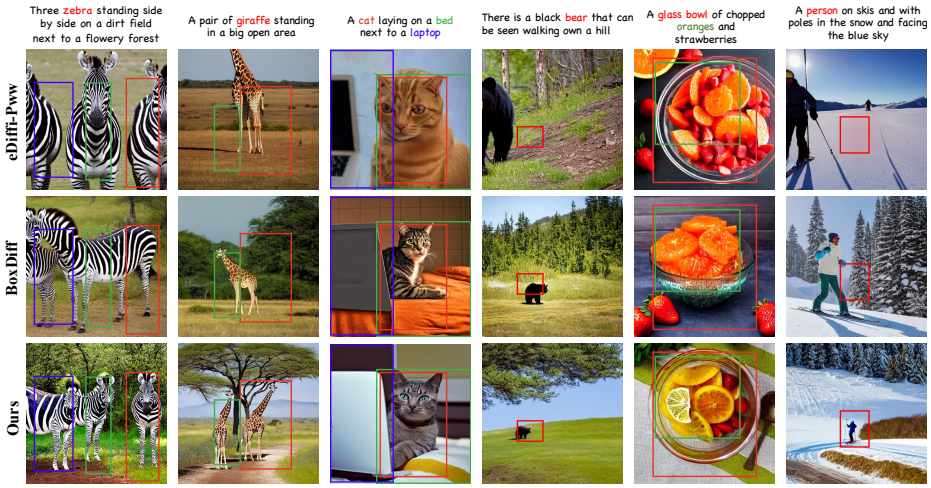

Fig. 2: Qualitative comparisons of methods that use bounding box entities as the spatial condition. Our results show better spatial alignments than other methods.

subsection. 5.1, there is a trade-off correlation between the image-score and align-score. Our results significantly outperform eDiffi-Pww and DenseDiff by over 16% and 22%, respectively, in the IOU metric while closely aligning with their image-score outcomes. The observed low image-score in MultiDiff [4] is attributed to its limited interaction with global captions, leading to a 'copy-paste-like' phenomenon. In contrast, our method integrates all layers and interacts with global captions after delineating object outlines in layers. Consequently, compared to MultiDiff [4], our method achieves a closer alignment with the baseline image-score, surpassing it by 1.6% in the mIoU metric. Furthermore, substantial enhancements in both AP_{50} and AP_{75} indicators by over 9% each signify the accurate alignment and positioning of generated objects within the specified mask area, validating the efficacy of our approach.

Qualitative comparison. Fig.3 presents qualitative comparisons between our method and others across both single and multi-object layouts. The results highlight our effectiveness in generating small-scale objects, which is still a challenge for DenseDiff [29]. Examples to justify our effectiveness are provided in the first and fifth columns of Fig.3 where the synthesised elephants and zebras are faithful to the shapes provided in the layout entity. MultiDiff [4] demonstrates high effectiveness in achieving precise spatial alignment within images. However, it exhibits limitations in harmonising the overall image composition, occasionally resulting in a 'copy-paste' result. This inadequacy is particularly evident in the first and last columns of the figure, where an elephant appears inserted into a tree and a toilet lacks seamless integration with the surrounding environment. Conversely, our method ensures precise image layout accuracy by employing visual guidance and achieves seamless integration throughout the entire image due to the fusion of all layers and interaction with the global captions.

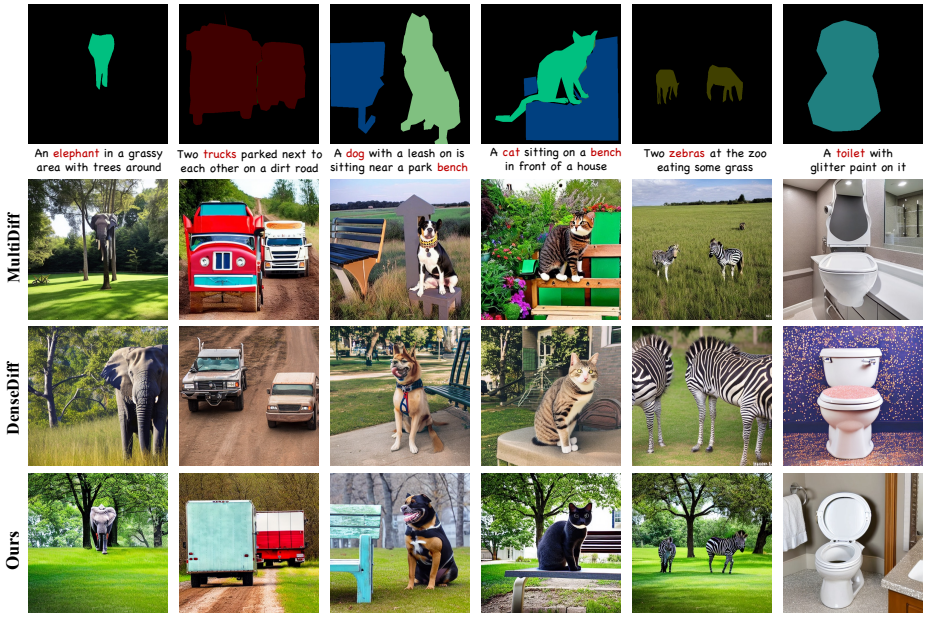

Fig. 3: Qualitative comparisons of methods that use instance mask entities as the spatial condition. Our results show better spatial alignments than other methods.

5.3 Ablation Study

In this section, we conduct experiments primarily to demonstrate the necessity of our designed vision guidance and the different impacts on different ways of calculating δ_l. Additionally, we showcase the impacts of our method across diverse t_0 values. Further ablation experiments and more results are provided in the supplementary material.

The necessity of vision guidance. To investigate the necessity of visual guidance, we conduct an experiment by removing the visual guidance within the target area and only suppressing the vision guidance outside the target area. Correspondingly, we modify Eq. 4 as follows:

$$\xi_{j,k,l} = -\delta_l \cdot (1 - \mathcal{M}_{j,k}). \tag{8}$$

To write concisely, we name the two ways of constructing $\boldsymbol{\xi}$ using Eq. 8 and Eq. 4 respectively as setting #1 and setting #2. First, the qualitative difference between the two ways can be found in Fig. 4 and Fig. 5. The results show that suppressing visual guidance outside the target region can effectively eliminate the tendency to generate objects. However, this does not relatively enhance the tendency to generate objects within the target region, as the features in the unmodified region may inherently lack the capability to generate objects. Interestingly, the use of setting #2, as demonstrated in Fig. 5, has a certain effectiveness on simple objects, such as pizza, where realism is lacking. This is attributed to the mask providing certain additional shape information. Furthermore, Tables 3 and 4 present a direct comparison between the results obtained with and without vision guidance in the target area. The outcomes strongly indicate that the absence of

Bounding Box	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	mAP↑	AP_{50}↑	AP_{75}↑
setting #1	0.1967	0.1642	1.3	3.8	0.7
setting #2	0.281	0.295	17.4	35.6	15.5

Table 3: These results depict the outcome when vision guidance is not applied within the target areas under the bounding box condition.

Instance Mask	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	AP_{50}↑	AP_{75}↑	IOU↑
setting #1	0.2599	0.2767	5.2	0.8	28.16
setting #2	0.280	0.295	35.0	15.4	49.06

Table 4: These results depict the outcome when vision guidance is not applied within the target areas under the instance mask condition.

such guidance renders the results nearly unusable. Considering the findings from these tables and figures, it can be concluded that vision guidance significantly impacts the effectiveness of the generated outcomes.

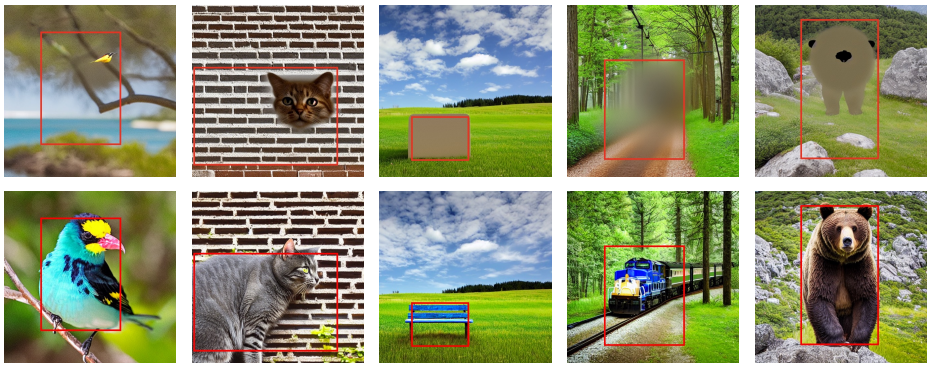

Fig. 4: The bounding boxes condition. The first row shows the results using setting #1 and the second row shows the results using setting #2.

Fig. 5: The instant mask condition. The different results of using setting #1 and setting #2.

Constant vs. Dynamic. We visually compared two approaches for calculating the vector $\boldsymbol{\delta}$ in vision guidance (Fig. 6). In constant mode, if the prompt does not specify colour (the first column), $\boldsymbol{\delta}$ defaults to brown, which is a common colour for horses. However, this may mislead, as brown becomes associated with roads instead of horses. The reason is that 'brown information' is not effectively associated with the concept of a horse without explicit cue. The constant mode excels when colour is explicitly mentioned (second column). On the contrary, the dynamic vector strategy consistently aligns with

the layout, making it preferable for real-world applications, and eliminating the need for human intervention in specifying object colours.

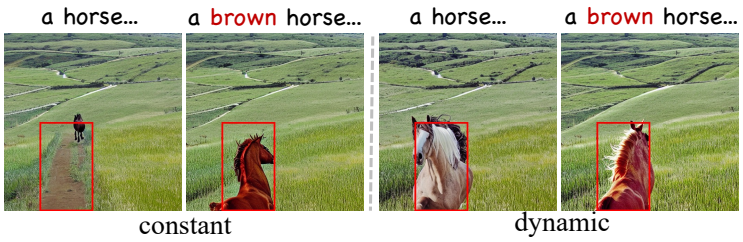

Fig. 6: The dynamic vector strategy shows high spatial alignment to the given layout, eliminating the need for human intervention .

Constructing δ with K random vectors. From the above experiments, we observe that dynamic vision guidance serves as an effective cue for providing layout information. Furthermore, we present an alternative implementation of δ in dynamic mode, utilising K random vectors. Specifically, for each position in the object area, we fill the position with a vector randomly sampled from set $\mathbf{S}$ from Eq. 5. We denote this as 'random-vectors', distinguishing it from the 'mean-vector' method mentioned in Sec. 4.1. Table 5

Bounding Box	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	mAP↑	AP_{50}↑	AP_{75}↑
random-vectors$_{10}$	0.278	0.293	17.0	30.0	13.9
mean-vector$_{10}$	0.284	0.298	17.2	33.2	14.6
random-vectors$_{20}$	0.272	0.287	17.1	32.9	14.5
mean-vector$_{20}$	0.281	0.295	17.4	35.6	15.5

Table 5: Comparison of 'random-vectors' and 'mean-vector' with bounding box entities. The grey line indicates the results in Table 1. The subscript denotes the choice of value K.

Instance Mask	Image-Score		Align-Score		
	T2I-Sim↑	CLIP↑	AP_{50}↑	AP_{75}↑	IOU↑
random-vectors$_{10}$	0.281	0.296	36.5	19.2	49.02
mean-vector$_{10}$	0.277	0.287	37.9	19.8	49.41
random-vector$_{20}$	0.278	0.292	40.5	21.0	49.54
mean-vector$_{20}$	0.280	0.295	35.0	15.4	49.06

Table 6: Comparison of 'random-vectors' and 'mean-vector' with instance mask entities. The grey line shows the results in Table 2. The subscript denotes the choice of value K.

and Table 6 highlight the remarkable similarity in results obtained through these two different approaches to constructing δ. The 'mean-vector' method better aligns with the input text compared to the 'random-vectors'. The results also show that K is set to 10, the image-score outperforms that achieved when K is set to 20. However, to ensure a balanced performance between image-score and align-score, we opt for the 'mean-vector' method with K set to 20. The 'random-vectors' approach exhibits advantages in achieving texture diversity among multiple objects of the same category within a scene. Illustrated in Fig. 7, an example clarifies the divergent visual effects of employing these two approaches. However, in most cases, the 'mean-vector' method suffices entirely.

The impact of different t_0. LRDiff divides the reverse diffusion process into two denoising sections: $[T, \cdots, t_0]$, $[t_0-1, \cdots, 1]$. As we analyse in Fig. 8, a longer first

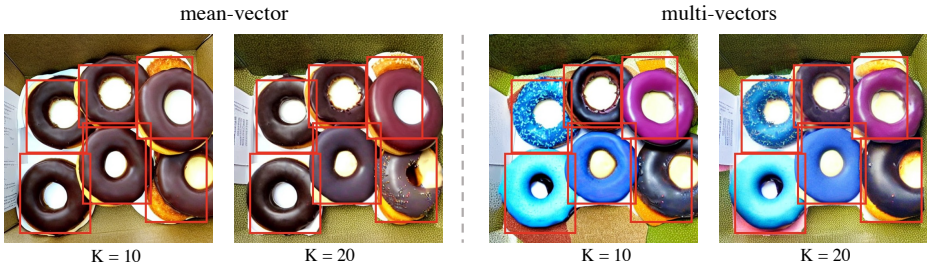

Fig. 7: The varying impacts of employing different construction methods. The generated images with 'random-vectors' are more diverse in terms of textures and patterns for different objects.

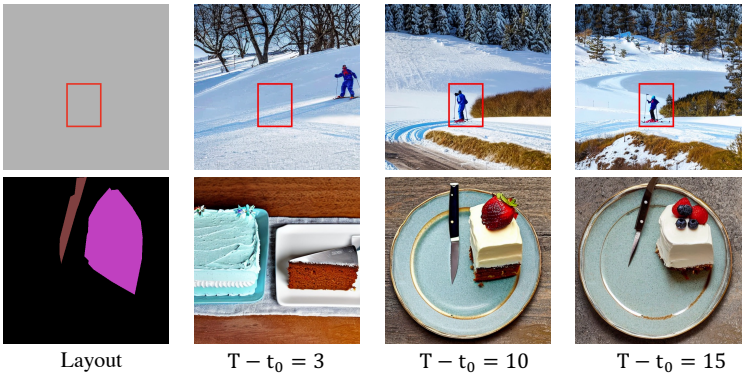

Fig. 8: The impacts of different t_0. A short first denoising period can result in less precise spatial alignment

denoising period (*i.e.*, $|T - t_0|$) will generate images of higher spatial aliment with the given spatial layout. However, a short first denoising period can result in less precise spatial alignment. We attribute this raised deviation to the ambiguity in object shapes during the early denoising steps, where they are particularly susceptible to noise from other layers. We observe that setting $|T - t_0| = 15$ usually gives some decent results. The lower part in Fig. 8 shows that when setting $|T - t_0| = 15$ the knife's shape fits better than the others.

6 Other applications

We extend our framework to controllable image editing. We employ the DDIM inversion technique [52] to obtain the noise latent representation. This inverted latent noise serves as the final layer (*e.g.*, the 3^{rd} layer in Fig. 1) in our pipeline, enabling further processing. As demonstrated in Fig.9, our method allows inserting or replacing objects of various sizes at different locations within a provided image. Thanks to our layered rendering technology, the edited images seamlessly fit the given prompts, while keeping other areas of content as unchanged as possible.

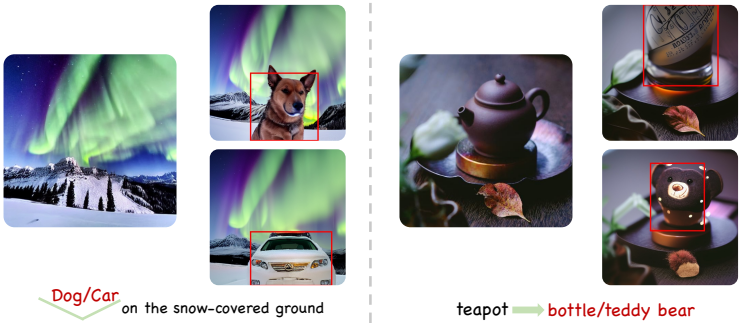

Fig. 9: Example results of controlled image editing. Insert or replace an object at a specific location.

7 Dissuasion and limitations

While the proposed framework has delivered refined controllability for layout-to-image tasks, certain limitations exist that necessitate further exploration. First, the spatial conditional inputs are limited to bounding boxes and object masks. We could extend our framework to include other layout conditions, such as key points refining human poses and depth maps providing 3D information about the scene. Additionally, we lack discussion of the sensitivity of the t_0 to different types of scenes, such as indoor and outdoor scenes. In our future work, we may address this by using learnable t_0 by optimising reward model [13] to enhance the robustness of our method across various scenes. In our experiments, we construct our framework with SD v1.5. Moving forward, we can explore our method with other versions of diffusion models with enhanced capabilities [41] or higher speed [50]. Another promising direction involves customised generation under additional layout conditions by combining with a personalisation method, such as DreamBooth [48].

8 Conclusion

In this paper, we introduce a universal framework designed to generate results aligned with the input spatial layout conditions while avoiding the blending of multiple visual concepts. The proposed framework, termed LRDiff, comprises a two-stage Layered Rendering Diffusion, establishing an image-rendering process with multiple layers. It incorporates an innovative concept known as vision guidance, playing a crucial role in achieving precise spatial alignment for individual objects in a zero-shot paradigm. The effectiveness of our method has been demonstrated through the experiments, highlighting the superior capabilities of our framework. Additionally, we explored various approaches to constructing vision guidance. Our technology finds applications in three domains: bounding box-to-image, instance mask-to-image, and controllable image editing. We also extended our framework to controllable image editing.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022) 3
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022) 2, 3, 8
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 843–852 (2023) 3
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. arXiv preprint arXiv:2302.08113 (2023) 3, 8, 9
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020) 8
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 3
7. Chen, K., Xie, E., Chen, Z., Hong, L., Li, Z., Yeung, D.Y.: Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. arXiv preprint arXiv:2306.04607 (2023) 3
8. Cheng, J., Liang, X., Shi, X., He, T., Xiao, T., Li, M.: Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. arXiv preprint arXiv:2302.08908 (2023) 3
9. Couairon, G., Careil, M., Cord, M., Lathuilière, S., Verbeek, J.: Zero-shot spatial layout conditioning for text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2174–2183 (2023) 2, 3
10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022) 3
11. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021) 3
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 3
13. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024) 14
14. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022) 3
15. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 3
16. Gu, J., Trevithick, A., Lin, K.E., Susskind, J.M., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In: International Conference on Machine Learning. pp. 11808–11826. PMLR (2023) 1
17. He, Y., Salakhutdinov, R., Kolter, J.Z.: Localized text-to-image generation for free via cross attention control. arXiv preprint arXiv:2306.14636 (2023) 3
18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 3

19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb 3
20. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 7
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 4
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 1
24. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005) 4
25. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 3
26. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) 3
27. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (2022) 3
28. Kim, S., Lee, J., Hong, K., Kim, D., Ahn, N.: Diffblender: Scalable and composable multimodal text-to-image diffusion models. arXiv preprint arXiv:2305.15194 (2023) 3
29. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023) 2, 3, 8, 9
30. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. arXiv preprint arXiv:2310.07771 (2023) 3
31. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023) 2, 3
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 7
33. Ma, W.D.K., Lewis, J., Kleijn, W.B., Leung, T.: Directed diffusion: Direct control of object placement through attention guidance. arXiv preprint arXiv:2302.13153 (2023) 3
34. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021) 3
35. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023) 3
36. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023) 3
37. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: Image editing via visual prompting. Advances in neural information processing systems (2023) 3

38. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 3, 4
39. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 3
40. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023) 3
41. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 14
42. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147 (2023) 3
43. Qu, L., Wu, S., Fei, H., Nie, L., Chua, T.S.: Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 643–654 (2023) 3
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 3
46. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 1, 3
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3, 6, 8
48. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023) 3, 14
49. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 3
50. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 14
51. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 3
52. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 7, 13
53. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 4
54. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7464–7475 (2023) 8
55. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems (2023) 1

56. Wynn, J., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4180–4189 (2023) 1
57. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023) 2, 3, 7, 8
58. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023) 3
59. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023) 3
60. Yang, Z., Liu, D., Wang, C., Yang, J., Tao, D.: Modeling image composition for complex scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7764–7773 (2022) 8
61. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022) 3
62. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 3
63. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22468–22478 (2023) 3
64. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1, 3
65. Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X.: Controllable text-to-image generation with gpt-4. arXiv preprint arXiv:2305.18583 (2023) 3
66. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv:2305.16322 (2023) 3
67. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023) 3