

LAW-Diffusion: Complex Scene Generation by Diffusion with Layouts

Binbin Yang¹ Yi Luo¹ Ziliang Chen² Guangrun Wang³ Xiaodan Liang¹ Liang lin^{1*}

¹Sun Yat-Sen University ²Jinan University ³University of Oxford

yangbb3@mail2.sysu.edu.cn, luoy97@mail3.sysu.edu.cn, c.ziliang@yahoo.com,

{wanggrun, xdliang328}@gmail.com, linliang@ieee.org

Abstract

Thanks to the rapid development of diffusion models, unprecedented progress has been witnessed in image synthesis. Prior works mostly rely on pre-trained linguistic models, but a text is often too abstract to properly specify all the spatial properties of an image, e.g., the layout configuration of a scene, leading to the sub-optimal results of complex scene generation. In this paper, we achieve accurate complex scene generation by proposing a semantically controllable Layout-Aware diffusion model, termed LAW-Diffusion. Distinct from the previous Layout-to-Image generation (L2I) methods that primarily explore category-aware relationships, LAW-Diffusion introduces a spatial dependency parser to encode the location-aware semantic coherence across objects as a layout embedding and produces a scene with perceptually harmonious object styles and contextual relations. To be specific, we delicately instantiate each object’s regional semantics as an object region map and leverage a location-aware cross-object attention module to capture the spatial dependencies among those disentangled representations. We further propose an adaptive guidance schedule for our layout guidance to mitigate the trade-off between the regional semantic alignment and the texture fidelity of generated objects. Moreover, LAW-Diffusion allows for instance reconfiguration while maintaining the other regions in a synthesized image by introducing a layout-aware latent grafting mechanism to recompose its local regional semantics. To better verify the plausibility of generated scenes, we propose a new evaluation metric for the L2I task, dubbed Scene Relation Score (SRS) to measure how the images preserve the rational and harmonious relations among contextual objects. Comprehensive experiments on COCO-Stuff and Visual-Genome demonstrate that our LAW-Diffusion yields the state-of-the-art generative performance, especially with coherent object relations.

*Corresponding author.

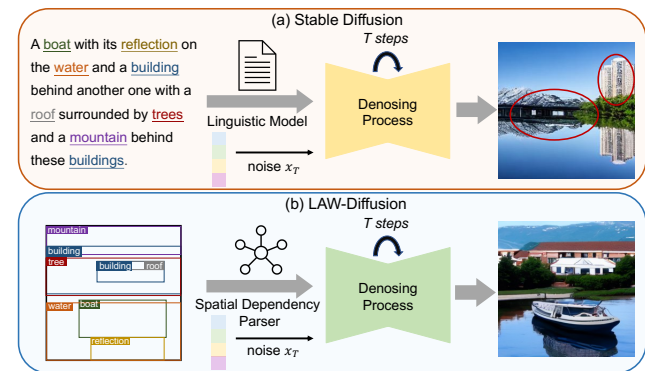


Figure 1. Illustration of complex scene generation by Stable Diffusion [28] (text-to-image model) and our LAW-Diffusion (layout-to-image model). Stable Diffusion relies on linguistic model and generates an unsatisfactory scene: the boat on the water is missed and the generated building and mountain are placed with undesired spatial relation according to the input description. By contrast, LAW-Diffusion introduces a spatial dependency parser to encode the spatial semantic coherence and produces the scene image with consistent contextual relations adhere to the layout configuration.

1. Introduction

Recently, astounding advances have been achieved in generative modeling due to the emergence of diffusion models [34, 13, 28, 42, 1, 6]. Despite the stunning generative performance in simple cases, e.g., single object synthesis, how to generate a complex scene composed of multiple visual concepts with their diverse relationships remains a challenging problem. A straightforward solution is to translate the scene into a text description and then resort to the state-of-the-art text-to-image (T2I) generative models [28, 6, 7, 31, 26]. However, text-to-image diffusion models, e.g., Stable Diffusion and its variants [28, 6, 7, 31, 26] fall short when it comes to the spatial composition of multiple objects in a scene. An underlying reason is that properly specifying all the spatial properties in an abstractive sentence is laborious and less accurate, usually resulting in unsatisfactory generated results. In addition, the linguistic model used in T2I model is incapable of accurately capturing the objects’

spatial relations whereas only providing a coarse-grained linguistic understanding from the text description. An example is shown in Fig. 1, in which we extract a sentence description from a scene layout configuration and compare the generated results of Stable Diffusion [28] and our model. From the result generated by Stable Diffusion in Fig. 1(a), we can observe that the spatial properties are not well preserved (*e.g.*, the generated mountain is besides the building while it should be behind the building) and some desired objects are missed (*e.g.*, the boat and its reflection). By contrast, our method generates the scene image by directly parsing the spatial dependency in the layout configuration.

Layout-to-image generation (L2I) is a very important task of controllable image synthesis, which takes a configuration of visual concepts (*i.e.*, objects’ bounding boxes with their class labels in a certain spatial layout) as the input. The scene layout precisely specifies each object’s size, location and its association to other objects. The key challenge for L2I lies in encoding the spatial dependencies among co-existing objects at each position, *i.e.*, the location-aware semantic composition, which is vital to eliminate the artifacts of spurious edges between spatial adjacent or overlapped objects [11]. Existing studies on L2I are usually developed based on the generative adversarial networks (GAN) [9, 37, 11, 38, 44]. These methods render the realism of image contents with instance-specific style noises and discriminators, and thus suffer from the lack of overall harmony and style consistency among things and stuffs in the generated scene. They have made a few attempts to capture the class-aware relationships in the generator by adopting LSTM [44] or attention mechanism [11]. Another type of approaches is based on transformer [16, 41], which reformulates the scene generation task as a sequence prediction problem by converting the input layout and target image into a list of object tokens and patch tokens. The transformer [40] is then employed to sequentially predict the image patches, which actually capture the sequential dependencies rather than scene coherence. Recently, generic T2I diffusion models, *e.g.*, LDM [28] and Frido [6] have been demonstrated that they can be extended to L2I by tokenizing the layout into a sentence-like sequence of object tokens and encoding them by linguistic model, following their standard T2I paradigm. Such brute-force solutions share some shortcomings inherent to the T2I diffusion models, *e.g.*, the aforementioned object leakage and unawareness of spatial dependencies in 1(a). But in fact, prior methods mainly exploit the location-insensitive relationships while overlooking the fine-grained location-aware cross-object associations.

To address the above issues, we present a novel diffusion model-based framework for L2I, termed *LAW-Diffusion*, for synthesizing complex scene images with mutually harmonious object relations. Unlike the traditional L2I methods

treating each object separately, our *LAW-Diffusion* learns a layout embedding with rich regional composition semantics in a delicate manner for better exploring the holistic spatial information of objects. Concretely, we first instantiate each object’s regional semantics as an object region map that encodes the class semantic information in its bounding box. Then, we split those region maps into fragments and propose a location-aware cross-object attention module to perform per-fragment multi-head attention with a learnable aggregation token to exploit the location-aware composition semantics. By regrouping those aggregated fragments according to their original spatial locations, we obtain a layout embedding encapsulating both class-aware and location-aware dependencies. In this way, when synthesizing a local fragment of image, such composed semantics faithfully specify whether objects are possibly overlapped at the certain location. Inspired by the effectiveness of text-to-image diffusion models [26, 31, 24], we employ the form of classifier-free guidance [14] to amplify the regional control from our layout embedding. To avoid losing objects’ texture details when leveraging a large guidance scale, we further propose an adaptive guidance schedule for the sampling stage of *LAW-Diffusion* to maintain both layout semantic alignment and object’s texture fidelity by gradually annealing the guidance magnitude. Furthermore, *LAW-Diffusion* allows for instance reconfiguration, *e.g.*, adding/removing/restyling an instance in a generated scene via layout-aware latent grafting. Specifically, we spatially graft an exclusive region outside a bounding box from the diffusion latent of the already generated image onto the target latent guided by a new layout at the same noise level. By alternately recomposing the local regional semantics and denosing these grafted latents, *LAW-Diffusion* can reconfigure an instance in a synthesized scene image while keeping the other objects unchanged.

The existing evaluation metrics for the L2I task basically focus on measuring the fidelity of generated objects while ignoring the coherence among objects’ relations in the scene context. Thus, we propose a new evaluation metric called Scene Relation Score (SRS) to measure whether the generated scenes preserve the rational and harmonious relations among contextual objects, which would facilitate the development of L2I research. We conduct both quantitative and qualitative experiments on Visual Genome [17] and COCO-Stuff [2], and the experimental results demonstrate that our *LAW-Diffusion* outperforms other L2I methods and achieves the new state-of-the-art generative performance, particularly in preserving reasonable and coherent object relations.

2. Related Work

Diffusion Models Diffusion models [34, 13, 21, 20, 32] recently emerges as powerful image generators due to their

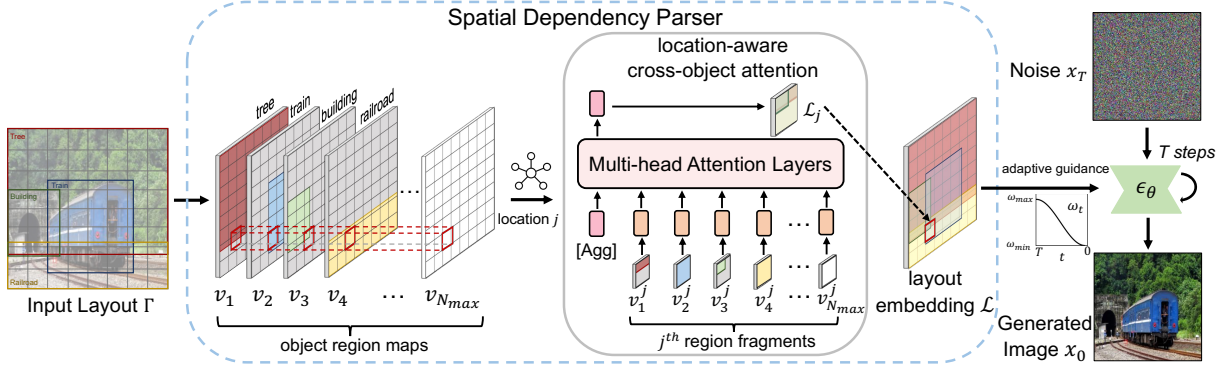


Figure 2. An overview of LAW-Diffusion. Given an input layout Γ , each object’s region map v_i is generated as its regional semantics by filling its class embedding into the region specified by its bounding box. The object region maps are split into patches of region fragments. For the region fragments at the location j , the location-aware cross-object attention module is used to aggregate them as \mathcal{L}_j via multi-head attention. In this way, \mathcal{L}_j encodes the spatial dependencies among objects at this location. Furthermore, the layout embedding \mathcal{L} is obtained by collecting all aggregated fragments and used to control the generation of LAW-Diffusion with an adaptive guidance schedule: the guidance magnitude ω_t gradually anneals from ω_{\max} to ω_{\min} during denoising process. Best viewed in color.

impressive generative performance. By training a noise estimator, the generative process of diffusion model is formulated as iteratively denoising from an image-level noise [13, 4]. With the introduction the techniques of classifier guidance [4] and classifier-free guidance [14], diffusion models are enabled to incorporate different types of conditional information during the sampling stage. Most recent progresses [1, 42, 6, 30, 8, 7, 31, 26] are made in the field of text-to-image (T2I) generation because the prevalence of Stable Diffusion [28]. However, those T2I diffusion models always fall short when it comes to the complex spatial semantic composition of multiple objects in a scene. In this paper, we manage to present a layout-aware diffusion model for complex scene image generation, by mining the spatial dependencies among co-existing objects in the scene layout.

Layout-to-Image Generation Image generation from a layout configuration (L2I) is a specific task of conditional image generation, whose input is a set of bounding boxes and class labels of the objects in a scene. It liberates people from racking their brains to formulate an accurate but complicated language description of a complex scene and rather provides a more flexible human-computer interface for scene generation. Layout2Im [44] generated objects’ features from noises and class labels, and fused them by LSTM [15]. LostGAN [37] further introduced mask prediction as an intermediate process and proposed an instance-specific normalization to transform the object features. OCGAN [38], Context-L2I [11] and LAMA [19] followed their training schemes and further improved objects’ representations and the quality of mask generation. Transformer based methods [16, 41] converted the layout and image into object tokens and patch tokens, which reformulating L2I as a sequence prediction task. Recently, T2I diffusion mod-

els [28, 6] are extended to L2I through encoding the list of object tokens by linguistic model and then regarding it as a T2I task. However, prior approaches merely mine the category-aware relationships while overlooking the location-aware cross-object associations. In this work, we present LAW-Diffusion by explicitly encoding the location-aware semantic compositions for the visual concepts in the scene.

3. LAW-Diffusion

3.1. Preliminaries

Diffusion Models Diffusion model is a type of likelihood-based generative models, consisting of a forward diffusion process and a backward denoising process. Formally, given an image sample $x_0 \sim q(x_0)$, the forward process is defined as a Markov chain with Gaussian transitions:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

where $\{\alpha_t \in (0, 1)\}_{t=1}^T$ is a decreasing sequence of the noise magnitudes in each step. From the property of Gaussian noise and Markov chain, we can directly derive the transition from x_0 to any latent variable x_t :

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. By re-parameterization, x_t can be written as the weighted sum of x_0 and a noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \quad (3)$$

A simple conclusion is that if the length of the Markov chain T is large enough, $\bar{\alpha}_T \approx 0$ and x_T will approximately follow a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

The generative process of diffusion model is defined as iteratively denoising from the Gaussian prior, *i.e.*, $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Due to the intractability of the reverse transition $q(x_{t-1}|x_t)$, another Markov process parameterized by θ , *i.e.*, $p_\theta(x_{t-1}|x_t)$ is learned to serve as its approximation and generate the denoised results $\{x_T, x_{T-1}, \dots, x_0\}$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (4)$$

Denoising diffusion probabilistic models (DDPM) [13] reveal that $\mu_\theta(x_t, t)$ derives from a noise estimator $\epsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t) \right). \quad (5)$$

By optimizing the re-weighted variational lower-bound (VLB) on $\log p_\theta(x_0)$ [13], the noise estimator $\epsilon_\theta(x_t, t)$ is trained to predict the noise ϵ in Eq. (3) and enables diffusion models to produce image samples:

$$\mathcal{L}_{\text{VLB}}(\theta) = \mathbb{E}_{t \sim [1, T], x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(x_t, t) - \epsilon\|^2]. \quad (6)$$

Conditional Diffusion Models Classifier-guidance [4] provides a way for diffusion model to achieve conditional generation by using the gradient of a separately trained classifier $p(y|x_t)$ during sampling. As a more efficient technique, classifier-free guidance [14, 24] replaces the noise estimator by a combination of conditional and unconditional model, without requirement of $p(y|x_t)$:

$$\tilde{\epsilon}_\theta(x_t, t|y) = \omega \epsilon_\theta(x_t, t|y) + (1 - \omega) \epsilon_\theta(x_t, t|\emptyset), \quad (7)$$

where y is the class label or text embedding from language model [24], $\omega \geq 1$ denotes the guidance scale and trivially increasing ω will amplify the effect of conditional input.

With the help of large-scale pre-trained CLIP [25] and other language models [31], diffusion models produce impressive results on text-to-image generation. However, their performance of complex scene generation are always unsatisfactory because the text embeddings from the linguistic models can not accurately capture the spatial properties, *e.g.*, objects' locations, sizes and their implicit spatial associations. Distinct from text prompts, we focus on the task of generating complex scene images from the structured layout configurations (L2I) and further propose a diffusion model-based method with flexibility and compositionality.

3.2. Layout-aware Diffusion Model

In this section, we propose a **Layout-AW**are diffusion model (LAW-Diffusion) to parse the spatial dependencies among co-existing objects and generate photo-realistic scene images with regional semantic alignment. The overview of our LAW-Diffusion is illustrated in Fig. 2 and we will elaborate the details following.

Layout-to-Image Generation Complex scene image synthesis from layout configuration, also known as layout-to-image generation, is specified by synthesising an image

$x \in \mathbb{R}^{H \times W \times 3}$ satisfying a layout configuration Γ consisting of N objects $\mathcal{O} = \{o_1, o_2, \dots, o_N\}$. Each object o_i is equipped with its bounding box $b_i = [r_x^i, r_y^i, h_i, w_i]$ and category c_i , where (r_x^i, r_y^i) is the left-top coordinate and (h_i, w_i) represents the object size.

Spatial Dependency Parser Unlike existing diffusion-based L2I solutions that depends on linguistic models [28, 6], LAW-Diffusion explores a distinctive way to explicitly harvest both location-aware and category-aware object dependencies in the compositional configurations by a spatial dependency parser. The parsing process is detailed below.

Aiming at condensing each object's spatial localization and class information, we first instantiate the regional semantics of object o_i as an object region map $v_i \in \mathbb{R}^{H \times W \times d_c}$, which shares the same spatial resolution as image x for spatial location alignment. Concretely, the rectangular region in v_i specified by the bounding box b_i is filled with a learnable class embedding $c_i \in \mathbb{R}^{d_c}$ (for brevity, symbol c_i is reused here), while the exclusive area is filled by a learnable background embedding $c_{bg} \in \mathbb{R}^{d_c}$. Since the number of objects N varies in different layout configurations, the set of region maps $\{v_i\}_{i=1}^N$ is padded to $\{v_i\}_{i=1}^{N_{\max}}$ using a learnable null region map $v_\emptyset \in \mathbb{R}^{H \times W \times d_c}$, where N_{\max} denotes the maximum number of objects.

In order to fully exploit the spatial dependencies among objects at each position, we propose a location-aware cross-object attention module to aggregate those disentangled object region maps $\{v_i\}_{i=1}^{N_{\max}}$ by their location-aware semantic composition. We split each object region map v_i into N_p patches of region fragments $\{v_i^j\}_{j=1}^{N_p}$, $v_i^j \in \mathbb{R}^{P \times P \times d_c}$ and perform multi-head self-attention (MHSA) for the set of region fragments at the same location. Formally, for the position of the j^{th} patch, we formulate $\{v_i^j\}_{i=1}^{N_{\max}}$ as an unordered set of N_{\max} objects' j^{th} region fragments and feed them into the stacked L multi-head attention [40] layers with a learnable aggregation token $v_{[\text{Agg}]} \in \mathbb{R}^{P \times P \times d_c}$:

$$z_j^0 = \text{concat}([v_{[\text{Agg}]}, v_1^j, v_2^j, \dots, v_{N_{\max}}^j]); \quad (8)$$

$$\tilde{z}_j^l = \text{MHSA}(\text{LN}(z_j^{l-1})) + z_j^{l-1}, l = 1, \dots, L; \quad (9)$$

$$z_j^l = \text{MLP}(\text{LN}(\tilde{z}_j^l)) + \tilde{z}_j^l, l = 1, \dots, L; \quad (10)$$

$$\mathcal{L}_j = \text{LN}(z_j^L)[0], \quad (11)$$

where \mathcal{L}_j is the composed regional semantics for the j^{th} patch. In this way, the per-fragment multi-head self attention in Eq. (9) serves as a location-specific permutation-equivariant interaction between different objects' representations. Furthermore, by regrouping $\{\mathcal{L}_j\}_{j=1}^{N_p}$ according to their original spatial locations, we obtain the layout embedding \mathcal{L} with abundant spatial dependencies among objects.

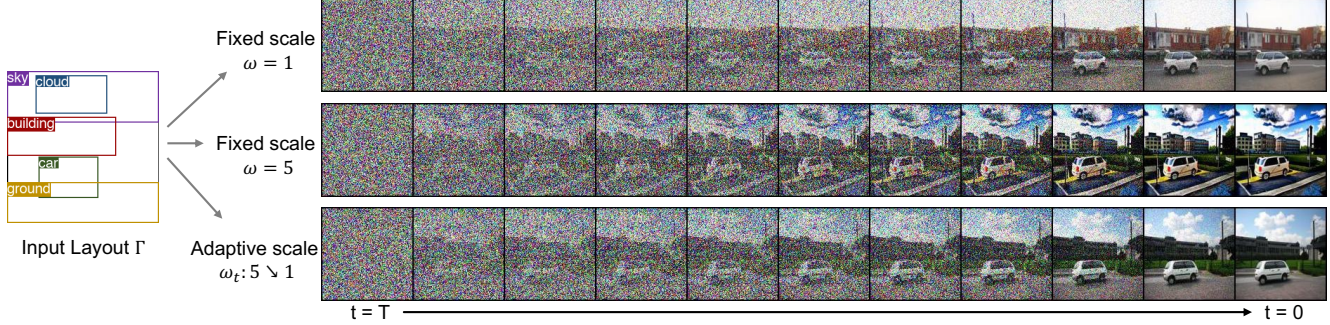


Figure 3. Illustration of the generation processes from the same input layout Γ using different guidance scales. A fixed small scale $\omega = 1$ for each denoising step provides insufficient semantic control, and the cloud is missed in the first row. In the second row, using a fixed large scale $\omega = 5$ leads to over-saturation and distortion of object texture. In the third row, using the adaptive guidance scale $\omega_t : 5 \searrow 1$ which anneals from $\omega_T = 5$ to $\omega_1 = 1$ maintains both semantic alignment and photo-realism. Best viewed in color.

Layout Guidance To develop a diffusion model with flexible control, we train LAW-Diffusion with the classifier-free guidance [14, 24] from the learned layout embedding \mathcal{L} , which contains regional composition semantics. Similar to Eq. (7), LAW-Diffusion learns a noise estimator $\tilde{\epsilon}_\theta(x_t, t|\mathcal{L})$ conditioned on the layout embedding \mathcal{L} :

$$\tilde{\epsilon}_\theta(x_t, t|\mathcal{L}) = \omega \epsilon_\theta(x_t, t|\mathcal{L}) + (1 - \omega) \epsilon_\theta(x_t, t|\emptyset), \quad (12)$$

where $\omega \geq 1$ denotes the magnitude of the layout guidance.

According to the spatial inductive bias of the image-level noise x_T introduced by diffusion models, we concatenate the noised latent code x_t and the layout embedding \mathcal{L} to align their spatial information, *i.e.*, $\text{concat}([x_t, \mathcal{L}]) \in \mathbb{R}^{H \times W \times (D+3)}$ and use it as the input of the conditional noise estimator in Eq. (12):

$$\epsilon_\theta(x_t, t|\mathcal{L}) = \epsilon_\theta(\text{concat}([x_t, \mathcal{L}]), t), \quad (13)$$

where ϵ_θ is implemented by a U-Net [29] and t is implemented as a sinusoidal time embedding following [13].

To this end, the layout embedding \mathcal{L} encapsulates location-aware semantic composition of the multiple visual concepts in the scene. By absorbing the nutrition from \mathcal{L} using the classifier-free guidance, LAW-Diffusion is able to generate a scene image with accurate regional semantics adhere to the input layout and coherent object relations.

3.3. Adaptive Guidance Schedule

As previously discussed, classifier-free guidance [24] provides a effective way to improve the semantic control during the sampling stage. The vanilla classifier-free guidance uses a fixed guidance scale ω in Eq. (12) for each denoising step t and has shown its effectiveness in a variety of application scenarios [24, 31, 26]. However, we empirically find in our experiment that the fixed ω will result in a dilemma of the trade-off between the layout semantic alignment and the photo-realism of generated objects. As shown

in Fig. 3, a fixed small guidance scale ($\omega = 1$) offers insufficient semantic control, *e.g.*, the cloud is missed, while a strong guidance ($\omega = 5$) leads to an over-saturated image where the cloud and car have over-smooth textures. Based on these observations, we can intuitively conclude that a large ω provides precise semantic compliance with the layout Γ while a small ω encourages photo-realistic textures for objects. Inspired by the human’s instinct of first conceiving the holistic semantics and then refining the details when drawing a picture, we propose an adaptive guidance schedule to mildly mitigate the aforementioned trade-off.

Specifically, our proposed adaptive guidance schedule is to gradually anneal the guidance magnitude ω_t during the sampling process of LAW-Diffusion: the generation starts with an initially large guidance scale $\omega_T = \omega_{\max}$ and it gradually anneals to a small magnitude $\omega_1 = \omega_{\min}$ with the annealing function $\phi(t)$ (t is decreasing from T to 1 in the sampling stage):

$$\omega_t = \omega_{\min} + \phi(t)(\omega_{\max} - \omega_{\min}). \quad (14)$$

For simplicity, here we specify $\phi(t)$ as the cosine-form annealing, due to its concave property in the early denoising steps:

$$\omega_t = \omega_{\min} + \frac{1}{2} \left(1 + \cos\left(\frac{T-t}{T}\pi\right) \right) (\omega_{\max} - \omega_{\min}). \quad (15)$$

In Fig. 3, it is evident that the adaptive guidance scale ω_t annealing from $\omega_T = 5$ to $\omega_1 = 1$ (denoted as $\omega_t : 5 \searrow 1$) combines the benefits of the fixed guidance with $\omega = 5$ and $\omega = 1$, thus enabling both accurate layout semantic alignment and preservation of photo-realistic textures.

3.4. Layout-aware Latent Grafting

To further explore the semantic controllability, we will showcase that LAW-Diffusion is capable of instance-level

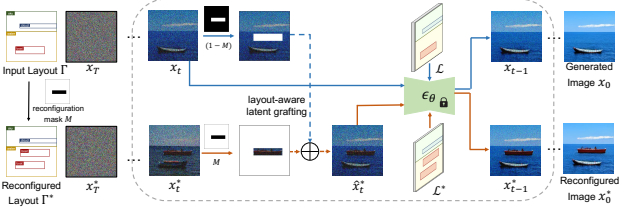


Figure 4. Illustration of our layout-aware latent grafting mechanism for instance reconfiguration (adding an object is taken as an example). Given an image x_0 generated from the layout Γ , reconfigured x_0^* is obtained by alternately grafting the region outside the object bounding box from x_t to x_t^* (\hat{x}_t^* is produced), and denoising the grafted latent \hat{x}_t^* to x_{t-1}^* with the guidance of a reconfigured layout Γ^* . Mask M indicates the region within the bounding box.

reconfiguration. Although LAW-Diffusion does not explicitly model each instance’s style by an individual noise like previous works [9, 37, 11, 38, 44], it allows for adding/removing/restyling an instance in the generated scene image by introducing a training-free layout-aware latent grafting mechanism. Fig. 4 illustrates the process.

Formally, suppose a scene image x_0 has been synthesized from the layout configuration Γ by learning its layout embedding \mathcal{L} , the process of instance reconfiguration can be formulated as generating an image x_0^* from another configuration Γ^* with layout embedding \mathcal{L}^* , where an object o^* within a bounding box b^* is added/removed/restyled while preserving the other objects in x_0 . Inspired by the grafting technique used in horticulture [23, 22] which connects the tissue of a plant to another plant and make them grow together, we aim to spatially graft the exclusive region outside b^* from the latents $\{x_t\}_{t=1}^T$ guided by \mathcal{L} onto the target latents $\{x_t^*\}_{t=1}^T$ guided by \mathcal{L}^* at the same noise level. The reconfiguration process is performed by alternately grafting from x_t to x_t^* and denoising \hat{x}_t^* to x_{t-1}^* :

$$\begin{cases} \hat{x}_t^* = x_t^* \odot M \oplus x_t \odot (1 - M), \\ x_{t-1}^* \sim p_\theta(x_{t-1}^* | \hat{x}_t^*, \mathcal{L}^*), \end{cases} \quad (16)$$

where \odot and \oplus denotes element-wise multiplication and addition, M denotes a rectangular mask indicating the region within the bounding box b^* , \hat{x}_t^* is the grafted latent, $p_\theta(x_{t-1}^* | \hat{x}_t^*, \mathcal{L}^*)$ denotes the layout-aware denoising process guided by \mathcal{L}^* , x_T^* is initialized as a Gaussian noise distinct from x_T . Since x_t^* is guided by holistic semantics from \mathcal{L}^* instead of only local control within b^* , LAW-Diffusion is able to yield a reconfigured scene with coherent relations.

4. Experiments

4.1. Experimental Settings

Datasets Following existing works on layout-to-image generation, our experiments are conducted on two benchmarks: COCO-Stuff [2] and Visual Genome (VG) [17].

COCO-stuff is an extension of the well known MS-COCO dataset with 80 *thing* classes and 91 *stuff* classes. Following [36, 44, 11], objects covering less than 2% of the image are disregarded and the images with 3 to 8 objects are used here ($N_{\max} = 8$). Then we have 74,777 training and 3,097 validation images of COCO-stuff. Different from COCO-stuff, Visual Genome is a dataset specifically designed for complex scene understanding and provides information of object bounding boxes, object attributes, and relationships. Each image in VG contains 3 to 30 objects from 178 categories. Consistent with prior studies [19, 36], small and infrequent objects are removed, resulting in 62,565 images for training and 5,062 for validation in the VG dataset.

Implementation Details Following [13, 4], we use $T = 1000$ and the noise magnitudes $\{\alpha_t\}_{t=1}^T$ of the diffusion process are set to linearly decrease from $\alpha_1 = 1 - 10^{-4}$ to $\alpha_T = 0.98$. Our LAW-Diffusion is trained by jointly optimizing the spatial dependency parser that generates the layout embedding \mathcal{L} , and the noise estimator $\tilde{\epsilon}_\theta(x_t, t | \mathcal{L})$ using the VLB loss defined in Eq. (6). We use the same diffusion training strategies and U-Net architectures as ADM [4]. Regarding the generation of layout embedding \mathcal{L} , we set the dimension of class embedding to $d_c = 32$ and the patch size of region fragments to $P = 8$. Then a two-layer MHSA with 8 attention heads is implemented as the fragment aggregation function in Eq. (9). Following [14, 31], we implement the conditional model $\epsilon_\theta(x_t, t | \mathcal{L})$ and unconditional model $\epsilon_\theta(x_t, t | \emptyset)$ in Eq. (12) as a single conditional model with 10% probability of replacing the conditional input \mathcal{L} by a learnable null embedding \emptyset . Due to the quadratic increase in computational overhead with the size of input images, directly generating 256×256 images can be prohibitively expensive. Hence, following [5, 28], we utilize a VQ-VAE to downsample 256×256 images to 64×64 , and perform our LAW-Diffusion in the compressed latent space. For the 64×64 and 128×128 images, we maintain the diffusion training on image pixels. Regarding the hyperparameters of our adaptive guidance in Eq. (15), we choose $\omega_{\max} = 3$ and $\omega_{\min} = 1$. Please refer to our supplementary materials for more implementation details.

Evaluation Metrics To comprehensively evaluate the performance of LAW-Diffusion, we adopt five metrics for quantitative comparison. Those metrics are: Inception Score (IS) [33], Fréchet Inception Distance (FID) [12], Classification Accuracy Score (CAS) [27], Diversity Score (DS) [43], YOLO Score [19] and our proposed Scene Relation Score (SRS). IS assesses the overall quality of images based on the Inception model [39] pre-trained on ImageNet [3]. FID measures the distribution distance between the synthesized images and the real ones. CAS measures the discriminative ability of generated objects and whether they can be used to train a good classifier. A ResNet [10] is trained on the objects cropped from generated images (5

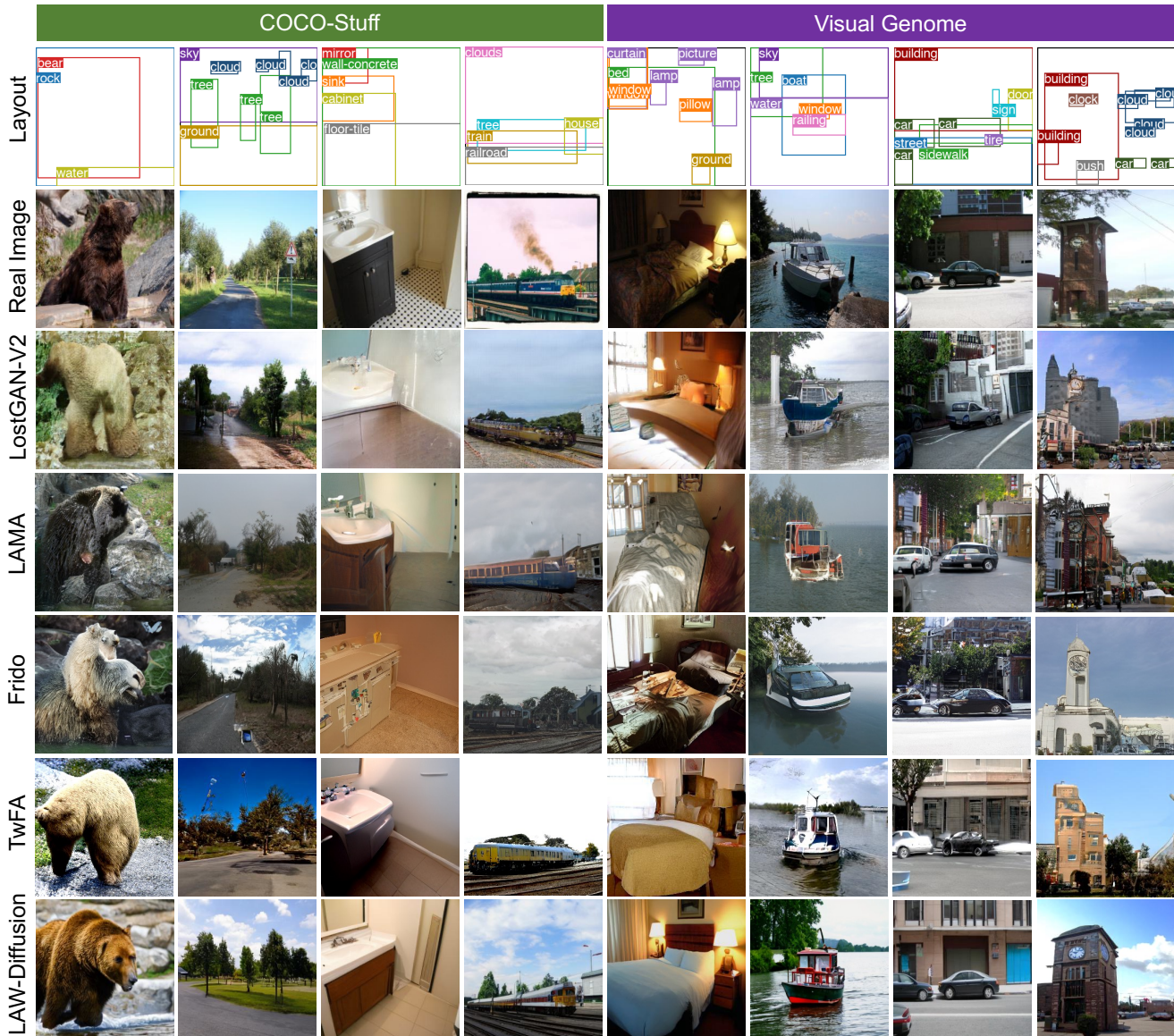


Figure 5. Examples of the 256×256 images generated by different layout-to-image methods on COCO-Stuff [2] and Visual Genome [17]. The first row shows the visualizations of layout configurations and the sampled images in the same column share a common input layout.

image samples are generated for each layout following [37]) and the classification accuracy on the real objects is reported as CAS. DS reflects the diversity of generated samples. YOLO score evaluates the localization alignment between the generated objects and input bounding boxes.

Scene Relation Score Here we propose Scene Relation Score (SRS) as a new metric for L2I to evaluate the rationality and plausibility of the object relations in the generated image. It is reasonable that a competent scene generator should implicitly capture the relationships among objects and the correct relations can be discovered from the synthesized images. Due to the availability of objects' bounding

boxes and labels, we use the predicate classification (PredCls) results predicted by a state-of-the-art scene graph generator to measure whether the correct relationships are captured by the image generator. Specifically, we resort to a publicly available scene graph generator, *i.e.*, VCTreeEB[35] pre-trained on Visual Genome and report the mean Recall@K(mR@K) as our Scene Relation Score (SRS).

4.2. Quantitative and Qualitative Comparisons

We compare our LAW-Diffusion with the state-of-the-art L2I methods, *i.e.*, Layout2Im [44], OC-GAN [38], Context-L2I [11], LostGAN-V2 [37], LAMA [19], TwFA [41],

Resolutions	Methods	FID ↓		Inception Score ↑		Diversity Score ↑		CAS ↑	
		COCO	VG	COCO	VG	COCO	VG	COCO	VG
64 × 64	Real Images	-	-	16.30±0.40	13.90±0.50	-	-	-	-
	Layout2Im [44]	38.14	31.25	-	-	0.15±0.06	0.17±0.09	27.32	23.25
	OC-GAN [38]	29.57	20.27	10.80±0.50	9.3±0.20	-	-	-	-
	Context-L2I [11]	31.32	33.91	10.27±0.25	8.53±0.13	0.39±0.09	0.40±0.09	-	-
	LAMA [19]	19.76	18.11	-	-	0.37±0.10	0.37±0.09	33.23	30.70
	LAW-Diffusion	17.14	16.44	14.81±0.23	12.64±0.32	0.45±0.10	0.46±0.10	35.29	33.46
128 × 128	Real Images	-	-	22.30±0.50	20.50±1.50	-	-	-	-
	LostGAN-V2 [37]	24.76	29.00	14.20±0.40	10.71±0.27	0.45±0.09	0.42±0.09	31.98	29.35
	OC-GAN [38]	36.31	28.26	14.60±0.40	12.30±0.40	-	-	-	-
	Context-L2I [11]	22.32	21.78	15.62±0.05	12.69±0.45	0.55±0.09	0.54±0.09	-	-
	LAMA [19]	23.85	23.02	-	-	0.46±0.09	0.47±0.09	34.15	32.81
	LAW-Diffusion	20.36	15.44	19.89±0.48	18.13±0.44	0.58±0.09	0.55±0.08	36.80	35.22
256 × 256	Real Images	-	-	28.10±1.60	28.60±1.20	-	-	-	-
	LostGAN-V2 [37]	42.55	47.62	18.01±0.50	14.10±0.38	0.55±0.09	0.53±0.09	30.33	28.81
	OC-GAN [38]	41.65	40.85	17.80±0.20	14.70±0.20	-	-	-	-
	LAMA [19]	31.12	31.63	-	-	0.48±0.11	0.54±0.09	30.52	31.75
	LDM [†] [28]	40.91	-	-	-	-	-	-	-
	Frido [†] [6]	21.67	17.24	-	-	-	-	-	-
	TwFA [41]	22.15	17.74	24.25±1.04	25.13±0.66	0.67±0.00	0.64±0.00	-	-
	LAW-Diffusion	19.02	15.23	26.41±0.96	27.62±0.67	0.63±0.09	0.64±0.09	37.79	36.82

Table 1. Quantitative results on COCO-stuff [2] and Visual Genome (VG) [17]. The models denoted by ‘†’ are fine-tuned from the ones trained on a significantly larger dataset, Open-Image [18]. ‘-’ indicates the results are not provided in their papers.

LDM [28] and Frido [6]. Tab. 1 reports the quantitative comparisons for different sizes of images, in terms of FID, Inception Score (IS), Diversity Score (DS) and Classification Accuracy Score (CAS). Besides, Tab. 2 provides the YOLO score and the Scene Relation Score (SRS) of different methods. For fairness, we report the performance of the compared methods from their original papers.

With regards to the image fidelity, LAW-Diffusion significantly outperforms the existing L2I methods, achieving a new state-of-the-art performance. Especially, we observe great improvements of FID and IS scores on both COCO and VG. The noticeable improvements of the challenging CAS further verify the photo-realism of generated objects by LAW-Diffusion, so that they can be used to train a discriminative model. The comparison of SRS in Tab. 2 shows that LAW-Diffusion is capable of synthesizing plausible scene images by capturing the relationships among objects.

Qualitative comparisons on COCO-Stuff and Visual Genome can be observed in Fig. 5, where the samples synthesized by different models using identical layout are presented. It is impressive that LAW-Diffusion produces perceptually appealing images with clear texture details and coherent scene relationships. Moreover, the images generated by our method faithfully complies with the spatial configurations, even in the case of large number of objects.

Resolutions	Methods	YOLO score ↑ AP/AP ₅₀ /AP ₇₅	Scene Relation Score (SRS) ↑ mR@20/50/100
128 × 128	Real Images	33.1 / 47.0 / 36.9	0.1652 / 0.1820 / 0.1821
	LostGAN-V2 [37]	5.5 / 9.2 / 5.8	0.1241 / 0.1307 / 0.1295
	LAMA [19]	7.9 / 12.0 / 8.9	0.1294 / 0.1482 / 0.1489
	LAW-Diffusion	14.1 / 20.6 / 17.8	0.1443 / 0.1603 / 0.1631
256 × 256	Real Images	42.9 / 60.2 / 48.2	0.1703 / 0.1927 / 0.1932
	LostGAN-V2 [37]	9.1 / 15.3 / 9.8	0.1241 / 0.1307 / 0.1295
	LAMA [19]	13.4 / 19.7 / 14.9	0.1260 / 0.1321 / 0.1333
	Frido [6]	- / 30.4 / -	0.1375 / 0.1535 / 0.1578
	TwFA [41]	- / 28.2 / 20.1	0.1407 / 0.1474 / 0.1487
	LAW-Diffusion	21.5 / 34.2 / 23.4	0.1485 / 0.1742 / 0.1750

Table 2. Comparisons of YOLO score and SRS.

Methods	IS ↑	Scene Relation Score (SRS) ↑		
		mR@20	mR@50	mR@100
$\omega = 1$	13.93±0.31	0.1271	0.1233	0.1316
$\omega = 3$	16.62±0.49	0.1419	0.1484	0.1488
$\omega = 5$	15.95±0.31	0.1324	0.1401	0.1438
$\omega_t : 1 \nearrow 3$	15.21±0.37	0.1319	0.1345	0.1370
$\omega_t : 1 \nearrow 5$	14.68±0.28	0.1302	0.1310	0.1334
$\omega_t : 3 \searrow 1$	18.13±0.44	0.1443	0.1603	0.1631
$\omega_t : 5 \searrow 1$	18.24±0.29	0.1392	0.1436	0.1544
$\omega = 1$ (w/o loc)	9.69±0.32	0.1168	0.1206	0.1257
$\omega = 3$ (w/o loc)	12.34±0.58	0.1235	0.1287	0.1298
$\omega = 5$ (w/o loc)	13.97±0.32	0.1214	0.1241	0.1260
$\omega_t : 1 \nearrow 3$ (w/o loc)	12.06±0.42	0.1198	0.1264	0.1278
$\omega_t : 1 \nearrow 5$ (w/o loc)	11.42±0.41	0.1190	0.1221	0.1259
$\omega_t : 3 \searrow 1$ (w/o loc)	14.78±0.33	0.1252	0.1315	0.1327
$\omega_t : 5 \searrow 1$ (w/o loc)	14.52±0.25	0.1263	0.1334	0.1358

Table 3. Ablation study on VG 128×128.

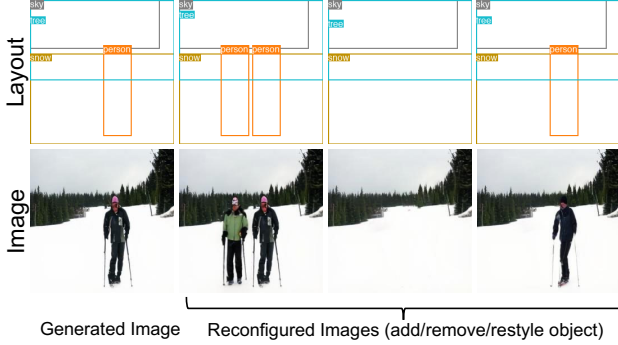


Figure 6. An example of instance-level reconfiguration by LAW-Diffusion. Three types of reconfiguration are shown in this figure (adding/removing/restyling a person in the generated image). Plausible results are obtained using layout-aware latent grafting.

4.3. Instance-level Reconfiguration

As presented in Sec. 3.4, a trained LAW-Diffusion has flexible instance-level controllability, involving the abilities of adding/removing/restyling an instance in the generated scene while preserving the other contents. An example of these three types of reconfiguration is given in Fig. 6. The reconfigured images look plausible and well preserve the coherence in the scene, thus verifying the effectiveness of our proposed layout-aware latent grafting mechanism.

4.4. Ablation Study

To verify the effectiveness of proposed location-aware cross-object attention and adaptive guidance schedule, we conduct ablation experiments on VG 128×128 in Tab. 3. Here, we first introduce a baseline variant of LAW-Diffusion, dubbed LAW-Diffusion (w/o loc), which replaces the location-aware attention by a location-agnostic but class-aware attention used in prior works [11, 41]. Specifically, we use the MHSA layers similar to what we used in Sec. 3.2 to augment each object’s class embedding c_i with only contextual class-aware information. Then the transformed object representations are filled into their bounding boxes and aggregated as the layout embedding using average pooling. In this way, it only captures class-aware relationships and is also used to guide the generation of diffusion model. Moreover, Tab. 3 shows the results of LAW-Diffusion and LAW-Diffusion (w/o loc) with different guidance strategies. For example, $\omega_t : 3 \searrow 1$ means LAW-Diffusion with cosine annealed guidance scale from $\omega_{\max} = 3$ to $\omega_{\min} = 1$, and $\omega = 3(\text{w/o loc})$ denotes LAW-Diffusion (w/o loc) using a fixed guidance scale $\omega = 3$. Similarly, $\omega_t : 1 \nearrow 3$ and $\omega_t : 1 \nearrow 5$ denotes the variants with increasing guidance scales.

By comparing IS and SRS between the variants of LAW-Diffusion and LAW-Diffusion (w/o loc) in Tab. 3, we can conclude that our location-aware cross-object attention can both improve the generated fidelity and capture the reason-

able relations among objects. Besides, it is clear that our adaptive guidance schedule promotes the improvement of the IS scores of generated images. Considering both image fidelity and rationality of the object relations, we select LAW-Diffusion with cosine annealing guidance $\omega_t : 3 \searrow 1$ as our final model. Please refer to our supplementary materials for more ablation studies and human evaluations.

5. Conclusion

In this paper, we present a semantically controllable Layout-Aware diffusion model, termed LAW-Diffusion to generate complex scenes from compositional layout configurations. Specifically, we propose a location-aware cross-object attention module to learn a layout embedding encoding the spatial dependencies among objects. Further, an adaptive guidance schedule is introduced for the layout guidance to maintain both layout semantic alignment and object’s texture fidelity. Moreover, we propose a layout-aware latent-grafting mechanism for instance reconfiguration on the generated scene. Furthermore, a new evaluation metric for L2I, dubbed Scene Relation Score (SRS) is proposed to measure how the images preserves rational relations. Extensive experiments show that our method yields the state-of-the-art generative performance, especially with coherent object relations.

Limitation and future work With regards to the limitation of our work, we only focus on the task of Layout-to-Image generation whose object categories are pre-defined, fixed, and closed-world. Additionally, current version fails to specify scene-level style and semantics with global scene description. In future, we aim to combine our LAW-Diffusion with RegionCLIP [45] to achieve open-vocabulary L2I generation, where the objects generated in the scene can belong to arbitrary novel categories and both object-level and scene-level fine-grained semantic controls can be achieved.

6. Acknowledgement

This work was supported in part by National Key R&D Program of China under Grant No.2021ZD0111601, National Natural Science Foundation of China (NSFC) under Grant No.61836012, U1811463, U21A20470, 62006255, 61876224, 62206314, 62206110, Guangdong Basic and Applied Basic Research Foundation under Grant No.2017A030312006, 2022A1515011835, 2023A1515011374 (GuangDong Province Key Laboratory of Information Security Technology)

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [6] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022.
- [7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022.
- [8] 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 preprint arXiv:2208.01618*, 2022.
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15049–15058, 2021.
- [12] 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. *Advances in neural information processing systems*, 30, 2017.
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [16] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *arXiv preprint arXiv:2105.06458*, 2021.
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [19] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021.
- [20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.
- [21] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.
- [22] Charles W Melnyk and Elliot M Meyerowitz. Plant grafting. *Current Biology*, 25(5):R183–R188, 2015.
- [23] Ken Mudge, Jules Janick, Steven Scofield, and Eliezer E Goldschmidt. A history of grafting. 2009.
- [24] 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 preprint arXiv:2112.10741*, 2021.
- [25] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [27] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

- synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [30] 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. *arXiv preprint arXiv:2208.12242*, 2022.
- [31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [32] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [35] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.
- [36] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019.
- [37] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5070–5087, 2021.
- [38] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2647–2655, 2021.
- [39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [41] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022.
- [42] 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. *arXiv preprint arXiv:2211.15518*, 2022.
- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [44] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.
- [45] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.