SSMG: Spatial-Semantic Map Guided Diffusion Model for Free-Form Layout-to-Image Generation

Chengyou Jia[1*], Minnan Luo[1†], Zhuohang Dang[1], Guang Dai[2,3], Xiaojun Chang[4,5], Mengmeng Wang[6,2], Jingdong Wang[7]

[1]School of Computer Science and Technology, MOEKLINNS Lab, Xi'an Jiaotong University
[2]SGIT AI Lab
[3]State Grid Corporation of China
[4]University of Technology Sydney
[5]Mohamed bin Zayed University of Artificial Intelligence
[6]Zhejiang University
[7]Baidu Inc
{cp3jia, dangzhuohang}@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn, {guang.gdai, cxj273}@gmail.com, mengmengwang@zju.edu.cn, wangjingdong@outlook.com

Abstract

Despite significant progress in Text-to-Image (T2I) generative models, even lengthy and complex text descriptions still struggle to convey detailed controls. In contrast, Layout-to-Image (L2I) generation, aiming to generate realistic and complex scene images from user-specified layouts, has risen to prominence. However, existing methods transform layout information into tokens or RGB images for conditional control in the generative process, leading to insufficient spatial and semantic controllability of individual instances. To address these limitations, we propose a novel Spatial-Semantic Map Guided (SSMG) diffusion model that adopts the feature map, derived from the layout, as guidance. Owing to rich spatial and semantic information encapsulated in well-designed feature maps, SSMG achieves superior generation quality with sufficient spatial and semantic controllability compared to previous works. Additionally, we propose the Relation-Sensitive Attention (RSA) and Location-Sensitive Attention (LSA) mechanisms. The former aims to model the relationships among multiple objects within scenes while the latter is designed to heighten the model's sensitivity to the spatial information embedded in the guidance. Extensive experiments demonstrate that SSMG achieves highly promising results, setting a new state-of-the-art across a range of metrics encompassing fidelity, diversity, and controllability.

Introduction

Despite notable advancements in Text-to-Image (T2I) generative models (Nichol et al. 2022; Saharia et al. 2022), text descriptions often struggle to adequately convey detailed controls, even when composed of lengthy and complex texts (as shown "Text Guided" in Figure 1). In contrast, methods of Layout-to-Image (L2I) generation have risen to prominence, enabling the rendering of realistic and complex scene images from user-specified layouts. Owing to its fine-grained control, L2I generation possesses enormous potential for a wide range of display applications. These range from enhancing the user experience in interactive design to revolutionizing content creation in film and animation (Zhu et al. 2016; Isola et al. 2017; Croitoru et al. 2023).

Recent methods adapt large-scale pre-trained diffusion models (Rombach et al. 2022) for the L2I generation, demonstrating superior performance compared to GAN-based approaches (Park et al. 2019; Richardson et al. 2021; Lee et al. 2020). These methods (Li et al. 2023; Yang et al. 2023; Zheng et al. 2023; Zhang and Agrawala 2023) can be broadly classified into two types, token-guided and image-guided. Token-guided methods, as illustrated in Figure 1, tokenize the layout of spatial and semantic information to embeddings, which are then integrated into T2I models via an attention mechanism. Semantic information "classes" is typically managed by a text encoder, *e.g.*, CLIP (Radford et al. 2021), while spatial position information is handled by various tokenization methods, *e.g.*, Fourier embedding in GLIGEN (Li et al. 2023) or learnable matrixes in ReCO (Yang et al. 2023). However, the tokenization of spatial information, akin to textualization, fails to effectively harness the inherent 2D spatial structure of the layout. Consequently, as demonstrated in Figure 1, it struggles to achieve fine-grained spatial control at the pixel level, leading to issues such as imprecise object boundaries and loss of objects. In contrast, ControlNet (Zhang and Agrawala 2023), a representative image-guided work, incorporates visual conditions, *e.g.*, the layout of boxes, into frozen T2I diffusion models to enable additional condition signals. ControlNet duplicates the weights of the diffusion model into a "trainable copy" and a "locked copy", with the trainable copy being trained to learn these visual conditions. This approach provides more accurate spatial controllability. Nevertheless, the semantic

*This work was completed during the internship at SGIT AI Lab, State Grid Corporation of China.
†Corresponding author.

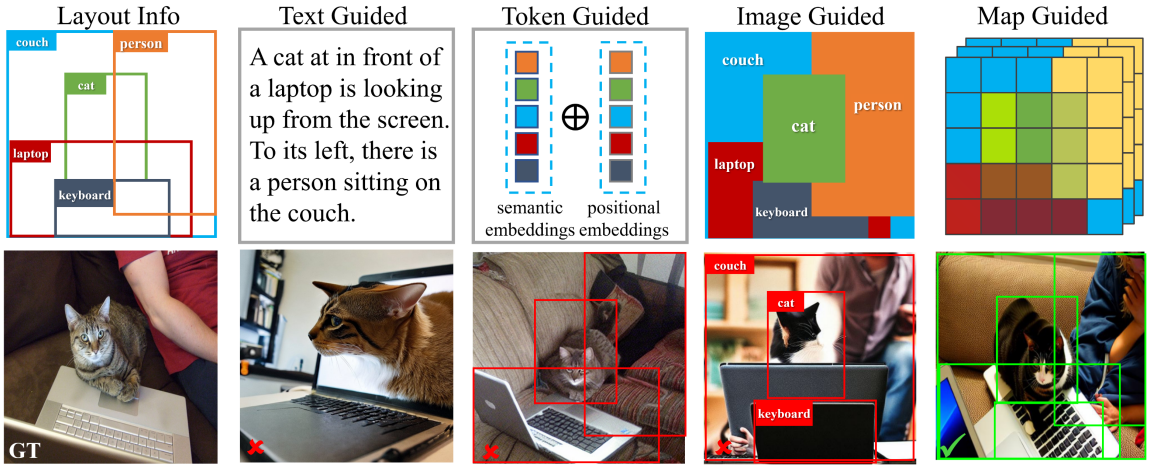

Figure 1: A comparison of image generation methods with different guidance. In contrast to the prevalent token-guided or image-guided methods, our map-guided method excels in providing superior control over both the spatial arrangement and semantic details of individual instances, thereby leading to higher-quality and more appropriate results.

control in ControlNet is solely derived from global image captions, resulting in a lack of control over the detailed semantics of individual instances. As shown in Figure 1, the objects within the box suffer from severe semantic ambiguity, resulting in erroneous or blurred objects.

Considering the aforementioned limitations, a pressing question arises: can a novel guiding strategy be developed that overcomes the shortcomings of both token-guided and image-guided strategies, thereby offering precise control over both the semantics and the spatial layout of the generated images? In response to this contemplation, we argue that a map-guided strategy could serve as an effective solution. On the one hand, the 2D feature map naturally inherits the spatial structure of layouts, which preserves the advantage of visual conditions as exemplified in ControlNet. On the other hand, feature maps offer a richer semantic dimension than the RGB space of image guidance, *e.g.*, $H \times W \times C$, where $C \gg 3$. This enriched semantic dimension allows for a more detailed content representation at each spatial pixel. As a result, it enables the effective control of positions and semantics of individual instances, achieving enhanced spatial accuracy and semantic richness.

In this paper, we present a novel Spatial-Semantic Map Guided (SSMG) diffusion model that adopts the feature map generated from layout as guidance. Specifically, we first initialize the spatial-semantic map based on the given layout info. By fully embracing the spatial structure inherent in the layout, we populate the corresponding position in the map with semantics learned in the textual encoder. Our initialization not only ensures the complete preservation of spatial information but also infuses the feature map with rich semantic content. Then, considering that the initial map processes each instance independently, we further propose Relation-Sensitive Attention (RSA) to establish the relationships among instances in the scene, as well as the relationship of each instance with the overall scene. RSA allows the spatial and semantic information of each instance to cross-reference all other instance or scene information, thereby providing a more nuanced understanding of the scene context. Finally, with the enhanced spatial-semantic map, we integrate it into the conditional generation process through the proposed Location-Sensitive Attention (LSA). This strategy empowers the model to warp the noise image features at each position according to the spatial-semantic map, heightening the model's sensitivity to the embedded spatial information. In such a way, our SSMG effectively enhances the model's controllability over both semantics and the spatiality of generated images. Moreover, our approach serves as a significant extension to traditional T2I methods. SSMG not only permits free-form textual descriptions for each instance but also supports a multitude of layout positional representations, transcending the limitations of bounding boxes. These advantages provide ample adaptability and flexibility for L2I generation, rendering it convenient for wider applications.

Experiments conducted on benchmark datasets demonstrate that our SSMG achieves highly promising results, setting a new state-of-the-art across a range of metrics encompassing fidelity, diversity, and controllability. Notably, SSMG verifies the superiority of the map-guided strategy in layout controllability, significantly improving previous state-of-the-arts YOLO scores from 30.5 to 37.6 on the COCO dataset. We also provide sufficient representative cases to highlight the distinctive features of our method.

Related Work

Layout-to-Image Generation Layout-to-image generation with given bounding boxes can be viewed as a reverse process of object detection. Early GAN-based methods adopt the encoder-decoder architecture to transform boxes into images. LAMA (Li et al. 2021), LostGANs (Sun and Wu 2021), and Context L2I (He et al. 2021) encoded the layout of boxes as a style feature, subsequently feeding this into an adaptive normalization layer. Adding to these

GAN-based methods, Taming (Esser, Rombach, and Ommer 2021) and TwFA (Yang et al. 2022) encode layout information as inputs to a transformer. They then employed an auto-regressive (AR) approach to predict the latent visual codes. Recently, diffusion-based methods (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022) show promising results in L2I. LDM (Rombach et al. 2022) employed a T2I model for box-to-image transfer, encoding layouts using a BERT model. Beyond that, (Li et al. 2023; Yang et al. 2023; Zheng et al. 2023) tokenized bounding boxes into embeddings by Fourier embedding (Mildenhall et al. 2021) or new trainable layers. These embeddings are then injected into pre-trained T2I models, facilitating the generation of scene images with positional information.

Note that existing L2I methods are limited to dealing with the form of bounding boxes. Other forms of layout, *e.g.*, key points, semantic masks in Figure 6, require specialized methods (Park et al. 2019; Richardson et al. 2021; Lee et al. 2020; Xue et al. 2023). In contrast, this paper introduces a novel map-guided approach that transforms given layout info into feature maps, which subsequently guide the model's learning process. Crucially, our approach is versatile and can be applied to a variety of forms of layout.

Diffusion Model Deep diffusion-based generative models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Dhariwal and Nichol 2021) have showcased their exceptional ability to generate high-quality and diverse samples. Building upon these works, LDM (Rombach et al. 2022) utilized VQ-VAE (Van Den Oord, Vinyals et al. 2017) to encode images into latent codes of smaller resolution, thereby reducing the computational overhead required for training a diffusion model. LDM also achieved impressive results in T2I generation by encoding text inputs into latent vectors, similar to approaches, GLIDE (Nichol et al. 2022), DALL-E 2 (Ramesh et al. 2022), and Imagen (Saharia et al. 2022).

Beyond text-guided generation, ControlNet (Zhang and Agrawala 2023) enabled additional image-guided conditions into frozen T2I diffusion models, *e.g.*, sketch, segmentation masks, canny edge, *etc*. It duplicates the weights of a large diffusion model into a "trainable copy" and a "locked copy" and the trainable copy is trained to learn these conditional signals. While ControlNet has demonstrated a commendable ability to control the shape or position of generated objects, it falls short in providing detailed semantic descriptions for individual instances, thereby failing to address L2I effectively. To overcome this, we introduce a novel map-guided approach that incorporates robust semantic and spatial awareness to enhance the model's controllability.

Methodology

Problem Definition

In Layout-to-Image (L2I) generation, we are presented with a set of instance entities $\boldsymbol{e} = \{(t_k, p_k)|k = 1, \cdots, n\}$, where each entity $\boldsymbol{e}_k$ consists of textual description t_k and positional information p_i. In addition to $\boldsymbol{e}$, a global textual description t_{global} is provided, serving as guidance for the overall style and content of the generated image. The L2I model can be conceptualized as a function $f(\boldsymbol{e}, t_{global}) = x$, aiming to generate a realistic image x from the input layout information of entities $\boldsymbol{e}$ and the global description t_{global}.

In this paper, we present an extension to the traditional L2I generation by introducing a novel method that renders both the textual description t_k and positional information p_k in a free-form manner. For t_k, traditional methods have typically restricted it to a predefined set of class descriptions, *e.g.*, "person" or "ball". Our method, however, broadens this definition to accommodate free-form long text, thereby offering more detailed semantic control over each instance. For example, instead of a generic "person" in Figure 2, we could specify "a football player with the number 10 on his blue jersey". What's more, our free-form L2I generation allows p_k to be represented in various forms of layouts, *e.g.*, boxes, masks, or keypoints. This flexibility breaks the constraints of previous methods that were limited to handling a single type of layout and thus makes it more convenient for various applications, *e.g.*, image-to-image translation (Wang et al. 2022), person image synthesis (Men et al. 2020).

Map-guided Diffusion Model

In this section, we present our Spatial-Semantic Map Guided (SSMG) diffusion model in detail. An overview of SSMG is illustrated in Figure 2. Our approach is characterized by three primary parts: (a) Initialize the spatial-semantic map that is based on the given positional information and textual descriptions. (b) Relation-Sensitive Attention (RSA) module that enhances the initial spatial-semantic map by modeling relationships among different instances and the overall scene. (c) Location-Sensitive Attention (LSA) module that enables conditional generation based on the designed spatial-semantic map, achieving sufficient integration of layout information while striving to retain the capabilities of the large-scale pre-trained T2I model. In the following sections, we will further elaborate on each of these steps and modules, shedding light on their contributions to our overall model.

Initialize Spatial-Semantic Map Our objective here is to generate a 2D feature map that is rich in both semantic information and spatial location information to serve as a guiding signal. As depicted in Figure 2, to capture semantic information for each instance, we leverage rich semantics learned in the pre-trained text-to-image model by inputting each instance's description into its textual encoder. Considering that the related layout dataset does not provide corresponding descriptions for each instance, we crop each instance according to its region. The cropped instance is then fed into a pre-trained caption model to obtain a corresponding region caption, which serves as its instance description.

Subsequently, for each pixel (i, j) associated with a specific instance $e_k = (t_k, p_k)$, we populate the corresponding position (i, j) in the feature map with the extracted semantics of that instance's t_k. We use $f_{text}(t_k) \in \mathbb{R}^C$ to represent the pre-trained textual encoder and the initialization process of feature map $\boldsymbol{F} \in \mathbb{R}^{H \times W \times C}$ is specified as

$$\boldsymbol{F}(i, j, :) = f_{text}(t_k), \text{ if } (i, j) \text{ in } p_k,$$

where t_k, p_k refers to the textual description and positional information of instance e_k. It's noteworthy that if a pixel

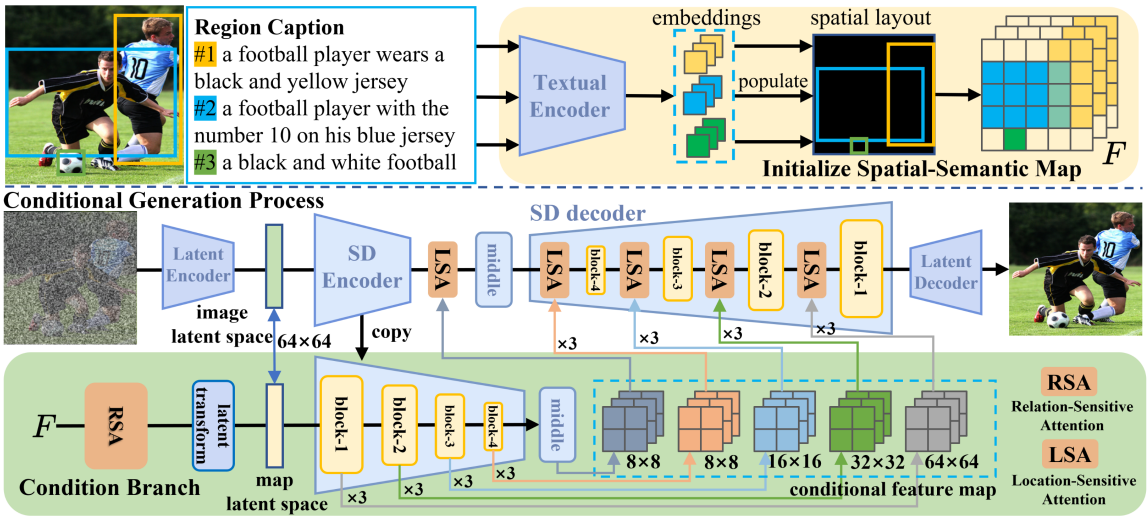

Figure 2: The overall architecture of the proposed SSMG. During the conditional generation process, we first leverage the VQ-GAN's latent encoder within Stable Diffusion (SD) to downsize the entire dataset of 512 × 512 images into the 64 × 64 latent space. To ensure consistency, we also transform the spatial-semantic map to the 64 × 64 latent space in the condition branch. Subsequently, we duplicate the structures and weights of the SD encoder and middle block as ControlNet. The latent map is then fed into the copied SD encoder and middle block to produce conditional feature maps at different scales. These conditional feature maps are then integrated with the corresponding blocks in the SD decoder and middle block through the proposed LSA.

(i, j) does not belong to any instance, the feature at this position remains unpopulated. In cases where a pixel is associated with multiple instances (due to overlap), the value filled in is the average of the embeddings of the multiple instances.

The above procedure not only guarantees the utmost preservation of positional information by fully inheriting the spatial structure from the layout but also integrates rich semantics into the feature map. Furthermore, this initialized way can be easily extended to other types of layouts. All that is required is to populate the corresponding semantics based on the new positional information p_k.

Relation-Sensitive Attention Evidently, the initial spatial-semantic map processes the semantic and spatial information of each instance independently, while neglecting the relationships among instances in the scene, as well as the relationship of each instance with the overall scene. This brings that some generated objects are incongruous with other objects or the entire scene, often appearing inappropriate especially when facing scenes with multiple objects or a high degree of overlap (as seen case (c) and (d) in Figure 4). To overcome this limitation, we propose the Relation-Sensitive Attention (RSA) module. This module allows the spatial and semantic information of each instance to cross-reference all other instance or scene information, thereby integrating contextual information into the feature representation of each instance.

Concretely, we first introduce an auxiliary scene token $\boldsymbol{g} = f_{text}(t_{global})$ to represent the global scene information. Then, inspired by (Yang et al. 2022), we modified vanilla self-attention with a relation matrix $\boldsymbol{M}$ as follows:

$$\text{Attention}(Q, K, V, \boldsymbol{M}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} \circ \boldsymbol{M}\right) V,$$

where $\circ$ indexes element-wise product and d is the dimension of queries and keys. The attention is performed over the concatenation of flattened feature map $\boldsymbol{F}$ and scene token $\boldsymbol{g}$ where $Q = \varphi_Q([\boldsymbol{F}, \boldsymbol{g}]), K = \varphi_K([\boldsymbol{F}, \boldsymbol{g}]), V = \varphi_V([\boldsymbol{F}, \boldsymbol{g}])$, and $\varphi_Q, \varphi_K, \varphi_V$ are linear projection layers. We then construct the relation matrix $\boldsymbol{M}$ in two ways, *i.e.*, instance-instance R_{inst}, and instance-scene R_{scene}, such that

$$\boldsymbol{M}[s_m, s_n] = \begin{cases} 1, & \text{if } R_{\text{inst}}(s_m, s_n) \text{ or } R_{\text{scene}}(s_m, s_n) \\ -\infty, & \text{else .} \end{cases}$$

As shown in Figure 3, we categorize two pixels (or tokens) s_m and s_n as instance-instance relative $R_{\text{inst}}(s_m, s_n)$, if they belong to distinct entities. In the case of overlapping, we prioritize treating them as belonging to different entities, thereby emphasizing the distinction and relationship between instances. We then define two pixels (or tokens) as instance-scene relative $R_{\text{scene}}(s_m, s_n)$, when one token corresponds to the scene token and the other is associated with an instance entity. The construction of $\boldsymbol{M}$ ensures that the relationships between instances, and their relationship with the overall scene, are adequately considered. Following the attention operation, the scene token is discarded, and the feature map reverts to its original shape.

Location-Sensitive Attention In the subsequent stage, we focus on integrating the previously generated condition (*i.e.*, enhanced spatial-semantic map) into the conditional generation process. Note that previous methods, such as channel-wise concatenation in Palette (Wang et al. 2022) or feature addition in ControlNet (Zhang and Agrawala 2023), encounter severe pixel-wise structural bias due to the randomness of noise features (Zhu et al. 2023). Such bias notably undermines the noise feature's sensitivity to positional

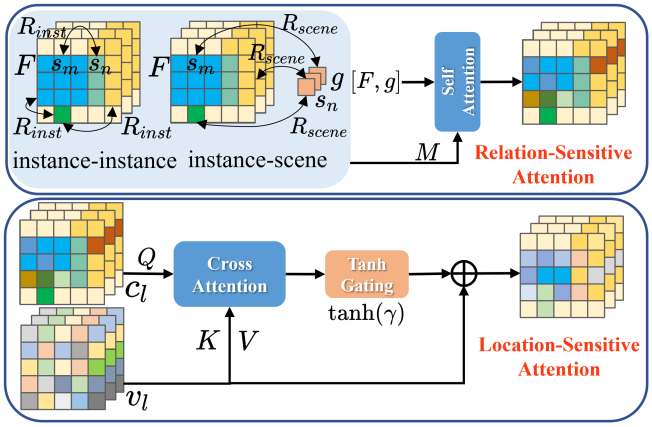

Figure 3: Illustrations of the RSA and LSA mechanisms. Feature maps are flattened before being fed into attention.

information within the conditions. Therefore, we introduce the Location-Sensitive Attention (LSA) to warp noise features according to the spatial-semantic map for avoiding bias. This mechanism is shown in Figure 3, formulated by

$$v_l = v_l + \tanh(\gamma) \cdot \text{Attention}\left(\varphi_Q(c_l), \varphi_K(v_l), \varphi_V(v_l)\right),$$

where v_l and c_l denote flattened noise features and spatial-semantic map at block l, respectively. It's worth noting that due to the scale consistency between v_l and c_l, LSA allows us to use spatial-semantic maps as queries and noise features as keys and values. In this way, the cross-attention can leverage similarities computed by $\varphi_Q(c_l) * \varphi_K(v_l)^T$ to re-weight noise features. This provides a learnable way to represent correspondence between noise features and positional information in c_l, effectively enhancing the generated noise feature's location sensitivity to spatial-semantic maps. Further drawing inspiration from (Alayrac et al. 2022), we also employ the gating mechanism by applying $\tanh(\gamma)$, where γ is a learnable scalar initialized to 0. This initialization ensures that the training starts from the pre-trained state, thereby preserving the integrity of the large-scale diffusion model and enhancing the stability of the training.

Model Fine-tuning We fine-tune the Stable Diffusion with the same LDM objective (Rombach et al. 2022), which is based on the layout information of instance entities $\boldsymbol{e}$, *i.e.*,

$$\min_{\theta} \mathcal{L} = \mathbb{E}_{z,t',\epsilon\sim\mathcal{N}(0,I)}\left[\left\|\epsilon - f_\theta\left(z_{t'}, t', f_{text}(t_{global}), \boldsymbol{e}\right)\right\|_2^2\right],$$

where z represents the latent code extracted from the input image; t' refers to the time step. We retain the text-guided signal $f_{text}(t_{global})$ in LDM, which also serves as the scene token $\boldsymbol{g}$. SSMG exclusively focuses on fine-tuning the conditional U-Net f_θ while keeping the text encoder and the SD model's latent encoder and decoder frozen during training.

Experiments

Experimental Setup

Datasets. We adopt widely recognized benchmarks COCO-Thing-Stuff (Lin et al. 2014; Caesar, Uijlings, and Ferrari 2018) for both training and evaluation. It consists of 118,287 training and 5,000 validation images, which are annotated with 80 thing/object classes and 182 semantic stuff classes. Following (Li et al. 2021), we disregard objects that occupy less than 2% of the entire image and only utilize images containing between 3 to 8 objects.

Evaluation Metrics. We adopt two widely recognized metrics, Inception Score (IS) (Salimans et al. 2016) and Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate the fidelity of the generated images. Additionally, to measure the diversity, we compute the Diversity Score (DS) between two images generated from the same layout by comparing the LPIPS (Zhang et al. 2018) metric in a deep neural network feature space. In line with metrics in previous studies, we adopt *YOLO score* mAP/mAP50/mAP75 to evaluate grounding alignment and semantic accuracy between the generated images and layouts.

Implementation Details. Our model is implemented based on the Stable Diffusion and ControlNet. During training, we take the AdamW as the optimizer within the PyTorch Lightning framework. We resize the input images to 512×512. The model is trained on 4 NVIDIA-A100 GPUs with a batch size of 64, requiring $\sim$ 2 days for 50 epochs. During inference, we use 20 DDIM (Song, Meng, and Ermon 2020) sampling steps with classifier-free guidance (Ho and Salimans 2022) scale of 9. For qualitative and quantitative evaluations on COCO, we use its caption annotations as t_{global}. For the free-form generation, we concatenate the text descriptions of each sub-instance to serve as t_{global}.

Quantitative Comparison

We benchmark our method against the state-of-the-art L2I methods with bounding boxes including LostGAN (Sun and Wu 2019), Con L2I (He et al. 2021), LAMA (Li et al. 2021), TwFA (Yang et al. 2022), ReCo (Yang et al. 2023), GLIGEN (Li et al. 2023), LayoutD (Zheng et al. 2023). We provide two versions built upon Stable Diffusion v15 and v21 (Patil et al. 2022), respectively. As demonstrated in Table 1, the proposed model outperforms all the competitors across FID, IS, and DS metrics, which attests to the high visual quality and diversity of the images synthesized by our method. It's noteworthy that SSMG exhibits a relatively minor improvement compared to GLIGEN and ReCO, as indicated by the FID (20.82 *vs.*21.04) and IS (32.18 *vs.*31.63). This is anticipated, given that the generative capabilities of these models are largely underpinned by the large-scale T2I models.

Crucially, our method demonstrates superior controllability compared to these methods with similar infrastructure. The YOLO score, a critical metric in our evaluation, is significantly higher for our method compared to the others. This metric not only reflects the model's spatial controllability over localization accuracy but also its semantic controllability to generate fine-grained details at specific locations. These results underscore the limitations of token-guided and image-guided methods as we discussed. Significantly, they highlight the effectiveness of our map-guided strategy in enhancing both spatial and semantic controllability.

Figure 4: Qualitative comparison with SOTA methods. The red boxes indicate unrecognizable or mispositioned instances.

Methods	FID↓	IS↑	DS ↑	YOLO score↑
LostGAN	42.55	18.01±0.50	0.45	9.1/15.3/9.8
Con L2I	29.56	18.57±0.54	0.65	10.0/14.9/11.1
LAMA	31.12	14.32±0.58	0.48	13.4/19.7/14.9
TwFA	22.15	24.25±1.04	0.67	-/28.2/20.1
LayoutD*	22.65	26.73±0.92	0.57	18.1/31.0/18.9
GLIGEN	21.04	-	-	22.4/36.5/24.1
Control*	28.41	28.85±0.85	0.65	25.2/46.7/22.7
ReCo*	27.47	31.63±0.72	0.62	30.5/56.3/29.9
Our-v15	22.31	**33.99±1.47**	**0.69**	35.9/57.0/38.5
Our-v21	**20.82**	32.18±0.85	0.68	**37.6/59.0/40.9**

Table 1: Quantitative comparison with state-of-the-art layout-to-image methods. '↑' stands for higher the better, '↓' stands for lower the better. All generated images are evaluated under 256 × 256 resolution for FID, IS, DS and 512 × 512 for YOLO score. Methods marked with '*' are re-evaluated using images generated from their official code.

Qualitative Results

Comparison Following our quantitative comparison, we further delve into a qualitative analysis of SSMG, presenting representative case studies to highlight the distinctive features of our approach. As shown in Figure 4 (a) and (b), our method not only preserves the overall realism of generated images but also ensures the fidelity of each individual instance. This stands in stark contrast to methods LayoutD (Zheng et al. 2023) and ControlNet, which exhibit distorted and unrealistic phenomena on local objects, despite utilizing similar network architectures. Meanwhile, state-of-the-art methods like GLIGEN (Li et al. 2023) and ReCo (Yang et al. 2023), although producing visually appealing images at first glance, continue to grapple with the challenges inherent to token-guided strategies, including imprecise boundaries and loss of objects. Our method, on the other hand, leverages the spatial control strength of our map-guided strategy and the Location-Sensitive Attention (LSA) module, demonstrating precise localization capabilities. As evidenced by all the cases in Figure 4, the objects generated by our method are well-confined within the layout, exhibiting an exceptional sense of spatial accuracy.

Especially when dealing with a set of multiple objects with complex relationships, *e.g.*, case (c) and (d), previous methods often struggle to generate recognizable objects at overlapping positions. However, our method, benefiting from the relationship modeling of Relation-Sensitive Attention (RSA), is capable of generating scene images that maintain complex object relationships. The overlap and interaction between objects within the image are well represented, demonstrating the model's nuanced scene understanding.

Free-form L2I Our approach can serve as an enhancement to traditional T2I methods, permitting free-form textual descriptions and a multitude of layout positional representations. As shown in Figure 5, SSMG is capable of leveraging free-form descriptions for each instance, allowing for the specific generation of particular styles or characteristics, *e.g.*, from "cup" to "a thermos cup with hello kitty". Moreover, training and inference with the free-form

Figure 5: Illustrations of free-form textual descriptions.

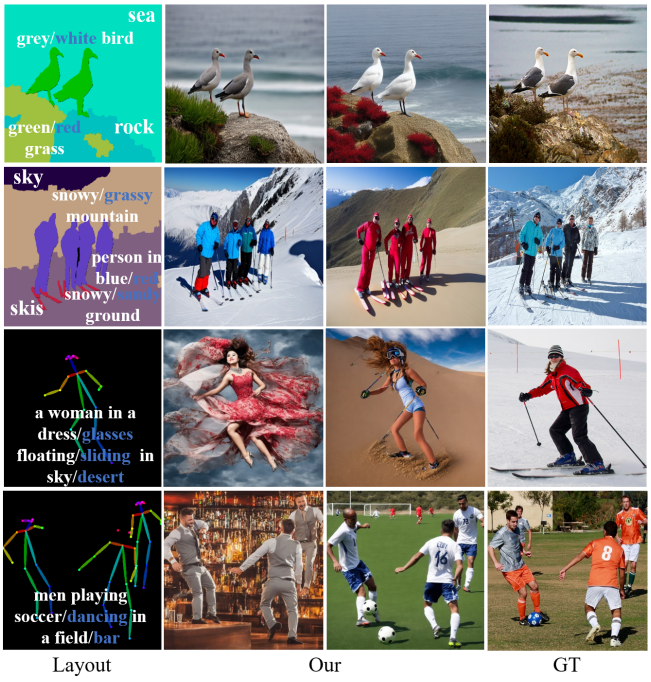

Figure 6: Illustrations of free-form positional layouts.

textual descriptions enable our method to inherit the zero-shot capabilities of T2I models. As illustrated in Figure 5, even though fine-tuned on COCO, our method can achieve out-of-distribution generation. For instance, the right column indicates that our method can generate novel content, *e.g.*, "bonsai", "chopsticks", and "scarves", which falls outside the scope of the COCO dataset.

We further demonstrate that SSMG can be applied to various positional representations, *i.e.*, masks and keypoints. Figure 6 demonstrates that our method can not only accurately generate instances at corresponding positions according to other layout forms but also control the diverse styles of instances through textual descriptions. This remarkable combination of controllability and flexibility significantly elevates the practical value of generative models, positioning L2I generation as a potent tool for various applications.

Ablation studies

To further validate the effectiveness of our proposed components, we conduct a series of ablation studies. These studies focus on the three key components of our model: the Map-Guided (MG) strategy, the mechanisms of Relation-Sensitive Attention (RSA) and Location-Sensitive Attention (LSA). We build the baseline based on image-guided ControlNet and add three components sequentially. As illustrated in Table 2, substituting the image-guided strategy with our map-guided strategy results in a noticeable enhancement in both the fidelity and grounding accuracy of the generated images. Similarly, the addition of the RSA and LSA mechanisms also results in a significant improvement. With the RSA mechanism, the model excels at accurately capturing the relationships between different objects within the scene, resulting in higher fidelity. With the LSA mechanism, the model's spatial awareness is greatly enhanced, leading to more accurate positioning of objects. These incremental studies clearly demonstrate the importance of each of these components to the overall performance. The map-guided strategy, coupled with the RSA and LSA, is integral to our model's ability to generate high-quality images with precise control over both fine-grained semantics and spatial layouts.

MG	RSA	LSA	FID ↓	YOLOScore ↑
			28.41 (+7.59)	25.2/46.7/22.7 (-14.3)
✓			25.12 (+4.30)	30.5/53.0/29.6 (-8.1)
✓	✓		21.41 (+0.59)	32.7/54.7/32.9 (-5.7)
✓		✓	23.49 (+2.67)	35.1/56.8/37.5 (-2.7)
✓	✓	✓	**20.82**	**37.6/59.0/40.9**

Table 2: Ablations of Map-Guided (MG), Relation-Sensitive Attention (RSA), and Location-Sensitive Attention (LSA).

Conclusions

This paper introduces a novel Spatially-Semantic Map Guided diffusion model that effectively addresses the limitations of previous token-guided and image-guided L2I methods. Our method, through its innovative map-guided strategy and bespoke attention mechanisms, delivers superior performance in terms of fidelity, diversity, and controllability. Furthermore, our method can be extended to be free-form, providing users with a more diverse and flexible way to describe layouts. We hope that this work will constitute an advancement in the field of layout-to-image generation, opening up new possibilities for future research and applications.

Ethics Statement

Our method's versatility opens up possibilities for its application across various domains of structured image generation. However, we must also pay attention to societal impacts that could arise from its misuse. For instance, the risk of data leakage could lead to privacy concerns and the model's capacity to produce highly realistic images could be exploited for illegal or harmful content. It's essential to follow clear usage guidelines and ensure responsible and ethical use.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3102600), National Nature Science Foundation of China (No. 62192781, No. 62272374, No. 62202367, No. 62250009, No. 62137002), Project of China Knowledge Center for Engineering Science and Technology, Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", and the K. C. Wong Education Foundation.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

He, S.; Liao, W.; Yang, M. Y.; Yang, Y.; Song, Y.-Z.; Rosenhahn, B.; and Xiang, T. 2021. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15049–15058.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Li, Z.; Wu, J.; Koh, I.; Tang, Y.; and Sun, L. 2021. Image Synthesis from Layout with Locality-Aware Mask Adaption. In *IEEE International Conference on Computer Vision (ICCV)*, 13819–13828. IEEE.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable Person Image Synthesis with Attribute-Decomposed GAN. In *Computer Vision and Pattern Recognition (CVPR)*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Patil, S.; Cuenca, P.; Lambert, N.; and von Platen, P. 2022. Stable Diffusion with Diffusers. *Hugging Face Blog*. Https://huggingface.co/blog/stable_diffusion.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Sun, W.; and Wu, T. 2019. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10531–10540.

Sun, W.; and Wu, T. 2021. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5070–5087.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, T.; Zhang, T.; Zhang, B.; Ouyang, H.; Chen, D.; Chen, Q.; and Wen, F. 2022. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Xue, H.; Huang, Z.; Sun, Q.; Song, L.; and Zhang, W. 2023. Freestyle Layout-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14256–14266.

Yang, Z.; Liu, D.; Wang, C.; Yang, J.; and Tao, D. 2022. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7764–7773.

Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490–22499.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 597–613. Springer.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4606–4615.