This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PLACE: Adaptive Layout-Semantic Fusion for Semantic Image Synthesis

Zhengyao Lv[1], Yuxiang Wei[2], Wangmeng Zuo[2,3], Kwan-Yee K. Wong[1](✉)
[1]The University of Hong Kong [2]Harbin Institute of Technology [3]Pazhou Lab, Guangzhou
{cszy98, yuxiang.wei.cs}@gmail.com wmzuo@hit.edu.cn kykwong@cs.hku.hk

Abstract

Recent advancements in large-scale pre-trained text-to-image models have led to remarkable progress in semantic image synthesis. Nevertheless, synthesizing high-quality images with consistent semantics and layout remains a challenge. In this paper, we propose the adaPtive LAyout-semantiC fusion modulE (PLACE) that harnesses pre-trained models to alleviate the aforementioned issues. Specifically, we first employ the layout control map to faithfully represent layouts in the feature space. Subsequently, we combine the layout and semantic features in a timestep-adaptive manner to synthesize images with realistic details. During fine-tuning, we propose the Semantic Alignment (SA) loss to further enhance layout alignment. Additionally, we introduce the Layout-Free Prior Preservation (LFP) loss, which leverages unlabeled data to maintain the priors of pre-trained models, thereby improving the visual quality and semantic consistency of synthesized images. Extensive experiments demonstrate that our approach performs favorably in terms of visual quality, semantic consistency, and layout alignment. The source code and model are available at PLACE.

1. Introduction

Semantic image synthesis aims to generate high-quality images that are aligned with given semantic maps. It provides users the flexibility to precisely control the spatial layout of synthesized images using semantic maps while having important applications in content creation [6, 7, 54], image editing [20, 21, 25], and data augmentation [49].

Earlier semantic image synthesis works [14, 19, 26, 37] mainly relied on Generative Adversarial Networks (GANs) [8] and trained a model using semantic maps as condition within specific domains. However, due to the limited scale of the training dataset, the quality and diversity of generated images are usually compromised. Recently, large-scale text-to-image models [30, 31, 33] have shown high-quality and diverse generation results with open-vocabulary textual prompts. Based on these pre-trained text-to-image models (*e.g.*, Stable Diffusion [31]), ControlNet [50] and T2I-Adapter [23] introduced an additional adapter to inject layout guidance for high-quality semantic image synthesis. Nevertheless, these adapters failed to integrate textual semantics with corresponding regions accurately, resulting in inconsistent layouts in generated results, as shown in Fig. 1.

Figure 1. Comparisons in terms of visual quality as well as layout alignment and semantic consistency. Zoom in for details.

To facilitate the layout consistency, FreestyleNet [48] proposed an RCA module that forces each intermediate image token to attend to the respective textual semantic, while fine-tuning the diffusion model with RCA. However, the semantic map used in RCA is directly adapted to the intermediate image features in latent diffusion, which is considerably smaller than its original size (*e.g.*, 64×64 compared to 512×512), leading to inevitable layout information loss. Moreover, the mechanism of RCA disrupts the global interaction between image and text tokens, impeding the synthesis of high-quality images.

To alleviate the aforementioned issues, we propose the adaPtive LAyout-semantiC fusion modulE (termed as

PLACE) as depicted in Fig. 2, which leverages pre-trained Stable Diffusion for high-quality and faithful semantic image synthesis. Firstly, inspired by the spatio-textual representation [1], we introduce the layout control map (LCM), which represents the layout information faithfully in low-resolution feature space. Specifically, we explore the proportion of each semantic component within the receptive field of each image token in the intermediate image features and utilize a vector composed of these proportions as the layout feature for this image token. Such a layout control map retains layout information accurately in the feature space and can be then incorporated with textual features to guide the semantic image synthesis.

Manually constraining the regions influenced by each semantic component with the layout control map allows for the manipulation of the layout of synthesized images. However, we noticed that it also restricts the interaction between image tokens and global text tokens, compromising the visual quality of synthesized details. To effectively integrate layout control maps, while preserving the beneficial interactions for better visual quality, we develop a timestep-adaptive layout-semantic fusion module. Specifically, for each fusion module, a time-adaptive fusion parameter is learned from time embedding. Subsequently, this parameter is employed to adaptively combine our layout control map with the original cross-attention maps which encapsulate the global semantics. The resulting adaptive fusion maps not only encompass faithful layout information but also maintain the influence of contextual textual tokens, thereby improving the visual quality of generated images.

Additionally, we propose effective Semantic Alignment (SA) loss and Layout-Free Prior Preservation loss to facilitate the fine-tuning. The SA loss constrains the weighted aggregation results of the adaptive fusion map and self-attention maps to be as close as possible to the original adaptive fusion maps. It enhances the internal interactions of image tokens within the same or related semantic region, consequently improving the layout consistency and visual quality. Due to the limited scale of datasets, priors of the pre-trained model are prone to perturbation while fine-tuning. Our proposed LFP loss helps preserve priors without involving layout annotation during fine-tuning. Specifically, we compute the denoising loss in a layout-free manner with the text-image pairs to preserve the semantic concepts embedded in the pre-trained model. Owing to the enhanced preservation and utilization of semantic priors, our method exhibits better visual quality and semantic consistency, even in new domains (as shown in Fig. 1).

The contributions of this work can be summarized as:

- We introduce the layout control map as a reliable layout representation and propose an adaptive layout-semantic fusion module to adaptively integrate the layout and semantic features for semantic image synthesis.
- We propose effective SA and LFP losses. The former enhances the layout consistency of generated images, while the latter helps preserve the semantic priors of pre-trained models with readily available text-image pairs.

2. Related Work

2.1. Semantic Image Synthesis

Semantic image synthesis aims to synthesize realistic images with given semantic masks. Previous works primarily achieved the layout control over generated images through Generative Adversarial Networks (GANs) [8].

Pix2pix [14] was the first to propose using an encoder-decoder generator and a PatchGAN discriminator for semantic image synthesis. Pix2pixHD [41] accomplished high-resolution image synthesis by employing a coarse-to-fine generator and multi-scale discriminators. SPADE [26] proposed using spatially adaptive transformations learned from semantic maps to modulate features and significantly improved the image quality. Subsequently, CC-FPSE [19] introduced predicting conditional convolution kernel parameters based on semantic layouts and utilized a feature pyramid semantic-embedding discriminator to encourage the generator to produce images with higher-quality details and better semantic alignment. More recently, SC-GAN [43] learned a semantic vector to parameterize conditional convolution kernels and normalization parameters. LGGAN [38] introduced the utilization of a local class-specific and global image-level generative adversarial network to individually learn the appearance distribution of each object category and the global image. OASIS [37] innovatively designed a segmentation network-based discriminator, providing the generator with more potent feedback, thereby generating semantically aligned images with higher fidelity. Besides, there are also some methods [22, 28, 35, 39, 44] exploring structural and shape information in semantic maps to enhance the quality of images.

Despite significant achievements of previous methods in semantic image synthesis, generated images still show limited quality and diversity due to constraints in the scale of training data and the representation of semantic layouts.

2.2. Layout Controllable Text-to-Image Synthesis

Text-to-image synthesis focuses on generating images conditioned on given text prompts. Benefiting from the powerful diffusion model [12, 36] and extensive text-image training data, text-to-image synthesis has achieved unprecedented successes in terms of image quality, diversity, and alignment with the provided text [24, 30, 31, 33]. Among them, the balance between efficiency and quality of LDM [31] has attracted significant attention, making it the foundational model for many controllable [50, 52] synthesis, customized image synthesis [4, 9, 17, 32, 45] and some other fields [13, 51].

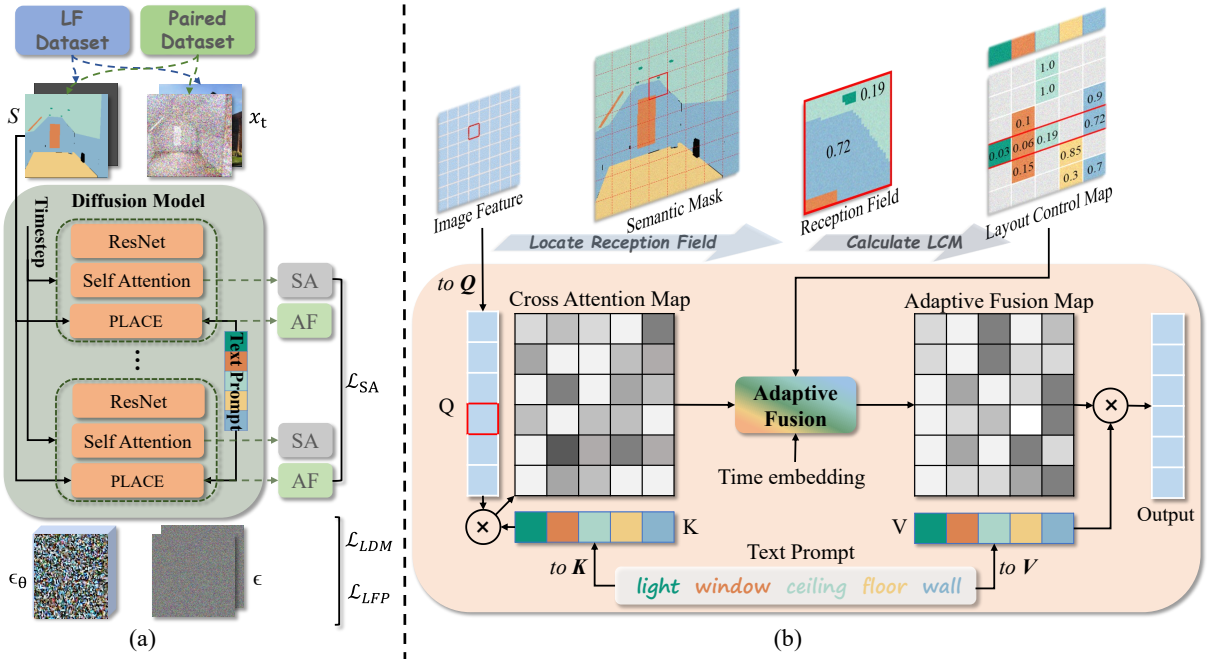

Figure 2. Overview of our method. (a) We utilize the layout control map calculated from semantic map S and PLACE for layout control. During fine-tuning, we combine the $\mathcal{L}_{LDM}$, $\mathcal{L}_{SA}$, and $\mathcal{L}_{LFP}$ as optimization objective. (b) Calculation of the layout control map and details of the adaptive layout-semantic fusion module. Each vector in the Layout Control Map encodes all the semantic components in the reception field. The adaptive layout-semantic fusion module blends the layout and semantics feature in a timestep-adaptive way.

Subsequent works investigated the utilization of pre-trained models to achieve layout-controllable text-to-image synthesis [1, 15, 27, 46, 47]. The eDiff-I [2] and the Two Layout Guidance [5] iteratively optimize the alignment between the constrained cross-attention map and the target layout. Nevertheless, they can merely roughly control the positioning of the synthesized objects. ControlNet [50] and T2I-Adapter [23] encode semantic maps with an additional layout encoder. However, constrained by the generalization capacity of the layout encoder, they fail to overcome the limitations of layout consistency. Another category of methods controls the layout of synthesized images in a training-free manner. FreestyleNet [48] introduces Rectified Cross Attention (RCA) to replace the cross attention module in Stable Diffusion, enabling each text token to interact exclusively with the corresponding image feature region. Subsequently, the pre-trained Stable Diffusion model is fine-tuned on specific domains to adapt to RCA. FreestyleNet has made progress in semantic consistency and layout alignment. However, due to the loss of layout information when utilizing semantic maps with RCA, the generated images lack sufficient layout alignment. Furthermore, due to the modifications in cross-attention and the limited scale of fine-tuning datasets, FreestyleNet is prone to losing priors in the pre-trained model and still exhibits limitations in visual quality and semantic consistency.

3. Proposed Method

Given a semantic map $S \in \mathbb{R}^{H\times W\times C}$ with C semantic classes, semantic image synthesis aims to synthesize photo-realistic images that are well aligned with S. The value of C is determined by the number of semantic categories specified by the user, rather than the cardinality of a pre-defined closed set. To enable controllable image synthesis with desired layouts, we first employ a faithful layout control map as layout representations in feature space. We then propose PLACE to adaptively integrate the layout and semantic features, as illustrated in Fig. 2 (b). During fine-tuning, we further introduce the semantic alignment (SA) loss to enhance the layout alignment and the layout-free prior preservation (LFP) loss to improve the performance of visual quality and semantic consistency. In the following subsections, we first give a concise introduction of the pre-trained text-to-image model we employed, namely Stable Diffusion [31]. We then provide the details of PLACE and learning objective.

3.1. Preliminary: Stable Diffusion

Stable Diffusion [31] is a text-to-image synthesis model based on the diffusion process in the latent space. It comprises two components, namely an autoencoder and a conditional latent diffusion model (LDM). The autoencoder $\mathcal{E}$ is designed to learn a latent space that is perceptually equivalent to the image space. Meanwhile, the conditional LDM ϵ_θ is parameterized as a U-Net with cross-attention and trained on a large-scale dataset of text-image pairs via:

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2], \quad (1)$$

where ϵ is the target noise, τ_θ and y are the pretrained CLIP [29] text encoder and text prompts, respectively, and

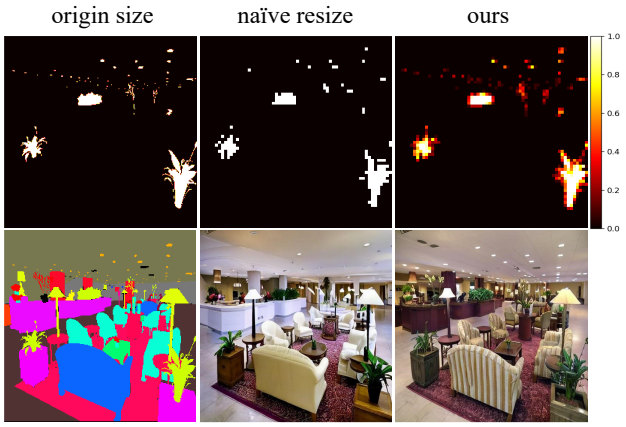

Figure 3. Comparison between naive resize and layout control map regarding information preservation (downsampling by 8 times). The $1st$ column displays the original mask of 'plants, lights' and full semantic map, the $2nd$ column shows the nearest resized mask and corresponding synthesized image, and the $3rd$ column presents the representation of the layout control map and its generated image. A higher value indicates a higher proportion of semantics within its patch. Ours preserves more details.

z_t is the noisy latent at timestep t.

In the conditional LDM, the text feature is integrated into the intermediate layers of the U-Net through the cross-attention modules:

$$Q = W_Q \cdot \phi_i,\ K = W_K \cdot \tau_\theta(y),\ V = W_V \cdot \tau_\theta(y),$$
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \quad (2)$$

where ϕ_i is the intermediate image features. W_Q, W_K, and W_V denote the learnable projection matrices of query, key, and value, respectively. In the cross-attention module, the interaction between each text token and image token significantly influences the layout of the generated image.

3.2. Adaptive Layout-Semantic Fusion Network

In this subsection, we will first introduce our proposed layout control map and then present the details of our adaptive layout-semantic fusion module.

Layout control map It has been noted that cross-attention maps in Stable Diffusion are closely related to the layout of the synthesized image [1, 2]. Specifically, $A^{ca}_{i,j}$ in the cross-attention map $A^{ca} \in \mathbb{R}^{(HW)\times N}$ determines the strength of the association between the i-th image token and the j-th text token, thus influencing the layout of synthesized images. Previous works roughly controlled the position of specific objects in the synthesized image by constraining the influence region of specific tokens in the cross-attention map. However, due to the significantly smaller size of the intermediate image features in LDM (less than or equal to 64×64) compared to that of given semantic layouts (512×512 or larger), simply resizing semantic maps to adapt to the size of intermediate features inevitably leads to distortion or even loss of details. For example, as illustrated in Fig. 3, even when the semantic map is resized to only $1/8$ of its original size with a naive nearest-neighbor interpolation, the details of the 'plants' are distorted, and some instances of 'light' are lost. Moreover, the image features from deeper layers have smaller dimensions, making it challenging to synthesize images that align precisely with the given semantic maps.

To address the above issue, we propose a layout control map that encodes layout information in the low-resolution feature space with less loss of layout information. For each token of the intermediate image features, we consider all the semantic components within its receptive field, along with the proportion occupied by each class. We then use a vector composed of these proportions as the layout feature for this token. As shown at the top of Fig. 2 (b), within the receptive field of the image token selected by the red border, there are four semantic categories, namely 'wall', 'ceiling', 'window', and 'light', and each corresponding to a different proportion. The vector formed by the proportions faithfully encodes the layout information within the receptive field of this image token. Given a semantic map $\hat{S} \in \mathbb{R}^{(HW)\times C}$ reshaped from $S \in \mathbb{R}^{H\times W\times C}$, the calculation of layout control map $L \in \mathbb{R}^{(hw)\times N}$ can be formulated as following:

$$L_{i,j} = \begin{cases} \frac{|\hat{S}_{RF(i),S(j)}=1|}{|RF(i)|}, & |\hat{S}_{RF(i),S(j)} = 1| \neq 0 \\ -\infty, & \text{otherwise}, \end{cases} \quad (3)$$

where $RF(i)$ denotes the receptive field of the i-th image token and $S(j)$ is the corresponding semantic channel of the j-th text token. $|\cdot|$ represents the number of elements in the set. As can be observed from Fig. 3, our layout control map encodes faithful layout details, including the branches and leaves of the plants, and subtle lighting, resulting in a synthesized image with richer and accurate details.

Adaptive fusion of layout and semantics Although manually restricting the influence region of each semantic component with the layout control map can manipulate the position of synthesized objects, the synthesis of specific objects fails to benefit from the global textual context. To appropriately incorporate our layout representations into the image synthesis process and synthesize high-quality images with desired layouts, we propose an adaptive layout-semantic fusion module (PLACE). As depicted in Fig. 2 (b), in each fusion module, the time embedding is fed into a linear layer to predict an adaptive fusion parameter α, which is then used to integrate the layout control map L and the cross attention map A^{ca} to produce the adaptive fusion map F and final output feature O:

$$F = \alpha\,(\text{softmax}(L \odot A^{ca})) + (1-\alpha)A^{ca},\ O = FV. \quad (4)$$

By adopting a timestep adaptive parameter α as the weight to fuse layout and semantic features, the global interactions

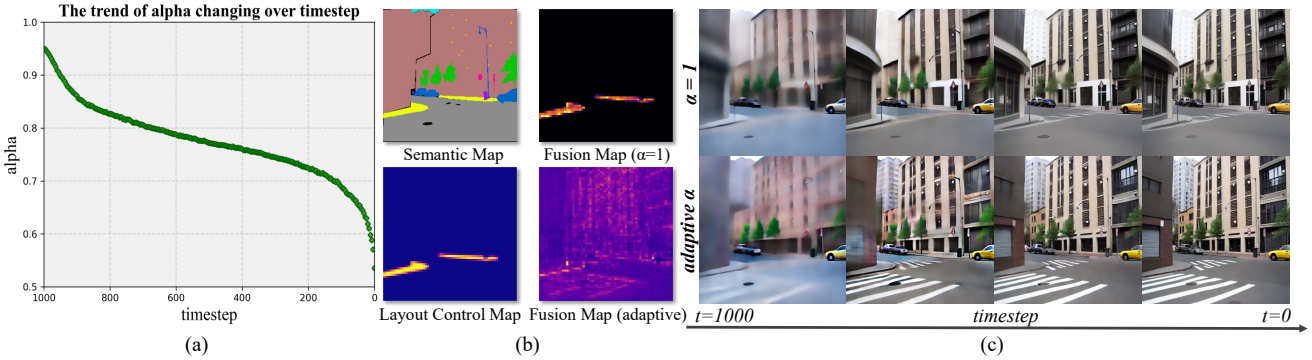

Figure 4. Analysis of adaptive fusion module. (a) shows the variation of adaptive α with respect to the timestep. α decreases as the timestep progresses. (b) presents the layout control map of the 'sidewalk' and the corresponding comparison of the fusion maps (at $t = 800/1000$) between fixed $\alpha = 1$ and adaptive α. (c) illustrates the variation of predicted $\hat{x}_0$ with respect to the sampling steps: one with a fixed α and the other with an adaptive α. The latter leads to the synthesis of more realistic details. Zoom in for details.

between image tokens and text tokens in the Stable Diffusion are maintained. The interactive mechanism allows each image token to access contextual information from a larger set of textual tokens. Such an adaptive integration of layout and semantics not only helps control the layout of the synthesized image but also facilitates the synthesis of high-quality details. The trend of learned adaptive α changing over timestep validates the effectiveness of our approach. As shown in Fig. 4 (a), the relatively large value of α during the early stages of sampling indicates the crucial role of the layout control map in determining the initial layout. However, in later stages, the adaptive α gradually decreases, suggesting that the influence of the layout control map diminishes as the process proceeds. This enables the image token to actively interact with global textual tokens, thereby synthesizing more realistic details and high-quality results. Fig. 4 (c) shows the variations of the predicted $\hat{x}_0$ during the image synthesis process for both fixed α ($\alpha = 1$) and adaptive α. One can see that in the early sampling phase, the layout of $\hat{x}_0$ is determined, while in the later stages, the model mainly synthesizes realistic details. Moreover, compared to the fixed $\alpha = 1$ case, adaptive fusion allows the image token to extract information from a greater set of text tokens, enabling synthesizing images with richer and more realistic details, such as the 'pedestrian crossing' in the 'road' shown in the figure. This is also corroborated in Fig. 4(b), where the text token 'sidewalk' not only influences its corresponding semantic region but also affects other contextual regions, such as the 'road' class. However, this kind of interaction could not be observed under the fixed α condition.

3.3. Learning Objective

During the fine-tuning stage, in addition to the original text-to-image denoising loss, we also introduce a semantic alignment (SA) loss and a layout-free prior preservation (LFP) loss to facilitate the learning.

Semantic Alignment Loss To further enhance the layout alignment of synthesized images, we propose the semantic alignment loss $\mathcal{L}_{SA}$. As illustrated in Fig. 5, we first utilize the adaptive fusion map $F \in \mathbb{R}^{N\times H\times W}$ as weights to aggregate the self-attention map $A^{sa} \in \mathbb{R}^{(HW)\times H\times W}$, resulting in weighted aggregation maps $W \in \mathbb{R}^{N\times H\times W}$. We then aim to minimize the difference between them and the original adaptive fusion maps, which can be formulated as:

$$W_i = \sum_j Reshape(F_i)_j \cdot A^{sa}_j,$$
$$\mathcal{L}_{SA} = \sum_i ||W_i - F_i||^2, \tag{5}$$

where $Reshape(\cdot)$ denotes the flatten operation and $Reshape(F_i) \in \mathbb{R}^{(HW)\times 1}$. $\mathcal{L}_{SA}$ effectively encourages image tokens to interact more with the same and related semantic regions in the self-attention module, thereby further improving the layout alignment of the generated images.

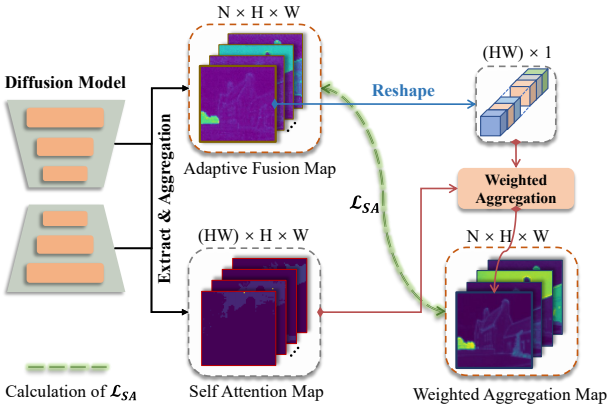

Figure 5. Calculation of the Semantic Alignment loss.

Layout-Free Prior Preservation Loss Due to the limited scale of the fine-tuning dataset, the model inevitably suffers from loss of semantic priors, resulting in suboptimal performance of semantic consistency and visual quality. Enlarging the scale of the fine-tuning dataset is one possible way to address this issue. However, obtaining a substantial number of real images annotated with semantic masks is non-trivial.

We introduce a Layout-Free Prior Preservation (LFP) loss to alleviate this issue. It relies solely on text-image data pairs to help preserve the prior knowledge of the pre-trained model, which is relatively more accessible. During each fine-tuning iteration, in addition to sampling regular paired training data with semantic mask annotations $< z_t, S, y, t >$, we also extract an additional set of text-image data pairs $< z'_t, y', t' >$ from the Layout Free (LF) dataset to feed into the network, as shown in Fig. 2 (a). Due to the absence of semantic masks S', we explicitly set the adaptive fusion parameter α to 0 when synthesizing the image. The original denoising loss $\mathcal{L}_{LDM}$ and our LFP loss $\mathcal{L}_{LFP}$ can be computed as follows:

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon,t,S}[||\epsilon - \epsilon_\theta(z_t, t, S, \tau_\theta(y))||_2^2, \quad (6)$$

$$\mathcal{L}_{LFP} := \mathbb{E}_{\mathcal{E}(x),y',\epsilon',t'}||\epsilon' - \epsilon_{\theta,\alpha=0}(z'_t, t', \tau_\theta(y'))||_2^2. \quad (7)$$

We collect approximately 300k text-image pairs from OpenaImages [16] and Laion-5b [34] datasets as the Layout Free dataset. More details about the implementation of the LFP loss can be found in supplementary materials.

Through employing the LFP loss, semantic concepts present in the pre-trained model are better preserved in the fine-tuning process, even without the involvement of semantic masks. Experimental results demonstrate that our model can generate diverse images and exhibit improved performance in visual quality and semantic consistency.

The optimization objective can be summarized by Eq. 5, Eq. 6, and Eq. 7 as follows:

$$\mathcal{L} = \mathcal{L}_{LDM} + \lambda_1 \mathcal{L}_{SA} + \lambda_2 \mathcal{L}_{LFP}, \quad (8)$$

where λ_1 and λ_2 are the weight coefficients, and they are set to 1 as default.

4. Experiments

4.1. Experimental Details

Datasets We conduct our experiments on two challenging datasets, namely ADE20K [53] and COCO-Stuff [3]. ADE20K consists of 150 semantic categories. It has 20,210 images for training and 2,000 images available for validation. COCO-Stuff contains 182 semantic categories covering diverse scenes. It comprises 118,287 training images and 5,000 validation images. During training, both the images and semantic maps are resized to 512×512. All the semantic classes present in the image are joined together with spaces to form the input textual prompt.

Implementation Details We utilize the pre-trained V1-4 Stable Diffusion model [31] as the initialization weights and fine-tune it with a learning rate of 5×10^{-6}. All experiments are conducted on a server with 4 NVIDIA V100 32G GPUs. We fine-tuned for about 300k iterations and the batch size is 4. During sampling, we employ 50 PLMS [18] sampling steps with a classifier-free guidance [11] scale of 2.

Methods	ADE20K		COCO-Stuff	
	mIoU ↑	FID ↓	mIoU ↑	FID ↓
pix2pixHD [41]	20.3	81.8	14.6	111.5
SPADE [26]	38.5	33.9	37.4	22.6
CC-FPSE [19]	43.7	31.7	41.6	19.2
LGGAN [38]	41.6	31.6	N/A	N/A
OASIS [37]	48.3	28.3	44.1	17.0
SC-GAN [43]	45.2	29.3	42.0	18.1
SAFM [22]	50.1	32.8	43.3	24.6
RESAIL [35]	49.3	30.2	<u>44.7</u>	18.3
ECGAN [39]	<u>50.6</u>	25.8	**46.3**	15.7
SDM [42]	39.2	27.5	40.2	15.9
PITI [40]	29.4	27.9	34.1	16.1
ControlNet [50]	36.9	31.2	N/A	N/A
T2I-Adapter [23]	N/A	N/A	20.7	16.8
FreestyleNet [48]	41.9	<u>25.0</u>	40.7	<u>14.4</u>
Ours	**50.7**	**22.3**	42.6	**14.0**

Table 1. Quantitative comparison on the ADE20K and COCO-Stuff. The upper row shows the results of GAN-based methods, while the lower row displays the scores of those based on diffusion models. ↑ (↓) indicates higher (lower) is better.

Methods	New Obj.		New Sty.	New Attri.
	mIoU (↑)	FID (↓)	Text-Alignment (↑)	Text-Alignment (↑)
ControlNet	18.2	27.4	0.274	0.284
FreestyleNet	24.6	20.4	0.260	0.269
Ours	**33.0**	**18.1**	**0.279**	**0.290**

Table 2. Comparison of out-of-distribution synthesis. ↑ (↓) indicates higher (lower) is better.

Evaluation Metrics Following prior works on semantic image synthesis [48], we quantitatively evaluate the results of in-distribution synthesis with Fréchet Inception Distance (FID) [10] and the mean Intersection over Union (mIoU). FID assesses the visual quality of generated images, while the mIoU measures semantic and layout consistency. Besides, benefiting from the superior prior of the pre-trained model, our method also shows the capability for out-of-distribution synthesis. We evaluate this ability from three perspectives, namely new object, new style, and new attribute. For new object synthesis, we employ the model fine-tuned on ADE20K to synthesize semantic categories that exclusively appear in COCO-Stuff (*i.e.*, categories not contained in the ADE20K). FID and mIoU are adopted to assess the quality and consistency of the results. Regarding new style and new attribute synthesis, we synthesize 260 images with 8 new global styles and 6 specific object attributes with the same model. We measure the consistency of the synthesized results with specific styles or attributes using CLIP [29] text-image similarity (*i.e.*, text alignment). More details can be found in supplementary materials.

semantic label SPADE CC-FPSE OASIS SCGAN SAFM FreestyleNet Ours

Figure 6. Visual comparisons on ADE20K ($1st \sim 3rd$ rows) and COCO-Stuff ($4th \sim 5th$ rows).

4.2. Evaluation of In-distribution Synthesis

Quantitative comparisons Table 1 reports the FID and mIoU performance of our approach compared to other competing methods. The upper rows present the results of GAN-based methods, while the lower rows display the scores of methods based on pre-trained text-to-image models. As shown, our method achieves FID scores of 22.3 and 14.0 on the ADE20K and COCO-Stuff datasets, respectively, which are 2.7 and 0.4 lower than the second-best scores. In terms of alignment, our method obtains results comparable to the state-of-the-art. On the ADE20K, our mIoU score reaches 50.7, while on the COCO-Stuff, our score is 42.6. The quantitative results indicate that our method not only achieves comparable performance in semantic and layout consistency to the current state-of-the-art works but also attains the highest image quality scores. The reliable layout representation allows our approach to demonstrate enhanced consistency in layout details compared to other methods based on pre-trained text-to-image models, on par with the most advanced GAN-based methods. Moreover, the efficient interaction between layout and semantic features in the adaptive layout-semantic fusion plays a vital role in synthesizing high-quality images.

Qualitative comparisons Fig. 6 illustrates the qualitative comparisons on the ADE20K and COCO-Stuff, from which the following observations can be made: (1) Our method produces synthesis results that exhibit higher fidelity to the semantic layout. For example, the 'clock' in the $3rd$ row demonstrates improved alignment with the semantic layout. (2) The images synthesized by our method demonstrate more realistic details. Notable examples include the 'bed' in the $1st$ row, the 'table' in the $2nd$ row, and the 'pedestrian crossing' in the $5th$ row. (3) Our method effectively preserves and utilizes the priors in the pre-trained model. An example can be seen in the $4th$ row with the preservation of the signage. More qualitative results can be found in supplementary materials.

4.3. Evaluation of Out-distribution Synthesis

Table 2 and Fig. 7 present the quantitative and qualitative comparisons of out-of-distribution synthesis results respectively. The evaluation of out-of-distribution synthesis comprises three aspects, namely new object, new style, and new attribute. As shown, our method achieves superior quantitative scores in all three aspects compared to both ControlNet and FreestyleNet. Especially in the new object category, our method achieves a significant mIoU improvement of 8.4 compared to FreestyleNet. From the visual comparison, it is evident that our method synthesizes out-of-distribution images with not only better semantic consistency with the given conditions (*i.e.*, new semantic, new style, and new attribute) but also maintains good performance in terms of layout alignment. For instance, in Fig. 7, the 'anime' style in the $1st$ row, the 'rainbow' in the $3rd$ row, and the 'bird' in the $4th$ row are all faithfully consistent with the provided conditions. The 'bird' and the 'bear' in the $4th$ row demonstrate strong layout alignment. More results can be found in supplementary materials.

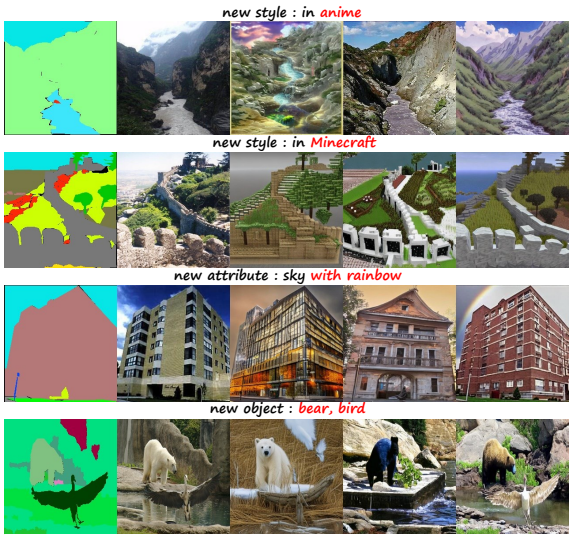

Figure 7. Visual comparison of out-of-distribution synthesis.

4.4. Ablation Study

We conduct the ablation study with variant models fine-tuned on the ADE20K dataset to validate the effectiveness of our method. Our baseline model employs a simple nearest resized semantic map to determine the region of influence for each text token based on given semantic maps. It does not involve adaptive fusion and is fine-tuned only with denoising loss. Besides, we utilize a more comprehensive vocabulary, wherein a greater number of synonyms are employed to represent the same semantic category.

The comparisons of quantitative and qualitative results are presented in Table 3 and Fig. 8, respectively. In Table 3, LCM denotes the Layout Control Map. Ada-α indicates the usage of the timestep-adaptive parameter during fusion, SA represents Semantic Alignment loss, and LFP refers to Layout-Free Prior Preservation loss. The ✓indicates the adoption of the corresponding module or strategy during the experiments. We compared the FID and mIoU scores of ADE20K and the new object classes (denoted by 'New Obj.') from COCO-Stuff. More qualitative results can be found in supplementary materials.

Layout control map From (1) and (2) in Table 3, the layout control map significantly improves the mIoU scores, increasing from 43.5 to 48.6 for in-distribution synthesis and from 25.3 to 28.5 for out-of-distribution synthesis. Besides, from Fig. 8, with the layout control map, our method can generate images that adhere closely to given layouts (*e.g.*, the 'table' and 'plants' in $1st$ row, and the 'street light' in $2nd$ row), demonstrating its effectiveness.

Adaptive α for fusion Referring to the (1) and (3) as well as the (2) and (5) in Table 3, with the adaptive layout-semantic fusion, the FID scores on the ADE20K decrease by 1.2 and 0.7, respectively. It is evident that the adaptive fusion enhances the quality of the synthesized images. In Fig. 8, the results using adaptive fusion exhibit more realistic details

Figure 8. Visual comparisons of different variants.

	Methods				ADE20K		New Obj.	
	LCM	Ada-α	SA	LFP	mIoU ↑	FID ↓	mIoU* ↑	FID ↓
(1)					43.5	24.2	25.3	20.2
(2)	✓				48.6	23.4	28.5	19.8
(3)		✓			46.2	23.0	26.5	19.4
(4)			✓		46.7	23.9	27.1	19.9
(5)	✓	✓			50.1	22.7	29.4	19.3
(6)	✓	✓	✓		**50.9**	22.8	29.9	19.4
(7)	✓	✓		✓	49.8	**22.3**	32.8	**18.1**
(8)	✓	✓	✓	✓	50.7	**22.3**	**33.0**	**18.1**

Table 3. Quantitative comparison of five variants on Ade20K in the ablation study. The mIoU* denotes the mIoU scores of semantic classes that are exclusively present in the COCO-Stuff dataset.

(*e.g.*, 'window' and 'road' in $1st$ and $2nd$ rows).

Semantic Alignment loss Both the (1) and (4) along with the (5) and (6) in Table 3 indicate that the semantic alignment loss contributes to the consistency of the layout. The mIoU scores increase by 3.2 and 0.8 on the ADE20K, individually. As shown in Fig. 8, the alignment loss also helps synthesize more realistic instances (*e.g.*, $3rd$ row).

Layout Free Prior Preservation loss The LFP loss better preserves the semantic priors in the pre-trained model, resulting in improved performance for visual quality and semantic consistency. The (5) and (7) in Table 3 show that the LFP loss leads to a 3.4 increase in the mIoU score and a 1.2 decrease in the FID score for new object synthesis. From experiments (6) and (8), the mIoU and FID scores of synthesized new objects improved by 3.1 and 1.3, respectively. With LFP loss, the $3rd$ row in Fig. 8 presents a realistic 'cat', even though it has never appeared in training dataset.

5. Conclusion

In this paper, we first present a novel layout control map for reliable representations of layout features. We further combine the semantic and layout features adaptively, resulting in the synthesis of high-quality images that are faithfully aligned with given semantic layouts. Additionally, we propose a semantic alignment loss to facilitate the layout alignment and a layout-free prior preservation loss to maintain semantic priors of pre-trained models for fine-tuning. Extensive quantitative and qualitative results demonstrate that PLACE exhibits remarkable visual quality, semantic consistency, and layout alignment for both the in-distribution and out-of-distribution semantic image synthesis.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 2, 3, 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3, 4

[3] 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. 6

[4] Yufei Cai, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hu Han, and Wangmeng Zuo. Decoupled textual embeddings for customized image generation. *arXiv preprint arXiv:2312.11826*, 2023. 2

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

[6] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 1

[7] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[9] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[10] 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. 6

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Tianyu Huang, Yihan Zeng, Zhilu Zhang, Wan Xu, Hang Xu, Songcen Xu, Rynson WH Lau, and Wangmeng Zuo. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. *arXiv preprint arXiv:2312.06439*, 2023. 2

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2

[15] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 3

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, 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. 6

[17] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. *arXiv preprint arXiv:2311.17461*, 2023. 2

[18] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 6

[19] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019. 1, 2, 6

[20] Wuyang Luo, Su Yang, Hong Wang, Bo Long, and Weishan Zhang. Context-consistent semantic image editing with style-preserved modulation. In *European Conference on Computer Vision*, pages 561–578. Springer, 2022. 1

[21] Wuyang Luo, Su Yang, Xinjian Zhang, and Weishan Zhang. Siedob: Semantic image editing by disentangling object and background. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1868–1878, 2023. 1

[22] Zhengyao Lv, Xiaoming Li, Zhenxing Niu, Bing Cao, and Wangmeng Zuo. Semantic-shape adaptive feature modulation for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11214–11223, 2022. 2, 6

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 3, 6

[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. 2

[25] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 394–411. Springer, 2020. 1

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 2337–2346, 2019. 1, 2, 6

[27] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 3

[28] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 2

[29] 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. 3, 6

[30] 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. 1, 2

[31] 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. 1, 2, 3, 6

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[35] Yupeng Shi, Xiao Liu, Yuxiang Wei, Zhongqin Wu, and Wangmeng Zuo. Retrieval-based spatially adaptive normalization for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11224–11233, 2022. 2, 6

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[37] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 1, 2, 6

[38] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020. 2, 6

[39] Hao Tang, Guolei Sun, Nicu Sebe, and Luc Van Gool. Edge guided gans with multi-scale contrastive learning for semantic image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 6

[40] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 6

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 6

[42] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 6

[43] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. *arXiv preprint arXiv:2109.07053*, 2021. 2, 6

[44] Yuxiang Wei, Zhilong Ji, Xiaohe Wu, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Inferring and leveraging parts from object shape for improving semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11248–11258, 2023. 2

[45] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[46] Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qingming Huang. R&b: Region and boundary aware zero-shot grounded text-to-image generation. *arXiv preprint arXiv:2310.08872*, 2023. 3

[47] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3

[48] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 1, 3, 6

[49] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *arXiv preprint arXiv:2310.15160*, 2023. 1

[50] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 3, 6

[51] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2

[52] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[54] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1