Training-free Composite Scene Generation for Layout-to-Image Synthesis

Jiaqi Liu[1], Tao Huang[1], and Chang Xu[1]

School of Computer Science, Faculty of Engineering, The University of Sydney
{jliu6979,thua7590}@uni.sydney.edu.au, c.xu@sydney.edu.au

Abstract. Recent breakthroughs in text-to-image diffusion models have significantly advanced the generation of high-fidelity, photo-realistic images from textual descriptions. Yet, these models often struggle with interpreting spatial arrangements from text, hindering their ability to produce images with precise spatial configurations. To bridge this gap, layout-to-image generation has emerged as a promising direction. However, training-based approaches are limited by the need for extensively annotated datasets, leading to high data acquisition costs and a constrained conceptual scope. Conversely, training-free methods face challenges in accurately locating and generating semantically similar objects within complex compositions. This paper introduces a novel training-free approach designed to overcome adversarial semantic intersections during the diffusion conditioning phase. By refining intra-token loss with selective sampling and enhancing the diffusion process with attention redistribution, we propose two innovative constraints: 1) an inter-token constraint that resolves token conflicts to ensure accurate concept synthesis; and 2) a self-attention constraint that improves pixel-to-pixel relationships. Our evaluations confirm the effectiveness of leveraging layout information for guiding the diffusion process, generating content-rich images with enhanced fidelity and complexity. Code is available at https://github.com/Papple-F/csg.git.

Keywords: Image generation · Layout-to-image synthesis · Diffusion models

1 Introduction

Substantial advancements have been achieved in large-scale text-to-image generative models [31–35, 37], which are now capable of converting complex text descriptions into visually compelling images with impressive accuracy and creativity. Despite these advancements, current models still struggle to comprehend layout descriptions contained within texts and to precisely outline and render detailed images. This is largely due to limitations in model capacity and the quality of training datasets. Consequently, in tasks such as art creation [10] and

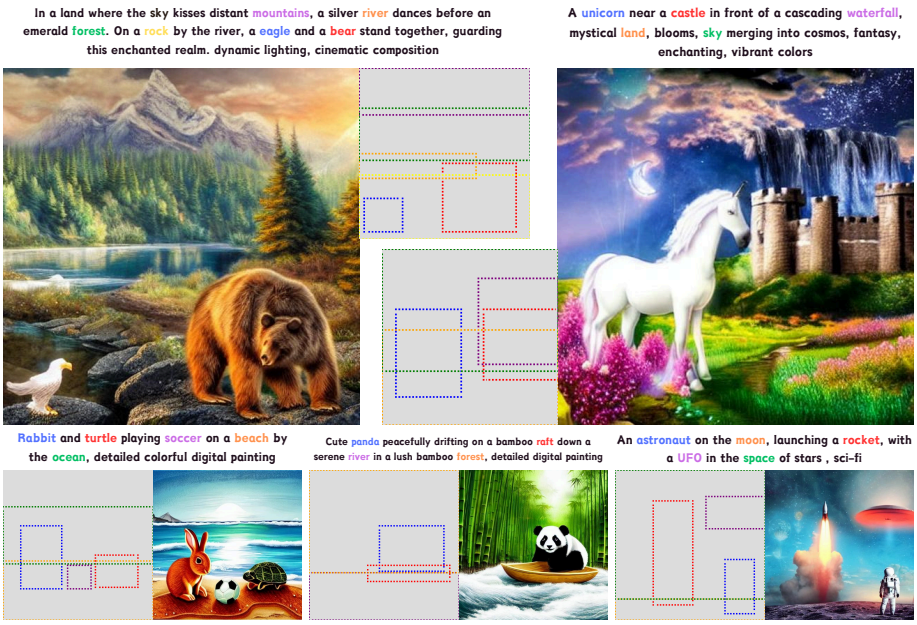


Fig. 1: Composite scene generation facilitates the blending of various foreground and background elements into an image based on layout details.

industrial design [28], where elements need to be precisely positioned, the models' intrinsic inability to comprehend user-defined spatial information limits their applicability.

Recent advancements in layout-to-image approaches have shown significant progress in achieving more controllable generation, employing both training-based [2, 6, 14, 23, 25, 43, 44] and training-free [3, 8, 22, 42] methodologies to transform detailed layouts into vivid, realistic images. While training-based methods have demonstrated promise, their effectiveness heavily relies on the availability of comprehensive and varied layout datasets. Assembling, annotating, and preprocessing these datasets for effective concept learning is challenging due to the high costs and resource-intensive nature of data acquisition, posing a substantial barrier for many research initiatives. In contrast, training-free methods utilize the intrinsic generative capabilities of models to dynamically guide the diffusion process, offering a significant efficiency advantage. As text-to-image models become more popular, there is an increasing demand for content generation that involves complex compositions of multiple objects and backgrounds. Existing layout-to-image approaches, which primarily focus on content positioning, often struggle with conflicts arising from intersecting similar concepts, especially when generating scenes with multiple elements. Moreover, the critical role of self-attention mechanisms in refining generated content with accurate textural details during the later stages of diffusion, as highlighted in [2], is often over-

looked. This oversight highlights a gap in current methodologies, which may fail to fully leverage the potential of diffusion models for producing detailed and contextually coherent images.

This study aims to generate composite scenes featuring multiple objects and backgrounds, utilizing bounding boxes for layout information. We propose enhancing the training-free backward guidance concept [7,8] with a novel selective sampling strategy. This strategy introduces a dropout mechanism that prioritizes attentions closely aligned with the current generation concept during the calculation of intra-token constraints. Such an approach not only improves content positioning accuracy but also ensures broader coverage of the targeted area, addressing key challenges in layout-to-image generation.

Our approach extends beyond individual token cross-attentions by implementing an inter-token constraint, evaluating attentions across tokens within a targeted region to ensure the prioritization of the correct concept. This technique aims to counteract semantic intersection—where overlapping concepts produce irregular textures and shape inaccuracies—thereby enhancing generation precision and reliability. Recognizing adversarial intersections' impact on pixel relationships, we employ a self-attention constraint for collective self-attention adjustment across the target region, critical in the later diffusion stages for maintaining coherent pixel interactions. Additionally, an attention redistribution method during forward diffusion corrects misaligned attentions, addressing refinement limitations and reducing semantic intersection effects, thereby improving overall generation accuracy.

Our comprehensive experimental evaluations demonstrate that our method significantly outperforms existing training-free layout-to-image generation approaches in terms of content localization accuracy and semantic fidelity. These advancements are vividly illustrated in the examples presented in Figure 1, showcasing our method's ability to maintain higher semantic correctness while accurately positioning content within the generated images.

2 Related Work

2.1 Text-to-image generative models

Recent advancements in text-to-image generative models, exemplified by Stable Diffusion [33], DALLE-3 [37], and Imagen [35], represent a significant leap forward beyond previous dominant techniques like generative adversarial networks [4,15,16,21]. These new models excel not only in image generation but also in enhancing the performance of tasks such as classification [1,20], action segmentation [12,27], and more. These models enable the generation of highly detailed visual content directly from textual descriptions, distinguishing themselves by producing contextually relevant and aesthetically pleasing images. Their ability to translate complex textual prompts into visual artworks demonstrates a remarkable proficiency, unlocking new possibilities across various applications, from digital art creation [10] to diverse content generation.

However, these models heavily rely on large-scale datasets (stable diffusion is trained using billions of images from LAION-5B [36]), which poses limitations in extending their capabilities to cover new tasks with learned concepts without incorporating additional task-specific datasets or extensive training. As previous studies have showcased the feasibility of integrating various novel tasks in a plug-and-play manner, including image editing [5,17,29,39], generation enhancement [7,24], layout-to-image [3,8,22,42], and more. Encouraged by these findings, we are motivated to delve deeper into whether the boundaries of layout-to-image tasks can be expanded further, enabling the model to handle scenarios involving multiple objects and backgrounds.

2.2 Layout-to-image generation

The reliance on purely linguistic methods limits a model's ability to decode specific layout details accurately. Various studies [2, 6, 14, 23, 25, 43, 44] have shown that models can be further trained with layout information to facilitate layout-to-image generation tasks. Nonetheless, these approaches also highlight the challenge of requiring datasets that pair image, text, and layout information—resources that are scarce and costly to compile. Some innovative strategies have aimed to integrate layout details without additional model training. For instance, MultiDiffusion [3] involves denoising and combining regions with corresponding text descriptions, while DenseDiff [22] directly modifies the attention probabilities to enhance focus on the targeted regions. BoxDiff [42] and Layout-control [8] both employs generative semantic nursing [7], which optimizing latents based on cross-attentions to achieve desired layouts.

Our approach adopts the backward guidance framework utilized by BoxDiff [42] and Layout-control [8]. However, while these existing methods are efficient for straightforward situations, their method narrowly focus on cross-attentions for individual tokens, overlooking the potential for semantic overlaps as the layout becomes more complex with the addition of multiple objects. This oversight can lead to the undesirable blending of features and disrupt the integrity of pixel relationships, which may prevent the accurate generation of targeted objects. Unlike these methods, our strategy takes a comprehensive perspective on the optimization process, integrating considerations of intra-token, inter-token cross attentions, and self-attentions. This broader approach demonstrates that a training-free method can not only adhere more closely to the intended layout but also enhance the quality of the generated images.

3 Preliminaries

3.1 Stable diffusion

Diffusion models [9,11,18,38], operates by gradually transforming a random noise distribution into a coherent image, guided by the semantics of the input text. This process involves a series of forward and backward steps, where the model

initially adds noise ϵ to an image $\boldsymbol{\mathcal{I}}$ and then learns to recover the original image from noise, conditioned on textual descriptions $\boldsymbol{c}$. Latent diffusion model [33] operates in latent space where for a given image $\boldsymbol{\mathcal{I}}$, it is first encoded as latent $\boldsymbol{z}$ by an encoder $\mathcal{E}$, and then reconstructed by a decoder $\mathcal{D}$ as $\hat{\boldsymbol{\mathcal{I}}} = \mathcal{D}(\boldsymbol{z}) = \mathcal{D}(\mathcal{E}(\boldsymbol{\mathcal{I}}))$ after the denoising process.

During the training stage of diffusion process, at given timestep t, with latent $\boldsymbol{z}_t$ and condition $\boldsymbol{c}$, the denoiser ϵ_θ learns to correctly predict the added noise ϵ through mean square error:

$$\mathcal{L}_{diff} = \mathbb{E}_{z\sim\mathcal{E}(\boldsymbol{\mathcal{I}}),\epsilon\sim N(0,1),\boldsymbol{c},t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(z_t, c, t)\|_2^2]. \tag{1}$$

During the inference stage, the denoising process gradually removes noise from a randomly sampled $\boldsymbol{z}_T$ with the noises predicted by diffusion model with condition $\boldsymbol{c}$, and guide the diffusion process through classifier [11] or classifier-free guidance [19].

3.2 Attention mechanism

The attention mechanism forms the cornerstone of transformer models [40], playing a pivotal role in natural language processing (NLP) tasks. This mechanism was subsequently integrated into the vision transformer architecture [13], thereby extending its applicability to tasks within the realm of computer vision. Within the framework of stable diffusion models, the intermediate features are processed through both self-attention and cross-attention layers. This processing facilitates the generation of an attention map at timestep t as follows:

$$A_t = softmax(\frac{QK^T}{\sqrt{d}}). \tag{2}$$

In this process, Q and K represent queries and keys, respectively, with d denoting the dimensionality of these query and key features. For cross-attention mechanisms, K is derived from the projection of text embeddings. These embeddings result from encoding the text conditioning c into a latent space, achieved through the use of a CLIP text encoder [30]. Conversely, Q corresponds to the projection of intermediate features sourced from a U-Net architecture. Within the self-attention framework, both Q and K are obtained from the projections of these intermediate features.

4 Methods

In this section, we explore the nuances of generating composite scenes. We begin in Section 4.1 by detailing our novel approach to attention selection, including the definition of mask representations and the application of our selective sampling strategy. This section also explains the development and role of intra- and inter-token constraints in influencing the diffusion process. In Section 4.2, we shift focus to our innovative handling of self-attentions, illustrating their pivotal role

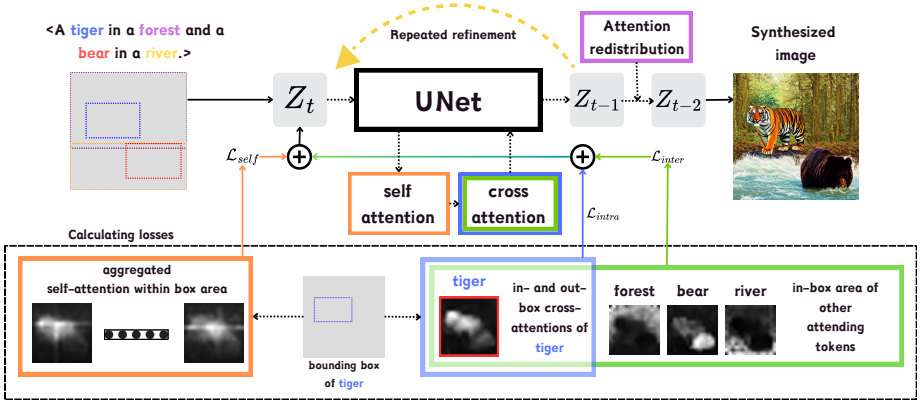


Fig. 2: In the workflow of Composite Scene Generation (CSG), at each refinement stage, we capture both self- and cross-attentions within a UNet structure. For self-attention, we aggregate self-attentions within each mask area $\boldsymbol{m}_i$ and calculate $\mathcal{L}_{self}$, which determines if pixel-level interaction is mostly constrained within the target area. For cross-attention, we first obtain $\mathcal{L}_{intra}$, a proportional measure of in-box and out-box cross-attentions for each attending token. Next, we assess the cross-attentions for all attending tokens within the same box area to obtain $\mathcal{L}_{inter}$, determining if the cross-attention of the current token is dominant within its own region. After a finite number of refinement steps, the latent is updated through the gradient of all three loss components. To further enhance the refinement process, we implement attention redistribution between each refinement stage.

in refining the diffusion process for enhanced image generation. Finally, Section 4.3 introduces our attention redistribution technique, employed during forward diffusion to further optimize generation outcomes through strategic constraint refinement and application.

4.1 Cross-attention constraints

To formulate constraints on cross-attentions, we begin by detailing the acquisition of attentions and layout information for constraint application. Leveraging latent noise manipulation is pivotal in aligning image-text attentions within designated spatial regions. Drawing on recent research [7, 42] which suggests that attentions at smaller scales capture semantic information more effectively during early diffusion phases, we employ averaged cross-attentions from user-defined attending objects at the 16×16 up-scaling block as our manipulation references. Formally, for N attending tokens, we define $\mathcal{A} = \{\boldsymbol{A}_i\}_{i=1}^{N}, \boldsymbol{A}_i \in \mathbb{R}^L$ as the set of cross-attentions relative to the tokens, with L indicating the number of image tokens in cross-attentions.

For representing layout information, we employ bounding boxes $\boldsymbol{\mathcal{B}} = \{\boldsymbol{b}_i\}_{i=1}^{N}$, which delineate the top-left and bottom-right corner coordinates for each object.

Correspondingly, each bounding box $\boldsymbol{b}_i$ is transformed into a binary mask $\boldsymbol{m}_i$, with pixels inside the box region marked as 1 and those outside as 0.

The mask is leveraged in our constraints to discriminate foreground and background pixels of the generation of each object. To mitigate the potential disruption of latent noise's natural Gaussian distribution by gradient-based backpropagation, while ensuring extensive attention coverage within the target region, we employ a `selective sampling` strategy $\mathcal{F}(\boldsymbol{m_i})$ for each mask $\boldsymbol{m_i}$. Specifically, we select the top K elements within the mask region for focused attention. To avoid concentrating excessively on a limited number of elements, we randomly retain M elements from those selected.

Intra-token constraint: For each attending token, we use its corresponding mask $\boldsymbol{m_i}$ and cross-attention probabilities $\boldsymbol{A}_i$ to compute our intra-token attention regularization loss. To make sure that the object is only generated in its own target region (*i.e.*, inside the bounding box), we encourage the cross-attention values associated with the target region to exceed those of the non-target region. Following [8], we define the intra-token constraint as:

$$\begin{aligned} \mathcal{L}_{intra} &= \sum_{i=1}^{N} \mathcal{L}_i^{intra} \\ \text{with} \quad \mathcal{L}_i^{intra} &= (1 - \frac{\sum_{j=1}^{L} (\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_i)_j}{\sum_{j=1}^{L} (\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_i)_j + \sum_{j=1}^{L} (\hat{\boldsymbol{m}}_i \cdot \boldsymbol{A}_i)_j})^2, \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{m}}_i = \mathcal{F}(\boldsymbol{m}_i)$ and $\hat{\boldsymbol{m}}_i = \mathcal{F}(1 - \boldsymbol{m}_i)$.

The constraint is designed to increase the proportion of attention values within the target region relative to those outside, enhancing focus on the intended area. By applying selective sampling to pivotal values, we address the issue of excessive attention concentration in confined areas. Furthermore, selective sampling from both inside and outside regions ensures a more equitable representation of high values, mitigating bias in the aggregation process and enhancing the fairness of their contribution to the loss function.

Inter-token constraint: The semantic intersection due to the conceptual similarity (*e.g.* a bear and a tiger shares considerable similarity in textual latent space), under the context of limited latent space, inevitably intertwine their cross-attentions and cause a degree of spatial intersection. Aiming to completely eliminate such adversarial effects may not be plausible due to the inherent dependence between image intermediate features are extracted from the same latent variable and difficulties to manipulate textual latent space without altering its original semantic meanings. Thus, we employ a conservative approach where we encourage attentions inside corresponding region not to be exclusive, but surpass over other tokens' attentions at the same location. To accomplish this, for a given token i, we first obtain the maximum attentions within target region for other tokens:

$$(\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_j)^{max} = \max_{j=1, j \neq i}^{N} (\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_j). \tag{4}$$

Following, we calculate the difference d between such attentions and attentions of target region of given token with a margin g.

$$d = \sum_{k=1}^{L}((\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_i)_k - g) - \sum_{k=1}^{L}(\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{A}_j)_k^{max}. \tag{5}$$

To achieve our goal of making the given token more prominent than others within the targeted region, we establish an inter-token constraint defined as follows:

$$\mathcal{L}_i^{inter} = \begin{cases} 0, d > 0 \\ d^2, d < 0 \end{cases}, \ \mathcal{L}_{inter} = \sum_{i=1}^{N} \mathcal{L}_i^{inter}. \tag{6}$$

The constraint on inter-token relations, utilizing the same elements from a given $\boldsymbol{m}_i$, focuses on a side-by-side comparison of cross-attention values with other tokens at identical spatial locations. This optimization is particularly aimed at scenarios involving intersections, encompassing both the semantic similarities previously discussed and intersections of user-specified regions. The goal is to promote the dominance of a single token's content within a specific area to ensure the generated image is both clear and coherent.

4.2 Self-attention constraints

The impact of cross-attention is critical in the initial diffusion stages, guiding the structural formation of content via textual cues. However, as [2] highlights, self-attention gains prominence in later diffusion stages, influencing object textures. An unintended effect of semantic intersection can misguide self-attention, causing inappropriate pixel correlations during early stages. To mitigate this, we propose to align self-attention within target regions in early diffusion stages, improving local coherence and maintaining necessary external interactions.

Following the same setup as section 4.1, we obtain averaged self-attentions $\boldsymbol{S} \in \mathbb{R}_{>0}^{L \times L}$ at 16×16 up-scaling blocks. Given mask $\boldsymbol{m}_i$, we aggregated self-attentions which lies within the mask region:

$$\boldsymbol{S}_i = \sum_{j=1}^{L}(\boldsymbol{m}_i \cdot \boldsymbol{S})_j, \tag{7}$$

where $\boldsymbol{S}_i \in \mathbb{R}_{>0}^{L}$.

Similar to the intra-token constraint, the self-attention constraint is defined as:

$$\begin{aligned} \mathcal{L}_{self} &= \sum_{i=1}^{N} \mathcal{L}_i^{self} \\ \text{with} \quad \mathcal{L}_i^{self} &= (1 - \frac{\sum_{j=1}^{L}(\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{S}_i)_j}{\sum_{j=1}^{L}(\tilde{\boldsymbol{m}}_i \cdot \boldsymbol{S}_i)_j + \sum_{j=1}^{L}(\hat{\boldsymbol{m}}_i \cdot \boldsymbol{S}_i)_j})^2, \end{aligned} \tag{8}$$

where $\tilde{\boldsymbol{m}}_i = \mathcal{F}(\boldsymbol{m}_i)$ and $\hat{\boldsymbol{m}}_i = \mathcal{F}(1 - \boldsymbol{m}_i)$.

The self-attention constraint is based on the assumption that pixels within an object have stronger connections to other pixels associated with the same object. Therefore, through selective sampling of averaged self-attention scores, we reinforce these connections by emphasizing self-attentions that are most relevant to the object. This method also leverages the principle that total self-attention probabilities equal to 1 along each channel, meaning while enhancing connections between pixels linked to the object, less relevant pixels are diverted outside the target area, thereby improving the object's coherence with its surroundings.

4.3 In-generation diffusion guidance

Given that the limited latent space can cause semantic-level overlaps, this results in unresolved spatial overlaps even with sufficient refinement steps. These errors accumulate during diffusion steps, potentially leading the generation results in incorrect directions. To address this, we introduce attention redistribution, a technique that reallocates cross-attentions with corresponding tokens during the diffusion process. For each token, its cross-attention is defined as the aggregated cross-attentions across all attending tokens within the bounding box area. After reallocation, we apply max normalization to all cross-attentions. The attention redistribution is defined as follows:

$$\boldsymbol{A}_i = \boldsymbol{m}_i \cdot \sum_{j=1}^{N} (\boldsymbol{A}_j) \tag{9}$$

The generative process is divided into refinement and guidance stages. During each guidance step, the latent variable is updated T_R times (the number of refinement steps per timestep). This process lasts for T_D steps (the total number of timesteps during the diffusion process with such recurrent updates).

For each refinement step, the overall constraint is defined as:

$$\mathcal{L} = \mathcal{L}_{intra} + \mathcal{L}_{inter} + \mathcal{L}_{self}, \tag{10}$$

and the latent is updated with a linear decay factor η_t as:

$$\boldsymbol{z}' = \boldsymbol{z}_t - \eta_t \cdot \nabla_{\boldsymbol{z}_t} \mathcal{L}. \tag{11}$$

Through this approach, at each step of the diffusion process, we refine the latent to achieve more focused attention on the targeted regions. This refinement process helps to resolve conflicts among tokens and aligns self-attentions, thereby improving both internal and external connections within the generated scenes.

5 Experiments

5.1 Experimental setup

Evaluation: In our quantitative analysis using the COCO 2014 dataset [26], we adopt a methodology akin to [3], filtering the dataset to include images with n

Table 1: Ablation study on proposed components.

$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{self}$	SS	AR	AP_{50}	AP	CLIP score
✓			✓	✓	32.1	11.4	0.3202
✓	✓				35.5	12.5	0.3196
✓	✓		✓	✓	40.2	16.1	0.3164
✓	✓		✓		38.1	14.5	0.3175
✓	✓			✓	39.1	14.3	**0.3206**
✓	✓	✓	✓		46.8	19.3	0.3200
✓	✓	✓	✓	✓	**50.3**	**20.4**	0.3201

distinct objects where each occupies at least 5% of the image area. We further refine the selection by excluding objects described by more than two words, as well as "person", resulting in a final set of 61 object classes. For performance evaluation, we employ YOLOv7 [41] for object detection, utilizing metrics such as YOLO-score (AP, AP_{50}) [25] to assess our method's effectiveness of locating and correctly generating objects. Object detection is conducted across all 80 COCO classes to ensure thoroughness. Additionally, we generate randomly sample and generate approximately 1,000 samples for subsets of the COCO 2014 dataset, categorized by the presence of 2, 3 and 2-4 distinct objects, using prompts structured as "`a {object`$_1$`} ... and a {object`$_n$`}`" to systematically evaluate our approach.

We utilize CLIP-score [30] for quantitatively assessing image-text compatibility, thereby evaluating the semantic accuracy of synthesized images. The text descriptions for images are formulated by appending a prefix of "a photo of" to the generation prompts.

Implementation: The experimental findings were derived using the CompVis1.4 model for text-to-image synthesis. We configured the model to perform 50 denoising steps with a constant guidance scale of 7.5, and produced synthetic images at a resolution of 512 × 512. The hyperparameters of selective sample is set to select the highest 80% of attention values and randomly keep 50%. the margin for the inter-token constraint was established at 0.1, a setting that resulted in achieving the highest scores for both YOLO and CLIP metrics.

5.2 Case studies

Quantitative ablation studies: The comprehensive evaluation of our proposed methods encompasses $\mathcal{L}_{intra}$, $\mathcal{L}_{inter}$, $\mathcal{L}_{self}$, selective sampling, and cross-attention redistribution, as summarized in the table 1. The results reveal that utilizing only $\mathcal{L}_{intra}$ with selective sampling and cross-attention redistribution yields subpar object generation and location, achieving a mere 32.1 AP_{50} and 11.4 AP. Introducing $\mathcal{L}_{inter}$ to emphasize the cross-attention of the token over others at the same location leads to a significant enhancement, with improvements of 7.9 in AP_{50} and 3.7 in AP. This highlights the presence of attention over-



Fig. 3: Visual ablation studies on various components of proposed method.

Table 2: Ablation study on margin with K = 80%, M = 50%.

margin	AP_{50}	AP	CLIP score
0.05	48.8	19.3	0.3179
0.1	**50.3**	**20.4**	**0.3201**
0.2	46.8	19.8	0.3167

lap, as previously mentioned, and underscores the effectiveness of our method in addressing this issue. The incorporation of $\mathcal{L}_{self}$ further boosts performance, resulting in 10.1 AP_{50} and 4.3 AP. This suggests that the presence of self-attention distraction compromises the positioning accuracy of layout-to-image methods. The concept of aggregating and treating self-attentions as a whole emerges as an effective strategy to mitigate such issues. Selective sampling and cross-attention redistribution also contribute quantitatively to improved efficacy.

In terms of the CLIP score, the introduction of $\mathcal{L}_{inter}$ appears to compromise the coherence of image semantics, causing a slight reduction from 0.3202 to 0.3164. However, with the addition of $\mathcal{L}_{self}$, our strategy of enhancing interconnection with object pixels while maintaining sufficient redundancy with other pixels proves to enhance semantic coherence.

Qualitative ablation studies: We conducted visual ablation studies to elucidate the progressive impact of our method on the generation process. In the latent space of textual representations, two objects, namely `a tiger and a bear`, inherently share similarities. According to figure 3, when only controlling $\mathcal{L}_{intra}$, these two concepts inevitably overlap and merge, with the more dominant concept exerting greater influence over the generated content. Introducing $\mathcal{L}_{inter}$ and AR helps mitigate this overlap to some extent. While $\mathcal{L}_{inter}$ enhances control over layout precision by competing for the same location, it does violate the fidelity of the resulting image. On the other hand, AR provides a more natural transition from objects to backgrounds. Finally, with the addition of $\mathcal{L}_{self}$, we further fine-tune location precision and re-establish the relationship between objects and backgrounds, aiming to recover some of the fidelity lost due to the impact of $\mathcal{L}_{inter}$.

Hyperparameter ablation studies: Hyperparameters of our work, including K and M for selective sampling and the margin in the inter-token con-

Table 3: Ablation study on K with margin = 0.1, M = 50%.

K	AP_{50}	AP	CLIP score
100%	46.6	18.5	0.3194
80%	**50.3**	**20.4**	0.3201
60%	43.6	16.0	0.3200
40%	37.2	13.3	**0.3207**

Table 4: Ablation study on M with margin = 0.1, K = 80%.

M	AP_{50}	AP	CLIP score
100%	48.2	17.7	**0.3201**
75%	48.9	19.1	**0.3201**
50%	**50.3**	**20.4**	**0.3201**
25%	45.7	16.5	0.3197

straint, play a critical role in the effectiveness of the proposed components and collectively impact the overall performance of our method. The detailed results of hyperparameter optimization are presented in Tables 2, 3, and 4. Specifically, for the inter-token constraint margin, as shown in Table 2, a margin of 0.1 provides the best results. This finding is somewhat counter intuitive, as one might expect that a larger margin would more effectively differentiate between token attentions. However, as discussed in Section 4.1. A plausible explanation for this observation might lie in the fundamentally intertwined nature of cross-attentions that intersect, influenced by text embeddings confined to a constrained latent space, preventing concepts from being represented distinctly, and the latent features that are shared among all tokens.

According to table 3, selecting a high-value portion with $K = 80\%$ results in optimal performance for both AP_{50} and AP metrics, though there is a slight decrease in CLIP score. This can be attributed to the fact that optimizing latents through gradient-based backpropagation can disrupt their naturally Gaussian distribution, potentially compromising image fidelity if not managed cautiously. With a smaller K, our method affects only a limited number of attention values, minimally directing the diffusion process. This approach preserves semantic integrity to some extent but sacrifices precision in object placement and generation. Conversely, using a higher K value, while including random drops, approximates the selection of highly-related attention values to random sampling, which harms performance.

Achieving the best performance with a margin of 0.1 and K at 80%, we find optimal results at M equal to 50% (table 4). This outcome aligns with the patterns observed in our K ablation studies, indicating a balanced exists in our selective sampling technique that prohibits incorporating either too many or too few attention values.

5.3 Qualitative analysis

When conducting a visual comparison (figure 4) using images generated from datasets featuring three distinct objects, we observe that previous methods often struggle with producing a coherent image representation or accurately placing objects. For example, while MultiDiffusion excels at positioning objects in specified regions, it frequently compromises the overall image coherence and object

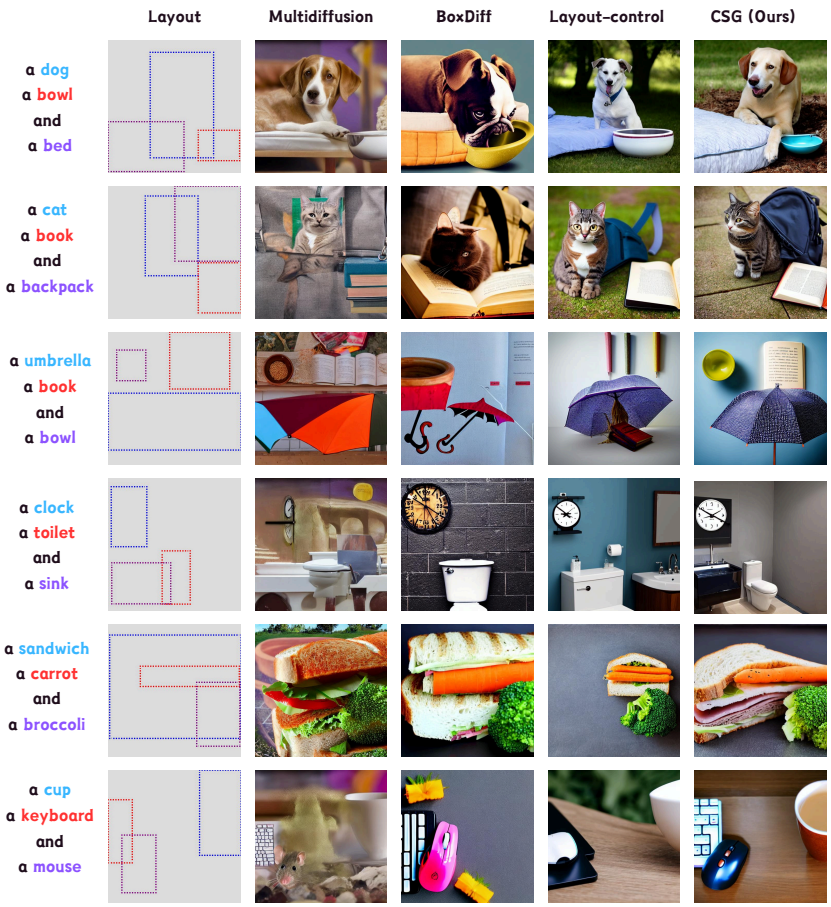


Fig. 4: Visual comparison with concurrent training-free methods including MultiDiffusion [3], BoxDiff [42] and Layout-control [8]. Layout information is sampled from COCO [26] with 3 distinct objects.

integrity, leading to visually jarring images with poor transitions between objects and the background. Conversely, methods like BoxDiff and layout-control, despite offering more coherent image presentations, tend to suffer from issues such as missing objects or the appearance of unintended objects. In contrast, our proposed CSG method successfully combines precise layout placement with semantically coherent image quality, addressing these shortcomings (More qualitative comparisons can be found in supplementary materials).

5.4 Quantitative analysis

In a quantitative evaluation against current training-free techniques such as MultiDiffusion [3], BoxDiff [42], and Layout-control [8], using Stable Diffusion version

1.4 [33], our approach surpasses these methods in metrics of AP_{50}, AP, and CLIP score across datasets with 2, 3, and 2-4 objects, according to table 5. Compared to MultiDiffusion, which demonstrates strong object placement and recognition with high AP_{50} and AP scores, it falls short in generating semantically meaningful content, as indicated by its lower CLIP score. Our method, however, not only improves upon these AP scores by 7.9, 6.6, and 5.8 respectively and AP scores by 1.9, 0.2, and 0.7, but also leads in CLIP score improvements by 0.0281 (7.3%), 0.0217 (7.16%), and 0.0206 (6.86%). When compared to BoxDiff, which shows comparable performance in CLIP scores, our method demonstrates superior AP improvements by 12.5, 11.8, and 6.1 and AP_{50} enhancements by 5.0, 3.2, and 5.6. These outcomes confirm the effectiveness of our method in achieving both accurate object generation and placement as well as semantic fidelity.

Table 5: Quantitative results on datasets with 2, 3 and 2-4 objects for comparison with concurrent training-free methods, including MultiDiffusion [3], BoxDiff [42] and Layout-control [8].

Methods	2 objects			3 objects			2-4 objects		
	AP_{50}	AP	CLIP	AP_{50}	AP	CLIP	AP_{50}	AP	CLIP
Layout-control [8]	24.4	7.84	0.3133	17.5	4.6	0.3181	23.1	7.01	0.3140
BoxDiff [42]	37.8	15.4	0.3124	30.3	11.3	0.3177	33.9	13.5	0.3132
MultiDiffusion [3]	42.4	18.5	0.2983	35.5	14.3	0.3029	42.2	18.4	0.3003
CSG (Ours)	**50.3**	**20.4**	**0.3201**	**42.1**	**14.5**	**0.3246**	**48.0**	**19.1**	**0.3209**

6 Conclusion

This paper introduces a training-free method that surpasses previous approaches in object placement and generation, effectively addressing issues of semantic overlaps and self-attention misalignment that were overlooked by prior research. We showcase artworks created by our method which composite several objects and backgrounds. Compared to these earlier methods, our approach demonstrates superior performance in assimilating layout information and generating high-fidelity images. Through extensive experiments, we show that our method significantly mitigates adversarial effects encountered during the diffusion process. Although our work currently utilizes only bounding boxes as layout information, the proposed method is designed to be compatible with various forms of layout data. Furthermore, given its training-free nature, it can be seamlessly adapted to enhance models pre-trained with layout information, promising improved outcomes. For further discussions, please refer to the supplementary material provided.

Acknowledgements

This work was supported in part by the Australian Research Council under Projects DP210101859 and FT230100549.

References

1. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
6. Chai, S., Zhuang, L., Yan, F.: Layoutdm: Transformer-based diffusion model for layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18349–18358 (2023)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
8. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)
9. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
10. Dehouche, N., Dehouche, K.: What's in a text-to-image prompt? the potential of stable diffusion in visual arts education. Heliyon (2023)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
12. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
16. Guo, T., Xu, C., Huang, J., Wang, Y., Shi, B., Xu, C., Tao, D.: On positive-unlabeled classification in gan. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 8385–8393 (2020)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Huang, T., Liu, J., You, S., Xu, C.: Active generation for image classification. arXiv preprint arXiv:2403.06517 (2024)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
22. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)
23. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
24. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Divide & bind your attention for improved generative semantic nursing. arXiv preprint arXiv:2307.10864 (2023)
25. Li, Z., Wu, J., Koh, I., Tang, Y., Sun, L.: Image synthesis from layout with locality-aware mask adaption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13819–13828 (2021)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
27. Liu, D., Li, Q., Dinh, A.D., Jiang, T., Shah, M., Xu, C.: Diffusion action segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10139–10149 (2023)
28. Liu, M., Hu, Y.: Application potential of stable diffusion in different stages of industrial design. In: International Conference on Human-Computer Interaction. pp. 590–609. Springer (2023)
29. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
36. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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)
37. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of captions. arXiv preprint arXiv:2006.11807 (2020)
38. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
39. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
41. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
42. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
44. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)