PSG-Adapter: Controllable Planning Scene Graph for Improving Text-to-Image Diffusion

Yi Gao[0009−0004−9168−1792]

Nanjing University of Science and Technology, Nanjing, CN
gaoyi@njust.edu.cn

Abstract. Significant progress in text-to-image generation has been driven by the application of diffusion models, highlighting their crucial role and exceptional impact on this field. However, diffusion models often fall short of comprehending spatial relationships within text. This limitation primarily stems from their challenge in constructing logical spatial relationships, such as distinguishing between foreground and background elements. Additionally, their limited text encoding capacity exacerbates inconsistencies in the generated images derived from textual prompts. In this paper, we introduce the **P**lanning **S**cene **G**raph Adapter (**PSG-Adapter**). Our approach employs **P**lanning **S**cene **G**raph (**PSG**) method to decompose the original text prompt into distinct subprompts containing spatial relationships. By leveraging the proposed **P**lanning **S**cene **G**raph ControlNet (**PSG-ControlNet**), additional spatial information is infused into original text embeddings. By fully exploiting the implicit spatial relationships within the text, our method achieves fine-grained control over the composition of the generated images. This enhancement is particularly notable in scenarios involving the generation of multiple objects and complex spatial relationships. Extensive experiments have been conducted to verify the efficacy of PSG-Adapter in generating spatially coherent images and complex scenes with multiple objects and relationships.

Keywords: Deep learning · Text-to-image generation · Diffusion model

1 Introduction

Recent years have seen tremendous advances in text-to-image generation, thanks mostly to diffusion models [7, 11, 19, 26, 29, 33]. By employing large training datasets [15, 31] and sophisticated neural network architectures, these models have attracted a great deal of interest for their ability to produce attractive images that match text descriptions. Even with their effectiveness, existing diffusion models have challenges accurately interpreting complicated text prompts, especially when there are several items involved and complex spatial interactions. These constraints frequently lead to inconsistent and low-fidelity images produced, emphasizing the necessity for more advancements in processing intricate and subtle textual inputs.

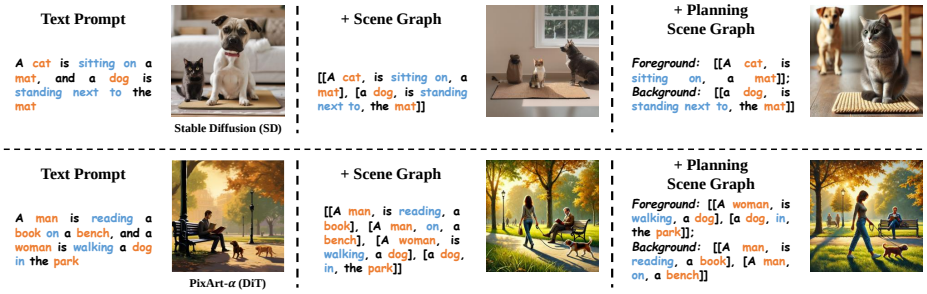


Fig. 1: Utilizing the Planning Scene Graph (PSG) method to embed explicit spatial relationships into the text prompt.

Certain studies encompass additional layouts or bounding boxes as conditions to enhance text-to-image synthesis [13, 14, 35, 36, 39]. For example, ALDM [14] proposes adversarial supervision using a segmentation-based discriminator and a multi-step unrolling strategy to enhance layout adherence and text controllability in layout-to-image generation models. BoxDiff [35] employs three spatial constraints (Inner-Box, Outer-Box, and Corner Constraints) to regulate object placement and scale in text-to-image diffusion models. Alternatively, some approaches focus on training-free learning strategies [6,37]. For instance, RPG [37] is a training-free framework that uses multimodal LLMs (MLLMs) to recaption text prompts, plan image compositions, and generate images with complementary regional diffusion. Pick-and-Draw [18] elevates text-to-image personalization without additional training by using appearance-picking guidance to extract and transfer visual features from a reference image and layout drawing guidance to maintain context diversity. Furthermore, there are methods dedicated to the refinement of text embeddings to facilitate text-to-image translation [16, 28, 32]. Specifically, ReDiffuser [16] enhances text embeddings for zero-shot image translation by generating and fusing rich prompts through cross-attention maps. PRedItOR [28] improves text embeddings by employing a diffusion prior model to create conceptual edits in the CLIP [24] embedding space.

Even though these methods have led to decent improvements, three significant limitations remain in the realm of text-to-image generation: (i) current layout-based or box-based solutions are hindered in effectively modeling the spatial relationships within the text prompt, resulting in images that are uncoordinated and exhibit overlapping elements; (ii) training-free methods lack the flexibility to handle text prompts containing multiple elements and complex relationships because of the no-additional training nature; (iii) text refinement approaches are governed by the maximum token capacity of the text encoders, which impedes their ability to fully harness the information contained in text prompts.

To address these limitations, we propose PSG-Adapter. Our approach achieves better text-to-image synthesis by leveraging the Planning Scene Graph (PSG) method to decompose the text prompt into different structural subprompts and

utilizing the PSG-ControlNet module to fuse additional spatial information into original text embeddings. As illustrated in Fig. 1, the images produced by our proposed PSG-Adapter exhibit distinct foreground and background relationships. We introduce two core strategies in PSG-Adapter:

Planning Scene Graph. We focus on transforming text prompts into detailed and structured prompts, which ensures better prompt understanding and semantic alignment in diffusion models. We use MLLMs to decompose the text prompt into distinct subprompts and recapitulate them with more informative descriptions that contain explicit spatial information.

PSG-ControlNet. While CLIP [24]/T5 [25] text encoders are broadly employed in diffusion models, their text encoding performance is limited. Some attempts have been made to increase the maximum number of tokens that text encoders can process [17], aiming to improve text-to-image generating performance. While there have been some advancements with these methods, the models' efficacy is still restricted when handling lengthy and intricate text instructions. To address the above challenge and circumvent the need for retraining or fine-tuning the text encoder, while accommodating diffusion models that use different text encoders as well, we introduce the PSG-ControlNet module. This module effectively extracts information from both the original text prompt and subprompts, integrating spatial relations from the subprompts into the original text embeddings. As a result, the diffusion model can better comprehend complex spatial relationships, allowing for precise control over the composition of generated images.

In summary, our contributions are as follows:

- We propose a novel text prompt parsing method named Planning Scene Graph (PSG) which unleashes MLLMs' impressive reasoning ability to convert the original prompt to informative subprompts explicitly containing spatial relations for steering diffusion models.
- We propose the PSG-ControlNet module to achieve fine-grained control over the composition of generated images, which refines text embeddings by integrating the original text embeddings with those of the subprompts.
- Both qualitative and quantitative results demonstrate that our PSG-Adapter surpasses state-of-the-art text-to-image methods in terms of visual quality and spatial control.

2 Related Work

2.1 Text-to-Image Generation

Modeling text-to-image generation with diffusion models has achieved widespread success, outperforming generative adversarial networks (GANs) in terms of photorealism and diversity, and sidestepping problems such as training instability and mode collapse [7]. Notably, models like Stable Diffusion (SD) [29] utilize a pre-trained autoencoder and the diffusion model to transform textual prompts

into detailed visual representations effectively. Models like GLIDE [19] and Imagen [30] have also demonstrated excellent results by incorporating pre-trained models such as CLIP [24], which enhances the semantic alignment between text prompts and generated images. DALL-E 2 [26], a progressive iteration from DALL-E [27], utilizes a hierarchical structure and the CLIP model [24] to generate highly detailed and contextually accurate images. Recent advancements like SDXL [23] and DALL-E 3 [2] continue to push the boundaries, offering improvements in both image fidelity and the complexity of generated scenes. Despite their impressive capabilities, they frequently struggle to understand implicit spatial information within text prompts. These models have difficulty establishing logical spatial relationships, such as foreground/background distinctions, and are constrained by limited text encoding capacity.

2.2 Training Free Strategy

Training-free approaches are premised on their speed and low cost, as they do not require any form of training throughout the entire process of text-to-image generation. This advantage allows for quicker and more economical generation of images without the need for extensive training phases. One of the most common and direct training-free strategies is attention modification, which enhances operations within the attention layers [9,21]. By concentrating on these modifications, these approaches effectively refine the process without the need for extensive training. While some works focus on regulating the latent space to generate spatially aligned images [8,37]. Despite advances in this field, these methods are constrained by the encoding capacity of text encoders. They cannot accurately control spatial relationships in the generated images, which hinders their ability to fully capture and render complex spatial arrangements. This limitation is critical, as accurate spatial control is essential for producing high-quality and accurate text-to-image translations.

2.3 Adapter

The integration of the adapters into the existing models has proven to be a significant innovation [13,38]. By applying a small, trainable neural network to the text-to-image process, these methods improve the performance of pre-trained text-to-image diffusion models. For instance, CAT [20] is a strategy that uses contrastive loss to fine-tune adapters, preserving the original model's knowledge and preventing degradation during personalization. ControlNet [38] operates as a plug-in module that enhances text-to-image diffusion models by adding conditional controls through zero-initialized convolution layers.

2.4 Scene Graph-to-Image Generation

Scene graphs are used to rectify inaccuracies in text embeddings resulting from "leakage issues", ensuring precise and contextually relevant image generation by

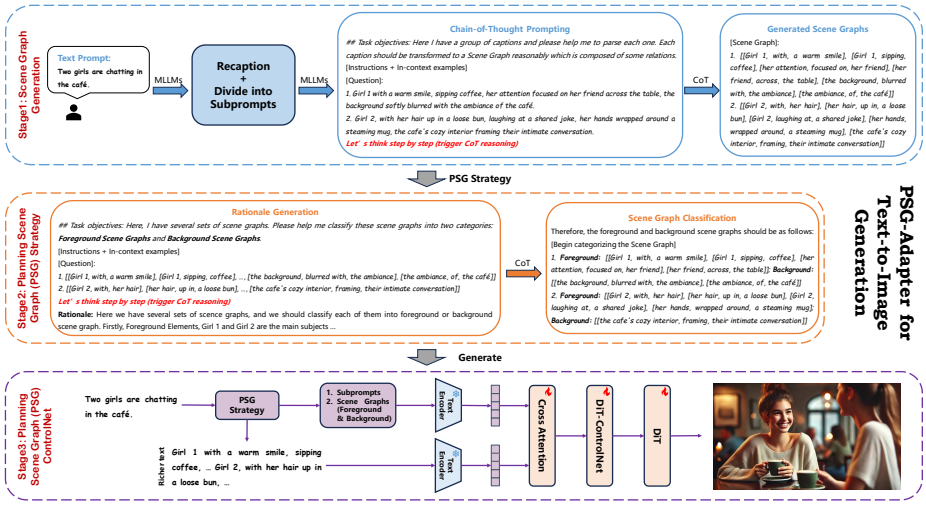


Fig. 2: Overview of PSG-Adapter for text-to-image generation.

leveraging structured relationships between entities [32]. However, constructing large-scale, high-quality scene graphs is a time-consuming and labor-intensive process. To address this problem, we can design appropriate prompt engineering strategies that leverage the powerful logical reasoning capabilities of MLLMs to construct text prompts that are structurally similar to scene graphs, while integrating more spatial information. This approach can streamline the process, making it more efficient while improving the accuracy and quality of the generated images.

Our paper introduces the Planning Scene Graph (PSG) method to enhance diffusion models' understanding of complex text prompts containing multiple elements and relationships. Additionally, we introduce the PSG-ControlNet module to address the challenges associated with attaining fine-grained control over the composition of generated images. This approach addresses a significant gap in the current landscape of text-to-image generation, enabling more accurate and detailed image synthesis from intricate textual descriptions.

3 Method

3.1 Scene Graph Guided Cross Attention

Scene Graph Triplets Attention Mask. A scene graph is a structured representation of an image, comprising entities (nodes) and their relationships (edges) in the form of ⟨subject, relation, object⟩ triplets [32]. Its explicit, non-fully connected structure allows for clear, non-linear associations, preventing ambiguity common in sequential text-based processing.

Unlike the causal attention mask in standard transformers [24], which restricts tokens to attending only to preceding tokens and often causes incorrect

contextual associations (e.g., relation leakage), the scene graph triplets attention mask limits interactions to tokens within the same triplet. This ensures accurate and contextually relevant text-to-image generation. Formally, for caption c of N tokens, the scene graph triplets attention mask $\mathbf{M}^{\tau}$ is an $N \times N$ matrix where each entry $\mathbf{M}^{\tau}_{ij}$ is defined as:

$$\mathbf{M}^{\tau}_{ij} = \begin{cases} 0, & \text{if } \tau(i) = \tau(j) \\ -\infty, & \text{otherwise} \end{cases}, \tag{1}$$

where each triplet $\mathcal{T}_k$ is indexed by k and $\tau(\cdot)$ is a token-to-triplet mapping function that maps each token i to its respective triplet index.

Cross Attention with Token-Triplet Mask. For each triplet $\mathcal{T}_k = \langle s_k, r_k, o_k \rangle$ in the scene graph, where s_k, r_k, and o_k denote the subject, relation, and object respectively. CLIP text encoder E_{CLIP} is applied to each component to obtain their embeddings. These embeddings are concatenated to form a composite triplet embedding:

$$\mathbf{e}_k = l\left(\text{concat}\left(E_{CLIP}(s_k), E_{CLIP}(r_k), E_{CLIP}(o_k)\right)\right), \tag{2}$$

where $l(\cdot)$ is a projection function that maps its input to the dimensionality D that matches the output of the CLIP text encoder. The final scene graph embedding $\mathbf{e} \in \mathbb{R}^{K \times D}$ is the concatenation of all the triplet embeddings.

To augment the text embedding $\mathbf{w}$ produced by the CLIP model with the scene graph information $\mathbf{e}$, SG-Adapter [32] uses a transformer module with a cross-attention layer. This layer calculates keys $\mathbf{K}$ and values $\mathbf{V}$ from the scene graph embedding $\mathbf{e}$, and queries $\mathbf{Q}$ from the text embedding $\mathbf{w}$. This process can be expressed as:

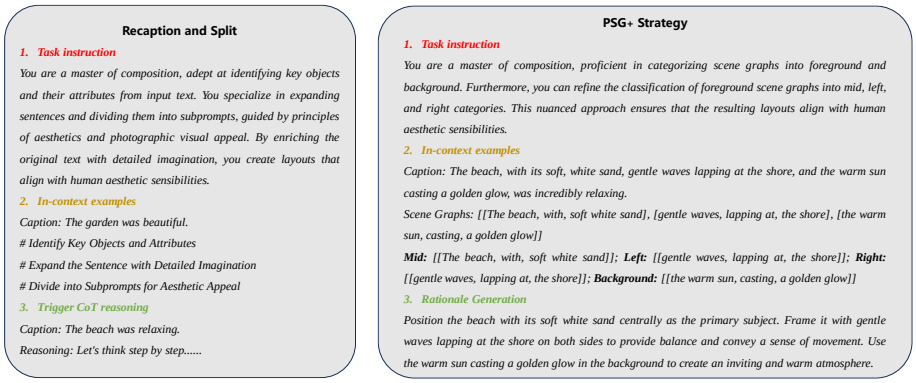
$$\mathbf{w}' = \text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}^{\mathbf{sg}}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}^{\mathbf{sg}}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{w}'$ represents the improved text embeddings resulting from the application of the cross-attention mechanism, and $\mathbf{Q} = l_Q(\mathbf{w})$, $\mathbf{K} = l_K(\mathbf{e})$, $\mathbf{V} = l_V(\mathbf{e})$ are derived using respective projection layers. The term token-triplet mask $\mathbf{M}^{\mathbf{sg}} \in \mathbb{R}^{N \times K}$ is designed to ensure each token embedding $\mathbf{w}_i$ attends exclusively to its corresponding triplet embedding. The mask is represented as:

$$\mathbf{M}^{\mathbf{sg}}_{ik} = \begin{cases} 0, & \text{if } \tau(i) = k \\ -\infty, & \text{otherwise} \end{cases}. \tag{4}$$

By implementing the aforementioned cross-attention mechanism with the token-triplet mask in the denoising process, the training loss can be expressed as:

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{x},t,\epsilon}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{w}')\|_2^2\right], \tag{5}$$



(a) Prompts for Reception&Split **(b)** Prompts for PSG$_+$ method

Fig. 3: The left section displays the prompts for "Reception and Split" stage in PSG. And the right section shows the prompts for PSG$_+$ method.

where $\mathbf{x}_t$ is derived from image $\mathbf{x}$ after adding noise over t time steps, ϵ_t represents the noise variable at time step t, and ϵ_θ denotes the noise prediction from the pre-trained diffusion model parameterized by θ.

Although SG-Adapter refines the text embeddings produced by the CLIP text encoder by embedding scene graph information through the above cross-attention mechanism, it still encounters two major drawbacks: (i) the scene graph lacks fine-grained categorization, limiting precise control over the composition of the generated image; (ii) the CLIP text encoder's maximum token length constraint leads to information loss with richer texts, resulting in generated images that deviate from the expected.

3.2 Planning Scene Graph Strategies

Planning Scene Graph method (PSG). To attain precise control over the composition of generated images, we propose the Planning Scene Graph (PSG) method. This method innovatively decomposes the original text prompt into distinct subprompts by integrating explicit spatial information through recaptioning. Fig. 3a shows the prompts used in the "Reception and Divide into Subprompts" step of the PSG method. By categorizing scene graphs into foreground and background, we refine prompt understanding and semantic alignment in diffusion models. Utilizing MLLMs, our approach breaks down the text prompt into detailed and structured subprompts, improving text-to-image synthesis and achieving the desired spatial accuracy.

Specifically, let c_o be the original text prompt which includes multiple entities with different attributes and relationships. We leverage the impressive language understanding and reasoning abilities of MLLMs and use MLLMs to reception and split original text prompts through Chain-of-Thought (CoT) prompting:

$$\hat{c} = \text{Reception}(c_o), \tag{6}$$

$$\{\hat{c}_i\}_{i=0}^{n} = \{\hat{c}_0, \hat{c}_1, ..., \hat{c}_n\} = \text{Split}(\hat{c}), \tag{7}$$

where n denotes the number of subprompts. Likewise, for each subprompt in $\{\hat{c}_i\}_{i=0}^{n}$, we use MLLMs to generate scene graphs and categorize them into foreground and background scene graphs using in-context and CoT prompting:

$$\text{SG}_i = \text{Generate}(\hat{c}_i), \tag{8}$$

$$\{\text{FG}_i, \text{BG}_i\} = \text{Categorize}(\text{SG}_i). \tag{9}$$

In this way, we enhance prompt understanding and semantic alignment in diffusion models by categorizing scene graphs into foreground and background.

Planning Scene Graph method Plus (PSG_+). Building on the PSG method, we introduce PSG Plus (PSG_+) to further enhance control over the composition and spatial relationships in generated images, thus better aligning them with human expectations. Fig. 3b displays the prompts to classify the scene graphs in PSG_+ method. Drawing inspiration from human visual attention patterns, we innovatively subdivide foreground scene graphs into detailed categories such as mid, left, and right scene graphs. This refined classification allows for more precise manipulation of spatial arrangements and visual emphasis, ensuring that the generated images are both accurate and contextually coherent. By leveraging these advanced techniques, PSG_+ further improves the quality and relevance of text-to-image synthesis, making the results more consistent with human perceptual and interpretive standards. Formally, the procedure for categorizing foreground scene graphs into mid, left, and right categories is outlined as follows:

$$\{\text{MG}_i, \text{LG}_i, \text{RG}_i, \text{BG}_i\} = \text{Categorize}_+(\text{SG}_i), \tag{10}$$

where $\text{MG}_i, \text{LG}_i, \text{RG}_i$ denotes mid, left and right scene graphs respectively.

3.3 Planning Scene Graph ControlNet (PSG-ControlNet)

Previous ControlNet-based approaches rely on additional conditions (canny edge, pose, depth, sketch, segments, *etc.*) from the image encoder [5, 12, 38] due to the text encoder's maximum encoding length, which limits text information extraction and image logic understanding. To address this challenge, we propose PSG-ControlNet, a text-driven ControlNet based on the Diffusion Transformer (DiT) [22]. This novel approach enhances the integration of textual and visual information, enabling more accurate and spatially coherent text-to-image generation.

To implement a more accurate and spatially coherent text-to-image generation, the text-driven PSG-ControlNet is designed. As illustrated in Fig. 2, in Stage 3, the original text prompt c_o is initially processed through the PSG, which results in an enriched text prompt $\hat{c}$ and a set of subprompts $\{\hat{c}_i\}_{i=0}^{n}$ as

outlined in Eq. (6) and Eq. (7) respectively. For each prompt in $\{\hat{c}_i\}_{i=0}^{n}$, the foreground and background scene graphs $\{\mathrm{FG}_i, \mathrm{BG}_i\}$ are generated using Eq. (9). Subsequently, the text embedding $\mathbf{e}_{\hat{c}}$ is derived from Eq. (11):

$$\mathbf{e}_{\hat{c}} = E_{T5}(\hat{c}), \tag{11}$$

where the fixed T5 text encoder [25] is employed. Meanwhile, for each subprompt $\hat{c}_i$ in $\{\hat{c}_i\}_{i=0}^{n}$, the subprompt and its associated foreground and background scene graph information are combined using Eq. (12):

$$\mathbf{e}_i^{sp} = \mathrm{concat}\left(E_{T5}(\hat{c}_i), E_{T5}(\mathrm{FG}_i), E_{T5}(\mathrm{BG}_i)\right). \tag{12}$$

The final subprompt embedding $\mathbf{e}^{sp} \in \mathbb{R}^{n\times 3D}$ is the concatenation of all elements in $\{\mathbf{e}_i^{sp}\}_{i=0}^{n}$, where n is the number of subprompts. This embedding is then to refine the embedding $\mathbf{e}_{\hat{c}}$ to enhance the model's spatial understanding and control over the generated images.

To fully exploit the spatial relationships within the text, we have designed a cross-attention module. At first, we employ two separate MLPs to map $\mathbf{e}_{\hat{c}}$ and $\mathbf{e}^{sp}$ to the same dimensional space:

$$\hat{\mathbf{e}}_{\hat{c}} = \mathrm{MLP}_1(\mathbf{e}_{\hat{c}}), \tag{13}$$

$$\hat{\mathbf{e}}_{sp} = \mathrm{MLP}_2(\mathbf{e}^{sp}), \tag{14}$$

where $\hat{\mathbf{e}}_{\hat{c}}, \hat{\mathbf{e}}_{sp} \in \mathbb{R}^D$. Then, within the cross-attention module, the modeling and comprehension of spatial relationships attention are markedly strengthened. This is accomplished by regarding $\hat{\mathbf{e}}_{\hat{c}}$ as the queries $\mathbf{Q}_{\hat{\mathbf{e}}_{\hat{c}}}$ and $\hat{\mathbf{e}}_{sp}$ as the keys $\mathbf{K}_{\hat{\mathbf{e}}_{sp}}$ and values $\mathbf{V}_{\hat{\mathbf{e}}_{sp}}$. Such interaction plays a crucial role in enabling the seamless integration of spatial information into the text embeddings, thus improving the coherence of the generated images. The cross-attention module f_{CA} is mathematically represented as:

$$\hat{\mathbf{c}} = f_{CA}\left(\hat{\mathbf{e}}_{\hat{c}}, \hat{\mathbf{e}}_{sp}\right) \tag{15}$$

$$= \mathrm{Attention}\left(\mathbf{Q}_{\hat{\mathbf{e}}_{\hat{c}}}, \mathbf{K}_{\hat{\mathbf{e}}_{sp}}, \mathbf{V}_{\hat{\mathbf{e}}_{sp}}\right) \tag{16}$$

$$= \mathrm{softmax}\left(\frac{\mathbf{Q}_{\hat{\mathbf{e}}_{\hat{c}}}\mathbf{K}_{\hat{\mathbf{e}}_{sp}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\hat{\mathbf{e}}_{sp}}, \tag{17}$$

where term $\hat{\mathbf{c}}$ is treated as the control signal in the DiT-ControlNet module. We adopt the general design principles of ControlNet employed in PixArt-α [5]. In particular, we freeze each DiT Block and generate a trainable copy, which is augmented with two zero-initialized linear layers inserted before and after the block.

During the training of PSG-Adapter, starting with an initial image $\mathbf{x}_0$, the image diffusion algorithm incrementally introduces noise, producing a progressively noisier image $\mathbf{x}_t$, where t signifies the number of noise application steps.

Table 1: Quantitative Evaluation of each method in terms of SG-IoU, Entity-IoU, Relation-IoU and FID.

Method	SG-IoU↑	Entity-IoU↑	Relation-IoU↑	FID↓
SDXL [23]	0.172	0.689	0.563	20.7
PixArt-Σ [4]	0.324	0.692	0.586	19.3
DALL-E 3 [2]	0.465	0.720	0.624	**18.4**
SG-Adapter [32]	0.623	0.812	0.753	26.2
Ours	**0.652**	**0.843**	**0.790**	28.3

The algorithm is conditioned on several factors, including the time step t, generated rich text prompt $\hat{c}$ in Eq. (6), and the control signal derived from Eq. (15). The goal is to train a network ϵ_θ to predict the noise present in the noisy image $\mathbf{x}_t$ through the following process:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,t,\hat{c},\hat{\mathbf{c}},\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \hat{c}, \hat{\mathbf{c}})\|_2^2\right], \tag{18}$$

where $\mathcal{L}$ is the overall learning objective of the entire diffusion model. Our training algorithm iteratively refines the parameters $\hat{\mathbf{c}}$ to minimize the objective function $\mathcal{L}$. This process leverages spatial relationships to attain fine-grained control over image composition, thereby improving the overall quality and fidelity of the generated images in alignment with the input descriptions and foreground and background scene graphs.

4 Experiments

4.1 Implementation Details

Our proposed PSG-Adapter is versatile and scalable, enabling the integration of diverse MLLM architectures and a variety of diffusion backbones within the framework. In our experiments, we employed GPT-4V [1] as the recaptioner and CoT planner, and used DiT [22] as the base diffusion backbone to develop our PSG-Adapter. To harness the CoT planning capabilities of MLLMs, we designed task-specific templates and high-quality in-context examples for few-shot prompting. For training DiT-ControlNet, we utilized the advanced Q-Align evaluation system [34] to select 0.5M image-text pairs from LAION-5B [31], based on image quality assessment (IQA) and image aesthetic assessment (IAA).

4.2 Qualitative Evaluation

In the qualitative evaluation, we compare our approach with the previous state-of-the-art text-to-image models, such as SDXL [23], PixArt-Σ [4] and DALL-E 3 [2], showcasing synthesized images from each method. The ability of our PSG-Adapter to accurately capture complex relational structures and entities is

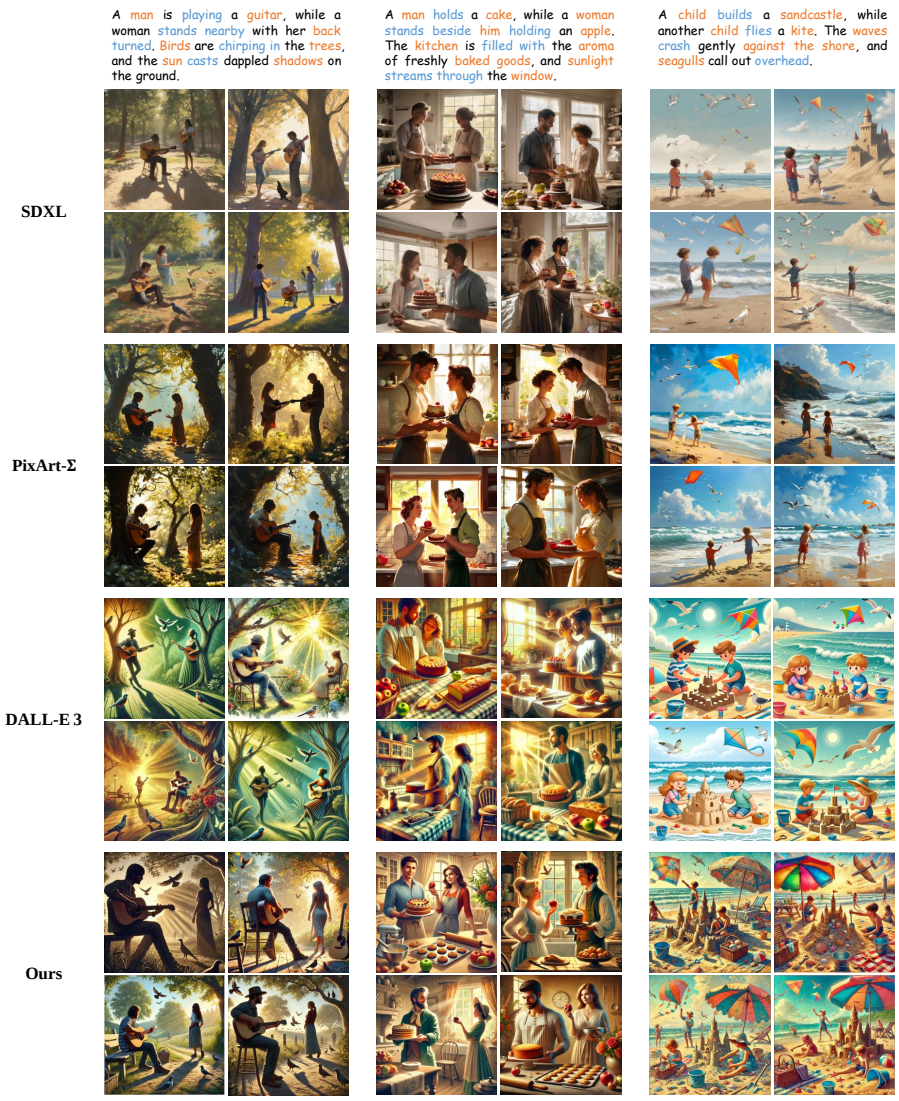


Fig. 4: Qualitative comparisons with the state-of-the-art text-to-image models are presented in the order of SDXL [23], PixArt-Σ [4], DALL-E 3 [2], and PSG-Adapter (Ours), from top to bottom. Beyond generating images with precise attribute binding, our PSG-Adapter method effectively produces images with distinct spatial relationships, clearly distinguishing foreground and background, even in complex scenarios involving multiple objects and relationships.

Table 2: Ablation study on Planning Scene Graph (PSG) method.

Method	SG-IoU↑	Entity-IoU↑	Relation-IoU↑	FID↓
SDv1.5 [29]	0.157	0.673	0.526	25.0
PSG-SDv1.5	0.230	0.682	0.549	25.4
SDXL [23]	0.172	0.689	0.563	20.7
PSG-SDXL	0.301	0.712	0.574	21.2

highlighted. Fig. 4 demonstrates its advantage in maintaining clear spatial relationships in generated images, particularly in multi-object and multi-relationship scenarios.

4.3 Quantitative Evaluation

Metrics. Due to the intricate and abstract nature of relational correspondence in images, traditional metrics like FID fail to capture this concept accurately. To comprehensively and fairly assess the accuracy of relational generation in each method, we employed three metrics proposed by SG-Adapter [32]: Scene Graph (SG)-IoU, Relation-IoU, and Entity-IoU. SG-IoU measures how accurately the relationships are generated in accordance with the input scene graph, whereas the other two metrics evaluate the presence and accuracy of each individual object and relation. Besides, we use the FID score [10] to measure the difference between the distribution of our generated images and a set of 5,000 validation images from the MS-COCO-Stuff dataset [3].

Quantitative Analysis. Tab. 1 illustrates that the PSG-Adapter consistently surpasses state-of-the-art (SOTA) methods in the metrics detailed in Sec. 4.3, with the exception of FID. Although fine-tuning pre-trained T2I models on relatively small datasets inevitably lowers FID, it is important to recognize that FID is not a suitable metric for evaluating spatial relationships in generated images. Notably, it shows significant improvements in SG-IoU and Relation Accuracy. This highlights the PSG-Adapter's strong ability to generate relations with precise correspondences. In contrast, the high Entity-IoU and Relation-IoU but low SG-IoU scores of other SOTA methods indicate that, while these methods can generate the necessary entities and relations, they struggle to accurately align them in complex multi-object, multi-relationship scenarios.

4.4 Ablation Study

To understand the impact of each component within the PSG-Adapter, we performed a detailed comparative analysis by systematically removing specific features. Our experiments focused on evaluating the roles of PSG, PSG_+, and PSG-ControlNet, providing insights into their individual contributions and overall effectiveness.

Fig. 5: Ablation study on PSG and PSG$_+$. Colored text denotes critical part.

Effect of Planning Scene Graph (PSG). We first assessed the impact of the Planning Scene Graph (PSG) on controlling spatial structures in generated images. The PSG method breaks down the original text prompt into distinct subprompts with spatial relationships. As demonstrated in Tab. 2, incorporating PSG into Stable Diffusion led to an improvement in the SG-IoU metric. Furthermore, as shown in Fig. 1, integrating foreground and background knowledge into text embeddings using PSG significantly enhanced spatial relationships in the generated images.

Effect of Planning Scene Graph Plus (PSG$_+$). Expanding upon the PSG method, PSG$_+$ aims to provide greater control over the composition and spatial dynamics of generated images. Specifically, PSG$_+$ subdivides foreground scene

Table 3: Ablation study on DiT-ControlNet.

Method	SG-IoU↑	Entity-IoU↑	Relation-IoU↑	FID↓
PixArt-Σ [4]	0.324	0.692	0.586	19.3
SG-Adapter [32]	0.623	0.812	0.753	26.2
w/o DiT-ControlNet	0.634	0.825	0.772	29.0
Ours	0.652	0.843	0.790	28.3

graphs into detailed categories such as mid, left, and right, allowing for more precise manipulation of spatial arrangements and visual emphasis. In this section, we validated the effectiveness of the PSG_+ method. As evidenced in Fig. 5, PSG_+ creatively segments foreground elements into three distinct parts: mid, left, and right. This segmentation allows for enhanced control over image composition, producing results that better match human expectations and possess increased aesthetic appeal.

Effect of DiT-ControlNet. We evaluated the significance of the DiT-ControlNet module within PSG-ControlNet for adjusting the control signal and its impact on refining the image generation process. We removed the DiT-ControlNet and directly used the control signal as the textual condition for image generation. As shown in Tab. 3, the performance of our method declined without the DiT-ControlNet but still outperformed the SG-Adapter. This is attributed to the superior performance of the T5 text encoder compared to the CLIP text encoder. Additionally, our method also surpassed PixArt-Σ due to the design of the PSG method and the incorporation of more spatial knowledge into the text embeddings.

5 Conclusion

In this paper, we introduced the PSG-Adapter, a novel approach designed to enhance text-to-image diffusion models by improving their understanding and manipulation of spatial relationships within text prompts. Leveraging the Planning Scene Graph (PSG) method and the PSG-ControlNet module, our approach enables fine-grained control over image composition and improves spatial coherence. Experimental results demonstrate that PSG-Adapter outperforms existing methods in visual quality and spatial alignment. Future work will focus on refining the model and exploring its application to more complex and realistic scenarios.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
4. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. arXiv preprint arXiv:2403.04692 (2024)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
6. Chen, L., Zhao, M., Liu, Y., Ding, M., Song, Y., Wang, S., Wang, X., Yang, H., Liu, J., Du, K., et al.: Photoverse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793 (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023)
9. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback. arXiv preprint arXiv:2404.07987 (2024)
13. 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)
14. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Adversarial supervision makes layout-to-image diffusion models thrive. In: The Twelfth International Conference on Learning Representations (2024)
15. 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)
16. Lin, Y., Zhang, S., Yang, X., Wang, X., Shi, Y.: Regeneration learning of diffusion models with rich prompts for zero-shot image translation. arXiv preprint arXiv:2305.04651 (2023)
17. Liu, Z., Liang, W., Liang, Z., Luo, C., Li, J., Huang, G., Yuan, Y.: Glyph-byt5: A customized text encoder for accurate visual text rendering. arXiv preprint arXiv:2403.09622 (2024)
18. Lv, H., Xiao, J., Li, L., Huang, Q.: Pick-and-draw: Training-free semantic guidance for text-to-image personalization. arXiv preprint arXiv:2401.16762 (2024)

19. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
20. Park, J.W., Park, S.H., Koh, J.Y., Lee, J., Song, M.: Cat: Contrastive adapter training for personalized image generation. arXiv preprint arXiv:2404.07554 (2024)
21. 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)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
23. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2023)
24. 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)
25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
27. 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)
28. Ravi, H., Kelkar, S., Harikumar, M., Kale, A.: Preditor: Text guided image editing with diffusion prior. arXiv preprint arXiv:2302.07979 (2023)
29. 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)
30. 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)
31. 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)
32. Shen, G., Wang, L., Lin, J., Ge, W., Zhang, C., Tao, X., Zhang, Y., Wan, P., Wang, Z., Chen, G., et al.: Sg-adapter: Enhancing text-to-image generation with scene graph guidance. arXiv preprint arXiv:2405.15321 (2024)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
34. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)

35. 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)
36. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14256–14266 (2023)
37. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Bin, C.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
38. 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)
39. 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)