Scene Graph-Grounded Image Generation

Fuyun Wang[1], Tong Zhang[1], Yuanzhi Wang[1], Xiaoya Zhang[1], Xin Liu[2], Zhen Cui[1]*

[1]Nanjing University of Science and Technology, China.
[2]Nanjing SeetaCloud Technology, China.
{fyw271828, tong.zhang, yuanzhiwang, zhangxiaoya, zhen.cui}@njust.edu.cn, xin.liu@seetacloud.com

Abstract

With the benefit of explicit object-oriented reasoning capabilities of scene graphs, scene graph-to-image generation has made remarkable advancements in comprehending object coherence and interactive relations. Recent state-of-the-arts typically predict the scene layouts as an intermediate representation of a scene graph before synthesizing the image. Nevertheless, transforming a scene graph into an exact layout may restrict its representation capabilities, leading to discrepancies in interactive relationships (such as standing on, wearing, or covering) between the generated image and the input scene graph. In this paper, we propose a Scene Graph-Grounded Image Generation (SGG-IG) method to mitigate the above issues. Specifically, to enhance the scene graph representation, we design a masked auto-encoder module and a relation embedding learning module to integrate structural knowledge and contextual information of the scene graph with a mask self-supervised manner. Subsequently, to bridge the scene graph with visual content, we introduce a spatial constraint and image-scene alignment constraint to capture the fine-grained visual correlation between the scene graph symbol representation and the corresponding image representation, thereby generating semantically consistent and high-quality images. Extensive experiments demonstrate the effectiveness of the method both quantitatively and qualitatively.

Introduction

Recently, text-to-image (T2I) diffusion models (Nichol et al. 2021; Saharia et al. 2022; Balaji et al. 2022; Luo et al. 2024) have revolutionized the image generation landscape. Current techniques, like Stable Diffusion (SD) (Rombach et al. 2022), depend on protracted and loosely organized natural language for image generation. However, owing to the constraints imposed by linguistic ambiguity and a linear sentence structure, articulating the semantics of intricate scenes along with the interplay and correlations among diverse objects proves to be challenging for existing methods. Scene graph offers a viable alternative for generating visual data descriptions through the use of graph syntax. Through the proposition of a robust and versatile method for delineating objects and their interactive relations within a given scene graph, the scene graph-to-image generation technique has garnered considerable public interest in recent years.

There exist two primary categories of current scene graph-to-image generation approaches. One category employs a two-stage generative pipeline that relies on predictive scenario layout and adversarial training to generate images. SG2IM (Johnson, Gupta, and Fei-Fei 2018) is the pioneering generative model designed to generate images from a scene graph. It initially predicts a rough layout from an input scene graph, which subsequently serves as the input for conditional Generative Adversarial Network (GAN) (Goodfellow et al. 2020) in the image synthesis process. The alternative approach eliminates the necessity for an intermediate layout and optimizes the scene graph-to-image alignment by pre-training graph encoders. A classic work in this area is SGDiff (Yang et al. 2022), which extracts global and local information from the scene graph during pre-training, utilizes the improved scene graph representation as the conditional signal, and generates images based on the diffusion model.

Despite yielding promising outcomes, fundamental issues exist with the existing methods. For the first approach, transforming the scene graph into a precise layout restricts its expressive capacity in terms of object interactions (Sortino et al. 2023). As illustrated in Fig. 1, the coarse scene layout is proficient in depicting spatial positional relationships but falls short of capturing the intricate non-spatial relationships between objects (e.g., standing on, wearing, covered in). Furthermore, image generation processes that rely on adversarial strategies may limit the potential of conditional image generation due to the instability of adversarial-based methods during training (Sortino et al. 2023). For the second approach, given the high diversity of object categories, layout, and relation dependencies present in the scene graph, the pre-trained scene graph embedding differs significantly from the CLIP text embedding. End-to-end fine-tuning of SD solely based on the final image reconstruction loss (i.e., noise prediction loss) hinders the accurate learning of aligning the scene graph signal with the SD text-to-image control, leading to object distortions or content generation failures.

To mitigate the above issues, particularly in synthesizing coherent objects and layouts while considering the interactive relations among multiple objects, we propose a Scene Graph-Grounded Image Generation (SGG-IG) method. Specifically, to enhance scene graph representation, we design a rela-

*Corresponding author.

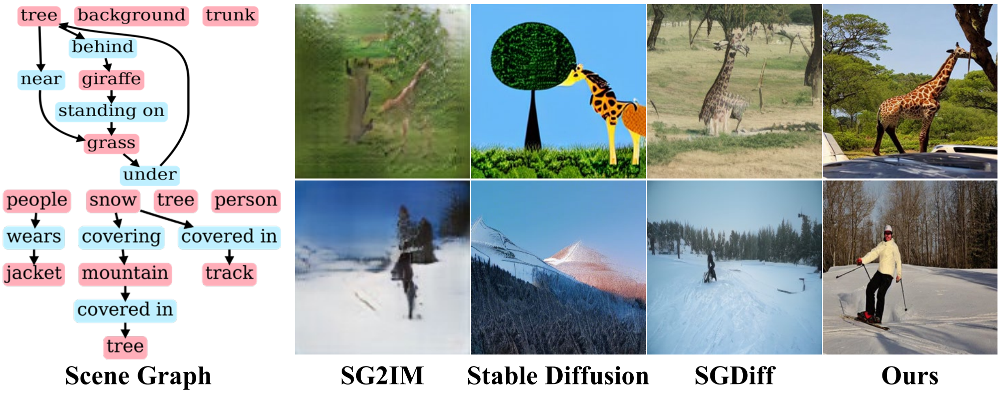

Figure 1: Compared to other image generation methods, SGG-IG yields results that faithfully adhere to the input scene graph, producing semantically coherent and high-fidelity images.

tion embedding learning module and a masked auto-encoder module to integrate structural knowledge and contextual information of scene graph with a mask self-supervised manner (He et al. 2022). We randomly mask a certain proportion of node objects and then predict the missing node objects with their related relationship embeddings and contextual clues from surrounding node objects. By this means, the pre-trained relation embedding module acquires a more profound comprehension of the semantic and structural details across various scene graphs. In particular, to bridge the scene graph with visual content, we introduce two constraints to capture the fine-grained visual correlation between the scene graph symbol representation and corresponding image representation in different stages. In the pre-training stage, we propose a spatial constraint to learn fine-grained visual correlation between scene graph and image. In the fine-tuning stage, we attempt to leverage a large-scale pre-trained text-to-image diffusion model for scene graph-to-image generation, which offers a generative prior with extensive semantic coverage. These models primarily rely on text embeddings, leading to notable semantic distinctions compared to the pre-trained scene graph embeddings. To improve the alignment between the scene graph and diffusion models, we propose an image-scene alignment constraint to refine the scene graph, text, and image embedding spaces, thereby generating semantically consistent and high-quality images.

The contribution can be summarized as follows:

- We propose a Scene Graph-Grounded Image Generation (SGG-IG) method, which designs a masked auto-encoder module and a relation embedding learning module to integrate structural knowledge and contextual information of the scene graph in a mask self-supervised manner.
- We introduce a spatial constraint and image-scene alignment constraint to capture the fine-grained visual correlation between the scene graph symbol representation and the corresponding image representation.
- Extensive quantitative and qualitative experiments demonstrate that the proposed SGG-IG method generates perceptually appealing images with distinct textures and coherent scene relationships, achieving superior performance.

Related Work

Controllable Text-to-Image Generation

Diffusion models have made notable strides in producing lifelike images promptly aligned with textual prompts (Ramesh et al. 2022; Saharia et al. 2022; Wang, Li, and Cui 2024; Wang et al. 2024; Luo et al. 2024). Recent efforts have been concentrated on integrating multiple conditions into the text-to-image (T2I) generation process, offering valuable insights into enhancing the controllability of image generation. Technically, efforts primarily concentrate on refining both the training and inference phases. One way to implement controllable image generation in training is to introduce additional modules. ControlNet uses a spectrum of conditional inputs to provide a high degree of control over the resulting image. Nevertheless, the incorporation of supplementary modules into the original backbone led to an expansion in the model's size, thereby incurring additional training and inference costs.

In contrast, strategies that apply generative control during the inference phase use reward functions to encourage faithful image generation conditioned on specific criteria; these are referred to as training-free methods. Layout Guidance Diffusion (Chen, Laina, and Vedaldi 2024) and BoxDiff (Xie et al. 2023) leverage predefined bounding boxes to backpropagate gradients to the latent space, steering cross-attention maps to emphasize regions corresponding to designated tokens. Similarly, Dense Diffusion (Kim et al. 2023) and Directed Diffusion (Ma et al. 2024) manipulate cross-attention maps to align generated images with predefined layouts. While significant advancements have been made in controllable image generation, further exploration is required to develop techniques that integrate scene graph signals. In this study, we investigate the potential of text-to-image diffusion models for image generation tasks guided by scene graphs.

Scene-Graph-based Image Generation

A scene graph captures the essence of a scene by encoding object instances, their properties, and inter-object relationships (Johnson et al. 2015). In the domain of scene graph-to-image generation, one prominent research direction adopts a two-stage generation pipeline: first, predicting scene layouts and then using adversarial training to generate images. SG2IM (Johnson, Gupta, and Fei-Fei 2018) employs a graph convolutional neural network (GCN) to extract objects and features from the scene graph. The predicted scene layout, represented by bounding boxes and segmentation masks of the objects, is subsequently transformed into an image using a cascading refinement network. Subsequent studies build upon this approach by enhancing graph comprehension (Ashual and Wolf 2019; Garg et al. 2021) or improving the quality of intermediate layouts (Sun and Wu 2019; Li et al. 2021; Yang et al. 2023). For instance, WSGC (Herzig et al. 2020) employs normalization to capture semantic equivalence, improving invariance before converting the scene graph into a scene layout. SIMSG (Dhamo et al. 2020) focuses on interactive image manipulation, enabling users to adjust the scene graph to control the composition outcome.

Since converting a scene graph into an exact layout may limit its representational capacity and cause inconsistencies between the generated image and the input graph, particularly in non-spatial relationships, a recent study (Sortino et al. 2023) focuses on directly aligning graph nodes and connections with image objects and their relationships to learn scene graph embeddings. SGDiff (Yang et al. 2022) introduces a mask auto-encoding loss and a contrastive loss for pre-training a graph encoder, enabling the extraction of both global and local information from the scene graph. Another recent study (Mishra and Subramanyam 2024) leverages a pre-trained text-to-image diffusion model and CLIP guidance to transform graph knowledge into images, while pre-training a graph encoder to align graph features with the corresponding CLIP image features.

Method

Preliminaries

Stable Diffusion. Stable Diffusion (SD) (Rombach et al. 2022) are adeptly trained to implement both the diffusion process and reverse sampling procedure within the latent space of images for faster and more stable training. Specifically, given an input image I, the encoder $\mathcal{E}$ translates it into a latent representation $\mathbf{z}$. Afterward, SD trains a well-designed denoising UNet ϵ_θ to perform denoising directly on the latent space. SD trains with a mean squared error loss as:

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

where $\mathbf{c}$ represents the contextual embedding of textual/visual condition obtained from CLIP (Radford et al. 2021).

Classifier-free Guidance. Classifier guidance offers a technique for diffusion models to achieve conditional generation by utilizing the gradient of the independently trained classifier $p(y|x_t)$ throughout the sampling process. A more efficient approach, known as classifier-free guidance, substitutes noise estimators with a blend of conditional and unconditional models, eliminating the need for $p(y|x_t)$:

$$\widetilde{\epsilon}_\theta(x_t,t|y) = \omega\epsilon_\theta(x_t,t|y)+(1-\omega)\epsilon_\theta(x_t,t|\emptyset) \tag{2}$$

where $\emptyset$ denotes that $\mathbf{c}$ is set to null text. y represents the class label or text embedding derived from a language model, while ω signifies the guidance scale. It is worth noting that incrementing ω will straightforwardly enhance the impact of the conditional input.

Overview

The overview framework of SGG-IG is shown in Fig. 2, which consists of three stages: pre-training (gray dotted lines), fine-tuning (blue solid lines), and inference (red solid lines). During the pre-training stage, we first employ the frozen CLIP text encoder to obtain node object embeddings and relational embeddings, which are then injected into the relational embeddings module and a masked auto-encoder module for aggregating triple relationships and enhancing scene graph representation with a masked self-supervised manner. At the same time, we introduce a spatial constraint loss to capture fine-grained visual correlations from the scene graph to the image and to optimize alignment between them. During the fine-tuning stage, we aim to leverage these embeddings to construct diffusion models for the scalable scene graph to image generation. We incorporate an additional image-scene alignment loss to enhance alignment between the scene graph representation and diffusion latent space at each denoising timestep. With the above well-trained models, we can infer spatial and semantic high-fidelity images according to given scene graph descriptions.

Enhanced Scene Graph Representation

SGs embody directed graphs, wherein nodes symbolize objects within a scene while edges depict the relationships interlinking these objects (Krishna et al. 2017; Herzig et al. 2018; Li et al. 2017; Tang et al. 2020; Suhail et al. 2021). Given a vocabulary $\mathcal{C}$ of object categories, a scene graph can be defined as a tuple $(\mathcal{O},\mathcal{R})$, where $\mathcal{O}=\{o_i|_{i=1}^n\}$ is a set of objects with $o_i\in\mathcal{C}$ and $\mathcal{R}=\{r_{i,j}|_{i,j=1,i\neq j}^n\}$ represents a set of relations among objects. Each o_i and $r_{i,j}$ could be mapped into a high-dimensional word embedding space by a CLIP text encoder as object embedding X^{o_i} and relationship embedding $X^{r_{i,j}}$. To enhance the scene graph representation, we build a relational embeddings module to model global scene graph embedding, as described below. Let $\mathcal{S}_{\text{out}}(o_i)$ (resp. $\mathcal{S}_{\text{in}}(o_i)$) denote the set of object to which o_i has an outgoing directed edge (resp. incoming directed edge), we compute embeddings for object o_i and relationship $r_{i,j}$ as follows:

$$X_{\text{in}}^{o_i} = g_{\text{in}}^{o}(X^{o_i},X^{r_{i,j}},X^{o_j})_{j\in\mathcal{S}_{\text{in}}(o_i)}, \tag{3}$$

$$X_{\text{out}}^{o_i} = g_{\text{out}}^{o}(X^{o_j},X^{r_{j,i}},X^{o_i})_{j\in\mathcal{S}_{\text{out}}(o_i)}, \tag{4}$$

$$X^{o_i} \leftarrow f_{\text{pool}}(\{X_{\text{in}}^{o_i}\cup X_{\text{out}}^{o_i}\}), \tag{5}$$

$$X^{r_{i,j}} \leftarrow g^{r}(X^{o_i},X^{r_{i,j}},X^{o_j}), \tag{6}$$

where g_{out}^{o}, g_{in}^{o}, and g^r are the graph convolution layers (Kipf and Welling 2016; Yang et al. 2022), and f_{pool} represents an average pooling operator. With the above extracted object and relationship embeddings, we can calculate a triple relationship embedding $X_{i,j}^{\text{rel}}$ that encapsulates each unique triplet within the scene graph, formulated as:

$$\begin{aligned} X_{i,j}^{\text{rel}} = {} & f^{o}(X^{o_i})+f^{r}(X^{r_{i,j}})+f^{o}(X^{o_j}) \\ & + \text{ConCat}(f_{\text{pool}}(\{X^{o_i}|_{i=1}^n\}), f_{\text{pool}}(\{X^{r_{i,j}}|_{i,j=1,i\neq j}^n\})), \end{aligned} \tag{7}$$

where f^o and f^r are two MLP layers to map the dimensionality of object embeddings and relationship embeddings into

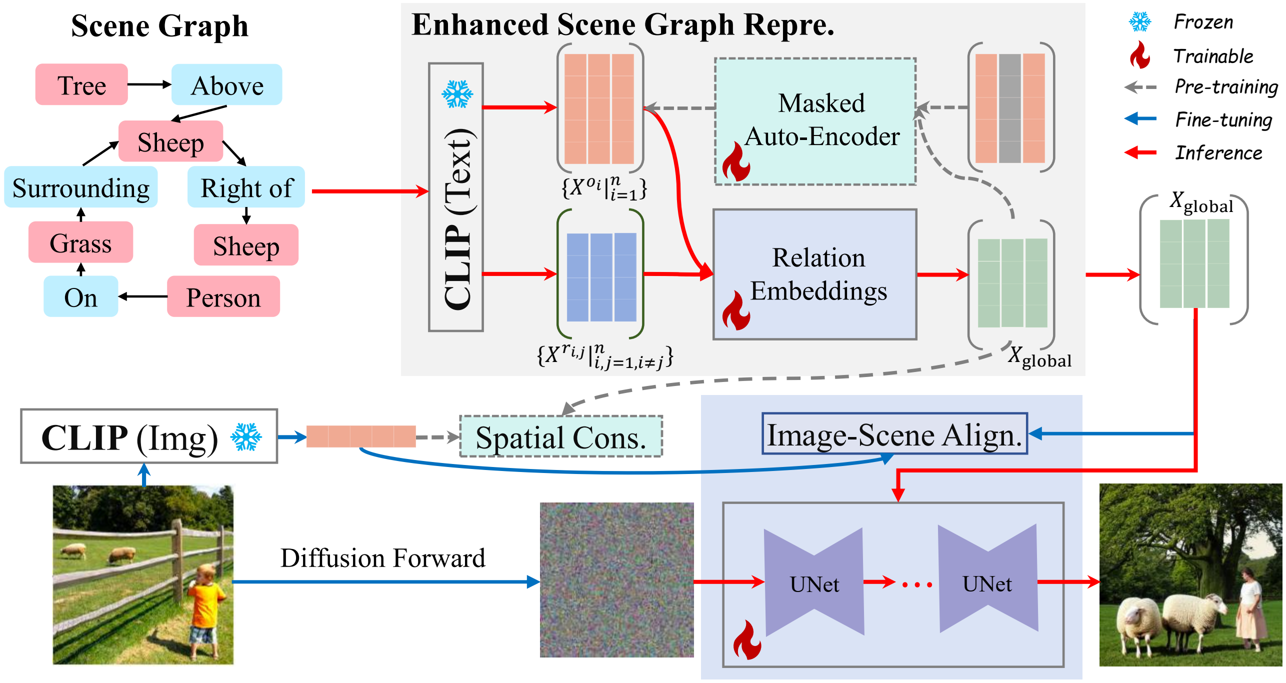

Figure 2: The overall framework of our proposed SGG-IG, which consists of three stages, pre-training (denoted by gray dotted lines), fine-tuning (denoted by blue solid lines), and inference (denoted by red solid lines). During pre-training, the frozen CLIP text encoder encodes the object embedding X^{o_i} and the relation embedding $X^{r_{i,j}}$, and they are fed into the relation embedding module for learning global scene graph representation X_{global}. We randomly discard a certain proportion of node objects, and perform masked self-supervised reconstruction based on the contextual cues of the scene graph and X_{global} to enhance the global scene graph representation. Meanwhile, we introduce a spatial constraint loss $\mathcal{L}_{\text{con}}$ to capture fine-grained visual correlations from the scene graph to the image and to optimize alignment between them. During fine-tuning, we aim to leverage X_{global} to construct diffusion models for the scalable scene graph to image generation. We incorporate an additional image-scene alignment loss $\mathcal{L}_{\text{ISA}}$ to enhance alignment between the scene graph representation and diffusion latent space at each denoising timestep.

an identical dimensional space. ConCat$(\cdot,\cdot)$ denotes the concatenation operation. Subsequently, we can obtain a global scene graph embedding by aggregating all triple relationship embeddings $\{X_{i,j}^{\text{rel}}|_{i,j=1,i\neq j}^{n}\}$ in scene graph, formulated as:

$$X_{\text{global}} = f_{\text{global}}(\text{ConCat}(\{X_{i,j}^{\text{rel}}|_{i,j=1,i\neq j}^{n}\})), \quad (8)$$

where f_{global} is a MLP layer used to extract global scene graph embedding. In training stage, to improve representation capability of relationships among objects for relational embeddings module, we introduce the masked self-supervised technique (He et al. 2022) and design a masked auto-encoder module to randomly mask some object embeddings, followed by incorporating the global scene graph embedding X_{global} to reconstruct the masked embeddings.

Implicit Spatial Layout Learning

With the above well-modeled global scene graph embedding X_{global}, we then utilize contrastive learning between X_{global} and image embeddings to bridge scene graph representation with image content. Specially, the InfoNCE (Oord, Li, and Vinyals 2018; Wang et al. 2023) technique is adopted to enable the maximization of mutual information between the scene graph representation and image embeddings:

$$\mathcal{L}_{\text{con}} = -\log\sigma(f_d(X_{\text{global}}, I_{\text{pos}}) - \log\sigma(1 - f_d(X_{\text{global}}, I_{\text{neg}})), \quad (9)$$

where f_d is a discriminator function implemented by dot product. The image embeddings I_{pos} and I_{neg} signify images that correspond to and deviate from X_{global}, respectively. Thus, $(X_{\text{global}}, I_{\text{pos}})$ and $(X_{\text{global}}, I_{\text{neg}})$ are the positive pair and negative pair, respectively. I_{pos} is sampled from the scene graph-image pair dataset, while I_{neg} is created by uniformly selecting an image that does not align with scene graph within the dataset. Intuitively, $\mathcal{L}_{\text{con}}$ discerns between positive and negative embeddings of images and scene graph, thus bridging scene graph representation with image content.

Scene Graph to Image Refinement

In the denoising (i.e., sampling) stage, however, as the inherent sampling errors in Stable Diffusion, the injected scene graph information of image embeddings by $\mathcal{L}_{\text{con}}$ may gradually deviate from the original scene graph semantics as the denoising timestep decreases. A question arises: *how to mitigate scene graph semantic gap during denoising stage?*

To address this issue, we design a image-scene alignment loss to constrain the learned global scene graph information aligned with image embeddings at each timestep. Taking the timestep t as an example, the image-scene alignment loss $\mathcal{L}_{\text{ISA}}$ at timestep t can be formulated as:

$$\mathcal{L}_{\text{ISA}} = \|1 - \frac{I(t) \cdot X_{\text{global}}}{|I(t)||X_{\text{global}}|}\|, \tag{10}$$

where $I(t)$ is a noisy image embedding at timestep t. Intuitively, $\mathcal{L}_{\text{ISA}}$ could constrain the X_{global} aligned with $I(t)$ at each timestep. To allow the pretrained SD model to adapt to the input X_{global}, we further fine-tune the whole SD model using the following diffusion loss $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0, X_{\text{global}}, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, X_{\text{global}}, t)\|_2^2 \right]. \tag{11}$$

Finally, the total loss $\mathcal{L}_{\text{total}}$ in the fine-tune stage is defined as:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{diff}} + (1 - \lambda)\mathcal{L}_{\text{ISA}}, \tag{12}$$

where λ denotes a hyperparameter used for adjusting the magnitude of the alignment task which is set 0.5 in our experiment. Consequently, $\mathcal{L}_{\text{total}}$ facilitates the alignment of scene graph features with images, resulting in a closer resemblance to text features. This alignment enables the unification of scene graph representation and images in a cohesive space, allowing refined scene graph embedding representation is better suited for SD image generation and enhancing the quality of the generated images.

Experiments

Datasets

Following previous works (Johnson, Gupta, and Fei-Fei 2018; Li et al. 2019; Ashual and Wolf 2019) on scene graph-to-image generation, we conduct our experiments on two standard benchmarks, Visual Genome (Krishna et al. 2017) and COCO-Stuff (Caesar, Uijlings, and Ferrari 2018).

- **Visual Genome** constitutes a dataset designed specifically for complex scene understanding contains 108,077 scene graph-image pairs, with additional annotations such as object bounding boxes, object attributes, and relationships. Each image within VG encompasses a spectrum of 3 to 30 objects spanning across 178 distinct categories. Following prior studies(Cheng et al. 2023; Zheng et al. 2023; Yang et al. 2023), we partition the data into training, validation, and test sets with proportions of 80%, 10%, and 10% respectively. Finally, small and uncommon objects were removed, yielding 62,565 images in the VG dataset for training, 5,062 for validation, and 5096 for testing.
- **COCO-Stuff** collects 164K images from COCO 2017, which contains 80 object categories and 91 object bounding boxes and pixel-level segmentation masks. Following (Yang et al. 2022; Zheng et al. 2023), we first employ footage from the COCO 2017 dataset to partition the challenge subset, comprising 40,000/5,000/5,000 images, correspondingly, for the training/validation/testing development sets. After that, we disregard objects occupying less than 2% of the image and filter out images that contain 3 to 8 objects, yielding 24,972 images for training, 1024 for validation, and 2048 for testing.

Evaluation metrics

We consider two standard evaluation metrics, i.e., Inception Score (IS) (Szegedy et al. 2016) and Fréchet Inception Distance (FID) (Heusel et al. 2017), that are used to measure the quality of scene graph-to-image generation.

- **Inception Score (IS)** uses an Inception-V3 (Salimans et al. 2016) pretrained on ImageNet network to compute the statistical score of the output of the generated images.
- **Fréchet Inception Distance (FID)** shows the overall visual quality and diversity of the generated image by measuring the difference in the distribution of features between real images and generated images on an ImageNet-pretrained Inception-V3 (Salimans et al. 2016) network.

Baselines and Implementation Details

Baselines. In the course of our research, we selected four established methods for scene graph-to-image generation as reference points: SG2IM (Johnson, Gupta, and Fei-Fei 2018), PasteGAN (Li et al. 2019), WSGC (Herzig et al. 2020) and SGDiff (Yang et al. 2022). All our experiments follow the evaluation setup and quantitative results reported in related work. For qualitative comparisons, we chose the classical two-stage GAN-based scene graph-to-image generation method (Johnson, Gupta, and Fei-Fei 2018) and the state-of-the-art diffusion-based scene graph-to-image generation method SGDiff (Yang et al. 2022). In addition, we convert scene graph inputs to textual description inputs, comparing the widely used text-to-image generation method Stable Diffusion (Rombach et al. 2022).

Implementation Details. SGG-IG consists of three stages: pre-training, fine-tuning and inference. In the pre-training stage, a standard multi-layer graph convolutional network is employed to implement the relation embedding module, where the node objects and relation edges of the input scene graph are with processed into 512-dimensionality vectors. For the pre-training process with mask self-supervision, we set the random mask ratio to 0.3, and for the pre-training process with spatial constraint, we use CLIP image encoder to encode the images. In the fine-tuning stage, we adopt stable-diffusion-v1-4 (Rombach et al. 2022) as the initialization parameter weights of SGG-IG, and the sampling process uses DDIM (Song, Meng, and Ermon 2020) sampling with 100 sampling steps. In both the pre-training and fine-tuning stages, we employ the Adam optimizer (Kingma and Ba 2014) with learning rates of 5e-4 and 1e-6, correspondingly. We set batch size to 2, and perform 700,000 iterations and 30,000 iterations in the fine-tuning stage. All experiments are conducted on the NVIDIA GeForce RTX 4090 GPUs. Our code will be available at our site[1].

Baseline Comparison Results

Quantitative Results As shown in Tab. 1, SGG-IG achieves state-of-the-art performance in both FID and IS metrics compared to other baseline methods at 128 × 128 and 256 × 256 resolution settings. In contrast to SGG-IG,

[1] https://github.com/fuyunwang/SGG-IG

Methods	COCO-Stuff 128×128 FID (↓)	COCO-Stuff 128×128 IS (↑)	COCO-Stuff 256×256 FID (↓)	COCO-Stuff 256×256 IS (↑)	Visual Genome 128×128 FID (↓)	Visual Genome 128×128 IS (↑)	Visual Genome 256×256 FID (↓)	Visual Genome 256×256 IS (↑)
SG2IM	93.3	7.1 ± 0.2	99.1	8.2 ± 0.2	82.7	6.1 ± 0.1	90.5	7.9 ± 0.1
PasteGAN	70.7	11.1 ± 0.7	79.1	12.3 ± 1.0	61.2	7.6 ± 0.7	66.5	8.1 ± 0.9
WSGC	108.6	5.1 ± 0.3	121.7	6.5 ± 0.3	80.4	7.2 ± 0.3	84.1	9.8 ± 0.4
SGDiff	30.2	14.6 ± 0.9	36.2	17.8 ± 0.8	20.1	11.4 ± 0.5	26.0	16.4 ± 0.3
SGG-IG	**26.2**	**19.3 ± 0.5**	**31.4**	**22.5 ± 0.3**	**16.3**	**15.2 ± 0.1**	**21.9**	**20.4 ± 0.4**

Table 1: SGG-IG achieves substantial performance gains on COCO-Stuff and Visual Genome datasets, surpassing previous

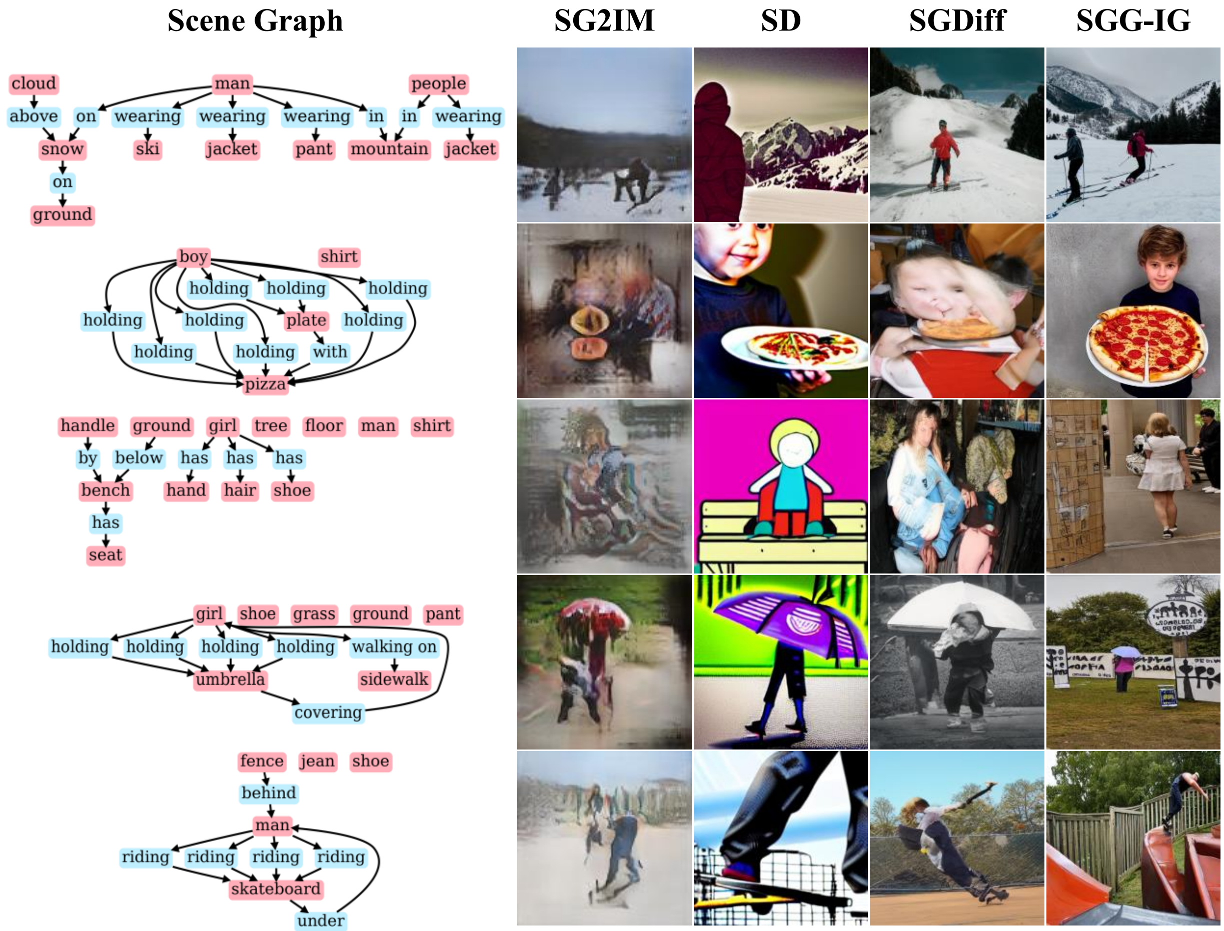

Figure 3: Visual comparisons of different methods on the Visual Genome. SD stands for Stable Diffusion. It can be seen that SGG-IG generates images that are semantically consistent with the scene graph and of high quality compared to previous work.

the outcomes achieved by current methods in generating images that exhibit both semantic coherence with the scene graph and high visual fidelity are found to be unsatisfactory. This limitation arises from the inability to effectively capture structural knowledge and contextual information within the scene graph and the gap between the scene graph and the visual content is not bridged. Consequently, the model struggles to comprehend the fine-grained semantics embedded in the scene graph, particularly in scenarios involving multiple objects and their complex interactions.

From the Tab. 1, we can observe that SGG-IG significantly outperforms the GAN-based two-stage scene-graph-to-image generation method due to the fact that the intermediate representation of the scene layout fails to optimize the semantic alignment of the scene-graph to the image, and falls short of expressing the interaction relations (e.g., wearing, covering,

MAE	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{ISA}}$	**COCO-Stuff**				**Visual Genome**			
			128 × 128		**256 × 256**		**128 × 128**		**256 × 256**	
			FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)
✗	✗	✗	66.2	8.0 ± 0.5	73.8	11.3 ± 0.5	56.7	7.8 ± 0.1	62.3	10.4 ± 0.2
✗	✗	✓	36.3	13.6 ± 0.5	40.2	15.5 ± 0.4	25.4	12.7 ± 0.2	33.9	14.6 ± 0.3
✗	✓	✗	43.5	9.1 ± 0.3	50.7	12.1 ± 0.4	33.9	8.8 ± 0.1	42.2	11.3 ± 0.3
✓	✗	✗	39.6	10.4 ± 0.3	45.2	13.2 ± 0.3	30.8	10.0 ± 0.2	39.5	12.4 ± 0.2
✓	✓	✗	34.4	15.7 ± 0.3	38.8	17.4 ± 0.3	23.3	13.3 ± 0.5	31.5	16.3 ± 0.5
✓	✗	✓	<u>28.8</u>	<u>18.2 ± 0.2</u>	<u>33.5</u>	<u>20.9 ± 0.1</u>	<u>21.6</u>	<u>14.2 ± 0.4</u>	<u>25.2</u>	<u>18.7 ± 0.2</u>
✓	✓	✓	**26.2**	**19.3 ± 0.5**	**31.4**	**22.5 ± 0.3**	**16.3**	**15.2 ± 0.1**	**21.9**	**20.4 ± 0.4**

Table 2: An ablation study was conducted to evaluate different combinations of components. MAE is an abbreviation for masked auto-encoder. The best performance is highlighted in **bold** and the second best performance is <u>underlined</u>.

holding). In contrast, SGDiff (Yang et al. 2022) uses a diffusion model trained from scratch and directly optimizes the scene-graph-to-image alignment to achieve decent generation results. However, it fails to utilize the strong generative prior of pre-trained text-to-image models, such as Stable Diffusion. Moreover, it only conducts initial alignment during the pre-training stage, neglecting the issue of injected scene graph information deviating from the original semantic content of the scene graph due to sampling errors in the denoising stage. In contrast, SGG-IG effectively mitigates these issues and attains superior performance compared to existing methods.

Qualitative Results Fig. 3 shows a qualitative comparison between the images generated by our method and those generated by publicly available models. Our comparisons include three representative methods, i.e., the GAN-based SG2IM (Johnson, Gupta, and Fei-Fei 2018), the text-to-image generation approach Stable Diffusion (Rombach et al. 2022), and the diffusion-based scene graph-to-image generation method SGDiff (Yang et al. 2022). The qualitative comparisons reveals the superior performance of our model. For instance, as illustrated in the second and third rows of Fig. 3, SGDiff synthesizes a distorted image that inaccurately represents the objects within the scene and SG2IM exhibits poor quality attributable to unstable GAN (Goodfellow et al. 2020) training and pattern collapse. In the first and third lines, SG2IM (Johnson, Gupta, and Fei-Fei 2018) and SGDiff (Yang et al. 2022) either fail to generate the node objects in the specified scene graph or there is an inconsistency in the number. One possible explanation is that these methods fail to maximize scene-graph-to-image alignment during the training or fine-tuning stage of the model. In addition, for all scene graph inputs, when we convert the scene graphs to text and generate them with the help of Stable Diffusion, it is difficult to generate images that correspond to the spatial and interactive relationships described by the scene graphs in a consistent manner due to the ambiguity of the text and the lack of understanding of the relationships among objects. Since SGG-IG excels in maximizing alignment with the input scene graph during image generation, the images synthesized by SGG-IG exhibit fidelity to the input scene graph structure, generating semantically coherent and high-fidelity images.

Ablation Study

In this section, we conduct extensive ablation studies on SGG-IG based on different generated image resolutions. Specifically, by removing the masked auto-encoder module, the spatial constraint loss, and the image-scene alignment loss, we constructed corresponding versions of six different model variants. As illustrated in Tab. 2, compared to solely removing the masked auto-encoder module and the spatial constraint, SGG-IG exhibits a notable decline in performance upon the exclusion of the image-scene alignment loss. This demonstrates the effectiveness of the image-scene alignment loss in maximizing scene graph-to-image alignment during denoising process. By aligning the global scene graph representation with image embedding at each time step through image-scene alignment constraints, the semantic deviation of the scene graph due to random sampling during the denoising process is avoided. As the second crucial module, the masked auto-encoder captures the structural knowledge and contextual information of a scene graph, enhancing the model's understanding of fine-grained semantics within the scene graph. Relatively speaking, the contribution of spatial constraints is minor but still beneficial. By bridging the semantic gap between the scene graph and image, spatial constraints offer an initial alignment from scene graph to image.

Conclusion

In this study, we propose SGG-IG, a novel approach for generating images from scene graphs. SGG-IG directly optimizes the alignment between the scene graph and the generated image, eliminating the need for intermediate rough layouts. By incorporating mask self-supervised learning and a spatial constraint loss based on contrastive learning, our model enhances its understanding of the structural and visual relationships within the scene graph, resulting in higher-quality and more precise images. To leverage the strong priors of pretrained text-to-image diffusion models, SGG-IG introduces an image-scene alignment loss during the denoising process to further improve scene graph-to-image alignment. The implementation consists of three phases: pre-training, fine-tuning, and inference, resulting in a streamlined and efficient tuning process. In the future, we aim to develop an end-to-end model that integrates these three stages.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants Nos. 62476133, 62272230), the Natural Science Foundation of Shandong Province (Grant No. ZR2022LZH003).

References

Ashual, O.; and Wolf, L. 2019. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4561–4569.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

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

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5343–5353.

Cheng, J.; Liang, X.; Shi, X.; He, T.; Xiao, T.; and Li, M. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*.

Dhamo, H.; Farshad, A.; Laina, I.; Navab, N.; Hager, G. D.; Tombari, F.; and Rupprecht, C. 2020. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5213–5222.

Garg, S.; Dhamo, H.; Farshad, A.; Musatian, S.; Navab, N.; and Tombari, F. 2021. Unconditional scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16362–16371.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Herzig, R.; Bar, A.; Xu, H.; Chechik, G.; Darrell, T.; and Globerson, A. 2020. Learning canonical representations for scene graph to image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 210–227. Springer.

Herzig, R.; Raboh, M.; Chechik, G.; Berant, J.; and Globerson, A. 2018. Mapping images to scene graphs with permutation-invariant structured prediction. *Advances in Neural Information Processing Systems*, 31.

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

Ivgi, M.; Benny, Y.; Ben-David, A.; Berant, J.; and Wolf, L. 2021. Scene graph to image generation with contextualized object layout refinement. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2428–2432. IEEE.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7701–7711.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, Y.; Ma, T.; Bai, Y.; Duan, N.; Wei, S.; and Wang, X. 2019. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems*, 32.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, 1261–1270.

Li, Z.; Wu, J.; Koh, I.; Tang, Y.; and Sun, L. 2021. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13819–13828.

Luo, J.; Wang, Y.; Gu, Z.; Qiu, Y.; Yao, S.; Wang, F.; Xu, C.; Zhang, W.; Wang, D.; and Cui, Z. 2024. MMM-RS: A Multi-modal, Multi-GSD, Multi-scene Remote Sensing Dataset and Benchmark for Text-to-Image Generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ma, W.-D. K.; Lahiri, A.; Lewis, J.; Leung, T.; and Kleijn, W. B. 2024. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4098–4106.

Mishra, R.; and Subramanyam, A. 2024. Scene Graph to Image Synthesis: Integrating CLIP Guidance with Graph Conditioning in Diffusion Models. *arXiv preprint arXiv:2401.14111*.

Mittal, G.; Agrawal, S.; Agarwal, A.; Mehta, S.; and Marwah, T. 2019. Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

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

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sortino, R.; Palazzo, S.; Rundo, F.; and Spampinato, C. 2023. Transformer-based image generation from scene graphs. *Computer Vision and Image Understanding*, 233: 103721.

Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eledath, J.; Medioni, G.; and Sigal, L. 2021. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13936–13945.

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

Sylvain, T.; Zhang, P.; Bengio, Y.; Hjelm, R. D.; and Sharma, S. 2021. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2647–2655.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

Wang, F.; Gao, X.; Chen, Z.; and Lyu, L. 2023. Contrastive Multi-Level Graph Neural Networks for Session-based Recommendation. *IEEE Transactions on Multimedia*.

Wang, Y.; Li, Y.; and Cui, Z. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Li, Y.; Zhang, X.; Liu, X.; Dai, A.; Chan, A. B.; and Cui, Z. 2024. Edit Temporal-Consistent Videos with Image Diffusion Model. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(12).

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Yang, B.; Luo, Y.; Chen, Z.; Wang, G.; Liang, X.; and Lin, L. 2023. Law-diffusion: Complex scene generation by diffusion with layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22669–22679.

Yang, L.; Huang, Z.; Song, Y.; Hong, S.; Li, G.; Zhang, W.; Cui, B.; Ghanem, B.; and Yang, M.-H. 2022. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*.

Zhao, B.; Meng, L.; Yin, W.; and Sigal, L. 2019. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8584–8593.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490–22499.