# R3CD: Scene Graph to Image Generation with Relation-aware Compositional Contrastive Control Diffusion

**Jinxiu Liu**[1], **Qi Liu**[1*]

[1]School of Future Technology, South China University of Technology
jinxiuliu0628@foxmail.com, drliuqi@scut.edu.cn

## Abstract

Image generation tasks have achieved remarkable performance using large-scale diffusion models. However, these models are limited to capturing the abstract relations (viz., interactions excluding positional relations) among multiple entities of complex scene graphs. Two main problems exist: 1) fail to depict more concise and accurate interactions via abstract relations; 2) fail to generate complete entities. To address that, we propose a novel Relation-aware Compositional Contrastive Control Diffusion method, dubbed as R3CD, that leverages large-scale diffusion models to learn abstract interactions from scene graphs. Herein, a scene graph transformer based on node and edge encoding is first designed to perceive both local and global information from input scene graphs, whose embeddings are initialized by a T5 model. Then a joint contrastive loss based on attention maps and denoising steps is developed to control the diffusion model to understand and further generate images, whose spatial structures and interaction features are consistent with *a priori* relation. Extensive experiments are conducted on two datasets: Visual Genome and COCO-Stuff, and demonstrate that the proposal outperforms existing models both in quantitative and qualitative metrics to generate more realistic and diverse images according to different scene graph specifications.

## Introduction

Scene Graph to Image Generation (SG2IM) (Johnson, Gupta, and Fei-Fei 2018) is a challenging task that aims to generate realistic and diverse images from graph-structured inputs. This is because most existing methods align nodes and connections in graphs with objects and their relations in images via image-like representations of scene graphs, which causes suboptimal alignment due to the manually crafted scene layouts. Nevertheless, scene graphs are rich in describing entities and their interactions to one another (Johnson et al. 2015)(Krishna et al. 2017). Therefore, to study image generation from scene graphs becomes necessary for users to specify through various types of specifications, e.g., labels, captions, and so on (Johnson, Gupta, and Fei-Fei 2018).

As mentioned above, one challenge for SG2IM is because of suboptimal alignment between nodes-connections in graphs and object-relations in images. To address that, some works introduced an additional layout prediction module to initialize the spatial arrangements of objects(Ashual and Wolf 2019; Herzig et al. 2020; Johnson, Gupta, and Fei-Fei 2018; Yang et al. 2022), while others encoded the entire scene graph via graph convolution network (GCN) to control the diffusion model more precisely (Yang et al. 2022). However, these methods are limited to generating abstract relations (excluding positional information) in scene graphs, such as eating, chasing shaking, etc., as demonstrated in Fig. 1. Although semantic-aware attention maps were introduced to guide the generation process, and yet still cannot perceive the abstract relations. This is because they focus more on learning the entity shapes and layouts to match the pixels, and yet it is hard to represent the abstract relations with pixels, especially for those interaction features between pixel regions of entities. Another problem is the compliance between the generated images and original scene graphs, which is due to the existing approaches using image-like representations of scene graphs to create coarse sketches.

To that end, a novel Relation-aware Compositional Contrastive Control Diffusion framework, named R3CD, is proposed, as shown in Fig. 3, which consists of twofolds: (1) a *SGFormer (Scene Graph transFormer)* to refine the entity and relation embeddings is initialized by a T5 model (Raffel et al. 2020), for capturing both local and global information; (2) *a relation-aware compositional contrastive control framework* with joint contrastive loss, which utilizes the attention maps and the denoising steps as contrastive factors to guide the image generation process. The proposed R3CD enables to facilitate the alignment between generated images and original graphs based on various specifications. We have evaluated the proposal on Visual Genome (Krishna et al. 2017) and COCO-Stuff (Caesar, Uijlings, and Ferrari 2018) datasets, where R3CD is superior to other competitors both in quantitative metrics (IS (Heusel et al. 2017), e.g., FID (Salimans et al. 2016)), and qualitative visualization results. We also conduct extensive ablation studies to verify the effectiveness of each module. Our main contributions are summarized as follows :

- We propose a novel SGFormer that encodes both nodes and edges of input scene graphs to capture both local and global information. Our ablation studies show that our model outperforms existing methods.

*Corresponding author


Figure 1: (a) R3CD versus existing layout-based methods. Layout-based models can generate images well with positional relations, e.g., "left to", "under", while they fail to depict more concise and accurate interactions via abstract relations, e.g., "eat", "play". (b) R3CD versus SGDiff. Regarding multiple entities in scene graph, SGDiff often ignores some entities, e.g., "boy", "helmet", while ours can generate complete ones (all 6 entities and 5 relations in the sample).

- We adopt a triplet-level approach for compositional image generation, which enables the images to not only reflect the visual interactions between two objects in a triplet but also address the problem of missing entities.
- We introduce the R3CD framework, which facilitates the alignment between generated images and original graphs based on various specifications. Our experimental results demonstrate the superior performance of our method.

## Related Works

**Compositional Image Generation** Diffusion models have achieved remarkable performance in this task by modeling the image generation process as a reverse diffusion process (Ruiz et al. 2023) (Saharia et al. 2022) (Nichol and Dhariwal 2021) (Ramesh et al. 2022). However, most existing methods rely on a single prompt to generate an image (Saharia et al. 2022) (Ramesh et al. 2021), which may not be sufficient to express the fine-grained details and structure of complex scenes. To address this issue, compositional generation techniques have been proposed to enhance the ability of Text-to-Image (T2I) diffusion models to synthesize features from multiple text segments without relying on additional bounding box inputs (Wang et al. 2023) (Cheng et al. 2023). For example, compositional diffusion (Wang et al. 2023) (Du et al. 2023) segments complex text descriptions

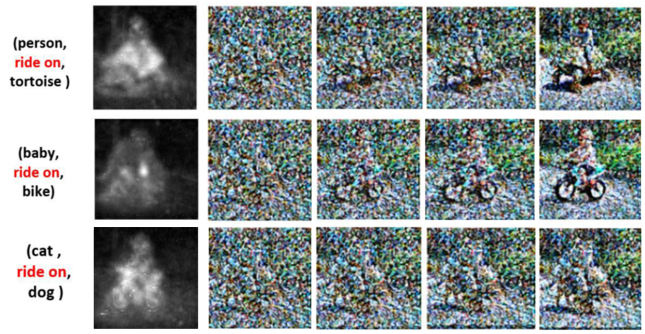


Figure 2: For images with same interactions and different objects, it is visualized that their attention maps and distributions of denoising steps look similar. Motivated by that, the proposal is developed to align the abstract relations.

into simpler parts that are easier to generate, and fuses the outputs of these parts into a coherent image. However, this method is restricted to conjunction and negation operators, and does not consider the specific pose features exhibited by objects when interacting with other objects, such as handshakes, hugs, riding, etc. In this paper, we propose a novel compositional generation method for scene graph generation based on diffusion models, which can integrate both abstract relations and entities to enhance the expressive ability of diffusion models for various information in scene graphs. Specifically, we assign a separate noise prediction module for abstract information generation, which leverages a Gaussian denoising module to improve the expressive ability and capture abstract information more accurately.

**Image Generation from Scene Graphs** Image generation from scene graphs (SGs) is a challenging task that aims to generate images from graphical representations (Johnson, Gupta, and Fei-Fei 2018) (Krishna et al. 2017). Existing methods for this task can be divided into two categories: layout-based methods (Johnson, Gupta, and Fei-Fei 2018; Ashual and Wolf 2019; Liu et al. 2022; Du et al. 2023) and graph-based methods (Feng et al. 2022; Yang et al. 2022; Li et al. 2022; Wang et al. 2023; Wu, Wei, and Lin 2023a). Layout-based methods first map the SG to a scene layout and then refine the scene layout into a realistic image using a generative model. However, they may suffer from relation ambiguity and entity missing problems, as not all connections in the SG can be accurately translated into spatial layouts. Moreover, they may introduce irrelevant information in the intermediate scene layout representations, which complicates the training of downstream generative models. In contrast, graph-based methods encode semantic information from the SG directly and avoid the limitations of scene layouts. But they still face challenges in generating abstract relations that are hard to express with pixels, such as eating and looking. Therefore, in this paper, we propose a scene graph global-awared node and edge encoding method and a novel method to relearn abstract relations in order to enhance the diffusion model's ability for complex SGs understanding. We do not directly use explicit image infor-

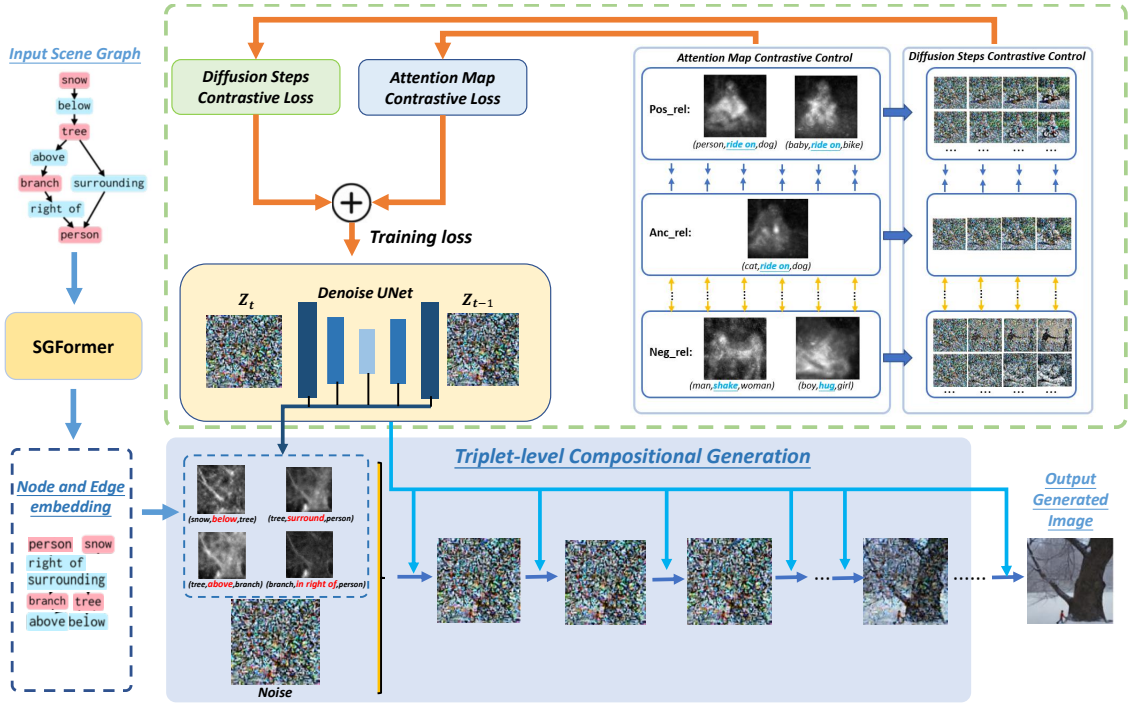


Figure 3: The whole pipeline of R3CD, where node and edge embeddings are encoded by the proposed SGFormer and then are fed to compositional generation model under the guidance of relation-aware contrastive control loss. The joint contrastive loss is designed on attention maps and diffusion steps, where "Pos_rel", "Anc_rel", "Neg_rel" denote positive, anchor, and negative samples' relationships, respectively.

mation to express relation information, but apply attention maps and denoising steps of a diffusion model controlled by contrastive control in compositional generation to represent abstract relation information and generate images from the complex scene graph.

## Method

In this section, we introduce the proposed R3CD in detail, including the SGFormer, the relation-aware contrastive control loss and triplet-level compositional generation, as shown in Fig. 3.

### SGFormer

As shown in Fig. 4, SGFormer comprises two components: (1) The graph attention layer to compute attention scores between node and edge features, and aggregate information from neighboring nodes and edges; (2) The graph update layer, to update the node and edge features based on the aggregated information.

The node and edge embeddings are initialized by a T5 model (Raffel et al. 2020). The T5 model is a pre-trained text-to-text transformer to generate natural language representations for various tasks. Let $G = (V, E)$ be a directed scene graph (SG), where $V$ is the set of nodes (entities) and $E$ is the set of edges (relations). Each node $v \in V$ and each edge $e \in E$ have a text label. The node and edge embeddings $h$ are defined as $h_v^0$ and $h_e^0$. For each layer $l$, the attention scores between nodes and edges based on their features and types are:

$$\alpha_{v\to e}^{l} = \frac{\exp\left(\sigma\left(W_{v\to e}^{l}[h_v^{l-1}; h_e^{l-1}; t_{v\to e}]\right)\right)}{\sum_{u\in N(v)} \exp\left(\sigma\left(W_{v\to u}^{l}[h_v^{l-1}; h_u^{l-1}; t_{v\to u}]\right)\right)} \tag{1}$$

$$\alpha_{e\to v}^{l} = \frac{\exp\left(\sigma\left(W_{e\to v}^{l}[h_e^{l-1}; h_v^{l-1}; t_{e\to v}]\right)\right)}{\sum_{f\in N(e)} \exp\left(\sigma\left(W_{e\to f}^{l}[h_e^{l-1}; h_f^{l-1}; t_{e\to f}]\right)\right)} \tag{2}$$

where $\sigma$ denotes an activation function, $\alpha_{v\to e}^{l}$ and $\alpha_{e\to v}^{l}$ are the attention scores from node $v$ to edge $e$ and from edge $e$ to node $v$, respectively. $N(v)$ and $N(e)$ are the sets of neighboring nodes and edges of $v$ and $e$, respectively. $W_{v\to e}^{l}$, $W_{v\to u}^{l}$, $W_{e\to v}^{l}$, and $W_{e\to f}^{l}$ are learnable weight matrices; $t_{v\to e}$, $t_{v\to u}$,$t_{e\to u}$ and $t_{e\to f}$ are vectors to represent different types of nodes and edges; [ ] denotes the concatenation operation.

Next, we use the attention scores to aggregate information from neighboring nodes and edges:

$$\widetilde{h}_v^l = \sum_{e\in N(v)} \alpha_{v\to e}^{l} h_e^{l-1}, \quad \widetilde{h}_e^l = \sum_{v\in N(e)} \alpha_{e\to v}^{l} h_v^{l-1} \tag{3}$$

where $\widetilde{h}_v^l$ and $\widetilde{h}_e^l$ are the aggregated node and edge features, respectively.

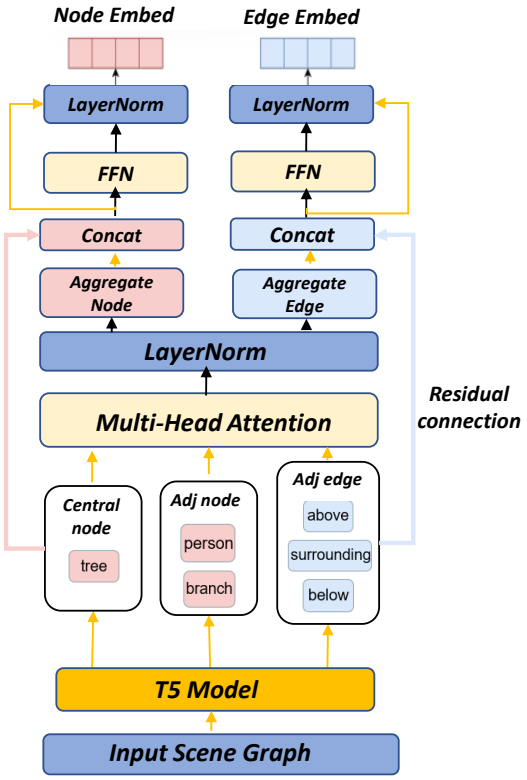


Figure 4: The architecture of SGFormer.

Finally, we use different feed-forward networks to update the node and edge features based on the aggregated information:

$$h_v^l = FFN_v(h_v^{l-1} + \widetilde{h}_v^l), \quad h_e^l = FFN_e(h_e^{l-1} + \widetilde{h}_e^l) \quad (4)$$

where $h_v^l$ and $h_e^l$ are the updated node and edge embeddings, respectively; and $FFN_v$ and $FFN_e$ are two different feed-forward networks with residual connections and layer normalization.

## Relation-aware Diffusion Contrastive Control

As discussed before, layout-based models suffer from relation ambiguity and entity missing problems. Motivated by that attention maps can provide free token-region associations in trained T2I models (Hertz et al. 2022), attention maps are utilized as position information. Attention maps are initially generated to indicate the positions of each entities in image. Then a joint contrastive loss based on attention maps and diffusion denoising steps is proposed to control the image generation process.

**Attention Map Contrastive Learning** The attention map represents the similarity between pixel in image and node in scene graph. It is expected that the attention map can capture the spatial layout and relative position of the entities in scene graph, regardless of different specific types. For example, the attention map for the triplet (cat, eat, apple) should be similar to the one for (dog, eat, banana), but different from the one for (cat, chase, dog), as shown in Fig 2. Therefore, we design a contrastive loss objective to minimize the similarity between attention maps that correspond to different relations, and maximize the similarity between those that correspond to the same relations. We apply the attention maps of triplets with the same abstract relation as positive samples, and ones with different relations but similar entities as negative samples.

Given a batch of $N$ triplets from different scene graphs, we first generate their corresponding attention maps $A_1, A_2, ..., A_N$ for each edge using the diffusion model based on scene graphs. Then, for each triplet $i$, we randomly sample another triplet $j$ from the same batch that has the same abstract relation as $i$, and use their attention maps $A_i$ and $A_j$ as positive samples. Similarly, we randomly sample another triplet $k$ from the same batch that has different relation but same or semantically similar entities and use their attention maps $A_i$ and $A_k$ as negative samples, which disentangles the layout and appearance. The intuition is that the attention maps of positive samples should have high similarity, while the attention maps of negative samples should have low similarity. The cosine similarity is applied as the metric to measure the similarity between two attention maps. The attention map loss for relation $i$ is defined as:

$$L_i = -\log \frac{\exp(\cos(A_i, A_j)/\tau)}{\exp(\cos(A_i, A_j)/\tau) + \sum_{k \neq i} \exp(\cos(A_i, A_k)/\tau)} \quad (5)$$

where $\tau$ is a temperature parameter that controls the sharpness of the distribution. The attention map loss can be minimized by increasing the cosine similarity of positive pairs and decreasing the cosine similarity of negative pairs. The total attention map loss is then computed as:

$$L_{att} = \frac{1}{N} \sum_{i=1}^{N} L_i \quad (6)$$

**Diffusion Steps Contrastive Learning** The diffusion model generates images by adding noise to an initial image at each time step, and then denoising it using a UNet (Ronneberger, Fischer, and Brox 2015). The noise reflects how much uncertainty there is about each pixel value at each time step. Ideally, the noise from the first few denoising steps should match the high-level information in the scene graph, which helps the model comprehend the abstract relations better than aligning the relation information with the pixels directly. To promote this property, we devise a contrastive loss objective that aligns the noise distributions predicted by the UNet with the abstract relation information. We use the predicted noise distributions for two corrupted images that share the same triplet as positive samples, and for two corrupted images that have different triplets as negative samples.

Given a batch of $N$ triplets embedding encoded by SGFormer, we first generate their corresponding images using a trained diffusion model. Then we corrupt each image by adding Gaussian noise at time step $t$, as a result of $n$ corrupted images $x_1, x_2, \ldots x_n$. The UNet predicts a noise dis-

tribution $\epsilon_\theta(z_t, t, c)$ for each pixel in each noisy image $\chi_n$, where $z_t$ is a latent variable representing a noisy image at time step $t$, and $c$ is a vector representing the triplet representation vector extracted from the scene graph. We then sample a new noisy image $x_n$ from the predicted noise for each $\chi_n$.

For each triplet $i$, we randomly sample another triplet $j$ from the same batch with the same abstract relation as $i$, and use their noises $\epsilon_\theta(z_t, t, c_i)$ and $\epsilon_\theta(z_t, t, c_j)$ as positive samples. Similarly, we randomly sample another triplet $k$ from the same batch with different relation but similar entities as $i$ and take their noise $\epsilon_\theta(z_t, t, c_k)$ as negative samples. We use the L2 norm as our similarity measure function. Therefore, we define $f(z_n, t, c) = \epsilon_\theta(z_t, t, c)$ and measures the similarity between two noise distributions as:

$$E(\epsilon_\theta(z_t, t, c_i), \epsilon_\theta(z_t, t, c_j)) = -\left\|\epsilon_\theta(z_t, t, c_i) - \epsilon_\theta(z_t, t, c_j)\right\|^2 \tag{7}$$

The L2 norm is inversely proportional to the Euclidean distance between the noise distributions, implying that a higher value indicates greater similarity. We propose a contrastive loss function that minimizes the L2 norm for positive pairs and maximizes it for negative pairs:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N}\sum_{t=T'}^{T}\sum_{i=1}^{N} log\frac{e^{E(\epsilon_\theta(z_t,t,c_i),\epsilon_\theta(z_t,t,c_j))/\tau}}{\sum_{k\neq i} e^{E(\epsilon_\theta(z_t,t,c_i),\epsilon_\theta(z_t,t,c_k))/\tau}} \tag{8}$$

Where $T' = T - \Delta t$, $T$ and $T'$ are the first few diffusion steps that can capture the high-level features of the image rather than the details. The contrastive loss function encourages the model to generate noise distributions that are similar for images with the same relation and dissimilar for images with different relations. Therefore, the model can learn to extract the abstract relation information from the scene graph and transfer it to the image domain.

**Total Loss Function and Optimization** The loss function of our model is composed of three terms: the diffusion loss, the attention map contrastive loss, and the diffusion steps contrastive loss. The diffusion loss is the standard loss function used for training diffusion models, it can be written as:

$$\mathcal{L}_{\text{diff}}(\theta) := E_{t,\mathbf{x}_0,\epsilon}\Big[\left\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \tau_\theta(c))\right\|^2\Big] \tag{9}$$

where $\theta$ is the whole model parameter, $t$ is a random time step sampled from a uniform distribution, $X_0$ is an original image sampled from the data distribution, $\epsilon$ is a standard Gaussian noise variable, $x_t$ is a noisy image obtained by adding $\epsilon$ to $x_0$. $\tau_\theta(c)$ is all of the relation representation vector extracted from the scene graph by SGFormer, and $\epsilon_\theta(x_t, t, \tau_\theta(c))$ is a predicted noise distribution conditioned on $x_t$, $t$, and $\tau_\theta(c)$.

The other two components are the attention map contrastive loss and the diffusion steps contrastive loss. We compute the total loss as a weighted sum of these components:

$$\mathcal{L}_{tot}(\theta) = \mathcal{L}_{diff}(\theta) + \lambda\mathcal{L}_{att}(\theta) + \gamma\mathcal{L}_{cont}(\theta) \tag{10}$$

where $\lambda$ and $\gamma$ are trade-off parameters to control the weights of the contrastive losses. The gradient of the total loss with respect to $\theta$ is computed by applying the chain rule:

$$\nabla_\theta\mathcal{L}_{tot}(\theta) = \nabla_\theta\mathcal{L}_{diff}(\theta) + \lambda\nabla_\theta\mathcal{L}_{att}(\theta) + \gamma\nabla_\theta\mathcal{L}_{cont}(\theta) \tag{11}$$

### Triplet-level Compositional Generation

Compositional generation is a technique that creates contents for different regions in an image, based on multiple layouts and corresponding prompts. Following (Wang et al. 2023), we use attention maps to control the generation of various scenes in each region. However, the existing works (Cheng et al. 2023; Wang et al. 2023) neglect the interaction relationships between objects, and they only use the single-object layout as guidance for the diffusion model, which makes them unaware of the inter-object relations, especially the abstract ones. Unlike them, we generate images at the triplet level rather than the object level. Specifically, we use edge embedding attention maps as layout guidance to generate the scene compositionally at the node-edge-node triplet-level during the diffusion model sampling process, where each triplet corresponds to an edge embedding attention map encoded by SGFormer and based on the two contrastive losses of UNet that we trained, which can capture the relationship information of adjacent nodes on the edge.

## Experiments

In this section, we evaluate the effectiveness of our model on two datasets: Visual Genome (Krishna et al. 2017) and COCO-Stuff (Caesar, Uijlings, and Ferrari 2018), where R3CD is superior to other competitors both in quantitative metrics (IS (Heusel et al. 2017), e.g., FID (Salimans et al. 2016)), and qualitative visualization results. We also conduct extensive ablation studies to verify the effectiveness of each module.

### Experiment Settings

We conduct experiments to compare R3CD with the state-of-the-art SG2IM method using diffusion model (Yang et al. 2022) and the previous GAN-based methods (Ashual and Wolf 2019; Herzig et al. 2020; Johnson, Gupta, and Fei-Fei 2018). We adopt their evaluation settings for all experiments on images of 256×256 resolution. We use Adam optimizer (Kingma and Ba 2014) to train diffusion models with a learning rate of 5e-5, a batch size of 16, and 700,000 iterations on RTX 3090. For the contrastive loss module, we choose the trade-off parameters as 0.01, and as 0.02.

### Quantitative Comparisons

Previous methods for image generation from scene graphs can be classified into three categories: GAN-based (Ashual and Wolf 2019; Herzig et al. 2020; Johnson, Gupta, and Fei-Fei 2018), layout prediction (Cheng et al. 2023), and scene graph encoding-based diffusion model (Yang et al. 2022). However, these methods have limitations such as mode collapse, high complexity, layout errors, and global information

unawareness. In contrast, our method achieves superior performance on COCO-Stuff and Visual Genome datasets in terms of IS and FID metrics, as shown in Table 1. We attribute this to our novel SGFormer encoder, which can capture both local and global information from the scene graph. Our contrastive control module enables the model to learn the abstract relations in the scene graph without introducing an extra layout prediction module, and triplet-level compositional generation which generates images by splitting them into triplets and learning each triplet separately instead of the whole image, thereby improving the training efficiency and the granularity and the flexibility of the generation.

| Method | COCO(IS)↑ | VG(IS)↑ | COCO(FID)↓ | VG (FID)↓ |
|---|---|---|---|---|
| Real Img | 30.7 | 27.3 | - | - |
| Sg2Im (Johnson, Gupta, and Fei-Fei 2018) | 8.2 | 7.9 | 99.1 | 90.5 |
| WSGC (Herzig et al. 2020) | 6.5 | 9.8 | 121.7 | 84.1 |
| SOAP (Ashual and Wolf 2019) | 14.5 | - | 81.0 | - |
| PasteGAN (Li et al. 2019) | 12.3 | 8.1 | 79.1 | 66.5 |
| KCGM (Wu, Wei, and Lin 2023b) | – | 11.6 | – | 27.4 |
| SGDiff (Yang et al. 2022) | 17.8 | 16.4 | 36.2 | 26.0 |
| **Ours**(*GCN*) | 16.5 | 16.0 | 38.6 | 28.6 |
| **Ours**(*SGFormer*) | 17.5 | 17.3 | 35.7 | 25.2 |
| **Ours**(*SGFormer+R3CD*) | **19.5** | **18.9** | **32.9** | **23.4** |

Table 1: Performance comparisons on COCO Stuff (COCO) and Visual Genome (VG) datasets via IS and FID metrics.

## Qualitative Evaluations

In this part, we present some qualitative results of our model on two aspects: abstract relation generation and graph-to-image generation. We compare our model with SGDiff (Yang et al. 2022), which is the state-of-the-art method based on diffusion models for SG2IM generation.

**Abstract Relation Generation** To demonstrate our model's ability to generate abstract relations, we present the attention maps and generated images from scene graphs with complex or ambiguous relations in Fig. 5. We also compare them with the results of SGDiff(Yang et al. 2022) to highlight the advantages of our model. In order to make SGDiff(Yang et al. 2022) consistent with our model in terms of node and edge initialization and facilitate its extension to unseen abstract relations, we modify the initial node and edge with T5 embedding. From the comparison of the attention maps, we observe that our model can capture the key interaction features among different entities in the image. Moreover, for a given relation, more attention weights are not only allocated to the adjacent nodes, but also to the global entities that are linked or semantically related to it by the scene graph, which reflects our model's global awareness. From the generated images, we discover that our model can produce clearer individual entities due to the compositional generation strategy, and generate images that reflect the interaction features with other entities, rather than just generating isolated entities. For instance, when generating (scene graph: a red cat rides on a yellow dog), SGDiff can only generate two dogs without capturing their relation, while our method solves this problem and can generate both entities and their abstract interactive relation.

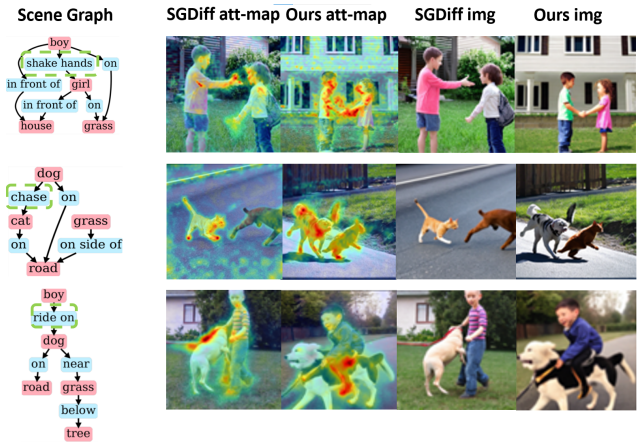


Figure 5: Visualization results of relation features. According to the heat map comparison, our method can capture more critical features that describe the relations than the SGDiff method.

**Graph-to-Image Generation** To evaluate our model on complex scene graphs with multiple entities and relations, we compare our generated images with SGDiff's in Fig. 6. Our model outperforms SGDiff in image quality and diversity, producing more realistic and detailed images that respect the scene graph specifications. Our model also captures the semantic and spatial information of the scene graphs better, such as the relative positions, sizes, orientations, colors, and shapes of different entities. For example, in the second image on the right side of Fig. 6, SGDiff misses the tree and the door entities, and ignores the "in front of" relation. It also fails to generate the (door, has, window) relation correctly. In contrast, our method generates all the entities and relations accurately.

## Ablation Study

In this subsection, we quantify the effectiveness of each component of our method. We conduct ablation studies to show the advantage of SGFormer in capturing global semantic relations over GCN-based methods that encode local relations at the graph level. We also highlight the importance of the two contrastive losses in R3CD for capturing abstract relations in scene graphs.

**Relation and Object Generation Accuracy** To demonstrate the improvement of our method , We evaluate our model in generating images that respect the relations and objects in the scene graphs. We compare with a GCN-based baseline (Yang et al. 2022) and ablate each module of our model: SGFormer, compositional generation, and R3CD. Table 2 shows the results. Our model outperforms the baseline on both tasks, indicating better semantic and spatial understanding of the scene graphs. Each component improves

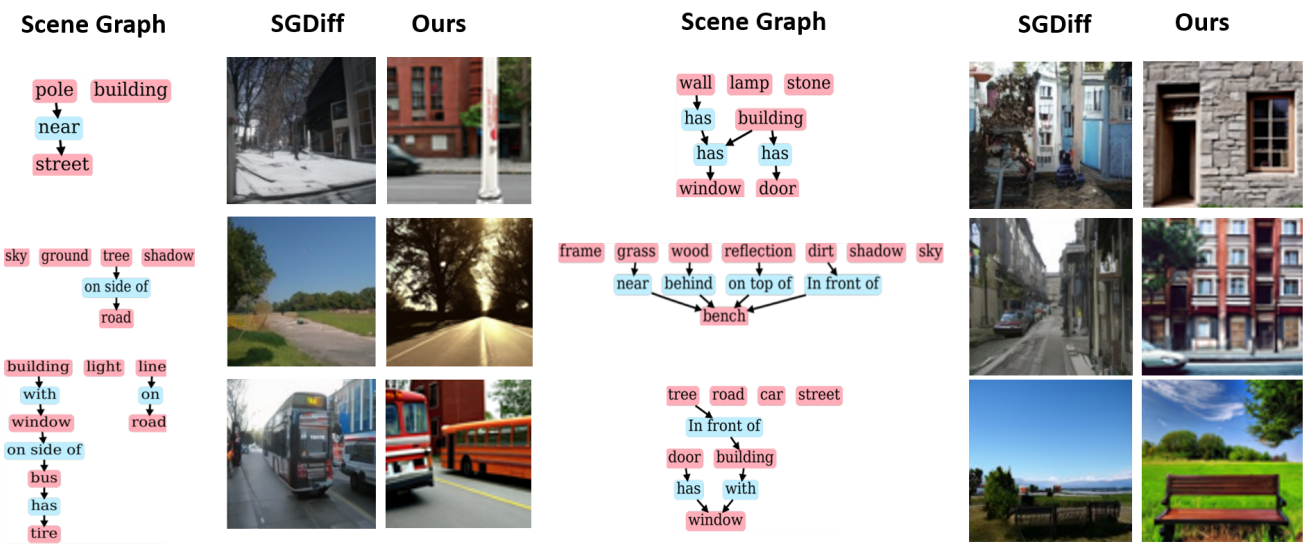


Figure 6: Visual examples of graph-to-image generation in complex scene.

| Methods | G2I-ACC ↑ | I2G-ACC ↑ |
|---|---|---|
| Obj (GCN) | 70.1 | 71.2 |
| Obj + Rel (GCN) | 72.4 (+2.3) | 72.3 (+1.1) |
| Obj + Rel *(SGFormer)* | 72.9 (+2.8) | **72.7 (+1.7)** |
| SGFormer*(Compositional)* | 73.2(+3.1) | – |
| SGFormer *(R3CD)* | **73.8(+3.7)** | – |

Table 2: Ablation study of SGFormer and R3CD on relation synthesis tasks. G2I-ACC stands for the average accuracy Graph-to-Image task, I2G-ACC stands for the average accuracy Image-to-Graph tasks.

the performance. SGFormer refines the node and edge embeddings with global and local information. Compositional generation improves the relation accuracy by generating and fusing each component. R3CD improves both relation and object accuracy by controlling the attention maps and diffusion steps with contrastive learning. These results confirm that our model generates images that comply with the scene graphs, and each module is necessary for this goal.

**Image Generation from Scene Graph** To evaluate the effectiveness of each module, we perform an ablation study on the task of image generation from scene graphs using FID and IS metrics. We compare with baselines of one-hot encoding, layout prediction, and GCN-based diffusion, as shown in Table 3. Table 3 shows that each component of our method improves the performance on both metrics and the complete method achieves the best result. SGFormer enhances the semantic encoding of nodes and edges with local and global information. Attention map contrastive loss ensures the spatial consistency of entities under the same relations by minimizing their attention map distance. Diffusion steps contrastive loss ensures the interaction consistency of entities under the same relations by aligning their noise distributions with relation embeddings. These results confirm that our method generates images from scene graphs effectively, and each module is necessary for this goal.

| Technique | IS ↑ | FID ↓ |
|---|---|---|
| One-hot | 9.9 | 87.1 |
| Layout | 13.1 | 52.7 |
| GCN Diffusion | 16.0 | 28.6 |
| **SGFormer** | 17.3 | 25.2 |
| **SGFormer + Att map Loss** | 17.5 | 24.6 |
| **SGFormer + Diff Loss** | 17.8 | 24.0 |
| **SGFormer + Att Map Loss + Diff Loss** | **18.9** | **23.4** |

Table 3: Ablation study of SGFormer and R3CD on generation fidelity and diversity.

## Conclusion

In this paper, we propose R3CD, a novel framework for image generation from scene graphs that leverages large-scale diffusion models and contrastive control mechanisms, which capture the interactions between entity regions and abstract relation in scene graph. Our method consists of two main components: (1) SGFormer, a transformer-based node and edge encoding scheme that captures both local and global information from scene graphs; (2) Relation-aware Diffusion contrastive control: a contrastive learning module that can align the abstract relation features and the image features across different levels of abstraction, and enhance the model to generate images that reflect the abstract relations. We have conducted extensive experiments on two datasets: Visual Genome and COCO-Stuff, and demonstrated that our method outperforms existing methods in terms of both quantitative and qualitative metrics. We have also shown that our method can generate more realistic and diverse images that respect the scene graph specifications, especially for abstract relations that are hard to express with entity stitching.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Fundamental Research Funds for the Central Universities under Grant 2023ZYGXZR085, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, and in part by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004.

## References

Ashual, O.; and Wolf, L. 2019. Specifying Object Attributes and Relations in Interactive Scene Generation. arXiv:1909.05379.

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.

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*.

Du, Y.; Durkan, C.; Strudel, R.; Tenenbaum, J. B.; Dieleman, S.; Fergus, R.; Sohl-Dickstein, J.; Doucet, A.; and Grathwohl, W. S. 2023. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, 8489–8510. PMLR.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626.

Herzig, R.; Bar, A.; Xu, H.; Chechik, G.; Darrell, T.; and Globerson, A. 2020. Learning Canonical Representations for Scene Graph to Image Generation. arXiv:1912.07414.

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.

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.

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

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, R.; Li, W.; Yang, Y.; Wei, H.; Jiang, J.; and Bai, Q. 2022. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *arXiv preprint arXiv:2210.09549*.

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. arXiv:1905.01608.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 423–439. Springer.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

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*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

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.

Wang, R.; Chen, Z.; Chen, C.; Ma, J.; Lu, H.; and Lin, X. 2023. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*.

Wu, Y.; Wei, P.; and Lin, L. 2023a. Scene Graph to Image Synthesis via Knowledge Consensus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2856–2865.

Wu, Y.; Wei, P.; and Lin, L. 2023b. Scene Graph to Image Synthesis via Knowledge Consensus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 2856–2865.

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*.