SPADE: Spatial-Aware Denoising Network for Open-vocabulary Panoptic Scene Graph Generation with Long- and Local-range Context Reasoning

Xin Hu[1], Ke Qin[1,2], Guiduo Duan[1,2], Ming Li[4], Yuan-Fang Li[3], Tao He[1] *
[1]The Laboratory of Intelligent Collaborative Computing of UESTC
[2]Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province
[3] Monash University, [4] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
{xh1m22,qinke,guiduo.duan}@uestc.edu.cn
ming.li@u.nus.edu, yuanfang.li@monash.edu, tao.he01@hotmail.com

Abstract

*Panoptic Scene Graph Generation (PSG) integrates instance segmentation with relation understanding to capture pixel-level structural relationships in complex scenes. Although recent approaches leveraging pre-trained vision-language models (VLMs) have significantly improved performance in the open-vocabulary setting, they commonly ignore the inherent limitations of VLMs in spatial relation reasoning, such as difficulty in distinguishing object relative positions, which results in suboptimal relation prediction. Motivated by the denoising diffusion model's inversion process in preserving the spatial structure of input images, we propose **SPADE** (SPatial-Aware Denoising-nEtwork) framework—a novel approach for open-vocabulary PSG. SPADE consists of two key steps: (1) inversion-guided calibration for the UNet adaptation, and (2) spatial-aware context reasoning. In the first step, we calibrate a general pre-trained teacher diffusion model into a PSG-specific denoising network with cross-attention maps derived during inversion through a lightweight LoRA-based fine-tuning strategy. In the second step, we develop a spatial-aware relation graph transformer that captures both local and long-range contextual information, facilitating the generation of high-quality relation queries. Extensive experiments on benchmark PSG and Visual Genome datasets demonstrate that SPADE outperforms state-of-the-art methods in both closed- and open-set scenarios, particularly for spatial relationship prediction. The project page is at here.*

1. Introduction

Panoptic Scene Graph Generation (PSG) represents a significant advancement in scene graph generation (SGG) by integrating pixel-level structural analysis with relation reasoning to achieve comprehensive scene understanding. By unifying instance segmentation [34, 38] and visual relationship detection [2, 11] within a single framework, PSG attempts to simultaneously segment all entities in an image and classify their relational interdependencies into subject–predicate–object triples.

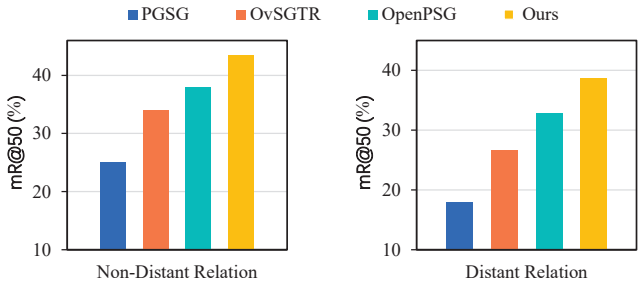

Figure 1. Mean Recall@50 of our proposed method and other VLM-based models [4, 31, 69] on the spatial predicate classification task on the PSG dataset [57]. Notably, we present the result in the two scenarios: distant and non-distant relation pairs.

In prior research, the majority of SGG [8, 13, 14, 16, 30] and PSG methods [51, 57, 68] have been developed from a closed-set perspective. Recent advancements, however, have introduced an open-vocabulary setting for SGG [17, 62], shifting research focus toward a more practical yet challenging problem. Within this paradigm, various methods [4, 31, 59, 65, 69] have been proposed to endow their models with open-world recognition capabilities. A key technique in these works is to leverage the powerful pre-trained Vision Language Models (VLMs) [26, 29, 37], enabling the recognition of unseen objects and relationships.

However, a suite of recent studies [3, 5, 25, 43, 56] have revealed that most state-of-the-art VLMs, such as CLIP

*Corresponding author.

[37], BLIP [26], and GLIP [28], struggle with spatial reasoning, particularly in understanding spatial relationships between objects. This limitation primarily arises because such VLMs are trained on internet-scale datasets composed of image-caption pairs, which often lack detailed spatial descriptions. Consequently, this raises concerns about whether VLM-based SGG or PSG models can still perform effectively in spatial relation prediction. To address this issue, we conduct a comprehensive investigation of several state-of-the-art VLM-based open-set PSG and SGG models, with a specific focus on their performance in spatial relation prediction. Detailed experimental protocols are provided in §4.3. The results, as illustrated in Figure 1, indicate that these VLM-based models underperform on the spatial relation, particularly when two objects are positioned at a considerable distance from each other within an image. We conjecture that increased inter-object distance complicates the VLMs to determine spatial relationships.

Motivated by these findings, we pose a question: *Can we integrate spatial knowledge into a VLM without compromising its inherent open-world recognition capabilities for PSG?* To address this question, we turn to diffusion models for two main reasons. First, diffusion models [1, 39, 40] exhibit remarkable compositional abilities that preserve both spatial and semantic consistency, enabling the injection of spatial knowledge without degrading other semantic information as much as possible. Second, recent work [50, 64] has demonstrated that the inversion process [45], widely utilized in image editing [22, 52], effectively preserves the spatial structure of original images. Thus, we believe that the cross-attention maps derived from the inversion process can serve as explicit spatial cues. However, directly leveraging the pretrained diffusion model does not yield the desired performance, as its preserved knowledge is not fully adapted to the PSG task.

Based on the insights, we present **SPADE** (SPatial-Aware Denoising-nEtwork), a novel two-stage approach for PSG. The proposed method addresses the challenges of spatial relation reasoning and open-vocabulary generalization through two key steps: (1) inversion-guided calibration for the UNet adaptation and (2) spatial-aware context reasoning via graph-based relation inference. In the first step, our core idea is to adapt a general pre-trained teacher diffusion model into a PSG-specific denoising network. Building on the observation that the Denoising Diffusion Implicit Models (DDIM) inversion [45, 50] inherently retains spatial information from input images, we leverage cross-attention maps derived during inversion as spatial priors. To preserve the generalization capabilities of the pre-trained UNet, we leverage a lightweight fine-tuning strategy LoRA [19] with only updating cross-attention layers.

In the second step, SPADE subsequently employs a Relation Graph Transformer (RGT) to capture long-range contextual dependencies and local spatial cues. The RGT operates on a spatial-semantic graph where nodes represent object mask proposals with geometric attributes and semantic embeddings. Through iterative graph propagation, the transformer layer aggregates multi-scale contextual features, resolving ambiguities in overlapping relations and sparse spatial configurations. This dual-context mechanism—combining long-range and local correlations—enables to generate high-quality relation queries.

To summarize, our contributions are as follows:

- We identify a shortcoming in the spatial relation prediction of current VLM-based PSG models. To address this limitation, we propose a two-step spatial-aware denoising network (**SPADE**).
- We introduce an inversion-guided calibration strategy that utilizes cross-attention maps from the reversion process to fine-tune the UNet in a LoRA fashion.
- We design a spatial-aware relation graph transformer that captures local and long-range pairwise contextual information, enabling robust spatial relation reasoning.
- Extensive experiments on benchmark datasets such as VG and PSG demonstrate that **SPADE** significantly outperforms state-of-the-art methods in both closed-set and open-set scenarios, particularly in improving relation prediction in open-vocabulary settings.

2. Related Work

Open-world Panoptic Scene Graph Generation. Panoptic Scene Graph Generation (PSG) or Scene Graph Generation (SGG) aims to represent an image through structured relationships among detected objects [24, 57]. Traditional methods generally focus on a closed set of categories, which limits their capacity to generalize to novel objects and relationships [7, 15, 18, 48, 63]. Recently, open-world SGG has emerged to address this challenge [7, 21, 48, 55, 60, 62, 63]. Earlier zero-shot approaches often leverage commonsense knowledge or structured knowledge graphs to transfer understanding from known to unknown relations [21, 60]. Later advancements integrate multi-modal models [59, 65]. However, they commonly ignore the limitation of VLMs in spatial relation reasoning, which hinders them to predict accurate spatial relationships. In this work, we turn to diffusion models [39] and imbue the spatial knowledge into the denoising network, enhancing its spatial reasoning ability.

DDIM Inversion Process. Denoising Diffusion Implicit Models (DDIM) have recently demonstrated remarkable success in generating high-quality images conditioned on text [1, 39, 40, 50]. These models operate via a forward process that progressively adds Gaussian noise to an image, and a reverse process that denoises the image. Formally, the forward (noising) process is defined as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$. The inversion process maps an input image x into a deterministic noise. During deterministic sampling [45], the image is reconstructed from the noise:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t), \tag{2}$$

where $\epsilon_\theta(x_t, t)$ denotes the predicted noise. In this work, the cross-attention maps derived from deterministic sampling are exploited to extract rich spatial and semantic priors. These priors are instrumental in guiding the calibration of the denoising network, thereby enhancing its suitability for open-world panoptic scene graph generation.

Graph Transformer. Graph transformers have emerged as a powerful paradigm for modeling node relationships by effectively capturing both global context and local dependencies within structured data [41, 42, 61]. This makes it well-suited for reasoning computer vision tasks, where understanding complex object relationships is crucial [36, 53]. This approach not only enhances the modeling of object interactions but also contributes to improved overall scene understanding [8, 46]. In this work, we capitalize on these strengths by integrating graph transformers into our framework to bolster relationship reasoning. Through this integration, our model is better equipped to capture intricate inter- and intra-object dependencies.

3. Methodology

The overall architecture of **SPADE** is illustrated in Figure 2. Briefly, SPADE is composed of two primary stages: inversion-guided calibration and spatial-aware relation context inference. The first stage focuses on imbuing the pre-trained denoising network with spatial and relation knowledge (§3.1), while the second stage, built on the calibrated UNet, performs spatial-aware context reasoning by a novel relation graph transformer (§3.2). Lastly, we present how we make the open-vocabulary relation prediction (§3.3).

3.1. Inversion-Guided Calibration for the UNet

Diffusion models have emerged as backbones for many downstream vision tasks [23, 32, 56, 58, 66]. Formally, given an input image-text pair (x, w), these models first construct a noisy image x_t and encode the text w using a pre-trained text encoder ϵ_t. Intermediate features f_t are then extracted from the UNet as follows:

$$f_t = \text{UNet}(x_t, \epsilon_t(w)). \tag{3}$$

In the context of PSG, however, an input text description w is not available at inference. Following [56, 58], one option is to replace $\epsilon_t(w)$ with a global image feature obtained from the CLIP [37] image encoder.

Nevertheless, we observe that this approach does not yield satisfactory performance for relation prediction. We hypothesize two main reasons for that: (1) although the alternative CLIP features effectively capture object- and pixel-level information, they are less proficient at discerning spatial and scene-level relation [47, 56]; and (2) the diffusion model is trained using long, context-rich descriptions, which may not align well with relation triple phrases [23, 32, 66], resulting in incompatibility with the PSG task.

To overcome these limitations, we introduce an inversion-guided calibration approach for the denoising network, leveraging the inversion process [45, 50, 64], which has been shown to effectively preserve spatial information from input images. Our calibration process aims to adapt a pre-trained diffusion model into a PSG-specific UNet by explicitly injecting spatial and relation knowledge.

Inversion for Denoised Feature Extraction. Inversion techniques [45, 50, 64] have been widely adopted in various applications, such as image editing [22, 52], owing to their ability to preserve the spatial information inherent in the original image. Leveraging this property, we first forward realistic images x (e.g., from the VG [24]) and obtain corresponding deterministic noise representations z through the inversion process. Next, we design a relation prompt of the form p = `[subject] is [predicate] [object]...`, where each placeholder (e.g., `[subject]`) corresponds to a specific entity. Subsequently, we input the deterministic noise z and the prompt p into a pretrained teacher deterministic sampling diffusion model (e.g., BELM [50]), which produces a cross-attention map A_i at the final timestep t to guide the calibration of our denoising network.

Calibration with LoRA. Our backbone network consists of an implicit text encoder and a denoising network UNet ϵ_ϕ. The implicit text encoder is implemented by the CLIP image encoder $\text{CLIP}_{\text{image}}$ with an adapter. Formally, the extracted feature by our backbone can be written as:

$$f_i = \epsilon_\phi(x_i, \text{MLP} \circ \text{CLIP}_{\text{image}}(x_i)), \tag{4}$$

where MLP represents the learnable adapter applied to the CLIP image encoder and x is the input image.

Learning large-scale parameters for the UNet is challenging and could damage its preserved knowledge. Previous works [10, 44, 44] have shown that weight updates often have a low intrinsic rank when adapted for specific tasks. Therefore, we employ a lightweight approach, LoRA [19], to fine-tune our UNet. The core of LoRA is to determine a set of weight modulations $\Delta\mathbf{W}$ added to the original weight $\mathbf{W}$. For example, for the weights $\mathbf{W}_k \in \mathbb{R}^{m_{in} \times m_{out}}$ in each cross-attention layer, we learn two smaller matrices $\mathbf{B} \in \mathbb{R}^{m_{in} \times r}$ and $\mathbf{D} \in \mathbb{R}^{r \times m_{out}}$, where r is the decomposition rank. The LoRA updating is defined as:

$$\mathbf{W}_k = \mathbf{W}_k + \Delta\mathbf{W}_k = \mathbf{W}_k + \mathbf{B} \times \mathbf{D}. \tag{5}$$

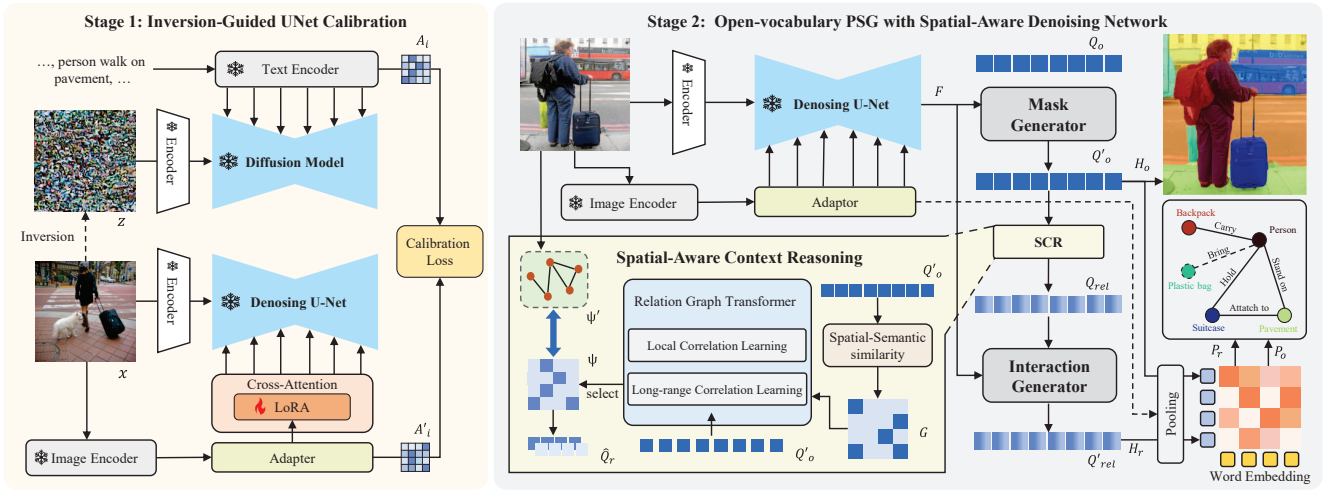

Figure 2. The overview of **SPADE**, which comprises two key steps: (1) inversion-guided calibration for the UNet to adapt a general pre-trained teacher diffusion model into a PSG-specific denoising network and (2) spatial-aware context reasoning (SCR) over relativeness and non-relativeness context through a relation graph Transformer to generate high-quality relation queries.

Similarly, we apply the same approach to learn the other projection layer $\mathbf{W}_v$.

Calibration Loss. At the finetuning, we use the training set of VG [24] and PSG [57] to calibrate our denoising network respectively. Specifically, we feed the realistic images into our denoising network, resulting in the corresponding cross-attention map A_i. Then, we align them with the inversion cross-attention map from the pre-trained diffusion model as:

$$\mathcal{L}_{cal} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda\|A_i - A_i'\|_1\right), \tag{6}$$

where we use the L_1 norm loss on the cross-attention maps and λ is a weighted factor. During the fine-tuning, we only update the MLP adapter of the CLIP' image encoder and the lower rank matrices $\mathbf{B}$ and $\mathbf{D}$ of cross-attention layers.

3.2. Spatial-aware Context Reasoning with RGT

As discussed previously, current open-vocabulary PSG and SGG models primarily focus on designing visual prompts [4, 17, 31, 69] to extract open-world knowledge from large multimodal models for novel relation prediction. However, these approaches often overlook the contextual information of relation pairs, particularly the spatial and semantic correlations among all entities within an image. This oversight limits their capability to handle complex scenes characterized by high spatial relation ambiguity. To address this challenge, we propose a spatial-aware relation graph transformer for enhanced spatial and semantic context reasoning. Our method integrates graph convolutional networks with self-attention mechanisms to capture both long-range and local correlations among subjects and objects, whether they are directly related or not. This integration enables the generation of high-quality relation queries that better capture the underlying spatial and semantic structure of the scene.

Spatial-semantic Graph Construction. The instance decoder Φ_{ins} takes the enhanced feature F from the calibrated UNet ϵ_ϕ and instance queries $Q_o \in \mathbb{R}^{N\times d}$ as input, where N is the number of queries and d is the feature dimension, producing a set of predicted masks $\{m_i\}_{i=1}^{N}$ with corresponding labels. The segmentation process is as follows:

$$\mathbf{H}_o = \Phi_{ins}(\mathbf{Q}_o, F), \tag{7}$$

where $\mathbf{H}_o$ is the output object features. Then, we construct a spatial-semantic graph $G \in \mathbb{R}^{N\times N}$ based on the spatial and semantic distances between instance pairs. The spatial distance is defined as 1 if two masks are neighboring and 0 otherwise. The semantic distance is set to 1 if the cosine similarity between the features of two objects (i.e., $\mathbf{H}_o$) exceeds a threshold and 0 otherwise. The final graph G is formed by summing the two distances; if the resulting value is 2, it is capped at 1.

Relation Graph Transformer. Intuitively, if two objects are connected in G, there is a higher likelihood of a relationship between them thanks to the closer initial distance on the spatial and semantics. To further capture long-range correlations among objects, we propose a spatial-aware Relation Graph Transformer (RGT) that iteratively captures pairwise global long-range and local correlations by aggregating messages from both connected and non-connected objects. The RGT architecture is shown in Figure 3.

Long-range Context Learning. In RGT, we design a long-range correlation reasoning block $\mathrm{RGT}_\mathrm{g}(.)$ to learn global context. Formally, given an object $\mathbf{q}_r \in \mathbf{H}_o$, we calculate

its self-attention with its neighbors as:

$$\underset{\mathcal{P}(r)^+}{\mathrm{RGT}_{\mathrm{g}}}(\mathbf{q}_r) = \mathrm{softmax}\left[\frac{(\mathbf{W}_K\phi[\mathcal{P}(r)^+])^\top(\mathbf{W}_Q\mathbf{q}_r)}{\sqrt{|\mathcal{P}(r)^+|}}\right]\mathbf{W}_V\mathbf{q}_r, \tag{8}$$

where $\mathcal{P}(r)^+$ denotes the neighbors of node r; $\phi[\mathcal{P}(r)^+]$ is a function that averages all neighbors' features of node r; and $\mathbf{W}_K$, $\mathbf{W}_Q$, $\mathbf{W}_V$ are projection layers. Meanwhile, we also calculate self-attention with non-neighbors as:

$$\underset{\mathcal{P}(r)^-}{\mathrm{RGT}_{\mathrm{g}}}(\mathbf{q}_r) = \mathrm{softmax}\left[\frac{(\mathbf{W}_K\phi[\mathcal{P}(r)^-])^\top(\mathbf{W}_Q\mathbf{q}_r)}{\sqrt{|\mathcal{P}(r)^-|}}\right]\mathbf{W}_V\mathbf{q}_r, \tag{9}$$

where $\mathcal{P}(r)^-$ denotes non-neighbors of node r.

The updated node feature $\mathbf{q}_r$ is computed through an element-wise sum to capture information from both connected and non-connected relations:

$$\mathbf{q}_r \leftarrow \mathbf{q}_r + \underset{\mathcal{P}(r)^+}{\mathrm{RGT}}(\mathbf{q}_r) + \underset{\mathcal{P}(r)^-}{\mathrm{RGT}}(\mathbf{q}_r). \tag{10}$$

Next, we use an MLP to fuse all features:

$$\mathbf{q}'_r \leftarrow \mathrm{MLP}[\mathbf{q}_r; \phi[\mathcal{P}(r)^+]; \phi[\mathcal{P}(r)^-]]. \tag{11}$$

Local Context Learning. In addition to global long-range relations, we propose a local relation learning block $\mathrm{RGT}_{\mathrm{l}}(.)$ to capture local correlations between object pairs. This process is defined as:

$$\hat{\mathbf{q}}_r = \mathrm{RGT}_{\mathrm{l}}(\mathbf{q}'_r) = \mathrm{GCN}(G, \mathbf{q}'_r; \mathbf{W}_l), \tag{12}$$

where $\mathrm{GCN}(.)$ is a graph convolution network and $\mathbf{W}_l$ represents learnable parameters.

Pairwise Relation Queries Construction. With the updated object features $\hat{\mathbf{q}}_r \in \hat{\mathbf{Q}}_r$ by the RGT, we then construct the pairwise relation queries for predicate prediction by selecting the close objects. Concretely, given the object feature $\hat{\mathbf{q}}_r$, the constructed relation queries Ψ_r is written as:

$$\Psi_r = \mathrm{Select}_{>}\{\mathrm{Dis}(\hat{\mathbf{q}}_r, \hat{\mathbf{Q}}_r), \eta\}, \tag{13}$$

where $\mathrm{Dis}(.)$ is a distance function, e.g. cosine distance, and $\mathrm{Select}_{>}\{.\}$ is a function to return the objects with a similarity score greater than a threshhold η. Then, we obtain all relation queries $\mathbf{Q}_{rel} \in \mathbb{R}^{M\times d}$, where M is the number of selected queries. At training, we add a auxiliary loss to optimize Ψ_r by:

$$\mathcal{L}_{rqc} = \left|\frac{\hat{\mathbf{Q}}_r\hat{\mathbf{Q}}_r^\top}{|\hat{\mathbf{Q}}_r||\hat{\mathbf{Q}}_r^\top|} - \Psi'\right|^2, \tag{14}$$

where $\Psi' \in \{0,1\}^{N\times N}$ is a ground truth pair indicator matrix.

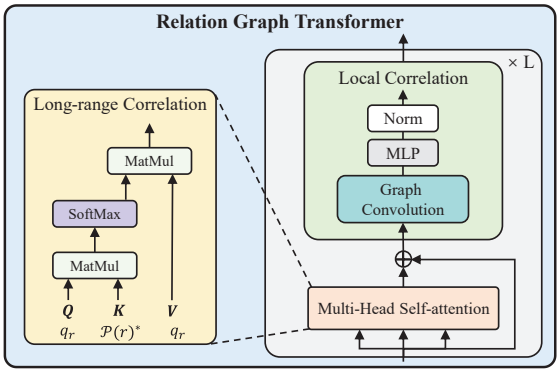

Figure 3. Overview of the proposed Relation Graph Transformer, which consists of long-range and local context reasoning components. It consists of L=8 identical blocks.

3.3. Open-vocabulary Relation Prediction

In this section, we introduce our proposed prompt-based open-world object and relation prediction. We concatenate all selected object pairs by Eq.(13) to form relation queries $\mathbf{Q}_{rel} \in \mathbb{R}^{M\times d}$ for the relation decoder Φ_{rel}. The relation prediction is formulated as:

$$\mathbf{H}_r = \Phi_{rel}(\mathbf{Q}_{rel}, F). \tag{15}$$

Following [4], we leverage a prompt-based classifier to calculate the object category distributions $\mathbf{P}_o$ and relation category distributions $\mathbf{P}_r$ formally,

$$\mathbf{P}_o = \mathrm{softmax}(\mathbf{H}_o^\top \cdot \tau/(\epsilon_t(\mathcal{T}_o))), \tag{16}$$

$$\mathbf{P}_r = \mathrm{softmax}(\mathbf{H}_r^\top \cdot \tau/(\epsilon_t(\mathcal{T}_r))), \tag{17}$$

where $\mathbf{H}_o$ is the output feature by the instance decoder Φ_{ins}, τ is a learnable temperature parameter, and $\mathcal{T}_o$ and $\mathcal{T}_r$ are the corresponding object and predicate category embeddings for the text encoder of CLIP, respectively.

Except for the prediction score from the diffusion model, we further leverage the pre-trained pretrained text-image discriminative visual feature in the first calibration stage to enhance the open-vocabulary object and relation classification. Specifically, for instance prediction, we extract the input image's feature by the implicit captioner $\mathrm{CLIP}_{\mathrm{image}}(x)$ and then compute a masked image feature that falls inside the predicted mask m_i via a pool operation as

$$\mathbf{H}'_o = \mathrm{Pooling}(\mathrm{MLP} \circ \mathrm{CLIP}_{\mathrm{image}}(x), m_i). \tag{18}$$

As for the relation prediction, we first combine the masks of the subject and object into a holistic mask via an element addition operation $\odot$. Then, we calculate the pooled image feature of the pair by:

$$\mathbf{H}'_r = \mathrm{Pooling}(\mathrm{MLP} \circ \mathrm{CLIP}_{\mathrm{image}}(x), m_s \odot m_o). \tag{19}$$

Methods	Close-Set		Open-Set (OvR)	
	R/mR50	R/mR100	R/mR50	R/mR100
Closed Methods				
MOTIF[63]	21.7 / 9.6	22.0 / 9.7	1.6 / 0.6	2.0 / 1.1
VCTree[48]	22.1 / 10.2	22.5 / 10.2	1.7 / 0.9	2.3 / 1.3
GPSNet[33]	19.6 / 7.5	20.1 / 7.7	2.2 / 1.5	2.9 / 1.8
PSGTR[57]	34.4 / 20.8	36.3 / 22.1	3.1 / 1.9	3.3 / 2.0
PSGFormer[57]	19.6 / 17.0	20.1 / 17.6	2.2 / 1.9	3.1 / 2.7
ADtrans[27]	29.6 / 29.7	30.0 / 30.0	2.4 / 2.0	3.6 / 3.4
PairNet[51]	35.6 / 28.5	39.6 / 30.6	3.2 / 2.5	6.5 / 5.1
HiLo[68]	40.7 / 30.3	43.0 / 33.1	3.7 / 2.8	4.8 / 3.6
DSGG[13]	41.6 / 36.8	47.3 / 43.4	3.9 / 3.5	5.6 / 5.0
Open PSG Methods				
OvSGTR†[4]	37.6 / 30.5	41.4 / 28.3	19.3 / 12.4	22.8 / 14.0
PGSG[31]	32.7 / 20.9	33.4 / 22.1	15.5 / 10.1	17.7 / 11.5
OpenPSG†[69]	42.8 / 38.9	49.3 / 47.5	21.2 / 19.8	25.1 / 21.4
SPADE	**45.1 / 41.2**	**54.3 / 51.7**	**26.7 / 23.3**	**31.8 / 25.8**

Table 1. Compared to the state-of-the-art PSG and SGG models on the **PSG** dataset in the close- and open-set scenarios. † represents the reproduced results based on their released code.

Methods	Close-Set		Open-Set (OvR)	
	R/mR50	R/mR100	R/mR50	R/mR100
Closed Methods				
IMP[54]	25.5 / 4.1	30.7 / 7.9	1.1 / 0.5	1.5 / 0.9
MOTIF[63]	32.5 / 6.6	36.8 / 7.9	2.2 / 1.3	2.9 / 1.7
VCTree[48]	31.9 / 6.4	36.0 / 7.3	1.3 / 0.4	2.4 / 1.8
SGTR[30]	24.6 / 12.0	28.4 / 15.2	2.6 / 1.3	3.9 / 2.1
SSR-CNN[49]	23.3 / 17.9	26.5 / 21.4	2.8 / 2.1	5.2 / 4.2
PE-Net[67]	26.5 / 16.7	30.9 / 18.8	2.2 / 1.4	4.3 / 2.6
EGTR[20]	30.2 / 5.5	34.3 / 7.9	2.4 / 0.4	3.6 / 0.8
DSGG[13]	32.9 / 13.0	38.5 / 17.3	3.5 / 1.4	5.2 / 2.3
Open Methods				
VS^3[65]	34.5 / —	39.2 / —	15.6 / 6.7	17.2 / 7.4
OvSGTR†[4]	35.8 / 7.2	41.3 / 8.8	22.9 / 4.6	26.7 / 5.7
PGSG[31]	20.3 / 10.5	23.6 / 12.7	15.8 / 8.2	19.1 / 10.3
OpenPSG†[69]	32.7 / 13.5	39.0 / 18.3	20.4 / 9.4	25.7 / 12.1
SPADE	**37.2 / 17.3**	**43.4 / 21.6**	**24.1 / 11.2**	**29.9 / 13.9**

Table 2. Compared to the state-of-the-art PSG and SGG models on the VG dataset in the close-set and open-set (OvR) scenarios. The second best results are underlined.

Then, we use the same way as Eq. (16) and Eq. (17) to calculate the new object score $\mathbf{P}'_o$ and relation score $\mathbf{P}'_r$, respectively. Finally, we leverage a geometric mean technique [9, 12, 56] to fuse the two scores. For example, the final object prediction score is computed as: $\mathbf{P}^o_{\text{final}} = \mathbf{P}^{\alpha}_o \cdot \mathbf{P}'^{(1-\alpha)}_o$, where α is a fixed weighting factor. Similarly, the enhanced relation prediction score is: $\mathbf{P}^r_{\text{final}} = \mathbf{P}^{\alpha}_r \cdot \mathbf{P}'^{(1-\alpha)}_r$.

Training Loss. In this stage, the overall training loss consists of three terms: mask prediction loss $\mathcal{L}_{maks}$ as [57], relation query construction loss $\mathcal{L}_{rqc}$ and relation Contrastive learning loss $\mathcal{L}_{rel}$ as [56, 58]. The overall loss is written as:

$$\mathcal{L} = \mathcal{L}_{\text{rel}} + \lambda_{\text{rqc}}\mathcal{L}_{rqc} + \lambda_{\text{mask}}L_{\text{mask}}. \tag{20}$$

where λ_{rqc} and λ_{mask} are weighting hyperparameters.

4. Experiments

4.1. Experimental Setup and Datasets

Datasets. We evaluate our model on two datasets: PSG [57] and Visual Genome (VG) [24]. The PSG dataset, derived from COCO, comprises $48,749$ annotated images, with $46,563$ allocated for training and $2,186$ for testing. It includes 80 object categories, 53 "stuff" background categories, and 56 relationship categories. In contrast, the VG dataset—widely used in SGG tasks—contains $108,077$ images with scene graph annotations. For our evaluation, we utilize the VG subset, which focuses on 150 object categories and 50 relation categories.

Evaluation Tasks and Metrics. We adopt the scene graph detection (SGDET) protocol to evaluate all baseline models. We report recall (R) and mean recall (mR) on the test sets of VG and PSG. Additionally, we assess the model across different evaluation settings: closed-set and open-set. For the open-set, following prior work [4, 17, 31, 69], we split the relations and objects into base and novel categories in a 7:3 ratio. Following [4], we report the open-set results on two scenarios: open-vocabulary relation (**OvR**) and open-vocabulary relation and object (**OvD+R**). Notably, for this open-set setting, at the calibration, we only use images whose object and predicate categories arise in the seen group to avoid data leakage problems.

Implementation details. We utilize a pre-trained diffusion model [39] as the diffusion backbone for calibration. Features are extracted from every three U-Net blocks, resized, and structured into a feature pyramid-like FPN. By default, we set the time step for the diffusion process to $t = 0$ for our denoising network. For the mask decoder, we follow the Mask2Former [6] architecture. Notably, for a fair comparison with baseline models, we use the VG and PSG datasets for the calibration respectively, instead of combining them. Both object and interaction decoders are standard three-layer Transformer decoders. The hyperparameter of α, η, λ_{rqc} and λ_{mask} are set to 0.34, 0.65, 0.6 and 1. During training, we apply the same data augmentation strategies as previous methods [6, 51]. The parameters of the diffusion model and CLIP remain frozen throughout. The AdamW optimizer is used with an initial learning rate of 10^{-4} and a weight decay of 10^{-4}. The model is trained for a total of 80 epochs, with the learning rate reduced to 10^{-5} at the 60-th epoch. All experiments are conducted on four A100 GPUs.

4.2. Main Results

Results on PSG. We first present the results on the PSG dataset, as shown in Table 1, for both the closed- and open-

Datasets	Methods	R@50	mR@50	R@100	mR@100
PSG	PGSG[31]	7.3	4.3	11.4	6.1
	OvSGTR†[4]	19.1	13.4	23.6	16.2
	OpenPSG†[69]	11.4	6.7	16.3	7.7
	SPADE	**22.7**	**17.8**	**26.2**	20.1
VG	Cacao[59]	1.2	0.5	3.3	1.4
	VS^3[65]	5.8	2.5	7.2	3.1
	PGSG[31]	5.3	2.8	8.1	4.4
	OpenPSG†[69]	9.5	6.9	14.6	8.9
	OvSGTR[4]	17.1	10.4	21.0	14.8
	SPADE	**19.4**	**13.1**	**24.7**	**16.4**

Table 3. Compared to the state-of-the-art PSG and SGG models on the VG and PSG dataset in the open-set (OvD+R) [4] scenario.

Methods	NDR		DR	
	R@50	mR@50	R@50	mR@50
PGSG[31]	35.6	24.5	28.3	18.1
OvSGTR[4]	40.3	33.4	34.5	26.7
OpenPSG[69]	43.7	39.5	37.1	32.8
SPADE	**46.5**	**42.8**	**42.3**	**38.7**

Table 4. Distant and non-distant relations detection performance of VLM-based PSG models on the PSG dataset.

set settings. For the open-set, we report results at two scenarios [4]: OvR (training with full objects and partial predicates) and OvR+D (training with partial objects and predicates). The results for the OvR+D on the PSG dataset are provided in Table 3. To ensure a fair comparison, we employ the Mask2Former [6] with the Swin-B backbone [35] as the object segmenter across PSG models.

The results indicate that, in the closed-set scenario, our proposed model, SPADE, achieves the highest performance on the PSG dataset. Specifically, SPADE also demonstrates superior performance in the open-set setting. Specifically, it outperforms OpenPSG [69] by approximately 4 points on average in the closed-set scenario and 3 points in the open-set setting. Notably, under the OvR+D scenario, our model achieves a 5.5% improvement in R@50 compared to the competitive OpenPSG model [69]. This significant performance gain highlights SPADE's strong capability in recognizing unseen objects and predicates.

Results on VG. To evaluate our model on the VG dataset, we adapted the mask prediction heads of our instance decoder to box prediction heads, following the approach [31, 69]. The remaining settings are consistent with those used in OvSGTR [4]. We report results for both close-set and open-set settings in Table 2. From the results, it shows that SPADE continues to achieve good performance on the VG dataset, surpassing many competitive open-set SGG models. In the open-set, SPADE demonstrates significant superiority over other models, achieving $\sim$ 8 points higher than OpenPSG in terms of OvD+R (R@100).

Strategies	OvR		OvD+R	
	R@50	R@100	R@50	R@100
w/o inversion	21.0	25.3	15.9	18.9
w/o LoRA	18.8	21.8	12.7	15.8
w/o calibration	15.3	18.9	10.1	12.5
Ours	**26.7**	**31.8**	**22.7**	**26.2**

Table 5. Ablation results with different fine-tuning strategies for calibration on the open-set PSG dataset.

4.3. Ablation Study

In this section, we investigate four crucial parts of **SPADE**: spatial relation prediction, fine-tuning strategy, the relation graph transformer and the open-vocabulary module.

Spatial Relation Prediction Analysis. To assess the spatial relation detection of VLM-based PSG models, we categorize relationships into spatial and non-spatial groups and focus only on spatial relationships in this analysis. We further define two evaluation scenarios: Distant Relation (DR) and Non-Distant Relation Pairs (NDR). DR includes object pairs where the distance between the centers of their bounding boxes exceeds one-third of the image width, while NDRP includes object pairs where this distance is less than one-third. The results are shown in Table 4. To ensure fairness, we use an equal number of DR and NDR samples during training to eliminate potential data bias. From the results, we observe that many LMM-based models perform better in the NDR scenario but struggle in the DR scenario. For instance, OpenPSG shows an 7% performance gap in R@50 between DR and NDR predicates. In contrast, our model achieves more balanced performance across both groups, outperforming prior models in both distant and non-distant relationship prediction. This suggests that existing VLM-based models have difficulty capturing spatial relationships.

Calibration Strategies. We examine different fine-tuning techniques: without inversion sampling (w/o inversion), i.e, random sampling a noise from a Gaussian distribution; fine-tuning all UNet parameters (w/o LoRA); and using the pre-trained UNet directly (w/o calibration). The results are presented in Table 5. From the results, we observe that w/o calibration yields the worst performance, confirming that the knowledge embedded in the pre-trained diffusion model differs significantly from that required for scene graph understanding. This highlights the necessity of our calibration operation. Similarly, w/o LoRA also performs poorly, likely because modifying all parameters of the denoising network disrupts the world knowledge preserved in the pre-trained diffusion model. Additionally, using a random sampling diffusion model does not yield strong performance, probably because the generated image features and spatial

LCNL	LCNNL	LCL	$\mathcal{L}_{rqc}$	R/mR@50	R/mR@100
				30.5 / 26.3	40.2 / 35.7
✓				35.6 / 32.2	46.5 / 43.7
✓	✓			37.1 / 34.4	47.9 / 44.8
✓	✓	✓		40.3 / 35.6	50.8 / 46.5
✓	✓	✓	✓	**45.1 / 41.2**	**54.3 / 51.7**

Table 6. Ablation study results of the Relation Graph Transformer on the close-set PSG dataset.

Diffusion	Pooling	OvR		OvR+D	
		R@50	R@100	R@50	R@100
	✓	12.5	14.6	8.5	9.2
✓		21.4	24.8	12.8	14.7
✓	✓	**26.7**	**31.8**	**22.7**	**26.2**

Table 7. Ablation results with open-vocabulary classification strategies on the PSG dataset in the open-set setting.

layout are inconsistent with the input real image. In contrast, our approach—LoRA with a deterministic sampling model—preserves the pre-trained parameters while learning two low-rank matrices specifically tailored for the PSG task. This provides a more effective fine-tuning strategy, balancing knowledge retention and task-specific adaptation.

Relation Graph Transformer. In this section, we analyze three key components of the relation graph transformer: long-range correlation neighbors learning (LCNL), long-range correlation non-neighbors learning (LCNNL), and local correlation learning (LCL). Additionally, we examine the impact of the auxiliary loss $\mathcal{L}_{rqc}$. Table 6 presents the ablation results, highlighting the contributions of each component. The baseline model, which excludes all components, achieves the lowest performance. Adding LCNL or LCNNL individually leads to noticeable improvements, while incorporating LCL further enhances the results. Notably, the full configuration—integrating LCNL, LCNNL, LCL, and $\mathcal{L}_{rqc}$ —achieves the highest scores. These findings confirm that both long-range (via neighbor and non-neighbor correlations) and local relation modeling, along with auxiliary relation query construction loss, are essential for robust relation prediction.

Open-Vocabulary Module. Table 7 presents the ablation results to assess the impact of incorporating diffusion features and pooled features on open-vocabulary relation prediction. The results indicate that integrating pooled features leads to notable improvements over the baseline, which relies solely on the predictions of the diffusion model. This suggests that diffusion-based representations alone may not be sufficient for predicting open-vocabulary relations and that pooled features provide complementary information that enhances the model's predictive capabilities. These findings highlight the importance of both diffusion-based features and pooled features for effective open-vocabulary relation prediction.

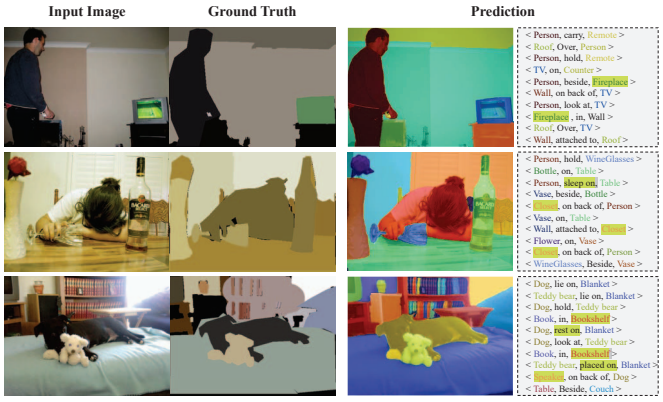

Figure 4. The visualization results of PSG by **SPADE**. The first column shows the input images, while the second column presents the ground truth. The other columns display the predictions by our model, where unseen predicted objects and relationships are highlighted on a yellow green background.

4.4. Visualization Results

In Figure 4, we present visualization results that illustrate our model's open-vocabulary capabilities. These visualizations reveal that our model not only successfully identifies and segments objects that are entirely unseen during training (highlighted in yellow-green) but also adeptly infers novel relationships among them. This indicates that SPADE is capable of generalizing beyond the set of categories encountered during training, effectively capturing both known and unknown object and predicate classes.

4.5. Conclusion

In this paper, we introduced SPADE, a novel spatial-aware diffusion-based framework for open-vocabulary PSG. SPADE addresses the limitations of VLM-based PSG models, particularly their weaknesses in spatial relation reasoning. Our approach consists of two key steps: inversion-guided calibration and spatial-aware context reasoning. First, we fine-tune a pre-trained teacher diffusion model into a PSG-specific denoising network using cross-attention maps from inversion, optimized with a lightweight LoRA-based calibration strategy. Second, we introduce a spatial-aware relation graph transformer that captures both local and long-range contextual dependencies, improving relation query generation. Extensive experiments on benchmark PSG and Visual Genome datasets demonstrate that SPADE achieves state-of-the-art performance in both closed-set and open-set scenarios. For future work, we will extend its applicability to broader multimodal tasks, e.g., human-object interaction detection.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (NSFC) (granted No. 62306064), the Central-Guided Local Science and Technology Development (granted No. 2023ZYD0165), and Sichuan Science and Technology Program (granted No. 2024ZDZX0011). We appreciate all the authors for their fruitful discussions. In addition, thanks are extended to anonymous reviewers for their insightful comments and suggestions.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022. 2

[2] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015. 1

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pages 14455–14465, 2024. 1

[4] Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Chang Wen Chen. Expanding scene graph boundaries: fully open-vocabulary scene graph generation via visual-concept alignment and retention. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024. 1, 4, 5, 6, 7

[5] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2025. 1

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 6, 7

[7] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 2

[8] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023. 1, 3

[9] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 6

[10] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 3

[11] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 41. BMVA Press, 2018. 1

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 6

[13] Zeeshan Hayder and Xuming He. Dsgg: Dense relation transformer for an end-to-end scene graph generation. In *CVPR*, pages 28317–28326, 2024. 1, 6

[14] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. Learning from the scene and borrowing from the rich: tackling the long tail in scene graph generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 587–593, 2021. 1

[15] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuan-Fang Li. Semantic compositional learning for low-shot scene graph generation. *arXiv preprint arXiv:2108.08600*, 2021. 2

[16] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15964–15973. IEEE Computer Society, 2021. 1

[17] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European Conference on Computer Vision*, pages 56–73. Springer, 2022. 1, 4, 6

[18] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Toward a unified transformer-based framework for scene graph generation and human-object interaction detection. *IEEE Transactions on Image Processing*, 32:6274–6288, 2023. 2

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2, 3

[20] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. Egtr: Extracting graph from transformer for scene graph generation. In *CVPR*, pages 24229–24238, 2024. 6

[21] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pages 466–482. Springer, 2021. 2

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2, 3

[23] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *CVPR*, pages 13883–13893, 2024. 3

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 3, 4, 6

[25] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807, 2024. 1

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[27] Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3145–3153, 2024. 6

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 2

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 1

[30] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, pages 19486–19496, 2022. 1, 6

[31] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *CVPR*, pages 28076–28086, 2024. 1, 4, 6, 7

[32] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, pages 7667–7676, 2023. 3

[33] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, pages 3746–3753, 2020. 6

[34] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR (CVPR)*, 2018. 1

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 7

[36] Luis Müller, Mikhail Galkin, Christopher Morris, and Ladislav Rampásek. Attending to graph transformers. *Trans. Mach. Learn. Res.*, 2024, 2024. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 6

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

[41] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics, 2018. 3

[42] Ahsan Shehzad, Feng Xia, Shagufta Abid, Ciyuan Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey. *arXiv preprint arXiv:2407.09777*, 2024. 3

[43] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, 2024. 1

[44] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *Trans. Mach. Learn. Res.*, 2024, 2024. 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3

[46] Renato Sortino, Simone Palazzo, Francesco Rundo, and Concetto Spampinato. Transformer-based image generation from scene graphs. *Computer Vision and Image Understanding*, 233:103721, 2023. 3

[47] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5198–5215. Association for Computational Linguistics, 2022. 3

[48] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 2, 6

[49] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *CVPR*, pages 19437–19446, 2022. 6

[50] Fangyikang Wang, Hubery Yin, Yue-Jiang Dong, Huminhao Zhu, Hanbin Zhao, Hui Qian, Chen Li, et al. Belm: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. *Advances in Neural Information Processing Systems*, 37:46118–46159, 2025. 2, 3

[51] Jinghao Wang, Zhengyu Wen, Xiangtai Li, Zujin Guo, Jingkang Yang, and Ziwei Liu. Pair then relation: Pair-net for panoptic scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 6

[52] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, pages 7677–7689, 2023. 2, 3

[53] Feng Xia, Shuo Yu, Chengfei Liu, Jianxin Li, and Ivan Lee. Chief: Clustering with higher-order motifs in big networks. *IEEE Transactions on Network Science and Engineering*, 9 (3):990–1005, 2021. 3

[54] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 6

[55] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 2

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pages 2955–2966, 2023. 1, 3, 6

[57] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 1, 2, 4, 6

[58] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16954–16964. IEEE, 2024. 3, 6

[59] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *ICCV*, pages 21560–21571, 2023. 1, 2, 7

[60] Xiang Yu, Ruoxin Chen, Jie Li, Jiawei Sun, Shijing Yuan, Huxiao Ji, Xinyu Lu, and Chentao Wu. Zero-shot scene graph generation with knowledge graph completion. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[61] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019. 3

[62] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 1, 2

[63] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 2, 6

[64] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. In *ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[65] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *CVPR*, pages 2915–2924, 2023. 1, 2, 6, 7

[66] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, pages 5729–5739, 2023. 3

[67] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *CVPR (CVPR)*, pages 22783–22792, 2023. 6

[68] Zijian Zhou, Miaojing Shi, and Holger Caesar. Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In *ICCV*, pages 21637–21648, 2023. 1, 6

[69] Zijian Zhou, Zheng Zhu, Holger Caesar, and Miaojing Shi. Openpsg: Open-set panoptic scene graph generation via large multimodal models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part X*, pages 199–215. Springer, 2024. 1, 4, 6, 7