Towards Open-vocabulary Scene Graph Generation with Prompt-based Finetuning

Tao He[1], Lianli Gao[2], Jingkuan Song[2], and Yuan-Fang Li[1†]

[1] Faculty of Information Technology, Monash University,
[2] Center for Future Media, University of Electronic Science and Technology of China
{tao.he, yuanfang.li}@monash.edu, lianli.gao@uestc.edu.cn,
jingkuan.song@gmail.com

Abstract. Scene graph generation (SGG) is a fundamental task aimed at detecting visual relations between objects in an image. The prevailing SGG methods require all object classes to be given in the training set. Such a closed setting limits the practical application of SGG. In this paper, we introduce *open-vocabulary* scene graph generation, a novel, realistic and challenging setting, in which a model is trained on a set of base object classes but is required to infer relations for unseen target object classes. To this end, we propose a two-step method which firstly pre-trains on large amounts of coarse-grained region-caption data and then leverages two prompt-based techniques to finetune the pre-trained model without updating its parameters. Moreover, our method is able to support inference over completely unseen object classes, which existing methods are incapable of handling. On extensive experiments on three benchmark datasets, Visual Genome, GQA and Open-Image, our method significantly outperforms recent, strong SGG methods on the setting of Ov-SGG, as well as on the conventional closed SGG.

Keywords: Open-vocabulary Scene Graph Generation, Visual-language Model Pretraining, Prompt-based Finetuning

1 Introduction

Scene Graph Generation (SGG) [40,53,38,39,49] aims at generating visual relation triples in a given image and is one of the fundamental tasks in computer vision. It has wide applications in a suite of high-level image understanding tasks, including visual captioning [47,48,46], visual question answering [41,1], and 3D scene understanding [2,54].

It was not until recently that the long-tail distribution in SGG datasets was identified [53]. Following this discovery, a number of works [40,6,12,21,43,25] endeavoured to reduce the impact of the biases in data by exploiting debiasing techniques [39,21,18]. Although remarkable improvements in the performance of unbiased SGG have been made, these state-of-the-art methods are limited to predicting relationships between pre-defined object classes only. In real-world

---

† Corresponding author.

scenarios, however, it is highly likely that an SGG model encounters objects of *unseen* categories that do not appear in the training set. In this more realistic and practical setting, the performance of these conventional SGG mod-
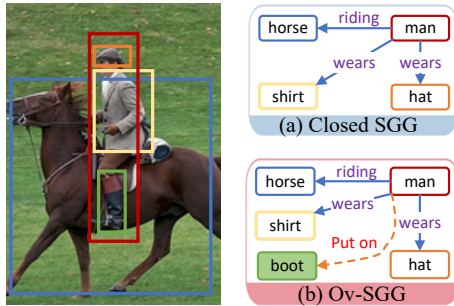

Fig. 1: An illustration of the conventional closed SGG *vs*. Ov-SGG. For the unseen target object `boot`, closed SGG methods such as EBM [38] and Motifs [53], cannot predict any relation regarding `boot` whilst our Ov-SGG method is able to.

els [53,39,18] degrades, and dramatically so when inferring over completely unseen object classes, as can be seen in Tab. 1 in Sec. 5). This motivates us to ask the question, *can we predict visual relationships for unseen objects?* Formally, we call this problem setting **Open-vocabulary Scene Graph Generation (Ov-SGG)**.

In Ov-SGG, the model is trained on objects belonging to a set of seen (i.e. base) object categories $\mathcal{O}^b$ and then predicts relationships on unseen (i.e. target) object categories $\mathcal{O}^t$, both of which are subsets of the open-vocabulary object class set $\mathcal{O} = \mathcal{O}^b \cup \mathcal{O}^t$. This is distinct from, and more challenging than, most existing zero-shot scene graph generation (Zs-SGG) settings [27] or weakly supervised scene graph generation (Ws-SGG) settings [50]. More specifically, Zs-SGG [27] is dedicated to predicting the relationship of objects whose combinations do not emerge in the training set, where the objects are all from seen categories. In contrast, in Ov-SGG, at inference time, not only object combinations may be novel, object categories themselves may not have been seen by the model during training. Zhong *et al.* [58] extends weakly supervised scene graph generation [50] to learning scene graphs from image-caption pairs and demonstrates their model's capability of open-set SGG. In their open-set configuration, the model is trained with a large object set $(4,273$ classes) and a predicate set $(677$ categories), and the object categories for testing are included in the training object set. In other words, both settings assume $\mathcal{O}^t \subseteq \mathcal{O}^b$ should be seen beforehand. In contrast, in our open-vocabulary setting, the model is presented with some unseen object classes at inference time, i.e., $\mathcal{O}^t \setminus \mathcal{O}^b \neq \varnothing$, and in the extreme case, completely unseen object classes, i.e., $\mathcal{O}^t \cap \mathcal{O}^b = \varnothing$. Learning scene graphs for unseen object categories has so far remained unexplored. Fig. 1 illustrates a comparison between Ov-SGG and its closed-world counterpart. Moreover, we propose a more realistic and more challenging setting, in which the test set contains novel relation predicates not seen during training.

We name this task as *general open-vocabulary* SGG, or gOv-SGG. For instance, in Fig. 1, the novel relation "put on" makes this task gOv-SGG.

The main challenge for Ov-SGG is the knowledge gap between the base and target object categories, i.e., how to leverage the learned visual patterns from the limited base categories for the target categories. To bridge this gap, we propose a two-step method of ***visual-relation pre-training*** and ***prompt-based finetuning***.

Firstly, we capitalize on a large number of visual-textual pairs to pre-train a cross-modal model to align the visual concepts with their corresponding unbounded textual descriptions. Different from visual-language models [56,37,33] that are pre-trained on whole images and their captions, we leverage the dense captions of Visual Genome that focus on detailed regional semantics, as we believe that the dense captions can provide more localised relational information. Secondly, we design two prompt-based learning strategies, hard prompt and soft visual-relation prompt. The pre-trained model makes predictions by *filling in the blank* in the designed prompt.

Finetuning has been widely employed to further reduce the knowledge gap between the pre-trained model and downstream tasks [37,57,23] . However, the standard finetuning practice does not lead to promising results on Ov-SGG, as the newly introduced task-specific prediction heads cannot well handle unseen scenarios, as observed recently [26]. Prompt-based learning [7,22,8] has enjoyed remarkable success in a variety of downstream tasks in natural language processing, including relation extraction [7], commonsense knowledge mining [8] and text generation [22]. This is achieved by learning a small amount of parameters for prompt generation without the need to update parameters of the large underlying pre-trained model. As a result, compared to standard finetuning, prompt-based learning suffers less from task interference and thus enjoys better zero-shot learning ability .

Our contributions can be summarized as follows.

- We propose the new, practical and challenging tasks of *open-vocabulary scene graph generation* (**Ov-SGG**), together with a more challenging setting, *general* open-vocabulary SGG (**gOv-SGG**). We believe that Ov-SGG and gOv-SGG represent a firm step towards the real application of SGG.
- We propose a two-step method that firstly ***pre-trains a visual-relation model*** on large amounts of available region-caption pairs aiming for visual-relation alignment with open-vocabulary textual descriptions, and secondly finetunes the pre-trained model by two ***prompt-based finetuning strategies***: hard prompt and soft visual-relation prompt. To our best knowledge, this is the first investigation of prompt-based finetuning for SGG.
- Extensive experiments on three standard SGG datasets, VG [19], GQA [14] and Open-Image [20], demonstrate our model's significant superiority over state-of-the-art SGG methods on the task of Ov-SGG. Moreover, our method is the only one capable of handling the more challenging zero-shot object SGG. Finally, our model also consistently surpasses all compared baselines on the standard closed SGG task.

2   Related Work

Scene graph generation (SGG) [5] aims to detect and localise visual relation between objects. Primary works [36,15] mainly view scene graphs as auxiliary information to improve the quality of image retrieval. Later on, a couple of following arts [47,46,41] demonstrate that scene graphs can be applied to various visual tasks and significantly improve their performance, particularly because scene graphs can provide those models with structured visual representations and facilitate image understanding. With the standard SGG benchmark dataset Visual Genome [19] coming forth for the public, a number of researchers [45,53,55,40,6,38] started put their efforts to SGG. Xu *et al.* [45] leveraged an iterative message passing technique derived from graph convolution network [16] to refine object features and improve the quality of generated scene graphs. Yang *et al.* [46] developed an auto-encoder network by incorporating image-caption into SGG. Zeller *et al.* [53] first pointed out the bias of relation distribution in the VG dataset and revealed that even using the statistic frequency can obtain the comparable SOTA performance at that time. Thus, a suit of subsequent works [39,40,6,25,38] started to work on tackling the tricky bias problem in SGG. However, many researchers ignore a common issue that is the SGG models' generalizability. Although some works [48,3,31] proposed the task of weakly supervised scene graph generation, they typically focus on generating scene graphs without localisaed clues, e.g., from image-caption pairs [48] or VCR [52].

Pretrained visual-language (VL) models have been widely applied for tremendous visual-language tasks [37,57,44,33]. Without of loss generality, we could divide those works into two stages: (1) a cross-modal model is first pretrained by natural supervision, e.g., image-caption pairs, to align visual features with their corresponding textual semantics in the common space by a contrastive learning [33,57] or masked token loss [37,56], and (2) developing a suite of visual-language fusion mechanisms, e.g., concatenation of both features [9], and then finetuning the pretrained VL model for downstream VL tasks, such as image caption [37] and visual dialog [44]. Recently, CLIP [33] used massive image-text data to train two encoders for images and texts solely by a simple contrastive loss and showed promising results on a wide-range of image classification tasks, especially in the zero-shot scenario.

Prompt-based learning [26] has gained extensive attention in natural language processing due to successful applications of large-scale pretrained language models such as Bert [10] and GPT-3 [4]. Early studies [34,29] typically focused on finetuning the pretrained language models to adapt the model to different downstream tasks by training new parameters via task-specific objective functions. With the advent of prompt-based learning [42], following works [7,22,8] turn to prompt engineering by designing appropriate prompts for the pretrained language model without adding task-specific training or modifying the language parameters. Chen *et al.* [7] develops an adaptive prompt finetuning strategy for entity relation extraction. Li *et al.* [22] proposes a lightweight prefix-tuning strategy which enables the mode to fix the parameters of the pretrained

language model during training while learns continuous task-specific vectors for text generation.

3   Problem definition

Scene graph generation (SGG) aims to detect and localize the relationship between objects in an image. *Open-vocabulary* scene graph generation (**Ov-SGG**) is a more challenging problem that works in a more generic scenario, i.e., trained on a set of base object but predicting the relationship for target objects. Formally, let us consider a set of base object categories $\mathcal{O}^b$ where c_k denotes a semantic label of an object. To model the realistic open-world scenario, we assume that there also exists another set of target object classes $\mathcal{O}^t$, where $\mathcal{O}^t \setminus \mathcal{O}^b \neq \emptyset$, i.e., the target set $\mathcal{O}^t$ contains *novel classes* not found in the base set $\mathcal{O}^b$. Moreover, all entities in relation triples in the training set are labeled only by the base object classes in $\mathcal{O}^b$, i.e., $\mathcal{D}^t = \{(x_i, \boldsymbol{y}_i)\}_{i=1}^M$, where M is the number of training images; x_i denotes the i-th image and $\boldsymbol{y}_i$ is the corresponding scene graph that comprises the set of n_i annotated relation triples, i.e., $\boldsymbol{y}_i = \{(s_j, r_j, o_j)\}_{j=1}^{n_i}$ where $s_j, o_j \in \mathcal{O}^b$ are the subject and object respectively; $r_j \in \mathcal{R}$ is the relation predicate; and $\mathcal{Y}$ is the set of all relation predicate words.

The goal of Ov-SGG is to train a model $\mathcal{M}$ on $\mathcal{D}^b$ so that during the inference stage, $\mathcal{M}$ can predict not only relations of $\mathcal{O}^b \times \mathcal{O}^b$ pairs, but also of $\mathcal{O} \times \mathcal{O}$, where $\mathcal{O} = \mathcal{O}^b \cup \mathcal{O}^t$ and $\times$ is the Cartesian product operation. At the same time, a derivative task, zero-shot object SGG (**ZsO-SGG**) requires a model $\mathcal{M}$ trained on $\mathcal{O}^b$ to predict over $\mathcal{O}^t \times \mathcal{O}^t$, where additionally, $\mathcal{O}^t$ contains novel classes only, i.e., $\mathcal{O}^t \cap \mathcal{O}^b = \emptyset$. We also define a more challenging setting, general Ov-SGG (**gOv-SGG**), in which the predication set $\mathcal{R}$ contains novel predicates in the testing stage that are not seen in training. I.e., $\mathcal{R} = \mathcal{R}^b \cup \mathcal{R}^t$ and $\mathcal{R}^t \setminus \mathcal{R}^b \neq \emptyset$.

4   Method

Fig. 2 shows the overview of our Ov-SGG framework, which is based on a pretrained visual-relation model (VRM) by tremendous visual-textual corpus, and finetuned on the base objects in a prompt-based learning manner. More specifically, we first train two transformer encoders for image and text, as in [33], to align the visual concepts with their open-vocabulary relational description. With the pretrained VRM, we devise two prompt-based finetuning strategies to reduce the knowledge gap between VRM and SGG by a fill-in-the-blank paradigm.

4.1   Pretrained context-aware visual-relation model

An intuitive idea to pretrain a visual-relation model is to use the relation triples subject-predicate-object (SPO) comprised of $\mathcal{O}^b$ to train such a visual-relation space. However, this practice could result in heavily overfitting issue due to the small number of $|\mathcal{O}^b|$. Thus, we consider to use unbounded vocabulary relation
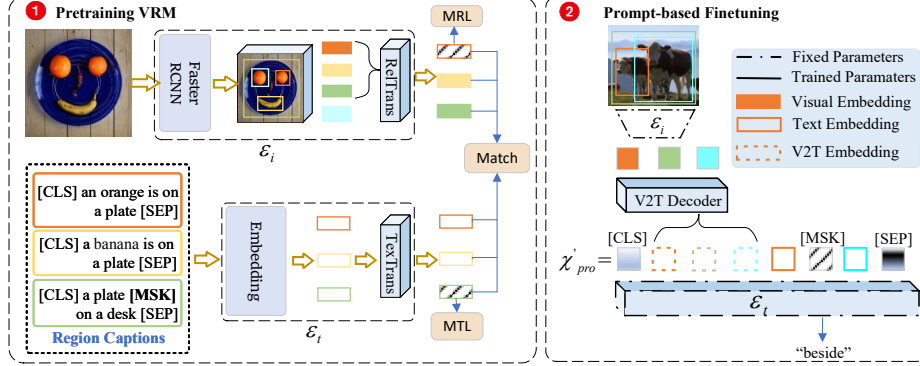
Fig. 2: The overview of our framework for Ov-SGG. **Left**: the architecture of pretraining VRM, which consists of two encoders for regions and dense captions in VG trained by three losses: masked region loss (MRL), match loss and masked token loss (MTL). **Right**: our proposed soft visual-relation prompt for finetuning, which leverages the off-the-shell VRM and optimizes a visual to textual (V2T) decoder network to produce soft prompts.

corpus to train the model. Additionally, the majority of VL models [33,13] totally attempt to align visual concepts with their global textual semantic from image-caption pairs in a fixed scheme, but the identical object in an image could have different relation in an SG, which depends on the other object in a pair. Hence, the pre-trained model should be able to map a visual component into various relation concepts according to its regional context. This ability is not available for the previous pretrained models [33,23,13]. Toward this goal, we propose to train a regional context-aware visual-relation model on the dense-caption of VG by two Transformer-based encoders for images and texts (left of Fig. 2).

Training data. We collect the dense-caption in the training set of Visual Genome (VG) as our pretrained data with ~ 2.6 million region-caption pairs with bounding box information. More specifically, each image on average consists of ~ 50 various dense region descriptions. Note that the dense caption is different from the image caption, e.g., MSCOCO [24] with a general global structured description, while the dense caption generally depicts an object's relation with another object within a region. In VG, the vast majority of region captions describe relationships. For more dataset statistic details, please refer to [19].

Image Encoder. The image encoder ($\mathcal{E}_i$ in Fig. 2) consists of two modules: a region proposal feature extractor (e.g., Faster-RCNN [35]) and a relation Transformer embedding module. The Transformer network takes the region proposals r as input visual tokens. Since our model attempts to encode the relation-specific regional context, we do not feed all regions into the encoder like in previous works [33,37,56]. Instead, we propose a union region based sampling strategy that samples feasible regions as the regional context. Concretely, we first randomly sample two anchor regions as the top left r_t and bottom right r_b region

and then select the other regions $[r_1...r_m]$ that overlap with the union region $\text{Union}(r_t, r_b)$ as the context. Here, we set an IoU threshold to select those regions. Briefly, we formulate the image encoding processing as follows:

$$\mathbf{h} = \text{RelTrans}(\boldsymbol{W}_1[r_t, r_1, \ldots, r_m, r_b] + \boldsymbol{l}) \tag{1}$$

where $\mathbf{h} = [\mathrm{h}_t, \mathrm{h}_1...\mathrm{h}_m, \mathrm{h}_b]$ are the embeddings for each visual token; $\text{RelTrans}(\cdot)$ is the relation transformer module; $\boldsymbol{W}_1$ is a learnable projection matrix; and $\boldsymbol{l}$ is the positional embedding [33] for each token.

Text Encoder. The text encoder ($\mathcal{E}_t$ in Fig. 2) is a parallel Transformer branch to the image encoder, which takes the corresponding region description as the input and produces their embeddings by:

$$\boldsymbol{e}_i = \text{TexTrans}(\boldsymbol{W}_2\,[\texttt{CLS}, w_1, w_2, \ldots, w_k, \texttt{EOF}] + \boldsymbol{l}') \tag{2}$$

where $\boldsymbol{c}_i = [w_1, w_2, \ldots, w_k]$ denotes k words in the dense-caption of region r_i; $\boldsymbol{l}'$ is the position embeddings for each token; and [CLS] and [EOF] are learnable special tokens, denoting the first and last word, respectively. Note that we choose the embedding of [EOF] as the embedding of $\boldsymbol{c}_i$.

Pre-trained Loss Function. The training loss is designed from two main perspectives: image-text matching loss [51] and masked token loss [23]. For the former, we adopt the cosine contrastive loss $\mathcal{L}_\mathrm{c}$ [33] to force the visual region embedding to match the embedding of its corresponding dense-caption. For the latter, regarding to RelTrans, we mask any region in $\mathbf{h}$, i.e., replacing it by the special token [mask], with 15% probabilities and let $\mathrm{h}_{[\mathrm{mask}]}$ be close to the embedding of its ground-truth caption by the contrastive loss, denoted as masked region loss $\mathcal{L}_{\mathrm{MRL}}$. As for the text encoder, we follow the standard masked language models [51,44] to train TexTrans by a cross-entropy loss, which we denote by $\mathcal{L}_{\mathrm{MTL}}$. The total pre-trained loss is defined as:

$$\mathcal{L}_{pre} = \mathcal{L}_{\mathrm{MRL}} + \mathcal{L}_{\mathrm{MTL}} + \mathcal{L}_\mathrm{c} \tag{3}$$

4.2   Prompt-based Finetuning for Ov-SGG

In this section, we introduce our proposed prompted-based finetuning method for Ov-SGG, which exploits the rich visual relationship knowledge in the visual-relational model (VRM) to equip the SGG model with the open-vocabulary capability.

Standard finetuning strategy. It is intuitive to utilise some standard finetuning techniques [37,10] to deploy a task-specific prediction head for relation prediction and update the pretrained VRM. Following this setup, we could naïvely design a finetuning strategy on the VRM.

Let r_{so} denote the union region of subject r_s and object r_o, and $r_1, \ldots, r_m$ are the object proposals overlapping with r_{so}. We feed them into the pretrained image encoder as in Eq.(1) to produce visual embeddings $\mathbf{h}$. Then, we simply

leverage two classifiers to predict the predicate and object label by a cross-entropy loss:

$$\mathcal{L}_f = \mathrm{CE}(\boldsymbol{W}_r \cdot \mathrm{h}_r) + \sum_{h \in \{\mathrm{h}_s, \mathrm{h}_o\}} \mathrm{CE}(\mathbf{W}_c \cdot \mathrm{h}) \tag{4}$$

where h_r is the embedding of the union region of r_s and r_o; $\mathrm{h}_r = \mathrm{LN}(\mathrm{h}_s, \mathrm{h}_{so}, \mathrm{h}_o)$ where $\mathrm{LN}(.)$ is a linear project function; and $\boldsymbol{W}_r$ is the randomly initialized relation classifier. Different from relation classification, we utilize a zero-shot classification setting to predict object labels, that is, using the fixed embeddings of object categories $\mathbf{W}_c$ by our pre-trained text encoder as the object classifier. In this way the model can predict unseen object categories; and CE is the cross-entropy function. At finetuning, all other parameters are updated, as in [37].

However, in practice, we find that the above finetuning strategy does not gain satisfying performance for SGG, particularly in the open-vocabulary scenario. The main reason is updating all parameters could modified the preserved knowledge in the VRM and damage the model's capability of generalization.

Prompt-based finetuning for Ov-SGG. Inspired by the success of prompt-based learning [42,30,22] on large-scale pre-trained language models such as GPT-3 [4], we propose two prompt-based finetuning strategies, hard prompt and soft visual-relation prompt (SVRP), for Ov-SGG based on our pre-trained VRM. A key advantage of our prompt-based tuning strategies is that they allow our pre-trained VRM to be optimized on task-specific data without updating its parameters. Doing so avoids altering the learned (open-vocabulary) knowledge during training, and thus supports predictions over unseen object labels at inference time.

The core of prompt-based learning is the design of *templates* to convert the input sequence, x_{in}, into a textual prompt $\mathcal{X}_{pro}$ with unfilled cloze-style slots. The VRM then makes a prediction by filling the slot with the candidate of the maximum probability in the label space. Without loss of generality, a prompt contains two key parts: a template $\mathcal{T}$ that can been manually designed [42] or learned from the training data [32]; and a label mapping function $\mathcal{M} : \mathcal{Y} \to \mathcal{V}$ that maps a downstream task's labels $\mathcal{Y}$ to the vocabulary $\mathcal{V}$ of a pre-trained model. Note that in this work, $\mathcal{Y} = \mathcal{R}$ is the set of relation labels $\mathcal{R}$ as defined in Sec. 3, while $\mathcal{V} = \{w_i\}_{i=1}^{|\mathcal{V}|}$ is the set of words in the dense captions.
Hard prompt based finetuning. Since relations in SGG are represented as SPO triples, we could readily formulate a relation prompt as follows:

$$\mathcal{X}_{pro} = \mathcal{T}(x_{in}) = \texttt{[CLS]}\, x_{\mathrm{s}}\ \texttt{[MASK]}\, x_{\mathrm{o}}\, \texttt{[SEP]} \tag{5}$$

where x_{s} and x_{o} denote the label of the subject and object, and $\texttt{[MASK]}$ is the slot for candidate predicate labels. Note that the labels of x_{s} and x_{o} are produced in an zero-shot manner describing in the second term of Eq. (4). Then, we could predict the relation label by:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\mathrm{argmax}}\, P\left(f_{\mathrm{fill}}\left(\mathcal{X}_{pro}, \mathcal{M}(y)\right) \mid \mathrm{h}_r; \boldsymbol{\theta}\right) \tag{6}$$

where $f_{\text{fill}}(.)$ is a function that fills the slot `[MASK]` in $\mathcal{X}_{pro}$ with a label word $\mathcal{M}(r) \in \mathcal{V}$, $h_r = \text{LN}(h_s, h_{so}, h_o)$ is defined as in Eq. (4), and $\boldsymbol{\theta}$ are the parameters of the linear projection function LN (updated) and of VRM (frozen). For the prediction score P, we also use Cosine similarity to calculate it:

$$P\left(p \mid h_r, \mathcal{X}_{pro}\right) = \frac{\exp\left(\cos\left(h_r, e_{\text{in}}(p)\right)\right)}{\sum\limits_{q} \exp\left(\cos\left(h_r, e_{\text{in}}(q)\right)\right)} \tag{7}$$

where $p = f_{\text{fill}}(\mathcal{X}_{pro}, \mathcal{M}(p))$ represents a filled prompt, $e_{\text{in}}(p) = \text{TexTrans}(p)$ denotes the textual embedding of p, and q ranges over all filled prompts.

Soft visual-relation prompt (SVRP) based finetuning. The above hard prompt employs a fixed template as shown in Eq. (5), where only object label information is used. In contrast, SVRP learns a *prefix* visual-to-textual vector as the context to complement the hard prompt, that is,

$$\mathcal{X}'_{pro} = \texttt{[CLS]}[x'_s, \ldots, x'_o], x_s \texttt{[MASK]} x_o \texttt{[SEP]} \tag{8}$$

where $[x'_s, \ldots, x'_o]$ denotes the prefix contextual vector.

To this end, we deploy a visual-to-textual decoder network $\mathbb{T}$ to decode the visual cues into the textual context, i.e., $[x'_s, \ldots, x'_o] = \mathbb{T}(\mathbf{h})$, where $\mathbf{h} = [h_s, h_1, \ldots, h_m, h_o]$ is produced by Eq. (1). Thus, we rewrite Eq. (8) as: $\mathcal{X}'_{pro} = \mathbb{T}(\mathcal{X}_{pro} \mid \mathbf{h})$. Similar to Eq. (6), the prediction can be formulated as:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\text{argmax}} \, P\left(f_{\text{fill}}\left(\mathcal{X}'_{pro}, \mathcal{M}(y)\right); \boldsymbol{\theta}'\right) \tag{9}$$

where $\boldsymbol{\theta}'$ are the trainable parameters of $\mathbb{T}$ and the fixed VRM. In this way, we view $\mathcal{X}'_{pro}$ as an input of a masked language model and feed it into the pre-trained TexTrans network to find the token that maximizes the probability of `[MASK]`. Thus, we could rewrite the prediction by:

$$\begin{aligned} P\left(r \mid \mathcal{X}'_{pro}\right) &= P\left(\texttt{[MASK]} = \mathcal{M}(r) \mid \mathcal{X}'_{pro}\right) \\ &= \frac{\exp\left(\cos\left(\mathbf{w}_r, e_{\texttt{[MASK]}}\right)\right)}{\sum\limits_{r' \in \mathcal{R}} \exp\left(\cos\left(\mathbf{w}_{r'}, e_{\texttt{[MASK]}}\right)\right)} \end{aligned} \tag{10}$$

where $\mathbf{w}_r$ is the embedding of r by $\mathcal{E}_t$, and $e_{\texttt{[MASK]}}$ is the output of the token `[MASK]` in $\mathcal{X}'_{pro}$ by $\mathcal{E}_t$. During finetuning, we inject the supervised examples $\{(\boldsymbol{h}, \boldsymbol{y})\}$ to the model and utilise cross-entropy loss to optimize Eq. (10).

5 Experiments

We evaluate our method on the task of SGG in both the new open-vocabulary and conventional closed-set settings on three benchmark datasets: Visual Genome (VG) [19], GQA [14] and Open-Image [20]. Implementation details can be found in the supplementary materials.

5.1   Datasets

Visual Genome (VG) is the mainstream benchmark dataset for SGG. Following previous works [53,39,45], we use the pre-processed VG with 150 object classes and 50 predicates [45]. VG consists of 108k images, of which $57,723$ images are used for training and $26,443$ for testing. Additionally, $5,000$ images make up the validation set.

GQA is a more challenging dataset derived from VG images. It contains $1,704$ object categories and 311 predicate words. Following [17], we have $66,078$ training images, $4,903$ validation images, and $10,055$ test images.

Open-Image(v_6) consists of 301 object categories and 31 predicate categories. Following the split of [11], the training set has $126,368$ images while the validation and test sets contain $1,813$ and $5,322$ images respectively.

5.2   Evaluation Settings

We evaluate model performance in Ov-SGG as well as two other related settings: closed-set SGG and zero-shot object SGG. Before training, we fist randomly split all the object classes into two groups, *base classes* and *target classes*, on each experimental dataset, with 70% objects for the base group whilst the remaining 30% for the target.

Closed-set SGG (Cs-SGG) is the conventional standard SGG evaluation protocol in which a model predicts the relation between base objects only, i.e., $\mathcal{O}^b \times \mathcal{O}^b$. We report the results on the two subtasks: Predicate Classification (PREDCLS) and Scene Graph Classification (SGCLS).

Open-vocabulary SGG (Ov-SGG) aims to evaluate the model's ability to recognize the relationship between open-vocabulary objects, i.e., $\mathcal{O} \times \mathcal{O}$. We discard the third subtask SGDET in this setting, as the object detection network [56] can not handle open-vocabulary object detection.

Zero-shot Object SGG (ZsO-SGG) is different from the previous Zero-shot SGG [39,38], which simply aims to evaluate a model's capability of predicting unseen object pairs, i.e., the combinations of subject and object classes has not appeared in the training set. In contrast, we set the zero-shot configuration on the object level, that is, to predict the predicate between two object classes completely unseen during training, i.e., $\mathcal{O}^t \times \mathcal{O}^t$. Thus, this setting can be regarded as a special case of Ov-SGG.

Metrics. For VG and GQA, we mainly report Recall@K (R@K). As the bias of R@K has been widely acknowledged [38,39], we further report the results on mR@K. For Open-Image, following the settings of [11], we report three metrics: Recall@50, weighed mean Average Precision (wmAP) and mean Average Precision (mAP) of triples.

Baselines. We compare with recent and strong SGG methods: VTransE [55], IMP [45], Motifs [53], VCtree [40], TDE [39], GCA [18], and EBM [38]. Since TDE and EBM are model-free, for a fair comparison, we choose VCTree as their baseline model to apply their techniques.

	Models	Cs-SGG(70%)				Ov-SGG(70%+30%)				ZsO-SGG(30%)			
		PredCls		SGCls		PredCls		SGCls		PredCls		SGCls	
		R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
VG	IMP [45]	46.93	48.12	28.20	28.91	40.02	43.40	23.86	25.71	37.01	39.46	-	-
	Motifs [53]	49.41	50.71	29.65	30.43	41.14	44.70	24.02	27.12	39.53	41.14	-	-
	VCtree [40]	50.13	52.50	30.09	31.02	42.56	45.84	25.24	28.47	41.27	42.52	-	-
	TDE [39]	45.21	46.03	27.14	27.64	38.29	40.38	22.56	24.22	34.15	36.37	-	-
	GCA [18]	51.15	53.38	29.89	32.10	43.48	46.26	25.71	27.40	42.56	43.18	-	-
	EBM [38]	52.81	54.91	31.64	32.95	44.09	46.95	26.03	28.03	43.27	44.03	-	-
Ablations	**SVRP**	**54.39**	**56.42**	**32.75**	**33.87**	**47.62**	**49.94**	**28.40**	**30.13**	**45.75**	**48.39**	**9.30**	**11.32**
	FT-p	48.13	50.05	28.82	29.12	42.13	45.10	24.47	26.64	41.02	44.10	-	-
	FT	51.42	53.51	29.13	30.46	44.73	46.49	26.24	28.02	42.86	46.25	5.16	7.42
	HardPro	46.83	54.12	30.93	32.41	46.34	48.02	26.87	29.41	43.42	47.05	6.61	8.13
	SVRP-d	47.05	54.27	31.05	32.20	46.96	48.29	27.68	29.89	44.31	47.53	7.84	10.25
GQA	IMP [45]	50.73	54.44	17.28	20.16	30.60	34.30	9.61	11.86	24.34	28.85	-	-
	Motifs [53]	51.74	56.10	18.93	21.02	32.24	36.04	12.37	13.40	28.24	31.34	-	-
	VCtree [40]	51.02	55.78	19.31	21.76	32.06	36.86	12.86	13.73	30.83	33.17	-	-
	TDE [39]	48.30	52.69	16.45	18.50	29.15	33.37	10.36	11.42	24.16	26.83	-	-
	GCA [18]	53.83	57.85	20.73	22.84	33.47	37.11	13.13	14.57	32.73	34.24	-	-
	EBM [38]	53.27	58.16	20.01	22.52	33.03	36.82	13.01	14.34	32.10	34.65	-	-
Ablations	**SVRP**	**55.85**	**61.21**	**22.54**	**24.38**	**35.26**	**39.03**	**15.16**	**16.42**	**34.70**	**36.79**	**1.72**	**2.35**
	FT-p	50.22	54.93	17.38	19.11	31.37	35.02	11.22	12.46	27.58	30.13	-	-
	FT	52.14	56.17	19.49	21.63	33.08	36.77	13.35	14.38	29.51	32.42	.19	.80
	HardPro	54.81	58.92	20.10	22.03	34.19	37.05	14.18	14.90	32.64	34.01	1.04	1.43
	SVRP-d	55.13	59.86	20.52	22.75	34.62	37.98	14.75	15.42	33.53	35.24	1.12	1.72

Table 1: The comparison Recall@K results of Visual Genome and GQA datasets with other state-of-the-arts SGG models on the tasks of closed SGG, ov-SGG and ZsO-SGG, where we use 70% object for training and the remaining 30% as the unseen objects for testing. All compared methods use the same backbone of ResNet-50 as in [11]. '-' indicates models incapable of obtaining the result. Note that following [17], we report the results of GQA under the setting of graph unconstraint.

5.3   Results and Analysis

Tab. 1 presents the comparison results on VG and GQA with other state-of-the-arts models on the three tasks: Cs-SGG, Ov-SGG and ZsO-SGG. It is worth noting that for the task of ZsO-SGG, the conventional SGG models are incapable of predicting object labels of unseen classes and thus we do not report their results. For GQA, we follow the setting of [17] and compute recalls without the graph constraint.

Cs-SGG. In this conventional setting, we could observe that our SVRP on both VG and GQA consistently surpasses all the compared baselines, even the recent, SOTA models GCA [18] and EBM [38]. For instance, SVPR on VG gains averagely about 1.55 and 1.02 points of improvements on the task of PredCls and SGCls when compared to EBM. Compared to the other models, e.g., IMP and Motifs, SVPR exceed them by even larger margins, e.g., about 6.32 points better than IMP on average.

Models	Cs-SGG(70%)			Ov-SGG(70%+30%)		
	R@50	mAP	wmAP	R@50	mAP	wmAP
IMP [45]	71.54	34.67	30.01	51.38	24.15	21.62
Motifs [53]	72.05	32.70	29.16	50.47	23.03	20.51
VCtree [40]	72.41	33.08	30.15	52.12	24.29	22.14
TDE [39]	70.52	30.10	31.68	49.50	21.38	18.33
EBM [38]	71.26	34.46	30.25	52.03	24.59	23.31
SVRP	**72.84**	**35.32**	**33.11**	**54.45**	**27.97**	**25.36**

Table 2: Results on Open-Image(v6) in the closed-set and open-vocabulary settings. All methods used the X152FPN network [11] pretrained on Open Images.

Ov-SGG. The middle block shows the results of the new, challenging Ov-SGG setting, in which all models suffer considerable performance drops when compared to Cs-SGG. However, our SVRP technique still obtains the best results, on average 3.91 and 2.71 points higher than GCA in terms of PREDCLS and SG-CLS, respectively. The conventional models such as Motifs and VCTree, which were designed for Cs-SGG, struggle on this task. We posit the main reason to be their reliance on manually-designed external knowledge, e.g., word embeddings from GLOVE [28]. Although beneficial for Cs-SGG, this knowledge damages the model's generalizability for open-vocabulary objects, as they do not see the embeddings of the target object during training, and are prone to overfitting on the base object classes. In contrast, through pre-training on massive dense-caption corpus, our pre-trained VRM directly learns to align the visual and relation knowledge, which can avoid the overfitting problem on the base classes. Additional, our prompt-based mechanisms allow us to finetune VRM without modifying its parameters, which makes it possible to extend the generalizability of VRM to Ov-SGG.

ZsO-SGG. The rightmost block shows the results of ZsO-SGG. Again, our SVRP is superior to all baseline models in PREDCLS on the two datasets. In particular, SVRP on average exceeds the strong EBM and GCA by about 3.42 and 4.20, respectively. Moreover, our method is the only one that is capable of handling SGCls in this setting. Tab. 2 shows the results on Open-Image. Interestingly, we could find that the performance of ZsO-SGG is significantly higher than the conventional Zs-SGG [39], possibly because the majority of predicates in the ZsO-SGG are frequent relationships and this is in favor of better results with the biased metric R@K.

Fully-closed scene graph generation. We further evaluate our technique for fully closed scene graph generation, as shown in Table 3. Specifically, we use all object classes in VG to train following the conventional setting [40]. We discard all the statistical bias prior information for all compared methods. From the results, we could observe that our SVRP surpasses all compared baselines, except for mR@100 on the task of PREDCLS, our SVRP has 0.7 points lower than EBM.

Models	PredCls		SGCls		SGDet	
	R@K	mR@K	R@K	mR@K	R@K	mR@K
IMP [45]	54.1/57.9	9.3/10.2	30.3/32.2	5.8/6.2	19.7/24.6	1.3/7.6
Motifs [53]	60.1/61.5	13.8/15.1	32.1/34.5	7.3/8.0	26.2/28.3	5.2/6.3
VCtree [40]	59.6/60.4	16.7/18.4	33.0/35.7	9.1/10.3	25.9/30.0	8.2/10.1
TDE [39]	50.2/55.8	20.3/24.2	27.2/30.4	10.4/12.5	22.6/25.9	8.6/10.5
GCA [18]	58.9/60.9	22.2/23.3	29.1/33.3	11.4/13.1	-/-	-/-
EBM [38]	59.8/61.7	24.1/**26.0**	32.0/34.6	13.1/14.9	26.8/33.7	9.2/11.6
SVRP	**60.2/62.3**	**24.3**/25.3	**33.9/35.2**	**12.5/15.3**	**31.8/35.8**	**10.5/12.8**
FT-p	56.0/60.5	20.1/23.6	31.7/32.1	10.1/11.4	26.7/32.5	7.3/9.4
FT	58.8/62.2	22.7/24.5	32.4/34.8	11.3/13.5	27.0/33.2	9.0/10.5
HardPro	60.5/63.1	23.2/25.0	32.1/34.5	13.2/14.9	30.2/34.1	9.8/12.0
SVRP-d	60.2/62.7	23.8/25.3	33.5/35.0	14.0/15.2	31.3/34.8	10.2/12.7

Table 3: The results of fully closed scene graph generation on VG and K is set to 50/100. All compared methods use the object detection network from [11] with ResNet-50 as the backbone.

Ablations. We evaluate the effectiveness of our model components in several variants: (1) FT-p uses the pretrained model from [56] instead of our VRM, i.e., without pre-training; (2) FT uses the standard finetuning strategy described in Sec.4.2; (3) HardPro uses hard prompt finetuning described in Sec. 4.2; and (4) SVRP-d removes the decoder network $\mathbb{T}$ for SVRP. The ablation results on the VG and GQA datasets are shown in Tab. 1 and 3, respectively. With the pre-trained VRM removed (FT-p *vs*. FT), we could find that our VRM can bring 2∼3 points improvements, which suggests that simply using the visual-language model [56] for SGG does not bring much benefit, possibly because the visual-language model trained on the global image-caption pairs focuses on the images' global semantics, but ignores the regional semantics. However, those regional cues play an important role in SGG. Furthermore, our two prompt-based finetuning techniques present clear superiority to the standard finetuning strategy, especially in the open-vocabulary scenarios, since the prompt-based strategies directly leverage knowledge preserved in the pre-trained model, which equips the downstream SGG model with the zero-shot capability, whilst the standard finetuning strategy updates the pre-trained model and therefore creates interference between tasks. With respect to the decoder network (SVRP-d), the results confirm solely feeding plain region embeddings to the language model does not generate good prefix contexts for prompts.

gOv-SGG. Additionally, we also provide a baseline for the challenging task of gOv-SGG, in which the model needs to make predictions on novel relation predicates during inference. In this setting, the model sees 70% objects and 70% predicates in the training stage. Specifically, the finetuning predicate set $\mathcal{R}$ in Eq. (10) only account for 70% of all predicates. The model then tries to predict the remaining 30% predicates words on the whole of objects. Since none of the

Models	PredCls				SGCls			
	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100
SVRP	**33.5**	**35.9**	**8.3**	**10.8**	**19.1**	**21.5**	3.2	**4.5**
FT-p	29.4	31.3	5.8	6.7	16.9	18.0	2.3	2.7
HardPro	30.0	32.6	6.9	8.1	17.2	18.6	3.0	4.1
SVRP-d	31.3	33.2	7.6	9.0	18.3	19.7	3.2	4.4

Table 4: The results of gOv-SGG on VG where K is set to 50/100.
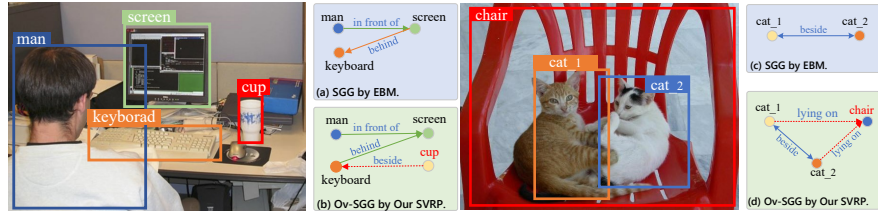
Fig. 3: Qualitative results of our SVRP and EBM [38] on the VG test set for Ov-SGG. EBM is unable to detect relations on the unseen 'cup' and 'chair' classes while our method can.

baseline methods can handle this task, we only report the results of our models and its ablated variants, as shown in Table 4. We could obviously see that our SVPR still achieves the best results over all metrics.

Qualitative analysis. On the task of Ov-SGG, we visualize the scene graphs produced by our technique as well as by the representative closed SGG model EMB [38], as shown in Fig. 3. For the left image, EBM cannot detect any relation regrading the unseen target object "cup", while our SVRP can predict the "beside" relationship between the keyboard and cup. Similarly for the right image, EBM is unable to make predictions about the unseen object class "chair" while our method can.

6   Conclusion

We propose the new practical and challenging open-vocabulary scene graph generation (Ov-SGG) setting and design a two-step method that firstly pre-trains a visual-relation model on large-scale region-caption data. We develop two prompt-based strategies to finetune the visual-relation model for the downstream Ov-SGG task without modifying the pret-rained model parameters. Our extensive experiments on three benchmark datasets show that our method significantly outperforms recent, strong SGG methods on the setting of Ov-SGG, gOv-SGG and the closed SGG. In the future, we hope to integrate open-vocabulary object detection into Ov-SGG.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: CVPR. pp. 2425–2433 (2015)
2. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5664–5673 (2019)
3. Baldassarre, F., Smith, K., Sullivan, J., Azizpour, H.: Explanation-based weakly-supervised learning of visual relations with graph networks. In: European Conference on Computer Vision. pp. 612–630. Springer (2020)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
5. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: Scene graphs: A survey of generations and applications. arXiv preprint arXiv:2104.01111 (2021)
6. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: CVPR. pp. 6163–6171 (2019)
7. Chen, X., Xie, X., Zhang, N., Yan, J., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: Adaprompt: Adaptive prompt-based finetuning for relation extraction. arXiv preprint arXiv:2104.07650 (2021)
8. Davison, J., Feldman, J., Rush, A.M.: Commonsense knowledge mining from pre-trained models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1173–1178 (2019)
9. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV (2021)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
11. Han, X., Yang, J., Hu, H., Zhang, L., Gao, J., Zhang, P.: Image scene graph generation (sgg) benchmark. arXiv preprint arXiv:2107.12604 (2021)
12. He, T., Gao, L., Song, J., Cai, J., Li, Y.F.: Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation. In: IJCAI (2020)
13. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
14. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
15. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
17. Knyazev, B., de Vries, H., Cangea, C., Taylor, G.W., Courville, A., Belilovsky, E.: Graph density-aware losses for novel compositions in scene graph generation. In: BMVC (2020)
18. Knyazev, B., de Vries, H., Cangea, C., Taylor, G.W., Courville, A., Belilovsky, E.: Generative compositional augmentations for scene graph prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15827–15837 (October 2021)

19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017)
20. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
21. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11109–11119 (2021)
22. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL (2021)
23. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
26. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)
27. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV. pp. 852–869. Springer (2016)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
29. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAC-ACL (2018)
30. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? In: EMNLP-IJCNLP (2019)
31. Peyre, J., Sivic, J., Laptev, I., Schmid, C.: Weakly-supervised learning of visual relations. In: Proceedings of the ieee international conference on computer vision. pp. 5179–5188 (2017)
32. Qin, G., Eisner, J.: Learning how to ask: Querying lms with mixtures of soft prompts. In: ACL (2021)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
34. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
36. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the fourth workshop on vision and language. pp. 70–80 (2015)
37. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: ICLR (2020)

38. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13936–13945 (2021)
39. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR (2020)
40. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6619–6628 (2019)
41. Teney, D., Liu, L., van Den Hengel, A.: Graph-structured representations for visual question answering. In: CVPR. pp. 1–9 (2017)
42. Trinh, T.H., Le, Q.V.: A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847 (2018)
43. Wang, T.J.J., Pehlivan, S., Laaksonen, J.: Tackling the unannotated: Scene graph generation with bias-reduced models. In: BMVC (2020)
44. Wang, Y., Joty, S., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.: Vd-bert: A unified vision and dialog transformer with bert. In: EMNLP (2020)
45. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5419 (2017)
46. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR. pp. 10685–10694 (2019)
47. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV). pp. 684–699 (2018)
48. Ye, K., Kovashka, A.: Linguistic structures as weak supervision for visual scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8289–8299 (2021)
49. Zareian, A., Karaman, S.: Bridging knowledge graphs to generate scene graphs. In: ECCV (2020)
50. Zareian, A., Karaman, S., Chang, S.F.: Weakly supervised visual semantic parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3736–3745 (2020)
51. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
52. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
53. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR. pp. 5831–5840 (2018)
54. Zhang, C., Yu, J., Song, Y., Cai, W.: Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9705–9715 (2021)
55. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5532–5540 (2017)
56. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR. pp. 5579–5588 (2021)

57. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)
58. Zhong, Y., Shi, J., Yang, J., Xu, C., Li, Y.: Learning to generate scene graph from natural language supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1823–1834 (2021)