Relation-aware Hierarchical Prompt for Open-vocabulary Scene Graph Generation

Tao Liu[1], **Rongjie Li**[1], **Chongyu Wang**[1], **Xuming He**[1,2*]

[1]ShanghaiTech University, Shanghai, China
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
liutao2023@shanghaitech.edu.cn, bugboy56@gmail.com, {wangchy, hexm}@shanghaitech.edu.cn

Abstract

Open-vocabulary Scene Graph Generation (OV-SGG) overcomes the limitations of the closed-set assumption by aligning visual relationship representations with open-vocabulary textual representations. This enables the identification of novel visual relationships, making it applicable to real-world scenarios with diverse relationships. However, existing OV-SGG methods are constrained by fixed text representations, limiting diversity and accuracy in image-text alignment. To address these challenges, we propose the Relation-Aware Hierarchical Prompting (RAHP) framework, which enhances text representation by integrating subject-object and region-specific relation information. Our approach utilizes entity clustering to address the complexity of relation triplet categories, enabling the effective integration of subject-object information. Additionally, we utilize a large language model (LLM) to generate detailed region-aware prompts, capturing fine-grained visual interactions and improving alignment between visual and textual modalities. RAHP also introduces a dynamic selection mechanism within Vision-Language Models (VLMs), which adaptively selects relevant text prompts based on the visual content, reducing noise from irrelevant prompts. Extensive experiments on the Visual Genome and Open Images v6 datasets demonstrate that our framework consistently achieves state-of-the-art performance, demonstrating its effectiveness in addressing the challenges of open-vocabulary scene graph generation.

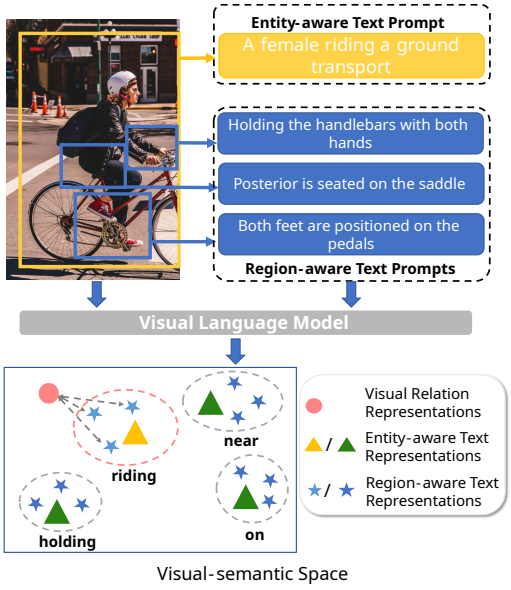

Figure 1: An illustration of RAHP for OV-SGG. RAHP generates entity-aware and region-aware hierarchical prompts to enrich the text representations of the relation, thereby enhancing OV-SGG.

1 Introduction

Scene Graph Generation (SGG) (Johnson et al. 2015; Zellers et al. 2018) is a fundamental task in computer vision, involving the construction of a structured representation of a scene by identifying the relations between entities depicted in an image. It has already demonstrated promising performance in various downstream tasks (Kamath et al. 2021; Lee et al. 2019; Chen et al. 2020; Li et al. 2021). Traditional SGG methods typically operate within a closed vocabulary, and due to the diversity of relational concepts that exceed existing data annotations, they face challenges in effectively modeling open-set relations. To address this challenge, Open-Vocabulary Scene Graph Generation (OV-SGG) (He et al. 2022; Zhang et al. 2023; Yu et al. 2023) has emerged as an active research area recently.

Previous studies (Yu et al. 2023; Liao et al. 2022; Chen et al. 2023) mainly leverage the image-text matching capabilities of pre-trained Vision-Language Models (VLMs) (Radford et al. 2021) to achieve open-vocabulary relation detection based on the similarity scores between relation features and text representations. However, these methods typically rely on a single, fixed form of text representation, which limits the diversity and accuracy of image-text matching, particularly in predicting novel relations. To address this, a promising strategy (Gao et al. 2023; Li et al. 2024b; Menon and Vondrick 2022) is to use generative category descriptions to expand the text representation space, thereby enhancing the flexibility and precision of image-text matching. Nonetheless, generating informative relation descriptions requires not only encoding triplet information (Li et al. 2024a), i.e.,

*Corresponding author

<subject, predicate, object>, but also capturing fine-grained interactions corresponding to different image regions. Given the quadratic growth in triplet combinations with increasing subjects and objects, incorporating all triplet information into representations becomes impractical. Moreover, previous methods often employ all the generated descriptions for matching with the entities or relations in the image, which usually includes many irrelevant descriptions for the input image. This introduces a significant amount of noise in text representation, thereby reducing prediction accuracy.

To address the above challenges, we introduce the Relation-Aware Hierarchical Prompting framework (RAHP), which integrates subject-object and regional relation information within a relational representation space. We focus on enhancing the textual representation in the visual-semantic space of VLMs. As shown in Fig. 1, we extend the range of relational text representations in the space, enabling the text representations to pair more effectively with visual representations. In open-vocabulary tasks, this approach can significantly enhance their consistency with visual representations and improve the model's generalization.

Specifically, RAHP contains a hierarchical prompt generation module that reduces triplet category space through entity clustering, lowering the complexity of encoding triplet information. This module also uses a large language model (LLM) to identify key regions for both subjects and objects, generating fine-grained region descriptions as text prompts by combining these regions. Those entity-aware and region-aware text representations more effectively captures contextual information in visual data, enhancing the model's understanding and generalization in complex interactive scenes. Additionally, RAHP implements a VLM-based dynamic selection mechanism that filters out completely irrelevant text representations based on visual concept, thereby improving matching accuracy.

We conduct extensive experiments to validate our approach on two SGG benchmarks: Visual Genome (Krishna et al. 2017) and Open Images-v6 (Kuznetsova et al. 2020) datasets, and it achieves state-of-the-art performance.

The main contribution of our work has three folds.

- We propose a relation-aware hierarchical prompting framework (RAHP) for OV-SGG that integrates entity-aware and region-aware text prompts, enhancing text representations and model generalization.
- We introduce a VLM-guided dynamic selection mechanism that adapts text prompts based on visual information, minimizing irrelevant content and enhancing the robustness of relation predictions.
- Experiments on two benchmark datasets demonstrate that our method achieves state-of-the-art generalization performance in OV-SGG.

2 Related Work

2.1 Scene Graph Generation

The Scene Graph Generation (SGG) task, initially proposed by (Johnson et al. 2015), traditionally relies on supervised methods (Xu et al. 2017; Gu et al. 2019; Tang et al. 2019; Zellers et al. 2018) that predict relationships using visual, spatial, and contextual cues. To reduce reliance on annotated scene graphs, some approaches (Suhail et al. 2021; Yang et al. 2019) use language supervision, extracting entity and relationship labels from image captions (Li et al. 2022; Shi et al. 2021; Zhong et al. 2021). However, these methods often use closed-set classifiers, limiting their ability to handle novel entities or relations. Recent studies have expanded SGG to open-vocabulary settings (He et al. 2022; Zhang et al. 2023; Yu et al. 2023). For example, SVPR (He et al. 2022) uses dense caption pre-training and prompt fine-tuning, while VS^3 (Zhang et al. 2023) aligns visual features with a pre-trained visual-semantic space for predicting new entities. Epic (Yu et al. 2023) introduces cross-modal entanglement, combining text and region embeddings to classify new predicates. OvSGTR (Chen et al. 2023) extends OV-SGG to open-vocabulary detection and relations-based scenarios.

Most of these methods rely on visual-text matching to classify novel relations, using fixed-form text prompts that limit recognition of novel relations in OV-SGG. To address this, we propose a hierarchical text representation enhancement method that enriches the text representation space by introducing text prompts at both subject-object and region levels, improving relationship recognition.

2.2 Open-vocabulary Methods

In recent years, researchers in the field of visual scene understanding, such as object detection (Gu et al. 2021; Zareian et al. 2021), have shifted their focus from traditional closed-set methods to more flexible open-vocabulary methods. A key driver of this evolution is the development and maturity of VLMs (Radford et al. 2021; Jia et al. 2021; Li et al. 2023). These models are typically pre-trained on large-scale image-text pairs, endowing them with strong cross-modal alignment capabilities. By leveraging natural language prompts (Wu et al. 2024), VLMs can compute similarities between images and language in open-vocabulary settings, facilitating category expansion (Gu et al. 2021; Ma et al. 2022).

Early research (Wang et al. 2023; Menon and Vondrick 2022) concentrates on simple prompts for open-vocabulary recognition. Other approaches (Wang et al. 2024, 2022) employ learnable prompts to enhance image classification. As research progresses, single prompts cannot adequately handle complex visual inputs, leading to the proposal of hierarchical prompting methods to better structure intricate query information. Models like (Ge et al. 2023) rely on object class hierarchies by WordNet (Miller 1995). RECODE (Li et al. 2024b) utilizes LLMs to generate hierarchical prompts from the perspectives of subject, object, and spatial levels, facilitating zero-shot relationship recognition. In contrast to RECODE, our work approaches the task from a regional perspective, enabling the generation of more detailed and specific relationship prompts. Additionally, we introduce a dynamic VLM-guided mechanism that adjusts prompts based on visual inputs, increasing the accuracy and flexibility of text representations.

3 Preliminary

3.1 Problem Setting

The goal of SGG is to create a descriptive graph $\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$ from an image I. This graph consists of N^v entities $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{N^v}$ and visual relationship triples $\mathcal{E} = \{\mathbf{v}_i, r_{i,j}, \mathbf{v}_j\}_{i \neq j}$, where $r_{i,j}$ represents the predicate category between them. Each entity $\mathbf{v}_i$ is represented as $(c_i^v, \mathbf{b}_i)$, where c_i^v denotes the label in the entity category space $\mathcal{O}^e$, and $\mathbf{b}_i$ represents its location through a bounding box in the image. The predicate category $r_{i,j}$ denotes the label in the category space $\mathcal{O}^r$. In the task of OV-SGG, the category spaces for entities and predicates are divided into two parts. Specifically, the predicate category space contains the base category space $\mathcal{O}_b^r$ and the novel category space $\mathcal{O}_n^r$, and it has $\mathcal{O}^r = \mathcal{O}_b^r \cup \mathcal{O}_n^r$. Similarly, the entity category space also has $\mathcal{O}^e = \mathcal{O}_b^e \cup \mathcal{O}_n^e$.

3.2 OV-SGG Pipeline

Most OV-SGG methods (Yu et al. 2023; Chen et al. 2023) can typically be decoupled into two steps: relationship proposal generation and predicate classification.

First, the model receives an image as input and feeds it into a proposal network, from which it extracts relationship proposals $\mathcal{P} = \{\mathbf{v}_i, \mathbf{v}_j\}_{i \neq j}$ and the corresponding relationship features $\mathbf{R} \in \mathbb{R}^{N \times d}$, where N is the number of relationship proposals and d is the dimension of the feature representation.

Then, the relationship features are fed into the predicate classifier as visual representations. The predicate classifier usually handles each predicate class using predefined text prompts, which generate text embeddings $\mathbf{T} \in \mathbb{R}^{\mathbb{C}_p \times d}$ through the text encoder $\mathrm{TextEnc}$ of a VLM, where $\mathbb{C}_p$ is the number of predicate categories. These text embeddings as the text representations replace the fixed predicate classifier weights, enabling the model to extend to new relationship categories that appear during the testing phase. The predicate classifier obtains the predicate classification scores $\mathbf{S} \in \mathbb{R}^{N \times \mathbb{C}_p}$ for each relationship proposal by calculating the similarity score between $\mathbf{R}$ and $\mathbf{T}$:

$$\mathbf{S} = \phi(\mathbf{R}, \mathbf{T}) = \frac{\mathbf{R} \cdot \mathbf{T}}{|\mathbf{R}| \cdot |\mathbf{T}|}, \tag{1}$$

where $\cdot$ is the dot product, we define this operation of calculating similarity as $\phi()$. During OV-SGG training, OV-SGG methods use a distillation loss to distill the knowledge of the VLM to maintain the model's generalization. The distillation loss ensures that the distance between the text embeddings and relationship features remains consistent across all pairwise classifications.

4 Method

4.1 Method Overview

We propose RAHP, a method that enhances the generalization of OV-SGG models on novel relations by using multi-level text prompts to strengthen visual relation text representations. Specifically, our framework is composed of three modules: hierarchical prompt generation (Sec. 4.2), visual relationship extraction (Sec. 4.3), and hierarchical relationship prediction (Sec. 4.4). Finally, we introduce the learning and inference pipeline of our method (Sec. 4.5).

4.2 Hierarchical Prompt Generation

The hierarchical prompt generation module enriches text representations by creating multi-level text prompts that include entity-aware and region-aware prompts. As shown in Fig. 2 (a), for the input vocabularies, we sequentially generate prompts at two levels.

- **Entity-aware text prompts**: These prompts include precise relationship content by combining predicate, subject, and object details. However, as the number of triplets grows cubically with subjects and objects, incorporating all triplet information into the prompts becomes impractical. To address this, we first cluster entities into super entities based on similarity. Similar to the approach in (Zhang et al. 2024), it can effectively reduce the triplet category space (more details can be found in the appendix). We then generate entity-aware text prompts by combining super entities with predicate categories.
- **Region-aware text prompts**: Building on entity-aware prompts, we create region-aware text prompts that capture finer visual details through a region-aware description mining strategy. As shown in Fig. 2 (b), we use an LLM to decompose key entity parts and naturally generate region-level visual relation descriptions by combining these parts' relationships. *For example, in the relationship triplet <male, sitting on seating, furniture> the "male" can be associated with specific body parts like the hips, thighs, and arms, while the "seating furniture" can be associated with components like the seat and backrest. The "sitting on" relationship is then represented by combining these elements in the LLM to provide extensive region-aware relation descriptions.* Following (Menon and Vondrick 2022), we design two cases for the LLM to learn from.

Hierarchical Prompt Encoding After generating the two levels of prompts, we encode them into text embeddings as text representations using the frozen VLM text encoder $\mathrm{TextEnc}$. As shown in Fig. 2 (a), we generate sentences for the entity-aware prompts through the template *"A photo of a/an [Subject] [Predicate] a/an [Object]"*. The entity-aware prompts are encoded by $\mathrm{TextEnc}$ into an entity-aware text embedding set $\mathcal{T}^e = \{\mathbf{T}_1^e, \mathbf{T}_2^e, ..., \mathbf{T}_{\mathbb{C}_{se}^2}^e\}$, where $\mathbf{T}^e \in \mathbb{R}^{\mathbb{C}_p \times d}$, $\mathbb{C}_p$ represents the number of predicate categories, and $\mathbb{C}_{se}$ denotes the number of super entity categories. Correspondingly, the region-aware prompts are generated sentences through the template *"A region that reflects [region descriptions]"*. The region-aware prompts are encoded into an text embedding set $\mathcal{T}^r = \{\mathbf{T}_1^r, \mathbf{T}_2^r, ..., \mathbf{T}_{\mathbb{C}_{se}^2}^r\}$, where $\mathbf{T}_j^r \in \mathbb{R}^{\mathbb{C}_p \times N_j^r \times d}$, N_j^r is the number of region-aware prompts, varying with the region descriptions per triplet.

4.3 Visual Relation Extraction

Following the process described in Sec. 3.2, the visual relationship extraction module is mainly designed to extract visual relation features. It employs a proposal network to extract relation proposals $\mathcal{P} = \{\mathbf{v}_i, \mathbf{v}_j\}_{i \neq j}$ from the visual input I, along with their corresponding relation feature representations $\mathbf{R}$. Then we merge the predicted subject and

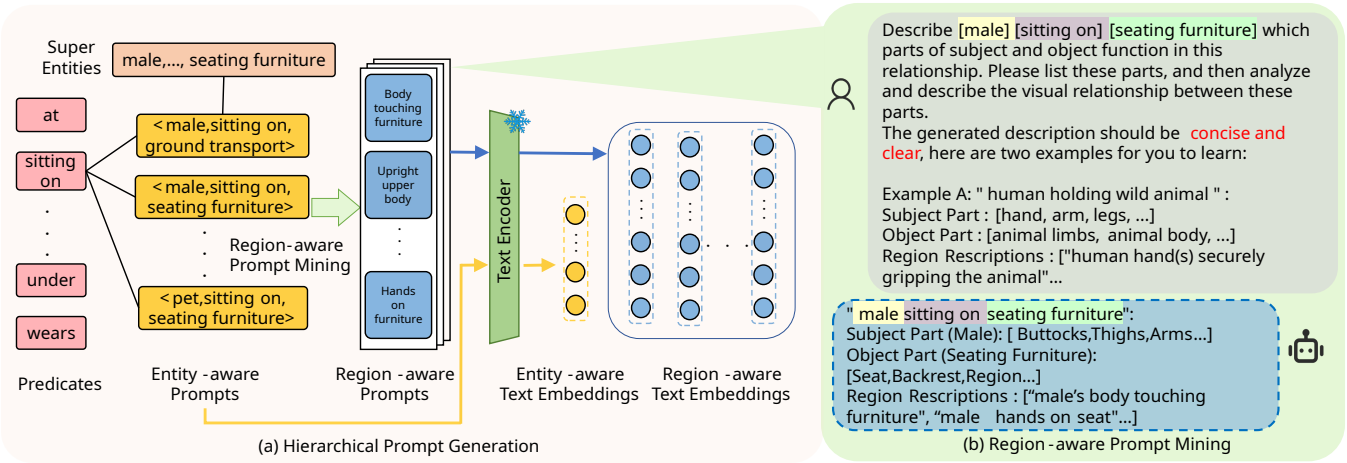

Figure 2: An overview of hierarchical prompt generation: (a) Predicates combine with super entities to create entity-aware prompts, which then expand into region-aware prompts. This process builds a rich textual representation space for extended relational triplets. (b) In region-aware prompt mining, our approach guides an LLM to decompose subjects and objects into distinct parts, enabling more detailed regional visual relationship descriptions.

object boxes in $\mathcal{P}$ into union boxes, and crop the corresponding region I^u from the image I. I^u is encoded into a unified feature $\mathbf{U} \in \mathbb{R}^{N \times d}$ by the VLM's visual encoder VisEnc, calculated as follows:

$$\mathbf{U} = \text{VisEnc}(I^u). \tag{2}$$

Both the relation features $\mathbf{R}$ and the union features $\mathbf{U}$ are input into the hierarchical relation prediction module.

4.4 Hierarchical Relation Prediction

The hierarchical relation prediction module predicts predicates by calculating the similarity between relation features $\mathbf{R}$ and two levels of text embeddings $\mathbf{T}$: entity-aware and region-aware. This module includes two key components: VLM-guided dynamic selection, which filters out irrelevant prompts, and hierarchical prediction aggregation, ensuring accurate predicate classification.

Image-guide Dynamic Selection The VLM-guided dynamic selection mechanism utilizes the image-text alignment capabilities of a VLM to match $\mathcal{T}^r$ with union features $\mathbf{U}$. The mechanism is aimed at filtering out region-text pairs that are completely irrelevant to the image, leveraging the robust object recognition capabilities of the VLM to achieve this goal. Specifically, for j^{th} in $\mathcal{T}^r$, upon receiving the unified feature $\mathbf{U}$, it computes the matching score $\mathbf{S}_j^{se} \in \mathbb{R}^{N \times N_j^r}$ between $\mathbf{U}$ and the region-aware text embeddings $\mathbf{T}_j^r$ as follows:

$$\mathbf{S}_j^{se} = \phi(\mathbf{U}, \mathbf{T}_j^r) \tag{3}$$

To capture core visual semantic information, we select the top k region-aware text embeddings with the highest matching scores and perform predicate classification. After performing VLM-guided selection on all region-aware prompts, we obtain the final region-aware text prompt embedding set $\mathcal{T}^{r\prime} = \{\mathbf{T}_1^{r\prime}, \mathbf{T}_2^{r\prime}, ..., \mathbf{T}^{r\prime}_{\mathbb{C}^2_{se}}\}$, where $\mathbf{T}^{r\prime} \in \mathbb{R}^{\mathbb{C}_p \times k \times d}$. This mechanism dynamically selects text prompts based on union features, prioritizing region-aware prompts with higher probabilities for subsequent predicate prediction, effectively reducing noise.

Hierarchical Prediction Aggregation After selecting region-aware text embeddings, we predict the final predicate scores by integrating entity-aware and region-aware embeddings. First, we calculate the similarity between $\mathbf{T}^e$ and $\mathbf{R}$ to derive the entity-aware predicate score $\mathcal{S}^e = \{\mathbf{S}_1^e, \mathbf{S}_2^e, ..., \mathbf{S}^e_{\mathbb{C}^2_{se}}\}$ for $\mathcal{T}^e$:

$$\mathbf{S}_j^e = \phi(\mathbf{R}, \mathbf{T}_j^e), \tag{4}$$

where $\mathbf{S}_j^e \in \mathbb{R}^{N \times \mathbb{C}_p}$. This score encapsulates the relation details between specific entity pairs. Next, at the region-aware level, we compute the region-aware predicate scores $\mathcal{S}^r = \{\mathbf{S}_1^r, \mathbf{S}_2^r, ..., \mathbf{S}^r_{\mathbb{C}^2_{se}}\}$ by evaluating the similarity between $\mathbf{T}_j^{r\prime}$ and $\mathbf{R}$:

$$\mathbf{S}_j^r = \frac{\sum_{m=1}^{k} \phi(\mathbf{R}, \mathbf{T}^{r\prime}_{j,m})}{k}, \tag{5}$$

where $\mathbf{S}_j^r \in \mathbb{R}^{N \times \mathbb{C}_p}$. The score emphasizes region relation features, providing additional text information to assist the model in understanding visual relationships. We then combine $\mathbf{S}_j^e$ and $\mathbf{S}_j^r$ to $\mathbf{S}_j^a \in \mathbb{R}^{N \times \mathbb{C}_p}$ using a weighted sum to produce the aggregated score:

$$\mathbf{S}_j^a = (1 - \alpha) \times \mathbf{S}_j^e + \alpha \times \mathbf{S}_j^r. \tag{6}$$

Finally, we select the highest scores from the $\mathbb{C}^2_{se}$ aggregated scores $\mathbf{S}^a$ as the final predicate prediction scores $\mathbf{S} \in \mathbb{R}^{N \times \mathbb{C}_p}$:

$$\mathbf{S} = \max(\mathbf{S}_1^a, \mathbf{S}_2^a, ..., \mathbf{S}^a_{\mathbb{C}^2_{se}}). \tag{7}$$

The prediction scores allow us to derive the probability for each predicate, enabling the determination of the predicate category. This multi-level prediction mechanism enhances RAHP by learning regional-level text representations, improving open-vocabulary capabilities, and enabling knowledge transfer to new relationship concepts.

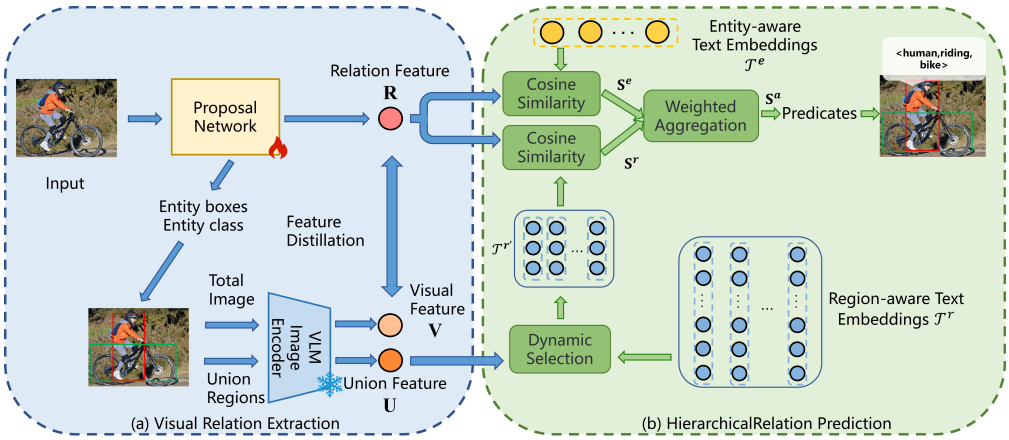

Figure 3: An overview of RAHP. (a) Visual Relation Extraction Module: The process begins with extracting relation proposals and their features from the image, which are then encoded into visual features using a VLM. (b) Hierarchical Relation Prediction Module: The visual features undergo a guided selection process, where the selected embeddings are combined with entity-aware embeddings to predict predicates.

4.5 SGG Learning and Inference

SGG Learning During the training stage, the model only receives information from the base classes. Similar to (Li, Zhang, and He 2024; Chen et al. 2023), we adopt a multi-task loss for our model training. Specifically, we use L1 loss and GIOU loss for entity bounding box regression to reduce the gap between the predicted bounding box $\mathbf{b}$ and the ground truth $\mathbf{b}_{gt}$:

$$\mathcal{L}_{bbox} = \|\mathbf{b} - \mathbf{b}_{gt}\|_1 + \mathrm{GIOU}(\mathbf{b}, \mathbf{b}_{gt}). \tag{8}$$

We also use a cross-entropy loss $\mathcal{L}_{ent} = \mathrm{CE}(c^v, c^v_{gt})$ to ensure the accuracy of the prediction c^v for entity classification against the ground truth category c^v_{gt}.

For predicate prediction, we use $\mathcal{L}_{pre} = \mathrm{FL}(r, r_{gt})$ to represent the Focal loss for predicate categories, where r_{gt} is the ground truth predicate category, and r is the predicted predicate category. In addition, we employ an L1 loss (Liao et al. 2022; Chen et al. 2023) to minimize the gap between the relation feature $\mathbf{R}$ and the visual features $\mathbf{V} \in \mathbb{R}^d$ extracted by the VLM visual encoder VisEnc. The goal is to align the relation features extracted by SGG with the VLM space, thereby enabling the prediction of novel predicates. It also acts as a form of regularization to prevent overfitting to the specific training data. For the i-th relation proposal, the distillation loss is designed as an L1 distance loss, defined as follows:

$$\mathcal{L}^d_i = \|\mathbf{R}_i - \mathbf{V}\|_1 . \tag{9}$$

The total training loss can be written as

$$\mathcal{L} = \mathcal{L}_{bbox} + \lambda_1 \mathcal{L}_{ent} + \lambda_2 \mathcal{L}_{pre} + \lambda_3 \mathcal{L}^d. \tag{10}$$

where the weights of each loss term $\lambda_1, \lambda_2, \lambda_3$ balance the learning progress and importance across different tasks.

SGG Inference To enhance the interpretability of novel relation triplets, we employ LLMs to generate informative visual descriptions before the inference phase. In the post-processing stage, we systematically eliminate invalid self-connected edges and exclude triplets where subject and object entities are identical. Subsequently, the remaining triplets are ranked based on the combined scores from entity predictions and predicate predictions. The top M relation triplets are then selected as the final output, providing comprehensive information in terms of subject entity probabilities, object entity probabilities, and predicate probabilities.

5 Experiment

In this section, we comprehensively evaluate our RAHP on the OV-SGG task. More results, including closed-set SGG, parameter sensitivity experiments and qualitative analysis, are provided in the Appendix.

5.1 Datasets and Experimental Settings

Datasets To evaluate the SGG task, we adopt two benchmarks: the VG150 version of the Visual Genome (VG) dataset (Krishna et al. 2017) and the Open Image v6 (OIV6) dataset (Kuznetsova et al. 2020).

Evaluation metrics We evaluate our method under two settings (Chen et al. 2023): Open Vocabulary Relation-based Scene Graph Generation (OVR-SGG) uses a closed vocabulary for objects and an open one for relationships, whereas Open Vocabulary Detection + Relation-based Scene Graph Generation (OVD+R-SGG) uses open vocabularies for both. We adopt the PredCLS and SGDet protocols (Xu et al. 2017) and report the performance on Recall @K (K=50/100) and mean Recall @mK (mK=50/100) for each setting.

Implementation Details We employ the GPT-3.5-turbo, as our LLM. We adopt CLIP (Radford et al. 2021) (ViT-B/32) as our VLM backbone. We categorize 150 entities into 30 super-class entities for VG and categorized 602 entities into 53 super-class entities for OIV6 (details can be found in the appendix). RAHP is applicable to both one-stage and two-stage models, therefore we select the one-stage methods SGTR† (Li, Zhang, and He 2024) and OvSGTR

S	T	B	D	M	Total (Relation) R@50/100	Total (Relation) mR@50/100	Base (Relation) R@50/100	Base (Relation) mR@50/100	Novel (Relation) R@50/100	Novel (Relation) mR@50/100
PredCLS	o	ViT	DETR*	PGSG	26.90/33.90	10.80/13.90	-	-	-	5.20/7.70
		R-101	DETR	SGTR†	36.32/41.51	13.30/17.60	40.50/46.67	19.57/24.96	0.00/0.00	0.00/0.00
				SGTR†+p	39.48/45.27	15.93/21.09	40.78/46.82	19.88/24.32	10.81/18.41	9.00/13.15
				SGTR†+RAHP	**39.92/46.03**	**16.88/22.18**	**41.29/47.65**	**20.51/25.18**	**15.46/20.37**	**11.82/15.46**
	t	R-50	Faster R-CNN	SVPR	33.50/35.90	8.30/10.80	-	-	-	-
		R-101		Epic	-	16.50/21.80	28.30/31.10	-	13.90/18.30	-
				PE-NET	58.79/61.23	19.18/20.97	63.62/67.09	23.18/25.79	0.00/0.00	0.00/0.00
				PE-NET+p	62.21/67.25	21.94/27.91	62.73/67.76	22.23/28.02	17.62/25.67	12.93/19.32
				PE-NET+RAHP	**64.70/69.11**	**24.50/28.25**	**65.15/70.54**	**24.99/30.19**	**20.79/29.00**	**15.70/23.73**
SGDet	o	Swin-T	DETR	OvSGTR	20.46/23.86	3.91/4.62	26.14/**30.16**	4.81/5.60	13.45/16.19	1.82/2.32
				OvSGTR+RAHP	**21.50/25.74**	**4.51/5.37**	**26.29/30.16**	**5.15/5.94**	**15.59/19.92**	**3.01/4.04**

Table 1: Experimental results of OVR-SGG on VG test set. - to signify methods that did not produce the result, p indicates the use of fixed-format text prompts, while DETR* denotes models with structural modifications. S is the SGG setting; T denotes the SGG model type, o means one-stage model, t is two-stage model; B is the backbone model; D is the object detector; M represents the model.

S	T	B	D	M	Total (Relation) R@50/100	Total (Relation) mR@50/100	Base (Relation) R@50/100	Base (Relation) mR@50/100	Novel (Relation) R@50/100	Novel (Relation) mR@50/100
SGDet	o	ViT	DETR*	PGSG	41.30/43.30	20.80/23.00	-	-	-	3.80/8.90
		R-101	DETR	SGTR†	36.10/38.40	11.00/16.70	-	-	-	0.00/0.00
				SGTR†+p	60.48/62.63	29.09/31.44	74.11/77.17	36.32/37.43	46.31/53.76	27.82/32.66
				SGTR†+RAHP	**62.42/64.86**	**30.79/34.46**	**78.21/80.27**	**37.72/38.66**	**49.61/56.28**	**29.43/34.16**

Table 2: Experimental results of OVR-SGG on OIV6 test set. - to signify methods that did not produce the result, p indicates the use of fixed-format text prompts, while DETR* denotes models with structural modifications. S is the SGG setting; T denotes the SGG model type, o means one-stage model; B is the backbone model; D is the object detector; M represents the model.

(Swin-T) (Chen et al. 2023), as well as the two-stage methods PR-NET (Zheng et al. 2023) and VS^3 (Swin-T) (Zhang et al. 2023) to validate the generality and high adaptability of RAHP. In each method, we retain the visual module and proposal network, replace the predicate classification part with RAHP's OV predicate prediction approach. To align relation and VLM features, we equip CLIP with a three-layer MLP of size 512. We set $k = 3$ to dynamically select and set $\alpha = 0.25$ to balance the weights of the two text prompts. For training losses, the weight of the entity detector is $\lambda_1 = 2$, the weight for predicate prediction is $\lambda_2 = 1$, and the weight for distillation loss is $\lambda_3 = 20$. All experiments are implemented in PyTorch and trained on 4 NVIDIA A40 GPUs.

5.2 Comparisons with OVR-SGG Methods

Setup We evaluate our design on the VG and OIv6 datasets, comparing it with OVR-SGG methods, including SVRP, Epic, PSGS, and OvSGTR (see Table 1). We adapt SGG methods (SGTR† and PE-NET) for OVR-SGG, as they perform well in closed-vocabulary settings. In each method, we retain the visual module and replace the predicate classifier with RAHP's OV predicate prediction module. Additionally, we use a fixed text prompt baseline for comparison, where the prompt only provides information about predicate categories. In the VG dataset's PredCLS setting, we follow Epic's predicate split, selecting 70% of the categories as base predicates and the remaining 30% as novel predicates. In the SGDet setting, we follow the OvSGTR predicate split. For the OIV6 dataset, we use the predicate split from PGSG. During training, only base relation annotations are available, with images lacking base relation annotations masked.

Visual Genome Compared to previous methods in Table 1, our approach demonstrates significant performance advantages. For instance, in the PredCLS task, our method improves the novel mR@100 by 7.76 over the one-stage method PGSG. In two-stage methods, compared to Epic, our method increases the novel R@100 by 1.75 in the PredCLS task. Whether in one-stage or two-stage methods, RAHP shows flexible generalizability, is capable of achieving open-vocabulary capabilities under different frameworks. Taking PE-NET as an example, our method outperforms baseline models with fixed text prompts, improving both base and novel predicate performance. This highlights RAHP's ability to enhance text representations and improve visual relation understanding.

Open Image v6 We compare our method with PGSG on the OIV6 dataset. As shown in Table 2, our approach achieves an improvement of 21.56 points in total R@100 and 25.26 points in novel mR@100. This demonstrates that the introduction of hierarchical text prompts can enhance text representation, leading to better visual-text matching.

5.3 Comparisons with OVD+R-SGG Methods

Setup We evaluate the performance of RAHP in a fully open vocabulary setting OVD+R-SGG on VG, where novel

S	T	B	D	M	Total		Novel (Object)		Novel (Relation)	
					R@50	R@100	R@50	R@100	R@50	R@100
SGDet	t	Swin-T	-	VS^3	5.88	7.20	6.00	7.51	0.00	0.00
				VS^3+RAHP	**12.66**	**15.39**	**13.01**	**14.82**	**3.75**	**5.12**
	o	Swin-T	DETR	OvSGTR	13.53	16.36	**14.37**	**17.44**	9.20	11.19
				OvSGTR+RAHP	**13.83**	**16.52**	12.45	15.38	**13.31**	**16.46**

Table 3: Experimental results of OVD+R SGG on VG test set. S is the SGG setting; T denotes the SGG model type, o means one-stage model, t is the two-stage model; B is the backbone model; D is the object detector; M represents the model.

#	EP	RP	DS	Total (Relation)		Base (Relation)		Novel (Realtion)	
				R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100
1				21.66/25.89	8.49/10.84	21.85/26.06	9.11/12.42	6.15/9.88	4.73/8.13
2	√			23.78/28.88	9.62/13.65	23.97/29.06	10.71/14.65	8.62/13.05	7.09/11.10
3		√		23.81/28.04	9.01/11.99	23.49/28.91	10.10/13.08	2.38/5.44	1.33/3.77
4		√	√	23.71/27.98	9.43/12.51	23.95/28.24	10.55/14.38	6.20/9.10	6.94/9.59
5	√	√		22.20/26.07	9.33/13.11	24.13/28.85	10.80/15.08	4.25/6.44	3.88/5.25
6	√	√	√	**24.25/29.17**	**10.09/13.85**	**24.43/29.35**	**11.46/15.39**	**9.25/13.44**	**8.88/11.35**

Table 4: Ablation study on model components of VG val set. EP: Entity-aware Prompt; RP: Region-aware Prompt; DS: Dynamic Selection. The first row represents the baseline with a fixed predicate text prompt.

object and relationship categories are excluded during the training phase. Additionally, we achieve fully open vocabulary capability on VS^3 by replacing the original closed-set predicate classifier with RAHP. We the performance of OvD+R SGG, covering results in three aspects: "Total" (i.e., all object and relationship categories), "Novel (Object)" (i.e., considering only novel object categories), and "Novel (Predicate)" (i.e., considering only novel predicate categories)."

Results Table 3 shows that the inclusion of RAHP significantly improved the performance of novel relation, whether in VS^3 or OvSGTR. RAHP expands the text representation space by dynamically selecting region-aware text prompts. This enhances the model's generalization ability, making it more effective in handling new relationship concepts. Compared the OvSGTR in line 3, RAHP increases R@100 by 5.27. We observe a performance drop on novel objects, likely due to the differing distillation methods of the two models: OvSGTR uses relation feature distillation from a pre-trained model, while RAHP employs visual feature distillation from VLM. These differences may lead to conflicts between the approaches.

5.4 Ablation Study

We conduct an ablation study to assess the impact of each part on the method's effectiveness and the validity of SGG training. We divide RAHP into three main components: the entity-aware prompt, the region-aware prompt, and the dynamic selection mechanism. We analyze their roles individually based on PE-NET under the OVR-SGG SGDet setting. The results are summarized in Table 4.

Replacing the fixed predicate text prompt with the entity-aware prompt results in a 3% performance improvement for both base and novel predicates, highlighting the effectiveness of incorporating entity information in enhancing text representation of relations. This underscores the importance of relation triplet information in relationship detection. Introducing the region-aware prompt further enhances text representation, improving alignment between visual and textual features. However, without a dynamic selection mechanism to filter region-aware prompts, performance on novel predicates declines due to noise interference from irrelevant prompts. Base predicates, with inherently higher prediction scores, are more robust against this noise. Implementing an image-guided filtering strategy effectively removes noise, improving prediction accuracy for novel predicates.

6 Conclusion

In this paper, we introduce the Relation-Aware Hierarchical Prompting framework (RAHP), designed to address the challenges of OV-SGG by enhancing text representations. By integrating entity-aware and region-aware relation text prompts, RAHP enhances text representation and enables more accurate and flexible image-text matching. Our dynamic selection mechanism further refines this process by adapting prompts based on visual information, reducing noise and improving the robustness of relation predictions. Through extensive experiments on the Visual Genome and Open Images v6 datasets, our method demonstrates state-of-the-art performance, and the demonstrated performance improvements—highlight the potential of RAHP to significantly advance the field of OV-SGG.

Discussion of Limitations: (1) *Effectiveness of Entity Clustering*. Clustering algorithms will struggle to maintain fine distinctions between diverse data categories, which can degrade the quality of text representations. (2) *Diversity of Generated Text Prompts*. Limited diversity in LLM-generated region descriptions can hinder model generalization for novel relationships.

Acknowledgments

This work was supported by NSFC 62350610269, Shanghai Frontiers Science Center of Human-centered Artificial Intelligence, and MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

References

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9962–9971.

Chen, Z.; Wu, J.; Lei, Z.; Zhang, Z.; and Chen, C. 2023. Expanding Scene Graph Boundaries: Fully Open-vocabulary Scene Graph Generation via Visual-Concept Alignment and Retention. *arXiv preprint arXiv:2311.10988*.

Gao, K.; Chen, L.; Zhang, H.; Xiao, J.; and Sun, Q. 2023. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*.

Ge, Y.; Ren, J.; Gallagher, A.; Wang, Y.; Yang, M.-H.; Adam, H.; Itti, L.; Lakshminarayanan, B.; and Zhao, J. 2023. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11093–11101.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1969–1978.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

He, T.; Gao, L.; Song, J.; and Li, Y.-F. 2022. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European Conference on Computer Vision*, 56–73. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

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

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

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

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Lee, S.; Kim, J.-W.; Oh, Y.; and Jeon, J. H. 2019. Visual question answering over scene graph. In *2019 First International Conference on Graph Computing (GC)*, 45–50. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Wang, Y.; Guo, X.; Yang, R.; and Li, W. 2024a. Leveraging Predicate and Triplet Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28369–28379.

Li, L.; Xiao, J.; Chen, G.; Shao, J.; Zhuang, Y.; and Chen, L. 2024b. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36.

Li, R.; Zhang, S.; and He, X. 2024. SGTR+: End-to-end Scene Graph Generation with Transformer. arXiv:2401.12835.

Li, X.; Chen, L.; Ma, W.; Yang, Y.; and Xiao, J. 2022. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4204–4213.

Li, Y.; Pan, Y.; Chen, J.; Yao, T.; and Mei, T. 2021. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3799–3802.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20123–20132.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14074–14083.

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

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

Shi, J.; Zhong, Y.; Xu, N.; Li, Y.; and Xu, C. 2021. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16393–16402.

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

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6619–6628.

Wang, H.; Yang, M.; Wei, K.; and Deng, C. 2023. Hierarchical prompt learning for compositional zero-shot recognition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 1470–1478.

Wang, Y.; Jiang, X.; Cheng, D.; Li, D.; and Zhao, C. 2024. Learning Hierarchical Prompt with Structured Linguistic Knowledge for Vision-Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5749–5757.

Wang, Z.; Wang, P.; Liu, T.; Lin, B.; Cao, Y.; Sui, Z.; and Wang, H. 2022. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. *arXiv preprint arXiv:2204.13413*.

Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. 2024. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10685–10694.

Yu, Q.; Li, J.; Wu, Y.; Tang, S.; Ji, W.; and Zhuang, Y. 2023. Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an Open World. *arXiv preprint arXiv:2303.13233*.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14393–14402.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5831–5840.

Zhang, C.; Stepputtis, S.; Campbell, J.; Sycara, K.; and Xie, Y. 2024. HiKER-SGG: Hierarchical Knowledge Enhanced Robust Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28233–28243.

Zhang, Y.; Pan, Y.; Yao, T.; Huang, R.; Mei, T.; and Chen, C.-W. 2023. Learning To Generate Language-Supervised and Open-Vocabulary Scene Graph Using Pre-Trained Visual-Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2915–2924.

Zheng, C.; Lyu, X.; Gao, L.; Dai, B.; and Song, J. 2023. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22783–22792.

Zhong, Y.; Shi, J.; Yang, J.; Xu, C.; and Li, Y. 2021. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1823–1834.