RA-SGG: Retrieval-Augmented Scene Graph Generation Framework via Multi-Prototype Learning

Kanghoon Yoon[1], Kibum Kim[1], Jaehyeong Jeon[1], Yeonjun In[1], Donghyun Kim[2], Chanyoung Park[1]*

[1] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology
[2] Department of Artificial Intelligence, Korea University
{ykhoon08, kb.kim, wogud405, yeonjun.in, cy.park}@kaist.ac.kr, d_kim@korea.ac.kr

Abstract

Scene Graph Generation (SGG) research has suffered from two fundamental challenges: the long-tailed predicate distribution and semantic ambiguity between predicates. These challenges lead to a bias towards head predicates in SGG models, favoring dominant general predicates while overlooking fine-grained predicates. In this paper, we address the challenges of SGG by framing it as *multi-label classification problem with partial annotation*, where relevant labels of fine-grained predicates are missing. Under the new frame, we propose **R**etrieval-**A**ugmented **S**cene **G**raph **G**eneration (RA-SGG), which identifies potential instances to be multi-labeled and enriches the single-label with multi-labels that are semantically similar to the original label by retrieving relevant samples from our established memory bank. Based on augmented relations (i.e., discovered multi-labels), we apply multi-prototype learning to train our SGG model. Several comprehensive experiments have demonstrated that RA-SGG outperforms state-of-the-art baselines by up to 3.6% on VG and 5.9% on GQA, particularly in terms of F@K, showing that RA-SGG effectively alleviates the issue of biased prediction caused by the long-tailed distribution and semantic ambiguity of predicates.

1 Introduction

SGG is a pivotal task in scene understanding, aiming to detect objects and predict their relationships within an image. SGG models are adept at capturing rich visual information that encompass both object-level and relation-level information. This compositional representation of the scene graph is beneficial for various vision applications (Hildebrandt et al. 2020; Schroeder and Tripathi 2020). However, existing SGG approaches have encountered two fundamental challenges:

1) Long-tailed Problem: The predicate classes in benchmark SGG datasets, e.g., Visual Genome, follow a severely skewed distribution, where certain predicates are heavily represented (head classes) while others are sparsely annotated (tail classes). This challenge arises from annotators' tendency to label general predicates such as "on" more frequently than fine-grained predicates such as "walking in." Consequently, the predictions of SGG models trained on datasets with such a skewed distribution tend to be biased towards head predicates, resulting in scene graphs dominated by general predicates. However, these scene graphs lack nuanced descriptions provided by tail predicates, which offer detailed relationship information. **2) Semantic Ambiguity**: Another challenge stems from the ambiguous semantic boundaries between predicate classes. For a relation instance, predicates such as "on," "walking on," and "walking in" may be hard-to-distinguish as these predicates share semantic meanings, and thus SGG models need to capture nuanced visual cues to distinguish between them. However, learning the detailed visual differences is challenging as the tail predicates, such as "walking in," rarely appear in the dataset whose predicate class distribution is usually skewed. Addressing these challenges is crucial for generating informative scene graphs with fine-grained predicates.

Seminal works have focused on generating fine-grained scene graphs by alleviating the issues posed by these challenges. Re-sampling, re-weighting, and various debiasing methods (Li et al. 2021; Lyu et al. 2022; Kaihua et al. 2020; Kim et al. 2024b) are utilized to generate the fine-grained scene graphs. Most recently, (Li et al. 2022; Zhang et al. 2022; Yu et al. 2023) propose the data enhancement approaches by re-annotating the general predicates to fine-grained ones. However, this refinement process results in a trade-off, where comprehension of general predicates is diminished. Moreover, we find that fundamental challenges persist in the prior framework. Current SGG models rely on single-label classification, where predicates compete against each other during prediction, forcing the model to select just one predicate while suppressing others. This approach is problematic because it ignores the nuanced nature of natural language, where multiple predicates convey the same single relationship. Hence, we cast the single-label classification with *multi-label classification with partial annotation problem* (Ben-Baruch et al. 2022). That is, we discover potential fine-grained predicates in the original training data and augment semantically similar predicates for relationships.

In this work, we propose a **R**etrieval-**A**ugmented **S**cene **G**raph **G**eneration (RA-SGG) framework that enhances single-label annotations to multi-labels with fine-grained predicates, addressing the challenges of the long-tailed predicate distribution and the semantic ambiguity. The core idea is to retrieve potential predicates that share semantics with

*Corresponding author

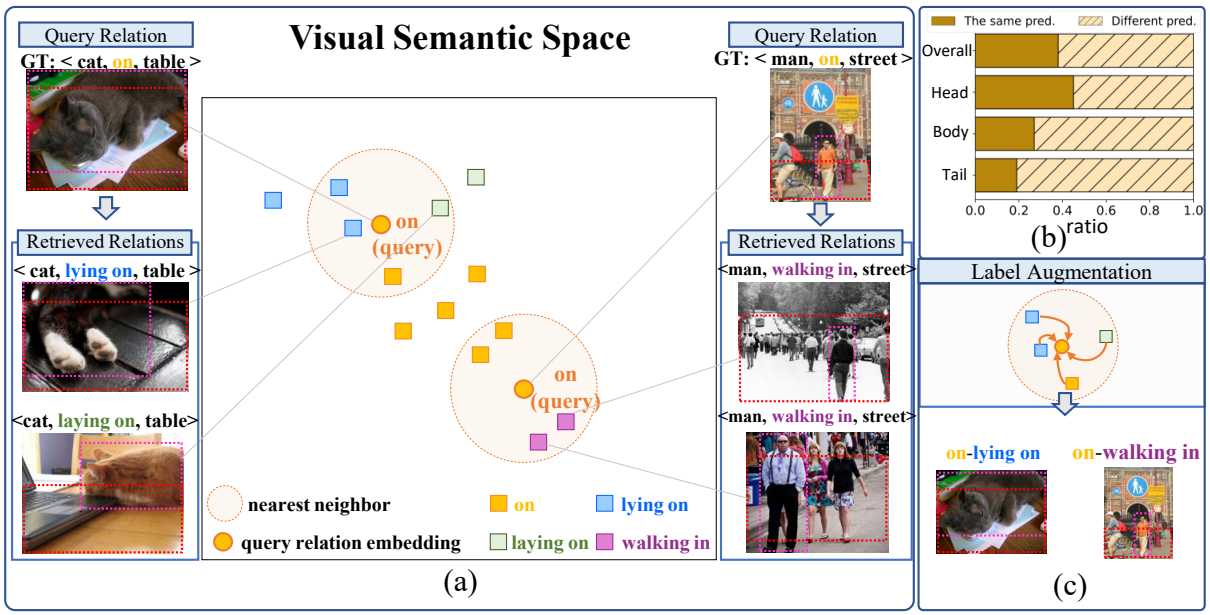

Figure 1: Illustration of the label augmentation of RA-SGG in SGG. (a) Examples of the query relation instances and their retrieved relation instances in the embedding space. **"GT"** denotes the ground truth. (b) The ratio of retrieved instances that have the same/different predicate as the query relation in the training data. (c) Fine-grained predicate label augmentation.

the original predicate and use them as pseudo-labels for the relation, as shown in Figure 1. Our approach is motivated by observations in the visual semantic space. Using the pre-trained SGG model, PE-Net (Zheng et al. 2023), for nearest neighbor retrieval on training instances, we discovered that query relations and their retrieved instances often share semantic meanings despite having different predicate labels. For instance, a relation with the ground-truth predicate "on" retrieves different fine-grained predicates: {"lying on", "laying on"} for ⟨cat,on,table⟩ or {"walking in", "walking in"} for ⟨man, on, street⟩. Notably, this pattern persists across head, body, and tail predicate classes (Figure 1.b), suggesting that *a relation instance that is currently annotated with a single predicate might actually co-reference multiple latent predicates with similar meanings.*

Building on these observations, RA-SGG aims to discover latent predicates of a relation instance by retrieving semantically similar relation instances from a memory bank. We introduce three key mechanisms: (1) *Reliable Multi-Labeled Instance Selection*: naively assigning pseudo-labels for all the relation instances would aggravate reliability. To obtain reliable pseudo-labels, we introduce a label inconsistency score which measures the discrepancy between the ground-truth predicate and its retrieved relation instances. (2) *Unbiased Augmentation of Multi-Labels*: we introduce the inverse propensity score-based sampling strategy to sample the less frequent (i.e., fine-grained) predicates among the retrieved relation instances. (3) *Multi-Prototype Learning with Multiple Predicates*: it ensures relation instance embeddings capture both original and discovered fine-grained predicate semantics. Our comprehensive experiments demonstrate that RA-SGG outperforms state-of-the-art SGG models, particularly in terms of F@K, which is a metric that emphasizes the ability to achieve high performance in fine-grained classes (i.e., tail classes) while minimally sacrificing the performance of general classes (i.e., head classes). **Our contribution** can be summarized as follows:

1. We highlight the problem of single-label classification in SGG and reformulate this in the light of *multi-label classification with partial annotation problem*.
2. We propose a novel retrieval-augmented SGG framework (RA-SGG), which simultaneously alleviates the long-tailed distribution and semantic ambiguity issues by discovering the latent predicates of a relation instance annotated with a single predicate.
3. Our extensive studies demonstrate that RA-SGG achieves the state-of-art performance in terms of F@K, implying that RA-SGG effectively generates fine-grained scene graphs without sacrificing the understanding of general predicates.

2 Related Works

Scene Graph Generation

SGG has garnered attention due to its capacity for dense understanding of a scene, bridging two modalities: vision and language. Early works have focused on capturing the contextual information within a scene using message-passing network and transformer architecture (Zellers et al. 2018; Yang et al. 2018; Li et al. 2021; Shit et al. 2022). Recent SGG studies have been centered on improving performance for fine-grained predicates (i.e., tail classes) to obtain informative scene graphs. However, it is challenging to predict

fine-grained predicates due to the fact that the predictions of SGG models easily collapse into general and trivial predicates, suffering from both the long-tailed distribution and semantic ambiguity issues. To address this, (Li et al. 2021; Lyu et al. 2022) focus training on tail predicate classes via resampling/reweighting, and (Kaihua et al. 2020; Dong et al. 2022; Jeon et al. 2024) proposed debiasing methods to mitigate the bias towards head classes. However, merely amplifying the weight of tail classes can lead to overfitting of SGG models to the limited number of tail instances (Zhai et al. 2022), as these methods do not contribute to increasing variations in the data for tail classes (S. Hasegawa and Umeda 2022). As an alternative, approaches that enhance the training data have emerged. Specifically, (Li et al. 2022) identifies noisy predicates in data by using a pre-trained model, and (Zhang et al. 2022) replaces general predicates with fine-grained ones or uses fine-grained predicates to fill in missing annotations in the training data. (Kim et al. 2024a) assigns pseudo-labels to the missing annotated instances.

However, the data enhancement approaches eliminate general predicates from the dataset, leading to a significant performance drop on these general predicates. This strategy overlooks a comprehensive understanding of general predicates to instead obtain fine-grained scene graph. Moreover, we argue that adjusting data based on a single prediction of model is risky in the SGG nature, where the models are trained on the extremely long-tailed distribution. To resolve these issues, we addresses the SGG problem in the light of the multi-label classification problem for comprehensive understanding across all predicate classes. Moreover, RA-SGG does not rely on a single model prediction but considers the distribution of the retrieved relation instances to accurately assign predicates as a multi-label.

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is a widespread approach in natural language processing (NLP). RAG employs a non-parametric memory to supplement intrinsic knowledge of a language model by retrieving relevant passages from knowledge bases (Borgeaud et al. 2022; Guu et al. 2020; Lewis et al. 2020). Similarly, in the realm of image synthesis, approaches like (Casanova et al. 2021; Blattmann et al. 2022) refer to examples of near neighbors to elevate the quality of generated images. Furthermore, RAG verifies the ability to handle rare examples within the training dataset. Re-Imagen (Chen et al. 2023) effectively generates the rare entities by referencing images with similar text descriptions. (Long et al. 2022) enhances the image classification performance on tail classes by drawing upon image features from a memory bank. In this work, we propose the retrieval-augmented scene graph generation framework to discover the latent multi-labels in the training dataset, which addresses the long-tailed problem in SGG.

3 Preliminaries

Scene Graph Generation Task

The SGG problem includes the detection of objects and the prediction on the classes of objects and their relationships. Formally, given an image, we aim at generating a scene graph $\mathbf{G} = \langle \mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i \rangle_{i=1}^{N}$, where N is the number of triplets in the image. $\mathbf{s}_i$ represents the subject, which includes the bounding box position and the object class, while $\mathbf{o}_i$ represents the object, which is similarly defined. $\mathbf{p}_i$ represents the predicate between $\mathbf{s}_i$ and $\mathbf{o}_i$.

Visual Semantic Embedding Space in SGG

(Zheng et al. 2023) proposes a novel SGG learning strategy, which reduces the distance between the prototype and visual relation embedding in the visual semantic space. This paradigm helps a model to discern between hard-to-distinguish predicates by measuring the distance in the visual semantic space.

Formally, let $\boldsymbol{v_s} \in \mathbb{R}^d$ and $\boldsymbol{v_o} \in \mathbb{R}^d$ denote the visual semantic features of subject and object, respectively, which are extracted from both the visual feature and the word embedding of entity classes. We use $\boldsymbol{u_p} \in \mathbb{R}^d$ to denote the relation embedding extracted from the visual features of the union box, and $\boldsymbol{t_p} \in \mathbb{R}^{d'}$ to denote the word embedding of the predicate class. The core idea of prototype learning is to match the subject and object instances to the predicate instances by $\mathcal{F}(\boldsymbol{v_s}, \boldsymbol{v_o}) \approx \mathbf{W}_p \boldsymbol{t_p} + \boldsymbol{u_p}$, where $\mathcal{F}$ is a fusion layer (Zheng et al. 2023) and $\mathbf{W}_p \in \mathbb{R}^{d \times d'}$ is a feature transformation matrix. To implement this, PE-Net considers the language semantic of predicates as a prototype (i.e., $\mathbf{W}_p \boldsymbol{t_p}$), and makes the relation embedding (i.e., $\boldsymbol{r} = \mathcal{F}(\boldsymbol{v_s}, \boldsymbol{v_o}) - \boldsymbol{u_p}$) close to its corresponding the predicate prototype. The prototype learning loss is defined as:

$$\mathcal{L}_{\text{proto}} = -\log \frac{\exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_{\text{gt}} \rangle)/\gamma}{\sum_{j=1}^{N_p} \exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_j \rangle)/\gamma}, \quad (1)$$
$$\text{where} \quad \bar{\boldsymbol{r}} = \mathsf{Proj}(\boldsymbol{r}) \text{and} \quad \bar{\boldsymbol{c}}_j = \mathsf{Proj}(\mathbf{W}_p \boldsymbol{t_p})$$

where N_p is the number of predicate categories, γ is a learnable temperature hyperparameter and Proj is a 2-layer MLP shared across both the relation embedding $\boldsymbol{r}$ and the prototypes. $\bar{\boldsymbol{c}}_{\text{gt}}$ denotes the class prototype of the ground truth predicates. This prototype-guided learning has demonstrated the effectiveness of producing discriminative relation embeddings in the visual semantic space, compared to the conventional classifier-based SGG models. Hence, our proposed method, RA-SGG, utilizes the relation embeddings of the training instances learned from the above objective to instantiate the retrieval-augmented approache for SGG.

4 Retrieval Augmented SGG

In this section, we introduce RA-SGG, a novel method that utilizes a retrieval process to discover and augment the latent and fine-grained predicates. We begin by formulating the SGG task as the multi-label classification with a partial (i.e., single) annotation problem, which aims to model the true unbiased predicate distribution based on the partially annotated samples. Subsequently, we explain how RA-SGG identifies relation instances that include potential multi-labels and augments the labels of these instances by assigning semantically similar predicates as additional pseudo-labels. Lastly, we outline the training procedure of RA-SGG, which employs a multi-prototype learn-

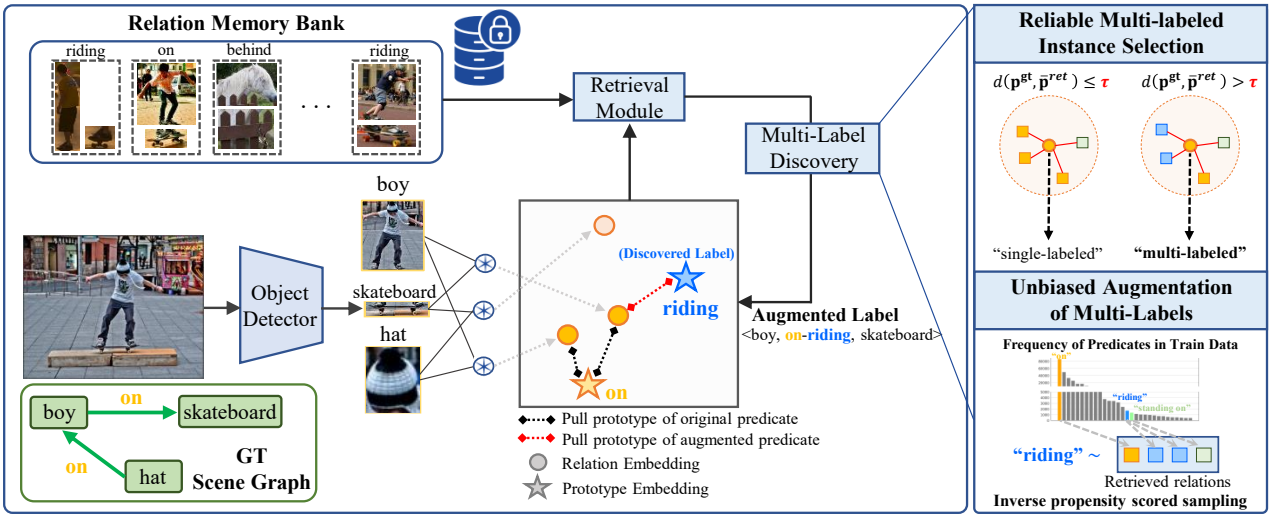

Figure 2: RA-SGG first uses the relation embeddings for querying instances in the memory bank to retrieve visually and semantically similar instances. Then, it augments multi-labels by assigning pseudo-labels to potentially multi-labeled instances.

ing strategy to effectively learn from the augmented relation labels. The pipeline of RA-SGG is illustrated in Figure 2.

Multi-Label Classification with Partial Annotation

Recall that existing SGG studies have relied on a single annotation for each relationship, overlooking the potential predicates that share similar semantic meanings. This oversight limits the model's ability to represent complex visual scenes, particularly affecting the representation of less common or tail predicates due to insufficient data variation. Herein, we first formulate the SGG task with multi-labeled classification problem with partial annotation to exploit the latent fine-grained predicates.

Let $\mathbf{y}^* \in \{0,1\}^{N_p}$ denote the ground-truth label vector of a relation instance from an unbiased true distribution, which is an ideal distribution without any latent unobserved annotations. $\mathbf{y}^*$ accommodates multi-labels including semantically relevant predicates, i.e., $\sum_{i=1}^{N_p} \mathbf{y}_i \geq 1$. In a common SGG setting, we deal with partially observed annotation with the observed label vector $\mathbf{y} \in \{0,1\}^{N_p}$, where $\mathbf{y}$ is single-labeled or not labeled, i.e., $\sum_{i=1}^{N_p} \mathbf{y}_i = 1$ or $\sum_{i=1}^{N_p} \mathbf{y}_i = 0$. Our goal is to minimize the loss function $\mathcal{L}(\mathbf{y}^*, \hat{\mathbf{y}})$ computed on the samples from the unbiased true distribution by utilizing observed predicate labels $\mathbf{y}$, where $\hat{\mathbf{y}}$ is the probability predicted by an SGG model. However, since obtaining samples from the unbiased true distribution is not feasible, we approximate the loss using the available observed samples. Hence, we introduce the following inverse propensity-scored loss:

$$\mathcal{L}_{ips} = -\sum_{i=1}^{N_p} \underbrace{P(\mathbf{y}_i = 1 | \mathbf{y}_i^* = 1)^{-1}}_{\text{inverse propensity score}} \mathbf{y}_i \log \hat{\mathbf{y}}_i \qquad (2)$$

where $\mathcal{L}_{ips}$ is the weighted cross entropy with the inverse propensity for each predicate class, and the propensity score is the label frequency (Jain, Prabhu, and Varma 2016; Chiou et al. 2021), which is the fraction of observed predicates among entire predicates in the unbiased true distribution. For simplicity, we follow (Jain, Prabhu, and Varma 2016) and use the frequency of each predicate class in the training dataset as the propensity score. Note that optimizing $\mathcal{L}_{ips}$ is approximately minimizing the loss using samples from the unbiased true distribution since $\mathcal{L}_{ips}$ serves as an unbiased estimator of the loss function $\mathcal{L}(\mathbf{y}^*, \hat{\mathbf{y}})$. However, directly minimizing the inverse propensity scored loss for SGG models is still restricted as it does not increase the variation of data instances of tail predicate classes, which incurs overfitting to a small number of tail predicates. Instead of optimizing the above $\mathcal{L}_{ips}$, we choose to augment a relation instance by sampling the predicate labels based on the inverse propensity score (Han et al. 2022), and assign the sampled predicate labels to the relation instance as pseudo-labels. We will discuss this in detail in Section 4.

Retrieval-Augmented Scene Graph Generation

We adopt a retrieval process to effectively train an SGG model based on the partially observed predicates. Specifically, we discover and augment the latent and fine-grained predicates of a relation instance by referring to the predicates of the retrieved instances. The process begins by retrieving semantically similar samples from the memory bank. We then leverage the predicates associated with these retrieved samples to identify relation instances that could be labeled with multiple predicates and to augment our dataset with these multi-labels. Although our method can be integrated into any SGG model, this paper specifically showcases its application through PE-Net as the backbone.

Relation Retrieval Module. The relation retrieval module finds semantically similar relation instances based on the relation embedding. Specifically, a frozen memory bank is established using a pre-trained SGG model prior to the training of RA-SGG. We store key-value pairs

$(k_1, v_1), ..., (k_M, v_M)$ in the memory bank, where the key is the relation embedding ($\boldsymbol{r}$ in Eqn. 1 for PE-Net), and the value is an associated one-hot vector representing the predicate class. Our approach limits memory storage upto 10 entries for each unique triplet, which only utilizes 8% of the training data. Note that the memory bank construction is executed during the pre-processing phase, without imposing additional computational costs in the training phase of RA-SGG. During the training phase, RA-SGG produces the relation embedding $\boldsymbol{r}$ for each subject-object pair in the training batch. The relation embeddings act as queries to retrieve K neighbor instances from the memory bank based on the cosine similarity. Among the retrieved instances, we exclude the first instance to avoid the trivial case of retrieving itself.

Reliable Multi-Labeled Instance Selection. Since some relation instances contain only a single semantic meaning, naively assigning pseudo-labels to all relation instances would aggravate reliability. Given the K nearest neighbor instances $\{(\boldsymbol{r}_1^{\text{ret}}, \mathbf{p}_1^{\text{ret}}), ..., (\boldsymbol{r}_K^{\text{ret}}, \mathbf{p}_K^{\text{ret}})\}$ of the query relation, RA-SGG first identifies whether the query relation is a single-labeled instance with only one predicate, or a multi-labeled instance that, despite potentially qualifying for multiple labels, currently has only one predicate. This selection is based on the label inconsistency between the ground-truth predicate of query relation $\mathbf{p}^{\text{q}}$ and the averaged one-hot predicates of retrieved instances $\bar{\mathbf{p}}^{\text{ret}} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{p}_i^{\text{ret}}$. By computing the label inconsistency distance $d(\cdot, \cdot)$ for all query relation instances in the training batch, we obtain the $\mathcal{D}_{\text{single}}$ and $\mathcal{D}_{\text{multi}}$ as follows:

$$\mathcal{D}_{\text{single}} \leftarrow \{(\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) | d(\mathbf{p}^{\text{q}}, \bar{\mathbf{p}}^{\text{ret}}) < \tau, \forall (\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) \in \mathcal{D}_{\text{Tr}}\} \quad (3)$$

$$\mathcal{D}_{\text{multi-}} \leftarrow \{(\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) | d(\mathbf{p}^{\text{q}}, \bar{\mathbf{p}}^{\text{ret}}) \geq \tau, \forall (\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) \in \mathcal{D}_{\text{Tr}}\}$$

where τ is the threshold hyperparameter and d is the distance metric. If the given ground-truth predicate of the query relation is significantly different from that of the nearest neighbor samples, i.e., large inconsistency, then the query relation is likely to co-reference additional potential predicate labels among the predicates of the neighbor samples.

Unbiased Augmentation of Multi-Labels. Having identified potential relation instances to be multi-labeled, we augment the fine-grained predicates by assigning specific multi-labels to the training instances within $\mathcal{D}_{\text{multi}}$. In this augmentation stage, we aim to obtain the augmented labels that follow the unbiased true distribution by adopting the inverse propensity score-based sampling. Specifically, we obtain the sampling weight $w \in \mathbb{R}^{N_p}$ by aggregating the inverse propensities of the retrieved predicates, i.e., $w = \mathsf{Softmax}(\sum_{k=1}^{K} s_k^{\text{ret}} \mathbf{p}_k^{\text{ret}})$, where s_1^{ret}, ..., $s_K^{\text{ret}} \in \mathbb{R}$ are the inverse propensities of retrieved predicates, which are pre-computed as the inverse of the predicate frequency in the training dataset. For each triplet $(\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i)$ in $\mathcal{D}_{\text{multi}}$, we augment the predicate labels using the sampling weight as:

$$\mathbf{p}_i \leftarrow \lambda_i \mathbf{p}_i^{\text{gt}} + (1 - \lambda_i)\mathbf{p}_i^{\text{aug}}, \text{where } \mathbf{p}_i^{\text{aug}} \sim \mathsf{Multinomial}(w) \quad (4)$$

where $\lambda_i \sim \mathsf{Beta}(\alpha, \beta)$ controls the degree of the mixing coefficient. Our augmentation strategy generates pseudo-labels, considering not only the overall semantic meaning of nearest neighbors but also the frequency of predicates. This approach effectively leverages partial annotations to create multi-labeled instances, which can address the long-tailed distribution problem and semantic ambiguity in SGG.

Multi-Prototype Learning with Multiple Predicates

With relation instances augmented with multiple predicate labels, we apply the multi-prototype learning strategy to train RA-SGG. For each relation instance, we minimize the distance between its relation embedding $\boldsymbol{r}$ and its corresponding prototypes of the multi-labeled predicates (i.e., $\bar{\boldsymbol{c}}_{\text{gt}}$ and $\bar{\boldsymbol{c}}_{\text{aug}}$) as follows:

$$\begin{aligned}\mathcal{L}_{\text{multi-proto}} = &-\lambda_i \log \frac{\exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_{\text{gt}} \rangle)/\gamma}{\sum_{j=1}^{N_p} \exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_j \rangle)/\gamma} \\ &- (1 - \lambda_i) \log \frac{\exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_{\text{aug}} \rangle)/\gamma}{\sum_{j=1}^{N_p} \exp(\langle \bar{\boldsymbol{r}}, \bar{\boldsymbol{c}}_j \rangle)/\gamma}\end{aligned} \quad (5)$$

where $\bar{\boldsymbol{r}} = \mathsf{Proj}(\boldsymbol{r})$ is the relation embedding after the projection layer is applied, $\bar{\boldsymbol{c}}_{\text{gt}}$ denotes the corresponding class prototype of the ground truth predicates, $\bar{\boldsymbol{c}}_{\text{aug}}$ denotes the prototype of augmented predicates. It is important to note that while the multi-prototype learning loss calculation considers up to two predicate classes per relation instance, our predicate augmentation approach based on sampling allows for incorporating more than two predicates by altering the augmented predicate at each batch iteration. Multi-prototype learning ensures that the embedding of a relation instance not only incorporates the original semantic of its ground-truth predicate but also embraces the semantic of the discovered latent fine-grained predicates.

To train RA-SGG, the final loss $\mathcal{L}_{\text{final}}$ is expressed as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{multi-proto}} + \mathcal{L}_{\text{reg1}} + \mathcal{L}_{\text{reg2}} \quad (6)$$

where $\mathcal{L}_{\text{reg1}} = \sqrt{\sum_{i \neq j} (\bar{\boldsymbol{c}}_i^{\top} \bar{\boldsymbol{c}}_j)^2}$ is the prototype similarity regularization (Zheng et al. 2023), which aims to minimize cosine similarities between all pairs of individual prototypes, and $\mathcal{L}_{\text{reg2}} = \max\{\gamma' - \frac{1}{N_p}\sum_{i \neq j} \|\bar{\boldsymbol{c}}_i - \bar{\boldsymbol{c}}_j\|_2^2, 0\}$ is the prototype distance regularization (Zheng et al. 2023), which increases the euclidean distance between prototypes, respectively. These regularizations serve to prevent the collapse of prototypes with semantically similar predicates.

Note that the retrieval process and multi-prototype learning strategy are exclusively applied during the training phase. During the inference phase, we calculate the cosine similarity between each relation embedding and the set of prototypes. The predicate corresponding to the highest cosine similarity is selected as the model's prediction.

5 Experiments

Experimental Settings

Datasets and Implementation Details. Following the prior approaches (Dong et al. 2022; Li et al. 2023, 2021; Zellers et al. 2018; Zhang et al. 2022), we used the benchmark datasets, VG (Krishna et al. 2017) and GQA (Hudson and Manning 2019). VG split contains the most frequent 150 object classes and 50 predicate classes. GQA is another vision and language benchmark, which includes top-200 object classes and 100 predicate classes. For all baselines, we adopt ResNeXt-101-FPN (Xie et al. 2017) and Faster R-CNN (Ren et al. 2015) as the object detector. For RA-SGG,

B	Methods	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
Specific	KERN(Chen et al. 2019)CVPR'19	65.8/67.6	17.7/19.2	27.9/29.9	36.7/37.4	9.4/10.0	15.0/15.8	27.1/29.8	6.4/7.3	10.4/11.7
	BGNN(Li et al. 2021)CVPR'21	59.2/61.3	30.4/32.9	40.2/42.8	37.4/38.5	14.3/16.5	20.7/23.1	31.0/35.8	10.7/12.6	15.9/18.6
	DT2ACBS(Desai et al. 2021)ICCV'21	23.3/25.6	35.9/**39.7**	28.3/31.1	16.2/17.6	**24.8/27.5**	19.6/21.5	15.0/16.3	**22.0/24.0**	17.8/19.4
	HL-Net(Lin et al. 2022)CVPR'22	67.0/68.9	- /22.8	- /34.3	42.6/43.5	- /13.5	- /20.6	33.7/38.1	- /9.2	- /14.8
	HetSGG(Yoon et al. 2023)AAAI'23	57.8/59.1	31.6/33.5	40.9/42.8	37.6/38.7	17.2/18.7	23.6/25.2	30.0/34.6	12.2/14.4	17.3/20.3
	SQUAT(Jung et al. 2023)CVPR'23	55.7/57.9	30.9/33.4	39.7/42.4	33.1/34.4	17.5/18.8	22.9/24.3	24.5/28.9	14.1/16.5	17.9/21.0
Motif	Motif(Zellers et al. 2018)CVPR'18	64.6/66.0	15.2/16.2	24.6/26.0	38.0/38.9	8.7/9.3	14.2/15.0	31.0/35.1	6.7/7.7	11.0/12.6
	TDE(Kaihua et al. 2020)CVPR'20	46.2/51.4	25.5/29.1	32.9/37.2	27.7/29.9	13.1/14.9	17.8/19.9	16.9/20.3	8.2/9.8	11.0/13.2
	DLFE(Chiou et al. 2021)MM'21	52.5/54.2	26.9/28.8	35.6/37.6	32.3/33.1	15.2/15.9	20.7/21.5	25.4/29.4	11.7/13.8	16.0/18.8
	NICE(Li et al. 2022)CVPR'22	55.1/57.2	29.9/32.3	38.8/41.3	33.1/34.0	16.6/17.9	22.1/23.5	27.8/31.8	12.2/14.4	17.0/19.8
	GCL(Dong et al. 2022)CVPR'22	42.7/44.4	36.1/38.2	39.1/41.1	26.1/27.1	20.8/21.8	23.2/24.1	18.4/22.0	16.8/19.3	17.6/20.6
	IETrans(Zhang et al. 2022)ECCV'22	54.7/56.7	30.9/33.6	39.5/42.2	32.5/33.4	16.8/17.9	22.2/23.3	26.4/30.6	12.4/14.9	16.9/20.0
	CFA (Li et al. 2023)ICCV'23	54.1/56.6	35.7/38.2	43.0/45.6	34.9/36.1	17.0/18.4	22.9/24.4	27.4/31.8	13.2/15.5	17.8/20.8
	ST-SGG(Kim et al. 2024a)ICLR'24	53.9/57.7	28.1/31.5	36.9/40.8	33.4/34.9	16.9/18.0	22.4/23.8	26.7/30.7	11.6/14.2	16.2/19.4
PE-Net	PE-Net† (Zheng et al. 2023)CVPR'23	64.9/67.2	31.5/33.8	42.4/45.0	39.4/40.7	17.8/18.9	24.5/25.8	30.7/35.2	12.4/14.5	17.7/20.4
	IETrans† (Zhang et al. 2022)ECCV'22	49.3/51.8	33.5/36.0	39.9/42.5	31.2/32.3	18.3/19.4	23.1/24.2	24.2/28.4	13.7/16.2	17.5/20.6
	CFA† (Li et al. 2023)ICCV'23	57.8/61.6	30.0/33.2	39.5/43.1	36.2/37.1	15.9/18.2	22.1/24.4	25.6/29.8	14.4/17.1	18.4/21.7
	RA-SGG	62.2/64.1	**36.2**/39.1	**45.7/48.6**	38.2/39.1	20.9/22.5	**27.0/28.6**	26.0/30.3	14.4/17.1	**18.5/21.9**

Table 1: Performance (%) of state-of-the-art SGG models on Visual Genome (Krishna et al. 2017). F@K is the harmonic mean of mR@50/100 and R@50/100. † denotes the result produced by us using their official code.

Models	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
VTransE(Zhang et al. 2017)CVPR'17	55.7/57.9	14.0/15.0	22.4/23.8	33.4/34.2	8.1/8.7	13.0/13.9	27.2/30.7	5.8/6.6	9.6/10.9
Motif(Zellers et al. 2018)CVPR'18	65.3/66.8	16.4/17.1	26.2/27.2	34.2/34.9	8.2/8.6	13.2/13.8	28.9/33.1	6.4/7.7	10.5/12.5
VCTree(Tang et al. 2019)ICCV'19	63.8/65.7	16.6/17.4	26.4/27.5	34.1/34.8	7.9/8.3	12.8/13.4	28.3/31.9	6.5/7.4	10.6/12.0
PE-Net† (Zheng et al. 2023)CVPR'23	54.3/56.0	26.2/27.1	35.4/36.5	26.2/27.0	11.2/11.5	15.7/16.1	19.5/22.9	10.3/11.9	13.5/15.7
PE-Net+RA-SGG	48.3/50.1	**35.4/36.8**	**40.9/42.4**	19.9/20.8	**16.4/17.2**	**18.0/18.8**	16.3/19.0	**12.9/15.0**	**14.4/16.8**

Table 2: Performance (%) of state-of-the-art SGG models on GQA (Hudson and Manning 2019)

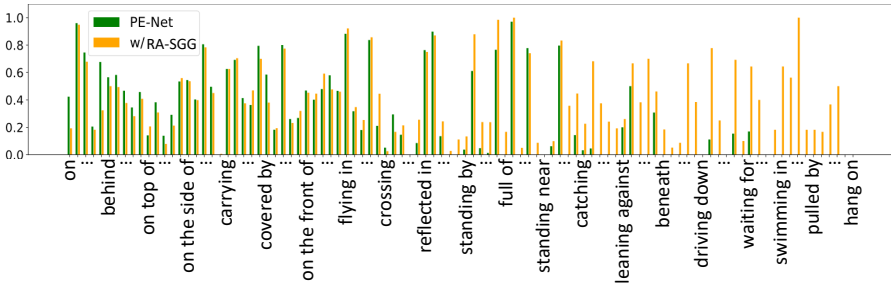

Figure 3: Class-wise comparison of RA-SGG with PE-Net on GQA under PredCls. Predicates are sorted by the frequency.

we set the threshold $\tau = 0.3$ at the reliable multi-labeled instance selection, and choose the number of retrievals K by performing the grid search in $\{1, 3, 5, 10, 20\}$.

Evaluation Protocol. We evaluate state-of-the-art models and RA-SGG on three conventional SGG tasks: (1) **Predicate Classification (PredCls)**, (2) **Scene Graph Classification (SGCls)**, and **Scene Graph Detection (SGDet)**. **PredCls** is the task that predicts the predicate classes given all ground-truth object bounding boxes and the object classes. **SGCls** aims at predicting the predicate classes given the ground-truth object bounding boxes. **SGGen** detects all entities and their pairwise predicates given an image.

Metrics. We evaluate SGG models on the three metrics: (1) Recall@K (R@K) calculates the proportion of top-K predicted triplets that are in ground truth. (2) Mean Recall@K (mR@K) calculates the average recall for each predicate class, which is designed to measure the performance of SGG models under the long-tailed predicate distribution. (3) Recent SGG studies suggest that there is a trade-off between R@K and mR@K (Kim et al. 2024a). Hence, recent works have focused on achieving the great F@K, where F@K calculates the harmonic average of R@K and mR@K.

Comparisons with State-of-the-art Methods

Results on Visual Genome. In Table 1, we compare RA-SGG applied to the PE-Net (Zheng et al. 2023) with state-of-the-art baselines. Based on the experimental result, we have summarized the following conclusions: **1) RA-SGG achieves state-of-the-art performance under conventional SGG tasks, particularly in terms of mR@K and F@K.** Our method not only shows better performance than the cutting-edge SGG models, such as HetSGG (Yoon et al. 2023) and SQUAT (Jung et al. 2023) by up to 6.2% in terms of F@K, but also achieves greater R@K. Compared to the backbone model, PE-Net, we greatly improve the performance in terms of mR@K, while slightly decreasing R@K. These results imply that RA-SGG's strategy that discovers the latent fine-grained predicates through the retrieval is effective in enhancing the model's generalization capability across all predicate categories without compro-

Task	Model	R@50/100	mR@50/100	F@50/100
PredCls	Vanilla PE-Net	64.9/67.2	31.5/33.8	42.4/45.0
	RA-SGG w/o select.	64.4/66.4	33.4/36.4	44.0/47.0
	RA-SGG w/o IPSS	64.6/66.7	32.9/35.1	43.6/46.0
	RA-SGG	62.2/64.1	**36.2/39.1**	**45.7/48.6**
SGCls	Vanilla PE-Net	39.4/40.7	17.8/18.9	24.5/25.8
	RA-SGG w/o select.	38.5/39.4	19.6/20.9	26.0/27.3
	RA-SGG w/o IPSS	38.6/39.5	18.6/19.8	25.1/26.3
	RA-SGG	38.2/39.1	**20.9/22.5**	**27.0/28.6**

Table 3: Ablation study of RA-SGG.

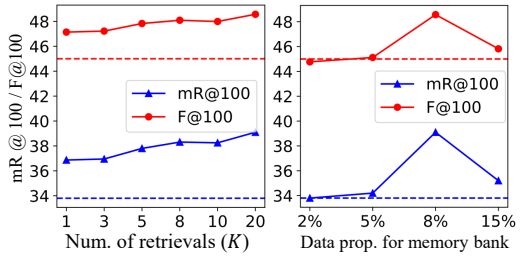

Figure 4: Sensitivity analysis on hyperparameters. The dotted line represents the result of PE-Net.

mising the recognition of general predicates. **2) Discovering the latent multi-label is a more effective strategy than data enhancement/augmentation approaches in addressing the challenges of SGG.** RA-SGG outperforms IE-Trans and CFA with the PE-Net backbone in terms of all metrics. We claim that it is attributed to the fact that while IE-Trans and CFA overlook the comprehensive understanding of general predicates to obtain the fine-grained predicates, our multi-label discovery effectively addresses both general and fine-grained predicates. This result indicates that the formulation of multi-label classification with partial annotation fundamentally addresses both the long-tailed problem and semantic ambiguity.

Results on GQA. In Table 2, we observed that RA-SGG greatly improves mR@K and F@K compared to the strong baseline, PE-Net, showcasing the effectiveness of RA-SGG on a large dataset, which includes numerous predicate categories. Similarly, Figure 3 demonstrates RA-SGG's ability to improve the recognition of fine-grained predicates, which PE-Net tends to overlook. These results imply that RA-SGG effectively discovers the latent fine-grained predicates even when using the dataset with numerous predicate categories.

Ablation Studies

Module Component Analysis. We conduct the ablation study of RA-SGG to evaluate the impact of each component. We ablate two components in RA-SGG. For RA-SGG w/o select., we skip the reliable multi-labeled instance selection, which determines whether a relation instance is single-labeled or multi-labeled. For RA-SGG w/o IPSS, we replace the inverse propensity score with a constant in the unbiased augmentation module. In Table 3, we obtained the following observations: 1) RA-SGG w/o select. still shows better performance than PE-Net in terms of mR@K and F@K. That is, the strategy, which roughly assigns pseudo-labels for all relations based on retrieved instances, can enhance the performance on fine-grained predicates. We argue that it is attributed to the superiority of the retrieval-augmented approach, which discovers semantically similar predicates using the memory bank. 2) RA-SGG w/o IPSS decreases mR@K and F@K compared to RA-SGG. This implies that inverse propensity score-based sampling is important for boosting the performance on fine-grained predicates. 3) Combining all components allows for the best generalization capability of RA-SGG.

Analysis on the Number of Retrievals. We analyze the effect of $K \in \{1, 3, 5, 8, 10, 20\}$ under PredCls task. In Figure 4.(a), we observe that RA-SGG consistently outperforms PE-Net, indicating that utilizing even one near neighbor is helpful for enhancing the performance of PE-Net. RA-SGG achieves the best performance at $K = 20$, implying that referring more neighbors is helpful for enhancing the performance on fine-grained predicates.

Analysis on the Size of Memory Bank. We anavlyze the impact of the memory bank size and vary the number of relation instances stored per unique triplet, testing with $\{1, 5, 10, 100\}$ instances for each unique triplet. These quantities represent approximately $\{2\%, 5\%, 8\%, 15\%\}$ of the entire training instances, respectively. In Figure 4.(b), we observed that RA-SGG generally outperforms the vanilla PE-Net, and attains the best performance when the memory bank comprises 8% of the training instances. This result suggests that leveraging a small portion of the training dataset can significantly improve the performance of SGG models. Moreover, the compact size of the memory bank not only enhances its efficiency but also contributes to the scalability of RA-SGG.

Performance of Retrieval Strategy. As there are no ground truth retrieval labels, it is difficult to measure the accuracy of retrievals. Hence, we conducted a human evaluation of the retrieval process on 100 sampled instances, and obtained 84.20% accuracy. This implies that our approach performs accurate retrieval and effectively augments the multi-labels.

6 Conclusion

Although research on fine-grained SGG has garnered its attention, fundamental challenges still remain in the prior framework. They either suffer from capturing the fine-grained predicates due to the SGG nature, which encompasses the long-tailed problem and semantic ambiguity, or overlook the general predicates by re-annotating the general predicates as fine-grained predicates. In this work, we first reformulate the SGG task as multi-label classification problem with partial annotation. To this end, we propose a novel RA-SGG, which enhances original labels to multi-labels through a retrieval process. RA-SGG effectively discovers latent fine-grained predicates in the training dataset, resulting in substantial improvements across overall predicates. Our works presents a new perspective in SGG research by considering the nature of natural language, where multiple predicates convey the same single relationship.

Acknowledgements

This work was supported by National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (NRF-2022M3J6A1063021, 10%, No. RS-2024-00341514, 10%), Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2023-00216011, Development of artificial complex intelligence for conceptually understanding and inferring like human, 79%) and (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University, 1%).

References

Ben-Baruch, E.; Ridnik, T.; Friedman, I.; Ben-Cohen, A.; Zamir, N.; Noy, A.; and Zelnik-Manor, L. 2022. Multi-Label Classification With Partial Annotations Using Class-Aware Selective Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4764–4772.

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-Augmented Diffusion Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 15309–15324. Curran Associates, Inc.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; De Las Casas, D.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J.; Elsen, E.; and Sifre, L. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2206–2240. PMLR.

Casanova, A.; Careil, M.; Verbeek, J.; Drozdzal, M.; and Romero-Soriano, A. 2021. Instance-Conditioned GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6163–6171.

Chen, W.; Hu, H.; Saharia, C.; and Cohen, W. W. 2023. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. In *The Eleventh International Conference on Learning Representations*.

Chiou, M.-J.; Ding, H.; Yan, H.; Wang, C.; Zimmermann, R.; and Feng, J. 2021. Recovering the Unbiased Scene Graphs from the Biased Ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 1581–1590. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Desai, A.; Wu, T.-Y.; Tripathi, S.; and Vasconcelos, N. 2021. Learning of Visual Relations: The Devil Is in the Tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15404–15413.

Dong, X.; Gan, T.; Song, X.; Wu, J.; Cheng, Y.; and Nie, L. 2022. Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation. In *CVPR*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Han, S.; Choi, E.; Lim, C.; Shim, H.; and Lee, J. 2022. Long-tail Mixup for Extreme Multi-label Classification. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, 3998–4002. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392365.

Hildebrandt, M.; Li, H.; Koner, R.; Tresp, V.; and Günnemann, S. 2020. Scene Graph Reasoning for Visual Question Answering.

Hudson, D. A.; and Manning, C. D. 2019. 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 (CVPR)*.

Jain, H.; Prabhu, Y.; and Varma, M. 2016. Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 935–944. New York, NY, USA: Association for Computing Machinery.

Jeon, J.; Kim, K.; Yoon, K.; and Park, C. 2024. Semantic Diversity-Aware Prototype-Based Learning for Unbiased Scene Graph Generation. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland.

Jung, D.; Kim, S.; Kim, W. H.; and Cho, M. 2023. Devil's on the Edges: Selective Quad Attention for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18664–18674.

Kaihua, T.; Yulei, N.; Jianqiang, H.; Jiaxin, S.; and Hanwang, Z. 2020. Unbiased scene graph generation from biased training. *CVPR*, 3716–3725.

Kim, K.; Yoon, K.; In, Y.; Moon, J.; Kim, D.; and Park, C. 2024a. Adaptive Self-training Framework for Fine-grained Scene Graph Generation. In *The Twelfth International Conference on Learning Representations*.

Kim, K.; Yoon, K.; Jeon, J.; In, Y.; Moon, J.; Kim, D.; and Park, C. 2024b. LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28306–28316.

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.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc.

Li, L.; Chen, G.; Xiao, J.; Yang, Y.; Wang, C.; and Chen, L. 2023. Compositional Feature Augmentation for Unbiased Scene Graph Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21685–21695.

Li, L.; Chen, L.; Huang, Y.; Zhang, Z.; Zhang, S.; and Xiao, J. 2022. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18869–18878.

Li, R.; Zhang, S.; Wan, B.; and He, X. 2021. Bipartite graph network with adaptive message passing for unbiased scene graph generation. *CVPR*.

Lin, X.; Ding, C.; Zhan, Y.; Li, Z.; and Tao, D. 2022. HL-Net: Heterophily Learning Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19476–19485.

Long, A.; Yin, W.; Ajanthan, T.; Nguyen, V.; Purkait, P.; Garg, R.; Blair, A.; Shen, C.; and van den Hengel, A. 2022. Retrieval Augmented Classification for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6959–6969.

Lyu, X.; Gao, L.; Guo, Y.; Zhao, Z.; Huang, H.; Shen, H. T.; and Song, J. 2022. Fine-Grained Predicates Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19467–19475.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*.

S. Hasegawa, A. N., M. Hiromoto; and Umeda, Y. 2022. Improving Predicate Representation in Scene Graph Generation by Self-Supervised Learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 2739-2748*.

Schroeder, B.; and Tripathi, S. 2020. Structured query-based image retrieval using scene graphs. *CVPR Workshops*, 178–179.

Shit, S.; Koner, R.; Wittmann, B.; Paetzold, J.; Ezhov, I.; Li, H.; Pan, J.; Sharifzadeh, S.; Kaissis, G.; Tresp, V.; and Menze, B. 2022. Relationformer: A Unified Framework for Image-to-Graph Generation. In *ECCV'2022*.

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.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 1492–1500.

Yang, J.; Lu, J.; Lee, S.; Batra, D.; and Parikh, D. 2018. Graph r-cnn for scene graph generation. *ECCV*, 128(7): 670–685.

Yoon, K.; Kim, K.; Moon, J.; and Park, C. 2023. Unbiased Heterogeneous Scene Graph Generation with Relation-aware Message Passing Neural Network. In *The 37th AAAI Conference on Artificial Intelligence*.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21560–21571.

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.

Zhai, R.; Dan, C.; Kolter, J. Z.; and Ravikumar, P. K. 2022. Understanding Overfitting in Reweighting Algorithms for Worst-group Performance.

Zhang, A.; Yao, Y.; Chen, Q.; Ji, W.; Liu, Z.; Sun, M.; and Chua, T.-S. 2022. Fine-Grained Scene Graph Generation with Data Transfer. In *ECCV*.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 (CVPR)*.