This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Mixture-of-Experts based Feature Decoupling for Open Vocabulary Scene Graph Generation

Yiming Li [1] Sisi You [1] Bing-Kun Bao [1,2,*]

[1]Hefei University of Technology, School of Computer Science and Information Engineering [2]PengCheng Laboratory

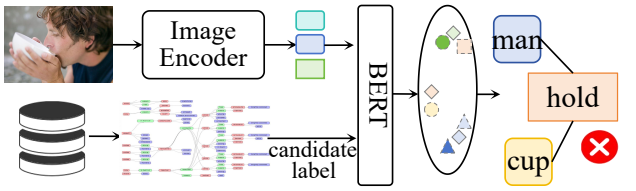
lymoon3129@gmail.com, ssyou.hfut@gmail.com, bingkunbao@hfut.edu.cn

Abstract

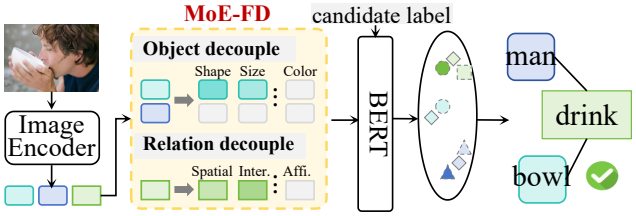
In recent years, while Scene Graph Generation has advanced significantly, mainstream methods remain constrained by predefined object and relationship categories, limiting generalization to open real-world scenarios. Inspired by open vocabulary object detection, recent efforts have expanded SGG to the open vocabulary domain. However, these models often rely on off-the-shelf VLMs, lacking discriminative attribute extraction and suffering from limited object-relationship semantic interaction, which leads to misclassification of novel categories. To address these issues, we propose the MoE Feature Decoupling (MoE-FD) framework for Open Vocabulary Scene Graph Generation. MoE-FD adaptively learns feature decoupling for objects and relationships via multiple experts, prioritizing critical features through gating network weights. Moreover, it models semantic interactions between objects and relationships using iterative cross-attention, enhancing relationship triple associations and visual-semantic alignment. The main contributions of MoE-FD are threefold: (1) A MoE-based feature decoupling framework that adaptively enhances discriminative feature representation for objects and relations. (2) Semantic interaction modeling between objects and relations to strengthen relationship triple associations and image-text alignment accuracy. (3) Extensive experiments demonstrate the effectiveness of MoE-FD on the Visual Genome dataset.

1. Introduction

The goal of Scene Graph Generation (SGG) is to parse an image into a structured representation that identifies objects and their semantic relationships as triplelets like $\langle subject, relationship, object \rangle$, typically formatted as a graph where nodes represent objects and edges denote the relationships between them. It is a promising way to represent the semantics of visual content, which can bridge the large gap between vision and natural language. In recent years, scene graph generation has attracted significant interest from researchers and demonstrated superior semantic representation capabilities in various downstream tasks, *e.g.*, image retrieval [13], image captioning [10, 19, 42, 43] and visual question answering [8, 9, 46].



(a) Framework of existing OVSGG method.



(b) Framework of the proposed method.

Figure 1. (a) Existing OVSGG methods simply utilize VLM models to align visual concepts with candidate labels, which easily leads to confusion between similar categories; (b) MoE-FD focuses on detailed information through feature decoupling, thereby enhancing the model's discriminative ability.

Although significant progress has been made in the field of scene graph generation, most mainstream methods are limited to predefined sets of object and relation categories, making SGG models difficult to generalize in open real-world scenarios. With the recent proposal of open vocabulary tasks, particularly inspired by the field of open vocabulary object detection [7, 17], some recent methods [3, 11, 16, 18] have begun to expand the SGG task from closed sets to the open vocabulary domain. In particular, Chen *et al.* [3] have explored fully open vocabulary scene graph generation, a challenging task where both objects and relationships are novel. However, as shown in Figure 1, most of these models only utilize off-the-shelf VLMs to

*indicates corresponding author:Bing-Kun Bao.

obtain visual features, lacking discriminative attribute extraction for novel objects and relationships, leading to the misclassification of novel classes as similar base classes. Meanwhile, these methods usually rely on similarity calculations in the semantic space to independently assign labels to objects and relationships, resulting in a lack of interaction between object-relationship semantics, leading to limited diversity and accuracy in image-text alignment.

To address the above issues, an intuitive approach is to perform attribute decoupling on the visual features of objects and relationships, selecting discriminative parts as feature representations to facilitate the distinction between similar categories. Meanwhile, the decoupled attributes also facilitate modeling interactions between objects and relations, for example, when distinguishing between "cup" and "bowl" in an image, decoupling attributes such as "shape" and "size" allows for accurate differentiation, and further enables modeling relational interactions like "man drinks from bowl ". However, in the real world, objects have diverse attributes, and abstract relational attributes are even more difficult to define, leading to difficulties for models in acquiring effective training data.

In response to the above challenge, we propose the **M**ixture-**o**f-**E**xperts **F**eature **D**ecoupling (MoE-FD) for Open Vocabulary Scene Graph Generation (OVSGG) framework, which adaptively learns the ability of feature decoupling for objects and relations by multiple experts, and focuses on the most important part by calculating the weight of each expert with a routing network. Moreover, we model the semantic interaction between objects and relationships after the feature decoupling with experts, thereby improving the prediction capability for novel classes without attribute annotations. In detail, after extracting visual and textual features using existing VLMs, we set up multiple expert layers to decouple object/relationship features and select salient attributes via a routing network, thereby enhancing the discriminability of object/relation features. In addition, we introduce structured semantic priors, *e.g.*, ConceptNet [33], into the routing network to guide our model to select more relevant experts. Subsequently, an iterative cross-attention mechanism is utilized to model the semantic interaction between objects and relationships, thus strengthening the semantic associations within relationship triples step-by-step. Finally, we align visual concepts with semantic labels in the semantic space. Furthermore, following OvSGTR [3], we also adopt relation-aware weakly supervised pre-training to improve the model's ability to perceive novel categories.

The main contributions are summarized as follows:

1. We propose a feature decoupling framework based on MoE, which could adaptively learn the attribute decoupling capabilities of objects and relationships, thus enhancing the discriminability of features.
2. The proposed MoE method could guide semantic interaction between objects, therefore strengthening the semantic association of relationship triples and improving the accuracy of image-text alignment.
3. The evaluation on the Visual Genome datasets verifies the effectiveness of the proposed MoE-FD.

2. Related Work

2.1. Scene Graph Generation

Since its proposal, Scene Graph Generation (SGG) has received widespread attention. Early works [4, 22, 36, 42] focused on traditional fully supervised settings, these approaches are limited by dataset annotations, making them difficult to apply to real-world scenarios. Therefore, some researchers [23, 35] have begun to explore extending SGG to a generalized filed. For instance, Tang *et al.* [35] attempted to achieve zero-shot relation inference through causal reasoning. However, these early explorations only tried to predict relation triples that do not fully appear in the training set, which means the elements in the triples are not completely novel. Due to this imprecise definition, their methods cannot substantially generalize to novel categories.

In recent years, with the proposal of open vocabulary tasks, especially the development of open vocabulary object detection, some researchers [3, 11, 16, 27, 47] have shifted their focus to open vocabulary scene graph generation (OVSGG). These works can be roughly divided into two categories: methods based on Vision-Language Models (VLM) and methods based on Multimodal Large Language Models (MLLM). Among them, VLM-based methods typically rely on the cross-modal mapping capabilities of VLMs, using models such as CLIP [29] or Grounding DINO [37] to map visual concepts and semantic labels into the same space, thereby enabling the prediction of novel classes. For example, He *et al.* [11] conduct OVSGG via visual-relation pre-training and prompt-based fine-tuning, Zhang *et al.* [47] construct a visual-semantic space through CLIP to align visual concepts and relational semantics. Furthermore, Li *et al.* [18] and Chen *et al.* [3] consider employ image caption to enhance the semantic context for OVSGG. Li *et al.* [16] further explore an interaction-aware fine-tuning strategy to focus on the most important objects in scenes. However, these methods only utilize pre-trained VLMs to process visual features and lack the ability to capture visual details, thus easily leading to confusion between similar visual concepts.

In contrast, MLLM-based models consider leveraging the sequence prediction capabilities of MLLMs (*e.g.,* BLIP [15] or LLaVA [25]) to transform scene graph generation into an image-based structured sequence prediction problem. For example, Zhou *et al.* [51] adopt an iterative prediction mechanism, which could reason fine-grained re-

lationships for OVSGG. Wang *et al.* [38] employ an instruction fine-tuning strategy based on LLaVA to unify text generation, object localization, and relation comprehension. However, these models require substantial computational resources, making them difficult to apply conveniently. Therefore, this paper focuses on VLM-based methods. Among these previous work, the most similar one to our proposed method is [3], where our main difference lies in the introduction of MoE-based feature decoupling, thereby enabling the active discovery of key details.

2.2. Mix-of-Expert Methods

Mixture of Experts (MoE) networks [12] have been proposed for decades and successfully applied to various network enhancement tasks, including image recognition [31, 40], video captioning [39], and person re-identification [5]. Their initial design consists of a set of expert networks, where each expert learns a subtask for the entire dataset. Each expert either shares the same weights or employs different gating schemes to selectively deactivate parts of the network. For example, partition and collaborative training strategies were proposed to enhance the learning capability of the original network [48]. In recent years, researchers have also attempted to apply MoE to scene graph generation. For instance, Zhou *et al.* [50] addressed unbiased scene graph generation through joint prediction by multiple experts.

The main difference between our proposed method and theirs is that we consider the more challenging fully open-vocabulary scene graph task and decouple visual features via MoE to improve the model's discriminative ability across different categories, rather than directly using MoE for prediction.

3. Method

As shown in Figure 2, to conduct the open vocabulary scene graph generation (OVSGG) for the given image I, we take I and possible object/relationship categories as input, then adopt the MoE feature decoupling module to learn the discriminative features. After that, we utilize the iterative cross-attention to model interactions between objects and relationship features. Finally, visual concepts and labels are aligned in a visual semantic space. We will provide the problem definition and detailed introduction of each component in the following.

3.1. Problem Definition

Given an image I, OVSGG aims to construct a graph structure $G = \{V, E\}$ to represent the visual concept within I, where V and E denote the object node and relationship edge set, respectively. Specifically, to generate scene graphs containing object and relationship categories beyond the training vocabulary, we define three critical settings based on vocabulary visibility: *OvD-SGG*, *OvR-SGG*, and *OvD+R-SGG*. In detail, with defining the training set and test set of objects and relationships as $V_{obj}^{train}, V_{rel}^{train}, V_{obj}^{test}$, and V_{rel}^{test}, *OvD-SGG* involves novel objects with fixed relations, which can be formulated as $V_{obj}^{test} \supset V_{obj}^{train}$, $V_{rel}^{test} = V_{rel}^{train}$, requiring detection of objects from V_{obj}^{test} while predicting relationship only from training set. OvR-SGG refers to novel relationships with fixed objects, defined by $V_{rel}^{test} \supset V_{rel}^{train}$, $V_{obj}^{test} = V_{obj}^{train}$, involving prediction of relationships from $V_{rel}^{test} \setminus V_{rel}^{train}$, using only training object categories. OvD+R-SGG entails novel objects and relations simultaneously, expressed as $V_{obj}^{test} \supset V_{obj}^{train}$, $V_{rel}^{test} \supset V_{rel}^{train}$, demanding handling of both $V_{obj}^{test} \setminus V_{obj}^{train}$ objects and $V_{rel}^{test} \setminus V_{rel}^{train}$ relations in scene graph generation.

3.2. Feature Extraction

MoE-FD takes the image and class prompt as input, where the class prompt is structured by concatenating all the possible object and relationship categories following [3, 17, 26]. Furthermore, during training, to avoid the difficulty in learning caused by an excessively large vocabulary set, we randomly sample negative words and fix the total number of positive and negative words to N. Then we extract visual and textual features with frozen image backbone and text encoder, *e.g.*, Swin Transformer [28] and BERT [6]. After obtaining embedded visual and textual features, a transformer-based context extractor is adopted to fuse information from both modalities and capture contextual information. Following DETR [1], we randomly initialize a query sequence to learn features of different objects in the visual scene, where the length of the object queries is fixed to K. Furthermore, we set a relation query for each potential relationship, which is used in the decoding stage of the context extractor to compute cross-attention with visual and text features, thereby learning relation-aware contextual information.

After passing through the context extractor, K object queries $\{v_i\}_K$ and relation query $\boldsymbol{r}$ capture the object and relationship context in the visual scene, and we then use them to construct features of nodes and edges in the scene graph. For the i-th object, its corresponding node feature can be constructed as $n_i = [v_i, p_i]$, where p_i is a 4-d position vector decoded by a bounding-box head similar as DETR, and $[,]$ is concatenate operation. Edge features are constructed by combining corresponding object features and the relation-aware query. For example, the feature of an edge connecting the i-th and j-th object can be constructed as $e_{i,j} = \varphi([n_i, n_j, \boldsymbol{r}])$, where $\varphi(\cdot)$ is a two-layer MLP.

3.3. Feature Decoupling via MoE

Traditional feature extraction methods typically treat features as a single representation and fail to capture fine-

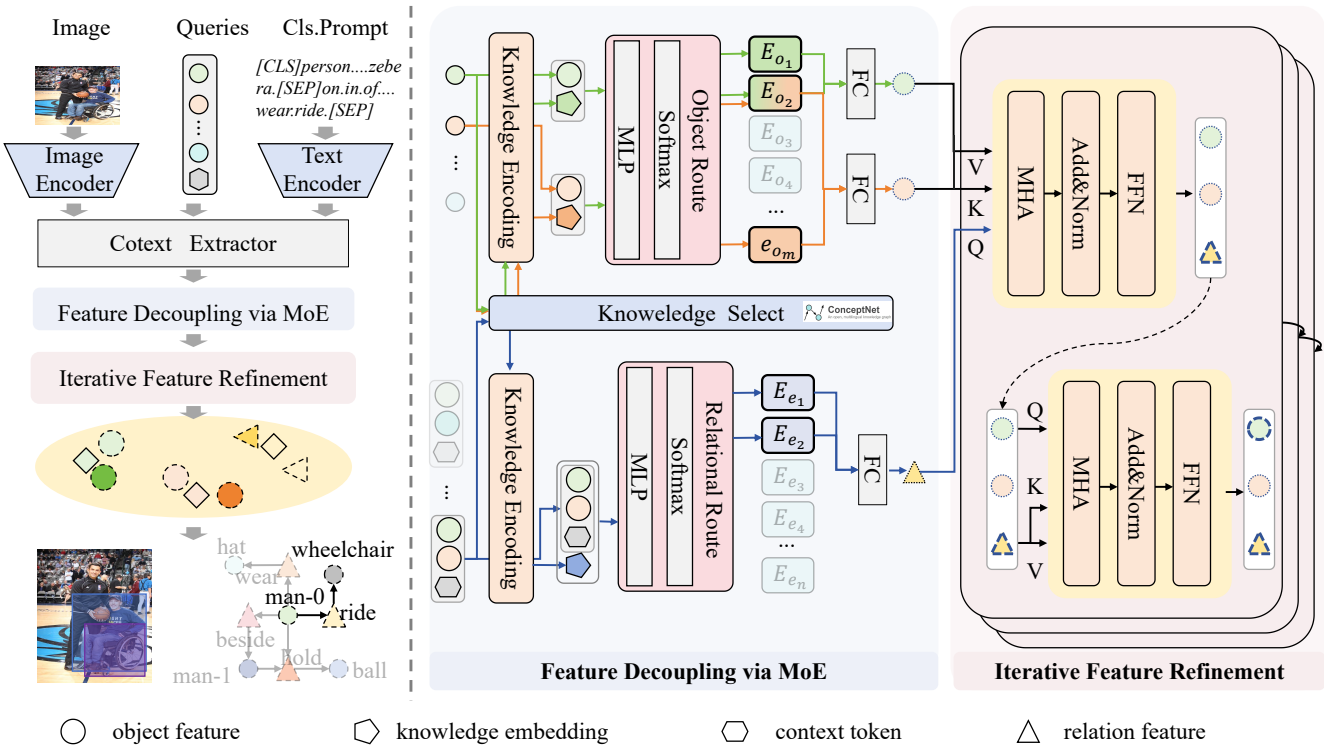


Figure 2. Left: the framework of proposed MoE-FD; Right: the detail of two main components, *i.e.*, Feature Decoupling via MoE and Iterative Feature Refinement.

grained visual semantic differences, such as object shape and texture or spatial and semantic relationships between objects. To address this limitation and enhance the model's generalization ability to novel categories, we propose a dynamic feature decoupling framework based on the Mixture of Experts (MoE) structure, named MoE-FD.

MoE-FD decouples features with two components: object feature decoupling and relationship feature decoupling. Specifically, we first design a routing network that adaptively selects experts based on input node (object) / edge (relationship) features and semantic prior, e.g., ConceptNet. This dynamic selection mechanism ensures that each feature type is processed by the most relevant expert, thereby improving feature specificity and discriminability.

Firstly, object feature decoupling disentangles object-specific visual-semantic attributes, *e.g.*, shape, texture, and functionality, via a knowledge-guided route network to capture diverse object characteristics. Given a node feature n_i, we first retrieve related nodes in ConceptNet with semantic similarity greater than ϵ based on the pseudo-labels provided by the VLM, extract the neighboring nodes and edges within two hops of these related nodes as the semantic prior subgraph $C_{k,i}$, and then encode $C_{k,i}$ into a 1024-dimensional continuous vector $\hat{C}_{k,i}$. Then we concatenate $C_{k,i}$, n_i, and relation query $\boldsymbol{r}$ to compute the expert weights as follows:

$$\begin{aligned} z_o &= \text{MLP}_{\text{Route}}([n_i, C_{k,i}, \boldsymbol{r}]), \\ \alpha_k &= \text{Softmax}(z_o)_k \quad \forall k \in \{1, ..., E_o\}, \end{aligned} \tag{1}$$

where E_o is the number of object experts, and α_k denotes the weight of the k-th expert. Specifically, the relation-aware query $\boldsymbol{r}$ is adopted to enable the routing network could perceive context information, while the semantic prior $\hat{C}_{k,i}$ is utilized to introduce knowledge with novel classes. The node feature is then computed as a weighted sum of expert outputs:

$$n_i^* = \sum_{k=1}^{E_o} \alpha_k \cdot \text{Expert}_k^{\text{obj}}(n_i), \tag{2}$$

where $\text{Expert}_k^{\text{obj}}$ is a 3-layer MLP with ReLU activation, specializing in a unique visual-semantic subspace, *e.g.*, shape, texture, or function. Secondly, relationship feature decoupling decomposes inter-object relational cues, such as spatial layout and semantic interactions, through expert collaboration to enhance context-dependent relationship understanding. Given the feature of relationship edge $e_{i,j}$, the routing network similarly outputs weights β_m for E_r rela-

tion experts and update the edge feature as follows:

$$\begin{aligned} z_r &= \text{MLP}_{\text{Route}}(e_{i,j}, \hat{C}_{k,ij}), \\ \beta_m &= \text{Softmax}(z_r)_m \quad \forall m \in \{1, ..., E_r\}, \\ e^*_{i,j} &= \sum_{m=1}^{E_r} \beta_m \cdot \text{Expert}^{\text{rel}}_m(e_{i,j}), \end{aligned} \tag{3}$$

where $\hat{C}_{k,ij}$ is the semantic prior extracted based on both n_i and n_j.

3.4. Iterative Feature Refinement

To further enhance the discriminative power of node/edge features, we design a cross-module interaction mechanism that enables dynamic information exchange between object and relationship experts, which is conducted in an iterative way. Moreover, for simplicity, here we still denote the reconstructed node and edge features as n_i and $e_{i,j}$

In the first stage, edge features are updated using cross-attention over node pairs. Given node feature $\{n_i\}_{i=1}^K$ and edge features $\{e_{i,j}\}_{i,j=1}^K$, we compute the attention weight between node i and node j to capture their pairwise interaction:

$$w_{i,j} = \text{Softmax}\left(\frac{\varphi_Q(n_i) \cdot \varphi_K(n_j)^T}{\sqrt{d}}\right), \tag{4}$$

where $\varphi_Q(\cdot)$ and $\varphi_K(\cdot)$ are linear projections for query and key, and d is the feature dimension. The updated edge feature $e'_{i,j}$ is then obtained by aggregating node i's value features weighted by $w_{i,j}$:

$$e'_{i,j} = e_{i,j} + \text{MLP}_{\text{edge}}\left(w_{i,j} \cdot \phi(n_i + n_j)\right), \tag{5}$$

where $\phi(\cdot)$ is the value projection, and MLP_{edge} is a projection layer. This step ensures edges encode fine-grained node interactions, *e.g.*, "person" and "skateboard" contribute to the "ride" relation edge.

In the second stage, node features are refined using the updated edge features from the first stage. For each node i, we compute attention weights over all adjacent edges $\{e'_{i,j}\}_{j=1}^K$ to model edge-to-node semantic feedback:

$$\gamma_{i,j} = \text{Softmax}\left(\frac{\varphi'_Q(e'_{i,j}) \cdot \varphi'_K(n_i)^T}{\sqrt{d}}\right), \tag{6}$$

where $\varphi'_Q(\cdot)$ and $\varphi'_K(\cdot)$ are edge-specific query/key projections. The refined node feature n'_i is computed by aggregating edge features weighted by $\beta_{i,j}$:

$$n'_i = n_i + \text{MLP}_{\text{node}}\left(\sum_{j=1}^{K} \gamma_{i,j} \cdot \phi'(e'_{i,j})\right), \tag{7}$$

where $V'(\cdot)$ and MLP_{node} are projection layers. This bidirectional update ensures nodes adapt to relational semantics. Furthermore, we iterate these two stages C times and use the final output for object/relationship classification.

3.5. Open-Vocabulary Classification for Objects and Relationships

Following previous work [3], we achieve open vocabulary scene graph generation by matching node/edge features with semantic concepts in the semantic space. For simplicity, here we still denote the refined node and edge features as n_i and $e_{i,j}$.

For object node classification, we first compute the similarity between each node feature n_i and the word embedding. Defining the word embedding as b_{g_k}, which is provided by pre-trained models like CLIP [29], the similarity is defined as $\text{sim}_{\text{obj}}(n_i, g_k) = \sigma\left(\langle b_{g_k}, n_i \rangle\right)$, where $\langle \cdot, \cdot \rangle$ denotes the dot product and σ is the sigmoid function. To handle multiple objects in a scene, we model classification as a bipartite graph matching problem between K predicted nodes $\{\tilde{n}_i\}_{i=1}^K$ and N ground-truth objects $\{g_k\}_{k=1}^N$, optimizing the binary mask $\boldsymbol{M} \in \{0,1\}^{K \times N}$ to maximize the total similarity $\max_{\boldsymbol{M}} \sum_{i=1}^K \sum_{k=1}^N \text{sim}_{\text{obj}}(\tilde{n}_i, g_k) \cdot \boldsymbol{M}_{i,k}$.

For relation edge classification, we align edge features $e_{i,j}$ with relation semantics via cross-modal scoring. Given an edge feature $e_{i,j}$ between nodes i and j, we first encode candidate relationship into contextual embedding t_{emb} using a pre-trained language encoder, then project these embeddings to the same space as $e_{i,j}$ via a linear transformation $t_{\text{proj}} = W_t t_{\text{emb}}$, where W_t is a learnable projection matrix. The alignment score between $e_{i,j}$ and the text query is computed as $s(e_{i,j}, t_{\text{proj}}) = e_{i,j}^T t_{\text{proj}}$, which is then normalized to a probability via sigmoid activation.

During training, we construct a balanced set of positive and negative samples, and optimize the binary cross-entropy loss $\mathcal{L} = -\mathbb{E}[y \log \sigma(s) + (1-y)\log(1-\sigma(s))]$, where $y \in \{0,1\}$ denotes the sample label. Furthermore, to alleviate data sparsity for rare relations, we follow [3] to incorporate a weakly supervised learning paradigm. In detail, relation triplets are automatically parsed from image captions using a dependency parser, then aligned with detected objects via the bipartite matching strategy.

4. Experiments

4.1. Experimental Settings

Dataset. We use the widely used VG150 [14] dataset to evaluate our model, which contains 150 object categories and 50 relationship categories, with a total of 108,777 images. Of these, 70% are used for training and 30% for testing. Since we use Grounding DINO [26] as the detector, to avoid information leakage, we follow previous work [3, 47] and exclude images used in pre-training, ultimately retaining 14,700 images for testing. In weakly supervised learning stage, we use scene graph parser [32] to extract relational triples from image descriptions, thereby obtaining 44k objects and 2.5k relational pseudo-labels based on the COCO caption training set, with a total of 117k images.

Method	Backbone	Base+Novel Obj.			Novel Obj.		
		R@20	R@50	R@100	R@20	R@50	R@100
IMP [41]	-	-	2.85	3.43	-	0.00	0.00
MOTIFS [45]	-	-	3.35	3.86	-	0.00	0.00
VCTREE [35]	-	-	3.56	4.05	-	0.00	0.00
TDE [35]	-	-	3.50	4.07	-	0.00	0.00
VS^3 [47]	Swin-T	-	15.07	18.73		10.08	13.65
OvSGTR [3]		-	18.14	23.20	-	12.06	16.49
Ours		**18.35**	**23.39**	**28.33**	**12.92**	**18.13**	**22.72**
OvSGTR [3]	Swin-B	-	21.35	26.22	-	15.58	19.96
Ours		**21.84**	**26.64**	**31.47**	**16.20**	**20.94**	**25.74**

Table 1. Experimental results of OvD setting on VG test set.

Method	Backbone	Base+Novel Rel.			Novel Rel.		
		R@20	R@50	R@100	R@20	R@50	R@100
IMP [41]	-	-	12.56	14.65	-	0.00	0.00
MOTIFS [45]	-	-	15.41	16.96	-	0.00	0.00
VCTREE [35]	-	-	15.61	17.26	-	0.00	0.00
TDE [35]	-	-	15.50	17.37	-	0.00	0.00
VS^3 [47]	Swin-T	-	15.60	17.30	-	0.00	0.00
OvSGTR [3]		-	20.46	23.86	-	13.45	16.19
RAHP [27]		-	20.50	25.74	-	15.59	19.92
ACC [16]		17.49	23.22	27.40	12.90	17.89	21.70
Ours		**20.02**	**26.23**	**30.66**	**16.01**	**21.24**	**25.40**
OvSGTR [3]	Swin-B	-	22.89	26.65	-	16.39	19.72
ACC [16]		18.77	24.81	29.28	14.72	20.04	24.66
Ours		**21.77**	**28.53**	**32.94**	**18.06**	**23.06**	**27.60**

Table 2. Experimental results of OvR-SGG setting on VG test set.

Metrics. Following previous work in scene graph parsing [4, 21, 22, 42, 45], we adopt the popularly used metrics of Recall@K (K=20,50,100). Moreover, to realistically simulate real-world scenarios, we only adopted the SG DET protocol and abandoned the Pred Cls and SG Cls protocols that use ground truth information in prediction. As introduced in the method, we evaluate MoE-FD under three settings, namely OvD-SGG, OvR-SGG, and OvD+R-SGG. Among them, OvD-SGG and OvR-SGG include novel objects and novel categories respectively during testing, while the OvD+R-SGG setting is more challenging, requiring the model to predict both novel objects and relationships from the training set during testing.

Implementation details. The proposed method is implemented by PyTorch, trained and evaluated on 8$\times$ NVIDIA 4090 GPUs. We initialize our model based on pre-trained Grounding DINO with frozen visual backbone (*i.e.,* Swin-T or Swin-B) and text encoder (*i.e.*, BERT-base), while other modules are randomly initialized. In training and testing, we set 100 object queries per image for pairwise relationship recognition. Following previous DETR-like methods [26, 52], we use L1 loss and GIoU loss [30] for bounding box regression. For object or relation classification, we use Focal Loss [24] as the contrastive loss between prediction and language tokens. In the MoE feature decoupling module, based on experiments, we set 8 experts for object decoupling and 6 experts for relationship decoupling, while the knoweledge selecting threshold ϵ is set to 0.7. In addition, to enhance the interaction between object features and relationship features, we iterated cross-attention four times, meaning both object features and relationship features are updated twice each. During the training stage, we use the SGD optimizer with an initial learning rate of 0.001 and decay the learning rate by multiplying it by 0.9 after every epoch. The momentum is set to 0.9, and the size of the mini-batch is set to 8.

Baselines. We compare our model with several state-of-the-art open vocabulry scene graph generation algorithms, *e.g.*, VS^3 [47], OvSGTR [3], and ACC [16]. In addition, we also compared some classic close-set scene graph genera-

Method	Backbone	Joint Base+Novel			Novel Obj.			Novel Rel.		
		R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
IMP [41]	-	-	0.77	0.94	-	0.00	0.00	-	0.00	0.00
MOTIFS [45]	-	-	1.00	1.12	-	0.00	0.00	-	0.00	0.00
VCTREE [35]	-	-	1.04	1.17	-	0.00	0.00	-	0.00	0.00
TDE [35]	-	-	1.00	1.15	-	0.00	0.00	-	0.00	0.00
VS^3 [47]	Swim -T	-	5.88	7.20	-	0.00	0.00	-	0.00	0.00
OvSGTR [3]		10.02	13.50	16.37	10.56	14.32	17.48	7.09	9.19	11.18
ACC [16]		12.61	17.43	21.27	12.48	17.16	21.10	11.38	15.90	19.46
Ours		**14.88**	**20.90**	**25.24**	**14.88**	**20.78**	**24.74**	**14.49**	**19.14**	**23.42**
OvSGTR [3]	Swim -B	12.37	17.00	21.03	12.63	17.58	21.70	10.56	14.62	18.22
ACC [16]		13.50	18.88	23.19	13.46	18.84	23.29	12.37	17.50	21.73
Ours		**17.35**	**23.15**	**26.97**	**16.99**	**23.01**	**26.64**	**16.61**	**20.90**	**24.19**

Table 3. Experimental results of OvD+R-SGG setting on VG test set.

Method	Backbone	Grounding	R@20	R@50	R@100
LSWS [44]	-	-	-	3.28	3.69
MOTIFS [45]	-	Li et al. [20]	5.02	6.40	7.33
Uniter [2]	-	SGNS [49]	-	5.80	6.70
Uniter [2]	-	Li et al. [20]	5.42	6.74	7.62
VS^3 [47]	Swin-T	GLIP-L [17]	5.59	7.30	8.62
OvSGTR [3]		Grounding DINO [26]	6.61	8.92	10.90
ACC [16]		Grounding DINO [26]	7.86	10.81	13.31
Ours		Grounding DINO[26]	**8.56**	**11.63**	**14.41**
OvSGTR [3]	Swin-B	Grounding DINO[26]	6.88	9.30	11.48
ACC		Grounding DINO[26]	8.28	11.61	14.33
Ours		Grounding DINO[26]	**9.14**	**12.41**	**16.40**

Table 4. Comparison with pre-training methods. All models are pre-trained on image-caption data and tested on VG150 test set directly. Our models trained on COCO captions are used as pre-trained models for OvD+R-SGG settings.

tion methods (*e.g.*, IMP [41], MOTIFS [45], VCTREE [34], and TDE [35]) to demonstrate the superiority of the open vocabulary methods in novel class prediction.

4.2. Comparision with State-of-the Art

OvD-SGG Setting. Following the previous work [3, 47], we select 70% object categories in VG150 as base classes and the remaining 30% as novel classes, where the training set contains 50,107 images after excluding novel objects. We report the performance of OvD-SGG on "Base+Novel" and "Novel" in Table 1. It can be observed that the proposed MoE-FD significantly outperforms previous methods, especially in the recognition of novel classes. Compared with OvSGTR, MoE-FD achieves an improvement of 5.36%/5.78% on R@50/R@100. We attribute this improvement to that the proposed object feature decoupling and refinement effectively enhance the discriminative ability of our model.

OvR-SGG Setting. Similar to OvD-SGG, we remove images containing novel relation categories from the training set and report the performance of OvR-SGG in Table 2. It can be observed that our model significantly outperforms previous methods, which proves that our method effectively captures the discriminative features of relationships. Meanwhile, it can be found that the performance difference when using different backbones is smaller than that in OvD-SGG. We attribute this to the fact that the OvR-SGG setup has less impact on object classification, and more accurate object classification obviously provides help for SGG.

OvD+R-SGG Setting. In this setup, we train the model using the base classes of OvD-SGG and OvR-SGG. Ultimately, the number of available training images in VG150 is 36,425. As shown in Table 3, our model outperforms previous methods in all three items (Joint Base+novel, Novel Obj., and Novel Rel.), especially in relationship classification, where it outperforms ACC by 4.24% (R@20-Novel Rel.). Moreover, we also the pre-training performance of MoE-FD. As shown in Table 4, with feature decoupling, MoE-FD consistently outperforms other methods, which

FD(Obj)	FD(Rel)	IFR	Joint Base+Novel			Novel Obj.			Novel Rel.		
			R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
✓		✓	16.09	21.91	25.58	15.48	21.60	25.51	15.21	19.42	22.65
	✓	✓	15.72	21.17	25.15	15.24	21.39	25.12	14.93	19.14	22.51
✓	✓		17.03	22.70	26.66	16.31	22.61	26.22	15.98	20.62	23.76
✓	✓	✓	**17.35**	**23.15**	**26.97**	**16.99**	**23.01**	**26.64**	**16.61**	**20.90**	**24.19**

Table 5. Experimental results of ablation feature decoupling experts of object (FD(Obj)), relationship (FD(Rel)), and iterative feature refinement(IFR) module.

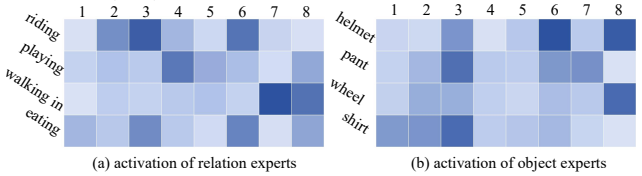


Figure 3. Visualized activation of relation/object experts.

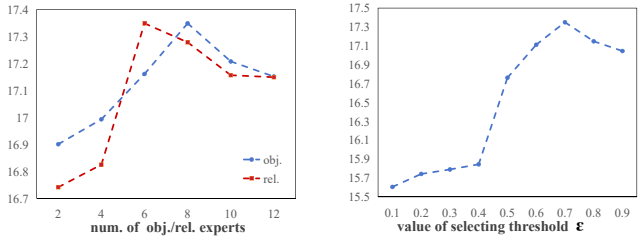


Figure 4. Experimental results of different numbers of feature decoupling experts and value of ϵ on OvD+R-R@20 setting.

demonstrates the effective generalization of our proposed MoE-FD in a fully open vocabulary setting.

4.3. Ablation Study

To demonstrate the effectiveness of each component, we further conducted ablation experiments as follows:

Impact of Feature Decoupling via MoE. To demonstrate the effectiveness of our proposed feature decoupling method, we evaluated the performance of MoE-FD after removing object/relationship feature decoupling MoE module. As shown in Table 5, the performance of MoE-FD significantly decreases when either obj.fd or rel.fd is removed, with a more pronounced decline observed when obj.fd is removed. We attribute this to the fact that the accuracy of object classification greatly influences relationship classification, which has also been mentioned in previous work [45].

Impact of Iterative Feature Refinement. To evaluate the impact of the Iterative Feature Refinement module on model performance, we also conducted experiments. As shown in Table 5, removing this module impaired the model's performance, which demonstrates that the interaction between object and relationship features is necessary for scene graph generation.

Impact of number of experts To show the impact of different numbers of experts on model performance, Figure 4 reports the performance variations of MoE-FD as the number of object/relationship experts changes. It can be observed that MoE-FD achieves the best performance when the number of object experts is 8 and the number of relationship experts is 6.

Impact of each expert To verify that different relationship/object experts are indeed effective in feature decomposition, we present the weight activations of different experts across several category samples in Figure 3. It can be observed that when processing different relationship/object categories, the knowledge-guided routing network exhibits significant differences in the activation of experts.

Impact of knowledge selection threshold ϵ We tested the influence of different ϵ values on model performance, as shown in Figure 4, where MoE-FD achieves the best performance when $\epsilon = 0.7$.

4.4. Discussion and Future Work

Although our proposed MoE-FD has significantly outperformed previous state-of-the-art methods, we have identified limitations in its design. Since the number of object queries and experts is fixed, MoE-FD may struggle to adapt to scenarios with a broader range of novel categories. Therefore, in future work, we plan to introduce learnable expert controllers. For example, selecting expert subsets via reinforcement learning methods or leveraging large language models to guide the dynamic selection of experts.

5. Conslusion

In this work, we propose a MoE-based feature decoupling approach, MoE-FD, specifically designed for OVSGG, which includes two key main components, *i.e.*, feature decoupling and iterative feature refinement. The proposed MoE-FD employs a set of specialized experts to extract and dynamically enhance the critical details in visual features via a gating network, and conduct adaptive interaction among object nodes and relational edges through iterative cross-attention mechanisms. We conduct comprehensive empirical experiments to show that the proposed method significantly outperforms existing SOTA methods, and systematic ablation studies further demonstrate that each core component within MoE-FD contributes substantially to its effectiveness.

Acknowledgment This work was supported by the National Natural Science Foundation of China under Grants (No.62502231, 62325206,62532003,62301276), the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, the China Postdoctoral Science Foundation (No.2025M771519, 2025M773507).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 7

[3] Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Chang Wen Chen. Expanding scene graph boundaries: Fully open-vocabulary scene graph generation via visual-concept alignment and retention. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024. 1, 2, 3, 5, 6, 7

[4] Weilin Cong, William Wang, and Wang-Chien Lee. Scene graph generation via conditional random fields. *arXiv preprint arXiv:1811.08075*, 2018. 2, 6

[5] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16145–16154, 2021. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3

[7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14084–14093, 2022. 1

[8] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. *arXiv preprint arXiv:2007.08751*, 2020. 1

[9] Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*, 2019. 1

[10] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *ICCV*, pages 10323–10332, 2019. 1

[11] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European conference on computer vision*, pages 56–73. Springer, 2022. 1, 2

[12] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015. 1

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 5

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[16] Lin Li, Chuhan Zhang, Dong Zhang, Chong Sun, Chen Li, and Long Chen. Interaction-centric knowledge infusion and transfer for open-vocabulary scene graph generation. *arXiv preprint arXiv:2511.05935*, 2025. 1, 2, 6, 7

[17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 1, 3, 7

[18] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28076–28086, 2024. 1, 2

[19] X. Li and S. Jiang. Know more say less: Image captioning based on scene graphs. *IEEE TMM*, 21(8):2117–2130, 2019. 1

[20] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4204–4213, 2022. 7

[21] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, pages 1347–1356, 2017. 6

[22] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, pages 1261–1270, 2017. 2, 6

[23] Yiming Li, Xiaoshan Yang, Xuhui Huang, Zhe Ma, and Changsheng Xu. Zero-shot predicate prediction for scene graph parsing. *IEEE Transactions on Multimedia*, 25:3140–3153, 2022. 2

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3, 5, 6, 7

[27] Tao Liu, Rongjie Li, Chongyu Wang, and Xuming He. Relation-aware hierarchical prompt for open-vocabulary scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5576–5584, 2025. 2, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 5

[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[31] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 3

[32] SceneGraphParser. Scenegraphparser. https://github.com/vacancy/SceneGraphParser, 2022. Accessed: 2025-08-02. 5

[33] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451. AAAI Press, 2017. 2

[34] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 7

[35] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 2, 6, 7

[36] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13688–13697, 2021. 2

[37] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 2

[38] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024. 3

[39] Xin Wang, Jiawei Wu, Da Zhang, Yu Su, and William Yang Wang. Learning to compose topic-aware mixture of experts for zero-shot video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8965–8972, 2019. 3

[40] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 3

[41] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 6, 7

[42] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene graph captioner: Image captioning based on structural visual representation. *JVCIR*, 58:477–485, 2019. 1, 2, 6

[43] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019. 1

[44] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299, 2021. 7

[45] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 6, 7, 8

[46] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*, 2019. 1

[47] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2915–2924, 2023. 2, 5, 6, 7

[48] Shuai Zhao, Liguang Zhou, Wenxiao Wang, Deng Cai, Tin Lun Lam, and Yangsheng Xu. Toward better accuracy-efficiency trade-offs: Divide and co-training. *IEEE Transactions on Image Processing*, 31:5869–5880, 2022. 3

[49] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834, 2021. 7

[50] Liguang Zhou, Yuhongze Zhou, Tin Lun Lam, and Yangsheng Xu. Context-aware mixture-of-experts for unbiased scene graph generation. *arXiv preprint arXiv:2208.07109*, 2022. 3

[51] Zijian Zhou, Zheng Zhu, Holger Caesar, and Miaojing Shi. Openpsg: Open-set panoptic scene graph generation via large multimodal models. In *European Conference on Computer Vision*, pages 199–215. Springer, 2024. 2

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6