Leveraging Predicate and Triplet Learning for Scene Graph Generation

Supplementary Material

In the supplementary material, we provide the following contents for the proposed Dual-granularity Relation Modeling (DRM) network which leverages predicate and triplet learning for Scene Graph Generation (SGG): (1) more implementation details of our method; (2) more comparison results, including comparisons on the M@K and F@K metrics and comparisons with VETO+MEET [17]; (3) more ablation studies, including the ablation on Dual-granularity Knowledge Transfer (DKT) strategy and dual-granularity learning; (4) hyper-parameter analysis; and (5) qualitative visualization. We will make the code publicly available upon acceptance of this paper.

A. Additional Implementation Details

We implement DRM using Pytorch [16] and the official code-base Scene-Graph-Benchmark.pytorch[1] with a NVIDIA A800 GPU. In the initialization of entity representations $\{\mathbf{v}_i\}_{i=1}^{N}$ and union features $\{\mathbf{s}_i\}_{i=1}^{N}$, we adopt the same strategy as VCTREE [18] and PE-Net [27], which involves a fusion of their visual and spatial features. We follow Xu *et al*. [21] to augment input images for model training. The expect number Q_i of each tail predicate in Equation 8 is equal to the count of the head predicates with the smallest number.

B. Additional Comparison Results

B.1. Trade-off Results between R@K and mR@K

Due to the imbalanced data distribution of Visual Genome [7], Open Image [6], and GQA datasets [4], there is a trade-off between Recall R@K and mean Recall mR@K metrics. To measure the trade-offs of the scene graph generation methods, Zheng *et al*. [27] introduce the Mean@K (M@K), which averages the R@K and mR@K, while Zhang *et al*. [25] propose the F@K, the harmonic mean of R@K and mR@K. Note that these two metrics only measure the trade-off between R and mR, and it is feasible for diverse methods, even with significant differences in R and mR, to still arrive at the same trade-off results. Evaluating either R@K or mR@K better aligns with the practical need to predict the highest number of relationships or to forecast relationships as uniformly as possible.

Visual Genome. Table 6 shows the performance of different methods in terms of M@50/100 and F@50/100 on VG150. Our DRM outperforms state-of-the-art methods at F@K measurements and DRM w/o DTK also achieves state-of-the-art performance at M@K. It indicates that although the recall of DRM degrades, the trade-off between the recall and the mean recall is the best in the state-of-the-art methods.

GQA. To evaluate the generalizability of our method across various datasets, we present the performance of M@50/100 and F@50/100 on GQA200. As shown in Table 7, DRM outperforms all of the state-of-the-art methods at both M@50/100 and F@50/100 metrics. These results demonstrate that our method remains effective in dealing with relation recognition, regardless of the variations of data distributions.

B.2. Comparison with VETO+MEET

The MEET [17] method assigns multiple relationships to each subject-object pair during inference. This is in accordance with the testing protocol termed "without graph constraint". The setting of "without graph constraint", as proposed by Zellers [24], permits the output scene graph to have multiple edges between the subject and object. Better performance is typically achieved without the graph constraint since the model is allowed to make multiple guesses for challenging relations. In the following, "ng-" denote the No Graph Constraint variant of the metric.

As shown in Table 8, we compare our DRM with VETO+MEET under the setting of "without graph constraint" on VG150 and GQA200 datasets. We have the following observations: 1) Compared to the performance with graph constraint, our method consistently exhibits promotion without graph constraint. 2) Our proposed DRM w/o DKT has considerably better performance compared to VETO+MEET. More specifically, our DRM w/o DKT outperforms VETO+MEET by an average of 11.9% and 4.9% at ng-R@100 and ng-mR@100, respectively. 3) Based on the proposed DKT strategy, DRM significantly outperforms VETO+MEET by 21.1%, 16.2%, 16.7% at ng-mR@100 on three tasks of VG150 datasets. It also surpasses VETO+MEET by 25.2%, 14.4%, 14.8% at ng-mR@100 on three tasks of VG150 datasets.

We also present the comparison results at ng-M@50/100 and ng-F@50/100 to demonstrate the trade-off performance under the setting of "without graph constraint". The results are shown in Table 9. Our DRM consistently and significantly outperforms the recent VETO+MEET in terms of both ng-M50/100 and ng-F@50/100 metrics. These results indicate the consistent effectiveness of our DRM under the setting of "without graph constraint".

[1] https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

Models	PredCls		SGCls		SGDet	
	M@50/100	F@50/100	M@50/100	F@50/100	M@50/100	F@50/100
IMP [20]$_{CVPR'17}$	36.1 / 37.5	18.6 / 19.9	21.9 / 22.5	10.6 / 11.1	15.1 / 18.3	7.2 / 9.1
VTransE [26]$_{CVPR'17}$	40.2 / 41.7	24.0 / 25.6	23.4 / 24.1	13.5 / 14.3	17.4 / 20.2	8.6 / 10.4
MOTIFS [24]$_{CVPR'18}$	40.3 / 41.9	23.9 / 25.6	23.6 / 24.2	13.3 / 14.0	18.8 / 21.9	9.4 / 11.5
G-RCNN [22]$_{ECCV'18}$	40.9 / 42.2	26.2 / 27.4	23.0 / 24.0	14.5 / 15.2	17.8 / 19.7	9.7 / 11.0
VCTREE [18]$_{CVPR'19}$	41.1 / 42.7	26.6 / 28.3	24.1 / 25.0	13.2 / 13.8	19.2 / 21.7	10.7 / 12.1
GPS-Net [12]$_{CVPR'20}$	40.2 / 41.9	24.7 / 26.6	23.2 / 24.2	13.9 / 14.8	19.0 / 22.3	11.0 / 13.9
RU-Net [14]$_{CVPR'22}$	- / 46.9	- / 35.9	- / 29.0	- / 21.8	- / 24.2	- / 16.8
HL-Net [13]$_{CVPR'22}$	- / 45.9	- / 34.3	- / 28.5	- / 20.6	- / 23.7	- / 14.8
PE-Net(P) [27]$_{CVPR'23}$	45.7 / 47.8	34.5 / 37.3	27.2 / 28.6	19.9 / 21.9	20.7 / 24.0	14.0 / 16.9
VETO [17]$_{ICCV'23}$	43.5 / 45.5	33.6 / 36.0	23.4 / 24.4	16.9 / 18.0	17.8 / 20.5	12.5 / 14.6
TDE$^{\diamond}$ [19]$_{CVPR'20}$	35.9 / 40.3	32.9 / 37.2	20.4 / 22.4	17.8 / 19.9	12.6 / 15.1	11.0 / 13.2
CogTree$^{\diamond}$ [23]$_{IJCAI'21}$	31.0 / 32.9	30.3 / 32.4	18.3 / 19.2	17.6 / 18.7	15.2 / 17.0	13.7 / 15.4
BPL-SA$^{\diamond}$ [3]$_{ICCV'21}$	40.2 / 42.1	37.5 / 39.5	23.3 / 24.3	21.3 / 22.4	18.3 / 21.3	17.0 / 19.7
NICE$^{\diamond}$ [8]$_{CVPR'22}$	42.5 / 44.8	38.8 / 41.3	24.9 / 26.0	22.1 / 23.5	20.0 / 23.1	17.0 / 19.8
PPDL$^{\diamond}$ [11]$_{CVPR'22}$	39.7 / 40.5	38.3 / 39.2	23.0 / 23.8	21.7 / 22.5	16.3 / 18.7	14.8 / 17.3
GCL$^{\diamond}$ [2]$_{CVPR'22}$	39.4 / 41.3	39.1 / 41.1	23.5 / 24.5	23.2 / 24.2	17.6 / 20.7	17.6 / 20.6
INF$^{\diamond}$ [1]$_{CVPR'23}$	38.1 / 42.9	33.4 / 39.4	23.4 / 25.6	20.0 / 23.0	16.7 / 19.4	13.5 / 16.3
CFA$^{\diamond}$ [9]$_{ICCV'23}$	44.9 / 47.4	**43.0** / **45.6**	26.0 / 27.3	22.9 / 24.4	20.3 / 23.7	17.8 / 20.8
EICR$^{\diamond}$ [15]$_{ICCV'23}$	45.1 / 47.2	42.8 / 45.0	27.7 / 28.6	**26.0** / **27.0**	**21.7** / **25.2**	**19.9** / **23.3**
BGNN [10]$_{CVPR'21}$	44.8 / 47.1	40.2 / 42.8	25.9 / 27.5	20.7 / 23.1	20.9 / 24.2	15.9 / 18.6
SHA+GCL [2]$_{CVPR'22}$	38.4 / 40.7	38.1 / 40.4	22.9 / 24.1	22.9 / 24.1	16.4 / 19.6	16.3 / 19.5
PE-Net [27]$_{CVPR'23}$	**48.2** / **50.5**	42.4 / 45.0	**28.6** / **29.8**	24.5 / 25.8	21.6 / 24.9	17.7 / 20.5
SQUAT [5]$_{ICCV'23}$	43.3 / 45.7	39.7 / 42.4	25.3 / 26.6	22.9 / 24.3	19.3 / 22.7	17.9 / 21.0
DRM w/o **DKT**	**46.8** / **48.9**	35.0 / 37.8	**28.9** / **29.9**	20.7 / 22.1	**21.5** / **25.1**	14.2 / 17.4
DRM	45.5 / 47.7	**45.4** / **47.6**	27.7 / 28.8	**27.6** / **28.8**	19.7 / 23.5	**19.7** / **23.5**

Table 6. Results in terms of M@K and F@K for three tasks on the VG150 dataset with graph constraints. "$\diamond$" denotes the combination of MOTIFS with a model-agnostic unbiasing strategy. The best and second best results under each setting are respectively marked in **red** and **underline blue**.

Models	PredCls		SGCls		SGDet	
	M@50/100	F@50/100	M@50/100	F@50/100	M@50/100	F@50/100
VTransE [26]$_{CVPR'17}$	34.9 / 36.5	22.4 / 23.8	20.8 / 21.5	13.0 / 13.9	16.5 / 18.7	9.6 / 10.9
MOTIFS [24]$_{CVPR'18}$	40.9 / 42.0	26.2 / 27.2	21.2 / 21.8	13.2 / 13.8	17.7 / 20.4	10.5 / 12.5
VCTREE [18]$_{CVPR'19}$	40.2 / 41.6	26.3 / 27.5	21.0 / 21.6	12.8 / 13.4	17.4 / 19.7	10.6 / 12.0
SHA [2]$_{CVPR'22}$	41.4 / 43.2	29.8 / 31.9	20.6 / 21.3	13.5 / 14.2	16.1 / 18.5	10.5 / 12.3
VETO [17]$_{ICCV'23}$	**42.9** / **44.1**	31.9 / 33.1	19.5 / 20.3	13.4 / 14.1	16.6 / 18.6	11.0 / 12.7
VTransE+GCL [2]$_{CVPR'22}$	33.0 / 34.9	32.8 / 34.7	19.8 / 20.5	19.2 / 20.0	15.0 / 17.2	15.0 / 17.2
MOTIFS+GCL [2]$_{CVPR'22}$	40.6 / 42.2	40.2 / 41.8	20.3 / 21.1	19.8 / 20.6	17.7 / 20.3	**17.6** / **20.2**
VCTREE+GCL [2]$_{CVPR'22}$	40.1 / 41.7	39.5 / 41.1	20.5 / 21.3	20.0 / 20.8	16.6 / 19.3	16.5 / 19.1
SHA+GCL [2]$_{CVPR'22}$	41.9 / 43.6	**41.8** / **43.6**	21.0 / 21.8	**21.0** / **21.7**	16.3 / 19.0	16.2 / 18.9
DRM w/o **DKT**	42.5 / 43.7	28.5 / 29.7	**21.8** / **22.3**	11.9 / 12.3	**18.8** / **21.5**	11.3 / 13.5
DRM	**42.6** / **44.0**	**42.5** / **43.9**	**21.6** / **22.3**	**21.5** / **22.2**	**18.8** / **21.4**	**18.7** / **21.3**

Table 7. Results in terms of M@K and F@K for three tasks on the GQA200 dataset with graph constraints. The best and second best results under each setting are respectively marked in **red** and **underline blue**.

C. Additional Ablation Studies

We conduct additional ablation studies to further evaluate the effectiveness of our dual-granularity learning and DKT strategy. We combine DKT with a recent state-of-the-art model, PE-NET [27]. As PE-Net is only concerned with modeling predicate features, we only transfer its predicate

Datasets	Models	PredCls		SGCls		SGDet	
		ng-R@50/100	ng-mR@50/100	ng-R@50/100	ng-mR@50/100	ng-R@50/100	ng-mR@50/100
VG150	VETO+MEET [17]$_{ICCV'23}$	74.0 / 78.9	42.0 / 52.4	41.1 / 44.0	22.3 / 27.4	28.6 / 34.0	10.6 / 13.8
	DRM w/o **DKT**	**85.8 / 92.0**	42.8 / 57.1	**53.6 / 56.8**	24.3 / 32.1	**37.4 / 43.9**	13.8 / 19.1
	DRM	68.1 / 80.4	**62.9 / 73.5**	43.2 / 50.1	**37.6 / 43.6**	26.1 / 33.6	**25.0 / 30.5**
GQA200	VETO+MEET [17]$_{ICCV'23}$	73.9 / 78.3	43.3 / 50.5	34.6 / 37.2	19.7 / 22.5	26.7 / 31.0	12.1 / 16.0
	DRM w/o **DKT**	**79.6 / 85.9**	44.8 / 59.4	**43.1 / 46.4**	17.9 / 24.1	**33.1 / 38.4**	13.2 / 18.6
	DRM	68.5 / 78.2	**65.6 / 75.7**	36.6 / 41.8	**32.1 / 36.9**	24.4 / 30.4	**26.0 / 30.8**

Table 8. Comparison results with VETO+MEET on the VG150 and GQA200 datasets without graph constraint. The metrics "ng-R@K" and "ng-mR@K" denote the No Graph Constraint Recall@K and No Graph Constraint Mean Recall@K, respectively. The best results under each setting are respectively marked in **bold**.

Datasets	Models	PredCls		SGCls		SGDet	
		ng-M@50/100	ng-F@50/100	ng-M@50/100	ng-F@50/100	ng-M@50/100	ng-F@50/100
VG150	VETO+MEET [17]$_{ICCV'23}$	58.0 / 65.7	53.6 / 63.0	31.7 / 35.7	28.9 / 33.8	19.6 / 23.9	15.5 / 19.6
	DRM w/o **DKT**	64.3 / 74.6	57.1 / 70.5	39.0 / 44.5	33.4 / 41.0	25.6 / 31.5	20.2 / 26.6
	DRM	**65.5 / 77.0**	**65.4 / 76.8**	**40.4 / 46.9**	**40.2 / 46.6**	**25.6 / 32.1**	**25.5 / 32.0**
GQA200	VETO+MEET [17]$_{ICCV'23}$	58.6 / 64.4	54.6 / 61.4	27.2 / 29.9	25.1 / 28.0	19.4 / 23.5	16.7 / 21.1
	DRM w/o **DKT**	62.2 / 72.7	57.3 / 70.2	30.5 / 35.3	25.3 / 31.7	23.2 / 28.5	18.9 / 25.1
	DRM	**67.1 / 77.0**	**67.0 / 76.9**	**34.4 / 39.4**	**34.2 / 39.2**	**25.2 / 30.6**	**25.2 / 30.6**

Table 9. Results in terms of ng-M@K and ng-F@K for three tasks on the VG150 and GQA200 datasets without graph constraints. The metrics "ng-M@K" and "ng-F@K" denote the No Graph Constraint Mean@K and No Graph Constraint Harmonic Mean@K, respectively. The best results under each setting are respectively marked in **bold**.

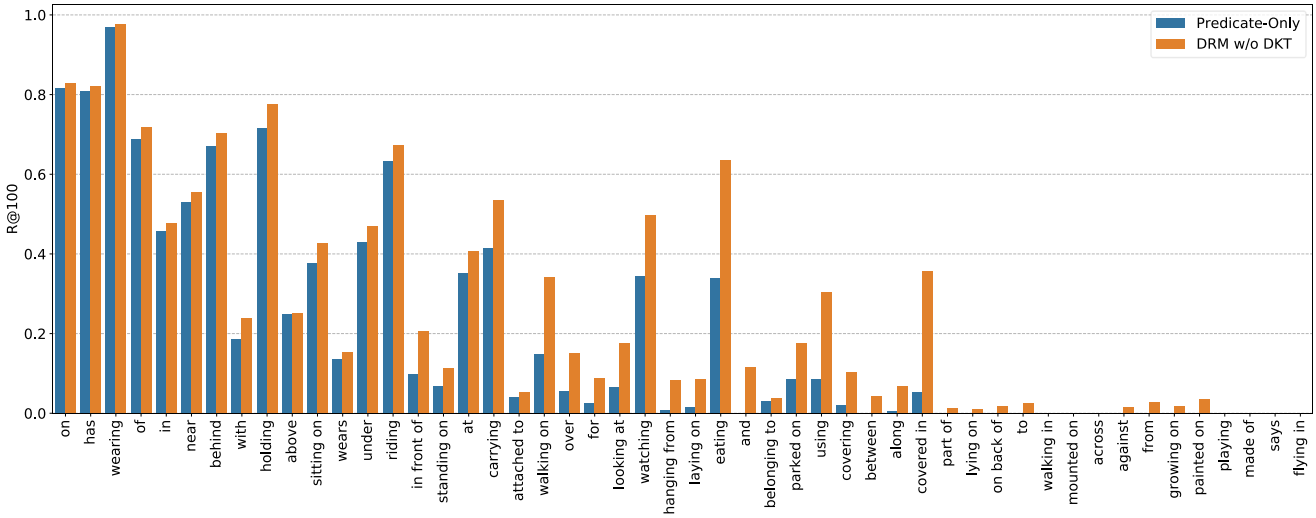

Figure 5. Results in terms of Recall@100 of all predicate classes of Predicate-Only and DRM w/o DKT on the PredCls task. Predicates are sorted according to their frequency.

knowledge from head to tail. As shown in Table 10, the incorporation of DKT substantially boosts the performance of PE-NET in three tasks at mR@K.

To further demonstrate the effectiveness of our method in modeling triplet clues, we provide the R@100 performance of our methods "predicate-only" and "DRM w/o DKT" on each predicate. As shown in Figure 5, our method outperforms the "predicate-only" baseline on each predicate. We also present the performance of the Recall@100 for the predicate "riding" and the predicate "eating" at the fine-

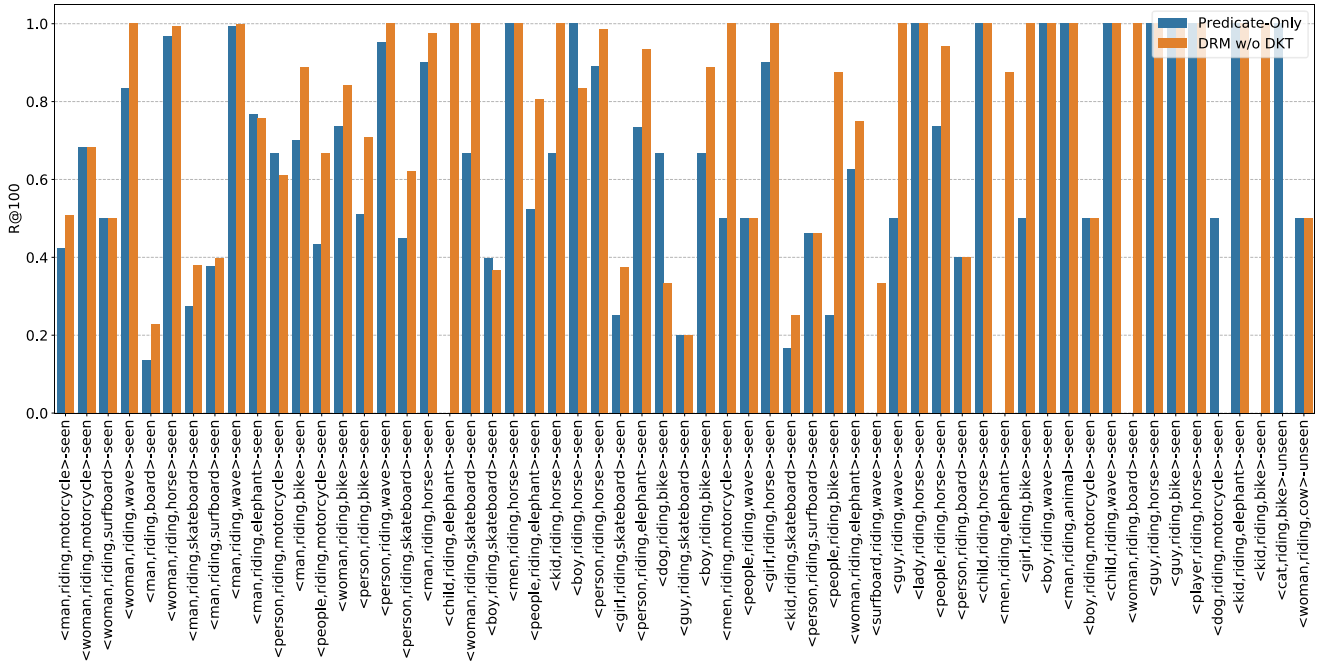

Figure 6. Results in terms of Recall@100 for triplets belonging to predicate "riding" of Predicate-Only and DRM w/o DKT on the PredCls task. The terms "seen" and "unseen" represent whether the triplets appear in the training set or not, respectively.

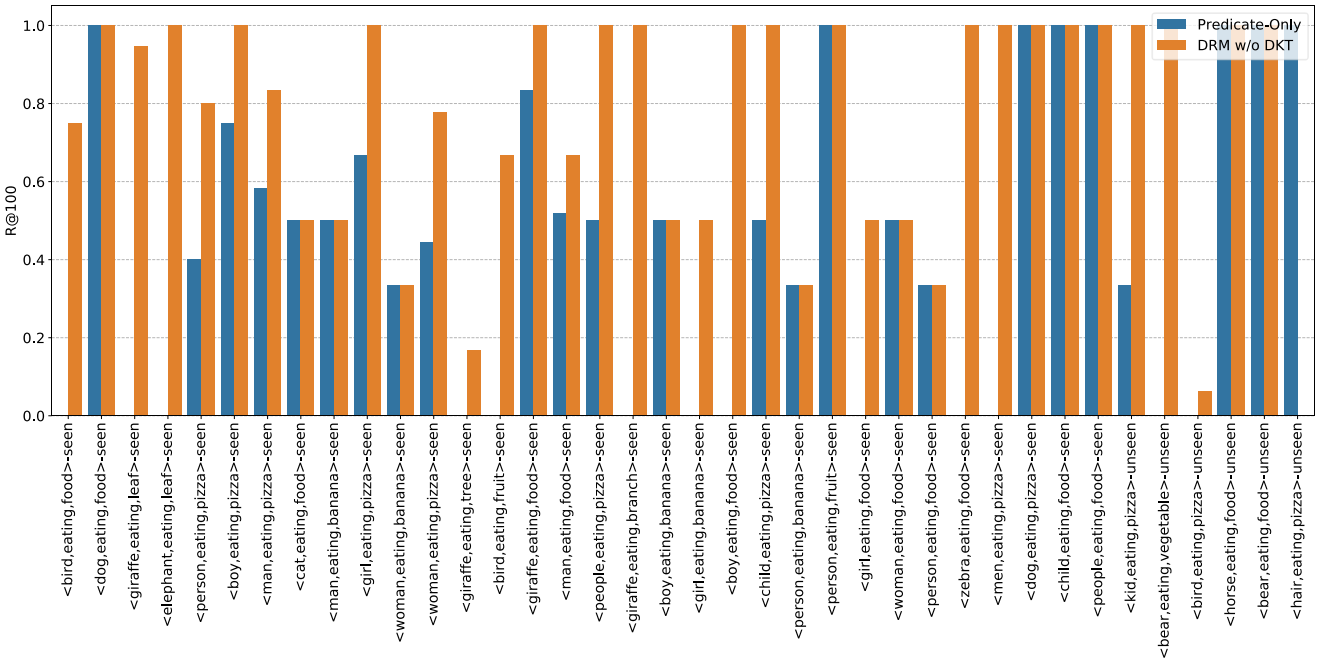

Figure 7. Results in terms of Recall@100 for triplets belonging to predicate "eating" of Predicate-Only and DRM w/o DKT on the PredCls task. The terms "seen" and "unseen" represent whether the triplets appear in the training set or not, respectively.

grained triplet level. The results are illustrated in Figures 6 and 7. We can observe that the performance is significantly improved in the seen triplets, especially for *<giraffe, eating, leaf>* and *<bird, eating, fruit>*, where our method can accurately predict multiple eating expressions that cannot be captured using only predicates. This suggests that our method is able to learn the triplet cues in the training, and utilize them to reason about the relationships under specific

Models	PredCls		SGCls		SGDet	
	mR@50	mR@100	mR@50	mR@100	mR@50	mR@5100
PE-Net [27]$_{CVPR'23}$	31.4	33.5	18.2	19.3	12.3	14.3
PE-Net + DKT-P	**44.3**	**48.4**	**24.9**	**27.0**	**15.0**	**17.7**

Table 10. The results on three tasks of PE-Net equipped with our DKT on VG150 dataset.

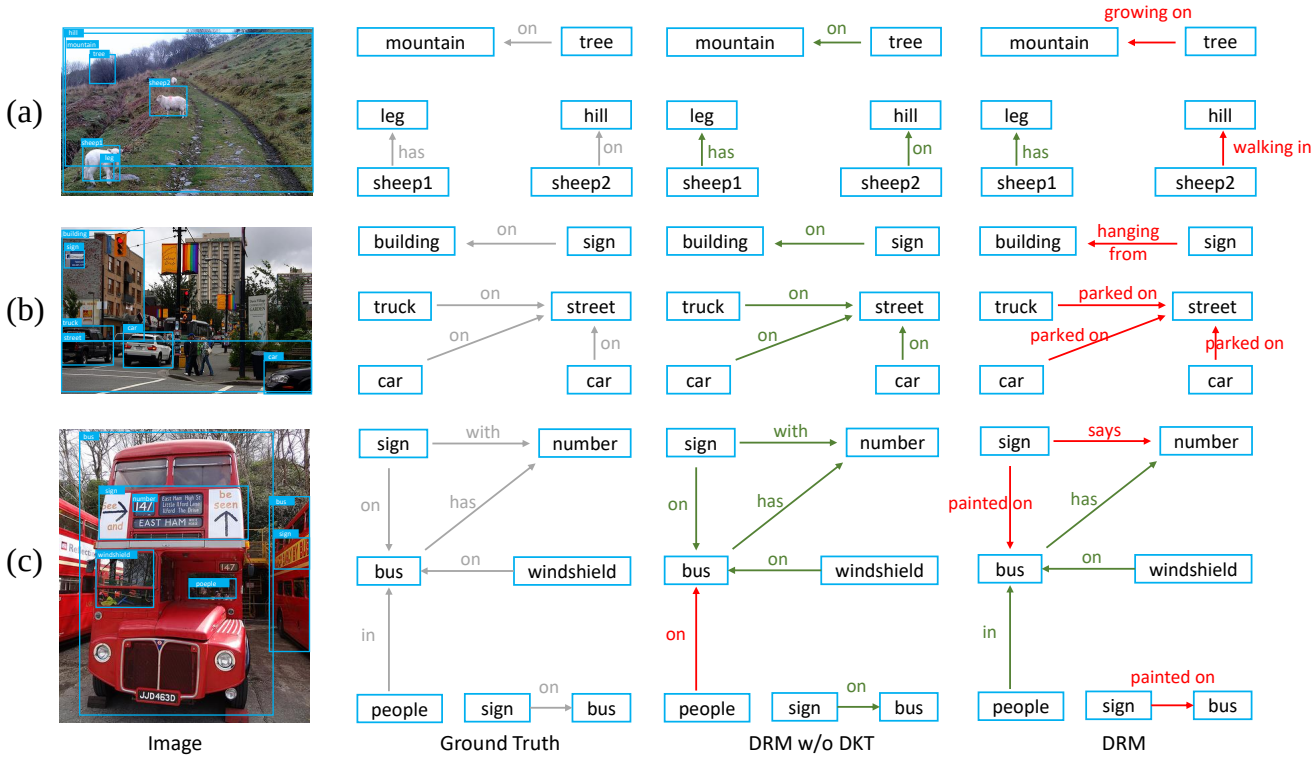

Figure 8. Scene graphs generated by our DRM w/o DKT and DRM in the PredCls Task. DRM tends to generate more precise fine-grained relations than DRM w/o DKT, which leads to the performance degradation on R@K. The use of green and red colors indicates whether the prediction matches the ground truth or not, respectively.

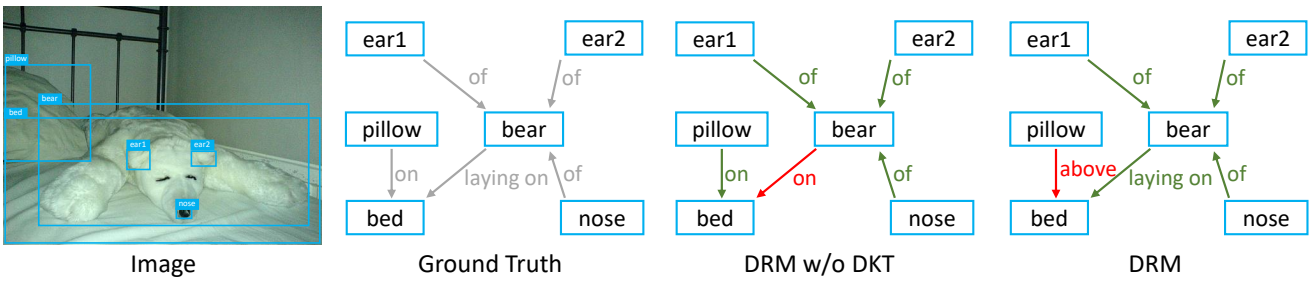

Figure 9. Scene graphs generated by our DRM w/o DKT and DRM in the PredCls Task. Triplet <*bear, laying on, bed*> only appears once in the VG150 training set. The use of green and red colors indicates whether the prediction matches the ground truth or not, respectively.

subject-object pairs during inference.

D. Hyper-parameters analysis

We first investigate the impact of hyper-parameters τ_p and τ_t in the Dual-granularity Constraints. The results are illustrated in Table 11. The decrease in τ_p and τ_t leads to

τ_p	τ_t	PredCls			
		R@50	R@100	mR@50	mR@100
0.1	0.1	70.1	71.9	22.6	24.9
0.2	0.1	**70.2**	**72.1**	**23.3**	**25.6**
0.3	0.1	70.1	72.0	22.5	24.9
0.2	0.2	70.2	72.0	23.0	25.3

Table 11. Hyper-parameters analysis of the temperature τ_p and τ_t.

λ_p	λ_e	SGCls			
		R@50	R@100	mR@50	mR@100
3	0.1	43.4	44.3	13.1	14.3
3	0.2	43.9	44.8	13.3	14.5
3	0.5	**44.3**	**45.2**	**13.5**	**14.6**
3	0.8	44.2	45.1	13.0	14.2
2	0.5	44.2	45.2	12.9	14.1
3	0.5	**44.3**	**45.2**	**13.5**	**14.6**
4	0.5	44.2	45.1	13.4	14.6

Table 12. Hyper-parameters analysis of the loss weights λ_p and λ_t.

the more compact predicate and triplet representations. We observe that the model achieves the best performance when $\tau_p = 0.2$ and $\tau_t = 0.1$, indicating a compact aggregation of the triplet compared to the predicate. This observation aligns with the fact that variations in triplet are significantly smaller than those in the predicate.

During pre-training, the predicate classification loss is much smaller than the entity classification loss due to the long-tailed distribution of predicates. To balance the scales between these losses, we set λ_p to be larger than λ_e. Experimental results in Table 12 show that the model performs best when $\lambda_p = 3$ and $\lambda_e = 0.5$.

E. Visualization Results

We present a visualization of the results of our DRM w/o DKT and our DRM on the PredCls task on the VG150 dataset. The results are shown in Figure 8 and 9. As the VG dataset is incompletely labeled, we focus our analysis only on labeled relations in the dataset. We observe that DRM w/o DKT can accurately predict the ground-truth relationships. However, our DRM often predicts these relationships differently. This difference arises from the tendency of our DRM to predict coarse-grained relations as more precise fine-grained relations. For instance, for the subject-object pair "tree-mountain", DRM generates the more accurate predicate "growing on" while the DRM w/o DKT predicts the coarse-grained predicate "on". These fine-grained predictions lead to a decrease in R@K and an increase in mR@K. This degradation in R@K is an inevitable result due to the large number of coarsely labeled relations in the test dataset.

References

[1] Bashirul Azam Biswas and Qiang Ji. Probabilistic debiasing of scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10429–10438, 2023. 2

[2] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 2

[3] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16383–16392, 2021. 2

[4] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 1

[5] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. 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*, pages 18664–18674, 2023. 2

[6] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 1

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

[8] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. 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*, pages 18869–18878, 2022. 2

[9] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21685–21695, 2023. 2

[10] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021. 2

[11] Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. Ppdl: Predicate probability distribution

based loss for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2022. 2

[12] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2

[13] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19476–19485, 2022. 2

[14] Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. Ru-net: Regularized unrolling network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19466, 2022. 2

[15] Yukuan Min, Aming Wu, and Cheng Deng. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13296–13307, 2023. 2

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[17] Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21882–21893, 2023. 1, 2, 3

[18] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 1, 2

[19] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 2

[20] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 2

[21] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 1

[22] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pages 670–685, 2018. 2

[23] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. In *International Joint Conference on Artificial Intelligence*, pages 1274–1280, 2021. 2

[24] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 1, 2

[25] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. In *European Conference on Computer Vision*, pages 409–424. Springer, 2022. 1

[26] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5532–5540, 2017. 2

[27] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792, 2023. 1, 2, 5