Vision-Language Interactive Relation Mining for Open-Vocabulary Scene Graph Generation

Supplementary Material

Figure 1. Selected pseudo-label examples demonstrate the model's capability to produce diverse and rich relation labels. The blue relation triplets indicate the pseudo labels generated by our model.

1. Details on OVSGTR

For further clarification on integrating RelGen with OVSGTR, we provide a formal description of OVSGTR, which utilizes a pre-trained model. Specifically, OVSGTR first pretrain a relation-aware model on large-scale image-caption datasets. This pre-trained model is then employed during training, where a knowledge distillation strategy is implemented to preserve consistency within the learned semantic space, and the distillation loss is defined as:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{N}|} \sum_{e \in \mathcal{N}} \left\| e^s - e^t \right\|_1, \tag{1}$$

where e^s and e^t refer to the student's and teacher's edge features, respectively. Thus the total loss $\mathcal{L}_{VL}$ for associate visual and text features can be given as:

$$\mathcal{L}_{VL} = \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{bce}} + \lambda \mathcal{L}_{\text{cons}}. \tag{2}$$

Besides, we use L1 regression loss and generalized IoU to improve localization accuracy. Finally, our overall loss function is given by

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{bce}} + \lambda \mathcal{L}_{\text{cons}} \\ + \text{GIOU}(\mathbf{B}, \mathbf{B}_{gt}) + ||\mathbf{B} - \mathbf{B}_{gt}||_1. \end{aligned} \tag{3}$$

At inference, since OVSGTR uses a pre-trained model, we directly generate the scene graph representation via the visual relation feature, ensuring the preservation of relation-aware knowledge from the pre-trained model.

2. Visualization of Pseudo-labels

To address the limitations of the SGG model trained on base relation categories, we utilize pseudo-labels generated by our model to supplement the relations in the images. Examples of these pseudo-labels are presented in Fig. 1. Our method demonstrates the ability to generate more diverse relation predicates while remaining consistent with the image content. Compared to the limited annotations in the base relation categories, pseudo-labels offer complementary annotations that encompass a broader range of objects within the image.

Method	Novel+Base		Novel	
	R@100	mR@100	R@100	mR@100
w/o fitting	16.24	6.62	10.23	4.36
Balanced init	15.48	7.48	10.25	4.99
Background init	**18.48**	**9.60**	**13.45**	**6.06**

Table 1. Experimental results of different settings for bias initialization.

Method	Novel+base			
	R@50	R@100	mR@50	mR@100
PGSG [1]	26.0	28.9	14.9	18.1
OpenPSG* [4]	31.6	36.7	24.0	25.4
DSFormer [2]	23.2	27.0	17.2	18.1
Ours+DSFormer [2]	27.1	32.3	18.2	22.5

Table 2. Results on the SGDet task on PSG dataset.

3. Other Results

Ablation Study for Bias Initialization. We evaluated the influence of constructing different values of the dynamic prior $\hat{r}_{i0}^{bias}$ in Tab. 3. "Balanced init" means we initialize the bias with equal probabilities to each class, i.e., $\hat{r}_{i0}^{bias} = 1/N_c$. "Background init" means we initialize the bias as "background", i.e., $\hat{r}_{i0}^{bias} = 0$. As shown in Tab. 3, initializing as "background" gains the best performance.

Results in panoptic scene graph generation: To achieve a more comprehensive understanding of scenes, PSG [3] is a dataset for panoptic scene graph generation, which replaces bounding boxes with panoptic segmentation masks to represent the objects. Since our method is built on the detection framework, we take DSFormer [2] as the alternative segmentation backbone on the PSG dataset. As shown in Table 2, we can see that constructing the generative relation model also improves the performance on the open-vocabulary panoptic scene graph generation task. As for the gap in the performance compared to the SOTA OpenPSG [4], we attribute it to the well-designed segmentation head in OpenPSG.

Results under various VLM backbones: For a more comprehensive evaluation, we report results with different VLMs in Tab. 3, in which our method significantly outperforms others when using GroundingDINO, whereas we had trouble reproducing for other baselines. Our method also achieves the best performance on other backbones, except R@50 on VG* with GLIP. We attribute this to the fact that *the generative pipeline requires more candidates to give full play to its open-ended advantages under base relations*, yet our method still improves the performance on novel relations by at least **0.7%** (14.2 *v.s.* 13.5) and up to **9.4%** (9.7 *v.s.* 0.3).

Method	Model	Dataset	Amount	Novel + Base		Novel	
				R@50	mR@50	R@50	mR@50
VS3	GLIP	VG	$<$ 57k	11.0	5.1	-	0.0
Ours (w/o $\mathcal{L}_{\text{distill}}$, $\mathcal{L}_{\text{cons}}$, PL)	GLIP	VG	$<$ 57k	11.4	4.8	0.0	0.0
Ours (w/o $\mathcal{L}_{\text{distill}}$, $\mathcal{L}_{\text{cons}}$)	GLIP	VG	$<$ 57k	12.2	6.1	0.0	0.0
Ours (w/o $\mathcal{L}_{\text{distill}}$)	GLIP	VG	$<$ 57k	13.7	**7.5**	10.1	**5.2**
SGTR	BLIP	VG	$<$ 57k	12.6	3.5	-	0.0
PGSG	BLIP	VG	$<$ 57k	15.8	6.2	-	3.8
Ours (w/o $\mathcal{L}_{\text{distill}}$)	BLIP	VG	$<$ 57k	**17.2**	7.1	**10.7**	4.9
VS3	GLIP	VG*	$<$ 57k	15.6	-	0.0	-
Ours (w/o $\mathcal{L}_{\text{distill}}$)	GLIP	VG*	$<$ 57k	14.3	8.7	9.0	-
OVSGTR (w/o $\mathcal{L}_{\text{distill}}$)	GDINO	VG*	$<$ 57k	17.7	6.4	0.3	-
Ours (w/o $\mathcal{L}_{\text{distill}}$)	GDINO	VG*	$<$ 57k	**18.6**	**9.2**	**9.7**	-
OVSGTR	GDINO	VG*	$\sim$ 569k	20.5	3.9	13.5	-
OVSGTR + Ours	GDINO	VG*	$\sim$ 569k	**21.1**	**4.2**	**14.2**	-
SGTR	BLIP	OIV6	$<$ 120k	36.1	11.0	-	0.0
PGSG	BLIP	OIV6	$<$ 120k	41.3	20.8	-	3.8
Ours	BLIP	OIV6	$<$ 120k	48.4	21.5	15.2	5.6
Ours	GLIP	OIV6	$<$ 120k	46.3	19.8	14.6	4.5
Ours	GDINO	OIV6	$<$ 120k	**49.5**	**23.1**	**16.1**	**6.2**

Table 3. Results on VG and OIV6 datasets. "GDINO" indicates using GroundingDINO as the base VLM and "w/o PL" denotes we do not use pseudo labels. "VG*" means testing under OVSGTR's setting with 30% novel relation categories, and the others are under PGSG's setting with 50% novel categories.

References

[1] 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. 3

[2] Julian Lorenz, Alexander Pest, and Daniel Kienzle et al. A fair ranking and new model for panoptic scene graph generation. In *ECCV*, pages 148–164. Springer, 2024. 3

[3] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 3

[4] 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. 3