Supplementary Material for “TextPSG: Panoptic Scene Graph Generation from Textual Descriptions”

Chengyang Zhao [1] Yikang Shen [2] Zhenfang Chen [2]
Mingyu Ding [3] Chuang Gan [2,4]
[1]Peking University [2]MIT-IBM Watson AI Lab
[3]UC Berkley [4]UMass Amherst

A1. Overview

This supplementary material provides additional information about the design of our TextPSG framework, the details of experiments, and more quantitative and qualitative results. In Sec. A2, we provide a more detailed explanation of our framework, including the region grouper, the entity grounder, the label generator, and the inference procedure. In Sec. A3, we provide further details on the dataset used for evaluation, the baselines developed, and the implementation process. In Sec. A4, we provide more ablation studies to demonstrate the effectiveness of our design, more diagnoses of our framework for a clearer understanding of the efficacy, additional visualization results for qualitative evaluation, and examples of the failure cases.

A2. More Details of TextPSG Framework

A2.1. More Details of Region Grouper

The region grouper follows the design of GroupViT [18]. The input scene image I is first split into N non-overlapping patches and projected to be initial image segments $\{\mathbf{s}_i^0\}_{i=1}^N$, which are then passed through K grouping layers $\{\mathbf{Grp}_k\}_{k=1}^K$ to be merged progressively. Each grouping layer $\mathbf{Grp}_k$ consists of H_k learnable grouping centers $\{\mathbf{c}_i^k\}_{i=1}^{H_k}$, a Transformer [16]-based block $\mathbf{Tfm}_k^I$ for communication between the centers $\{\mathbf{c}_i^k\}_{i=1}^{H_k}$ and the segments $\{\mathbf{s}_i^{k-1}\}_{i=1}^{H_{k-1}}$, and an attention-based block $\mathbf{Att}_k$ for assigning the segments to different centers and merging the segments corresponding to the same center into $\{\mathbf{s}_i^k\}_{i=1}^{H_k}$. Within $\mathbf{Grp}_k$, the grouping is performed as

$$\{\mathbf{s}_i^k\}_{i=1}^{H_k} = \mathbf{Att}_k(\mathbf{Tfm}_k^I(\{\mathbf{c}_i^k\}_{i=1}^{H_k}, \{\mathbf{s}_i^{k-1}\}_{i=1}^{H_{k-1}})).$$

Note that $H_0 = N$. Especially, the updated image segments $\{\hat{\mathbf{s}}_i^0\}_{i=1}^{H_0}$ from the communication block $\mathbf{Tfm}_1^I$ in the first grouping layer $\mathbf{Grp}_1$ will be further used by the label generator for the label prediction, as introduced in the following.

A2.2. More Details of Entity Grounder

In the entity grounder, meaningful region-entity alignment can be reached automatically during training, serving as pseudo labels for the learning of the segment merger and the label generator. Here we provide a further explanation of the automatic meaningful alignment.

In the entity grounder, the total fine-grained contrastive loss $\mathcal{L}_{fine}^k$ consists of two symmetry components $\mathcal{L}_{fine}^{k,I\to T}$ and $\mathcal{L}_{fine}^{k,T\to I}$. Minimizing $\mathcal{L}_{fine}^k$ equals to minimizing $\mathcal{L}_{fine}^{k,I\to T}$ and $\mathcal{L}_{fine}^{k,T\to I}$ simultaneously.

Here we take $\mathcal{L}_{fine}^{k,I\to T}$ as an example while the other remains the same. In each batch, we assume that for each region in each image, there is at most one corresponding entity in the corresponding caption, while all the other entities in the caption and all entities in the other captions are mismatched with the region.

To minimize $\mathcal{L}_{fine}^{k,I\to T}$, for each image I_i in the batch, the model needs to maximize $p^{k,i\to i}$ and minimize all other $p^{k,i\to j}$ where $j \neq i$.

To minimize $p^{k,i\to j}$, with $p^{k,i\to j}$ denoting the mean value of $p_l^{k,i\to j}$ and l for the index of the region in the image I_i, the model needs to minimize all $p_l^{k,i\to j}$. Since $p_l^{k,i\to j}$ denotes the max cosine similarity between the l-th region and all entities in T_j, minimizing $p_l^{k,i\to j}$ equals pushing the l-th region and all entities in T_j apart in the shared feature space.

To maximize $p^{k,i\to i}$, the model needs to maximize all $p_l^{k,i\to i}$. A global maximum is that the l-th region is close to the corresponding entity in T_i and far from all the other entities in the shared feature space.

By minimizing $p^{k,i\to j}$ and maximizing $p^{k,i\to i}$ at the same time, the model tends to pull similar region-entity pairs to be closer and push dissimilar pairs apart in the shared feature space, thus reaching a meaningful region-entity alignment automatically.

A2.3. More Details of Label Generator

Here we provide more details about the prompt-embedding-based technique (PET) used in the label generator.

To predict the object semantics, for each image mask $\mathbf{m}_i^k$, the label generator takes the updated image tokens $\{\hat{\mathbf{p}}\}_{i=1}^N$, *i.e.*, $\{\hat{\mathbf{s}}_i^0\}_{i=1}^{H_0}$, and the mask $\mathbf{m}_i^k$ as input, using a prompt

$$\textbf{a photo of }[\mathbf{ENT}]$$

to guide the object generation, where the $[\mathbf{ENT}]$ token is expected to be the pseudo label b_i^k.

To predict the relation predicates, for each mask pair $(\mathbf{m}_i^k, \mathbf{m}_j^k)$, the label generator takes $\{\hat{\mathbf{p}}\}_{i=1}^N$, the image masks $\mathbf{m}_i^k$ and $\mathbf{m}_j^k$, and the learnable positional embeddings $\mathbf{f}_{sub}$, $\mathbf{f}_{obj}$, $\mathbf{f}_{region}$ as input. For each mask pair, an additional region mask $\mathbf{m}_r^k$, *i.e.*,

$$\mathbf{m}_r^k = Rec(\mathbf{m}_i^k \cup \mathbf{m}_j^k) - (\mathbf{m}_i^k \cup \mathbf{m}_j^k),$$

is used to indicate the complement region of the relation, where Rec denotes the enclosing rectangle. The $\mathbf{f}_{sub}$, $\mathbf{f}_{obj}$, $\mathbf{f}_{region}$ are added to $\{\hat{\mathbf{p}}\}_{i=1}^N$ according to $\mathbf{m}_i^k$, $\mathbf{m}_j^k$, $\mathbf{m}_r^k$ respectively before decoding to indicate the different regions in the image tokens. With the enhanced image tokens and the union mask $\mathbf{m}_i^k \cup \mathbf{m}_j^k \cup \mathbf{m}_r^k$, the label generator uses a prompt

$$\textbf{a photo of }[\mathbf{SUB}]\textbf{ and }[\mathbf{OBJ}]$$

$$\textbf{what is their relation }[\mathbf{REL}]$$

to guide the relation generation, where the $[\mathbf{SUB}]$ and $[\mathbf{OBJ}]$ tokens are embedded by the pseudo labels b_i^k and b_j^k, and the $[\mathbf{REL}]$ token is expected to be the relation predicate between (b_i^k, b_j^k) with b_i^k as subject and b_j^k as object in the text graph.

Note that to reduce the noise in the pseudo object and relation labels from the caption-parsed text graphs, we change all pseudo labels into their lemma form for the generation.

A2.4. More Details of Inference

Different from the training procedure, during inference, the framework only takes a scene image I as input without its caption, so that the entity grounder is not used. With the given target concept sets of object semantics $\mathcal{C}_o$ and relation predicates $\mathcal{C}_r$, the goal for inference is to generate a PSG with its object and relation labels selected from $\mathcal{C}_o$ and $\mathcal{C}_r$.

During inference, an inference stage index l_{inf} is specified to generate the candidate image segments. The model firstly uses the region grouper to partition I into $H_{l_{inf}}$ segments $\{\mathbf{s}_i^{l_{inf}}\}_{i=1}^{H_{l_{inf}}}$, which are then merged by the segment merger based on the similarity matrix $\mathbf{Sim}_{l_{inf}}$. Ideally, after swapping rows and columns, $\mathbf{Sim}_{l_{inf}}$ should be a block diagonal matrix in $\{0,1\}^{H_{l_{inf}} \times H_{l_{inf}}}$ with a low rank, and the merging of segments can thus be formulated as a spectral clustering problem. However, $\mathbf{Sim}_{l_{inf}}$ is actually a noisy matrix in $[0,1]^{H_{l_{inf}} \times H_{l_{inf}}}$. To reduce the noise and perform a more accurate clustering, we employ a matrix recovery method [6] to recover the low-rank subspace structure of $\mathbf{Sim}_{l_{inf}}$, *i.e.*, by solving a convex optimization problem

$$\min_{\mathbf{Z}_{l_{inf}}, \mathbf{E}_{l_{inf}}} \|\mathbf{Z}_{l_{inf}}\|_* + \lambda \|\mathbf{E}_{l_{inf}}\|_{2,1},$$

$$\text{s.t. } \mathbf{Sim}_{l_{inf}} = \mathbf{Sim}_{l_{inf}} \mathbf{Z}_{l_{inf}} + \mathbf{E}_{l_{inf}},$$

where $\mathbf{Z}_{l_{inf}}$ denotes the recovered low-rank matrix, $\mathbf{E}_{l_{inf}}$ denotes the noise matrix, $\|\cdot\|_*$ denotes the nuclear norm, and $\|\cdot\|_{2,1}$ denotes the $l_{2,1}$ norm. λ is a hyperparameter that is set to 0.4 in our experiments.

Then the recovered matrix $\mathbf{Z}_{l_{inf}}$ is applied the normalized cut [13] for clustering, where the segments with similar object semantics tend to be merged into the same cluster. After this step, D merged segmentation masks $\{\hat{\mathbf{m}}_i\}_{i=1}^D$ are obtained.

For each merged mask $\hat{\mathbf{m}}_i$, the label generator uses a similar PET to predict the object label in $\mathcal{C}_o$, which are then be used to predict the relation label in $\mathcal{C}_r$. Different from training, here, the object labels and the relation labels are predicted in a cascaded manner. To select the label in $\mathcal{C}_o$ and $\mathcal{C}_r$, each candidate label is embedded into the prompt (at the $[\mathbf{ENT}]$ or the $[\mathbf{REL}]$ token) to compute its generation probability, which is then used in ranking to select the most probable as the final prediction. Here we use a greedy strategy in implementation to reduce the computation cost. Following the training procedure, all target concepts in $\mathcal{C}_o$ and $\mathcal{C}_r$ are changed into their lemma form for the generation.

A3. More Details of Experiments

A3.1. More Details of Datasets for Caption-to-PSG

In our experiments, we use the Panoptic Scene Graph dataset [19] for the evaluation of the problem Caption-to-PSG. Compared with this dataset, the commonly-used dataset Visual Genome (VG) [3] has three limitations that make it less suitable for our evaluation. Firstly, VG only uses bboxes for object location in scene graphs with no fine-grained segmentation masks provided. Secondly, the scene graphs in VG are not panoptic, in which only a few objects in the scenes are covered. Thirdly, the standard concepts [17] of object semantics and relation predicates in VG (*i.e.*, 150 objects and 50 relations) are not well-defined enough, where some similar and ambiguous concepts exist, such as *man*, *men*, *woman*, *person* for objects and *wears*, *wearing* for relations. In contrast, the Panoptic Scene Graph dataset not only provides object location in the form of both bboxes and segmentation masks, but also contains a more clear, more informative, more coherent class system with comprehensive and panoptic annotations, which is more suitable for the evaluation of Caption-to-PSG.

The original Panoptic Scene Graph dataset contains 133 object semantics and 56 relation predicates. However, in the original 133 object semantics, there are still some ambiguous classes not well-defined, such as *window-blind* and *window-other*, *floor-wood* and *floor-other-merged*. To reduce the ambiguity during evaluation, we further merge the ambiguous object semantics with their corresponding annotations, *i.e.*, *window-blind*, *window-other* into *window*; *floor-wood*, *floor-other-merged* into *floor*; *wall-brick*, *wall-stone*, *wall-tile*, *wall-wood*, *wall-other-merged* into *wall*. After merging, 127 object semantics and 56 relation predicates are obtained for our evaluation.

Note that the final set of 127 object semantics consists of 80 thing classes, which represent object classes that can be individually recognized and segmented in an image, and 47 stuff classes, which represent object classes that usually have a homogeneous texture or pattern and are difficult to be segmented individually. In the Panoptic Scene Graph dataset, objects belonging to stuff classes are not segmented individually, with each stuff class having only one mask at most. To accommodate this approach, during the evaluation of our method and the baselines on Caption-to-PSG, the predicted objects with the same stuff class are merged into a single object.

A3.2. More Details of Baselines for Caption-to-PSG

Firstly, we design four baselines that strictly follow the constraints of Caption-to-PSG for a fair comparison. In these baselines, objects in scenes are located by bbox proposals generated by selective search [15], which requires no location priors or supervision. For each scene image, 50 proposals are generated.

- **Random** predicts all object semantics and relation predicates fully randomly, where the score for each label is randomly selected from $[0, 1]$.
- **Prior** augments **Random** by predicting labels based on the statistical priors in the training set. Specifically, during inference, the model collects the distribution of the target concepts $\mathcal{C}_o$ and $\mathcal{C}_r$ in the training set, then follows the distribution frequency to predict the score in $[0, 1]$ for each label.
- **MIL** performs the alignment between proposals and textual entities, using a multiple instance learning [8] strategy to match the proposals and the entities in captions implicitly. The object label prediction is formulated as a classification problem in a large pre-built vocabulary. Specifically, similar to [20], the model builds a large object vocabulary with the most frequent 4,000 entities in the captions in the training set, and the training procedure for object prediction is a 4000-class classification problem. During inference, the model employs WordNet [9] to match the 4000 classes with the target concepts $\mathcal{C}_o$. Once the object labels are predicted, the relation labels in $\mathcal{C}_r$ are predicted with the statistical prior, similar to **Prior**.
- **SGCLIP** employs the pre-trained CLIP [10] to predict both object semantic labels and relation predicate labels. Specifically, for objects, the model uses a prompt

  $$\textbf{a photo of a } [\textbf{ENT}]$$

  to obtain the embedding for each object label in $\mathcal{C}_o$, and assigns the label with the highest cosine similarity to the proposal as the prediction. For relations, the model uses a prompt

  $$\textbf{a photo of a } [\textbf{SUB}]\ [\textbf{REL}] \textbf{ a } [\textbf{OBJ}]$$

  to obtain the embedding for each relation label in $\mathcal{C}_r$ for each object pair, and assigns the label with the highest cosine similarity as the prediction.

By gradually removing the constraints of Caption-to-PSG, we set two additional baselines to further benchmark the performance of our framework, based on the previous work [20] for weakly-supervised scene graph generation.

- **SGGNLS-o** [20] is built without the constraint of no location priors. It extracts object proposals with a detector [11] pre-trained on OpenImage [4]. Following [20], on average, 36 object proposals are extracted for each image. It formulates the label prediction as a classification problem within a large pre-built vocabulary, where a 4,000-class object semantics vocabulary and a 1,000-class relation predicate vocabulary are built from the most frequent 4,000 entities and 1,000 relations in the captions in the training set. During inference, the model employs WordNet [9] to match the 4000 object classes with the target concepts $\mathcal{C}_o$ and 1,000 relation classes with $\mathcal{C}_r$.
- **SGGNLS-c** [20] is built without the constraint of no location priors and no pre-defined concept sets, based on **SGGNLS-o**. It uses the same proposals as **SGGNLS-o**. In **SGGNLS-c**, the target concept sets for inference are known during training. It formulates the label prediction as a classification problem within $\mathcal{C}_o$ and $\mathcal{C}_r$, where all entities and relations from captions in the training set are pre-mapped to $\mathcal{C}_o$ and $\mathcal{C}_r$ through an accurate human-refined mapping as pseudo labels during training.

A3.3. More Details of Implementation

In TextPSG, the input image resolution for training is 384×384, and the resolution for inference is 512 for the shortest side. The patch size of the region grouper is 16.

The filtering threshold in the entity grounder is set to -0.5. We train TextPSG on the COCO Caption dataset [2] for 100 epochs. We use a batch size of 1,728, a learning rate of 0.0001, and the AdamW optimizer [7] with weight decay as 0.05.

A4. More Results on Caption-to-PSG

A4.1. More Ablation Studies

Here we conduct additional ablation studies to further evaluate the effectiveness of two design choices in our framework.

Positional Embeddings in PET. In Tab. 1, we compare the different strategies for indicating the different regions in the image tokens in PET. Based on the full PET in TextPSG (row 3), we first remove the region embedding $\mathbf{f}_{region}$ (row 2) and further remove the subject embedding $\mathbf{f}_{sub}$ as well as the object embedding $\mathbf{f}_{obj}$ (row 1). The results show that the design of $\mathbf{f}_{sub}$ and $\mathbf{f}_{obj}$ is very important to the generation, without which the model will suffer a significant performance drop. And the design of $\mathbf{f}_{region}$ can further improve the performance by indicating the compliment region information in the image tokens.

$\mathbf{f}_{sub}$	$\mathbf{f}_{obj}$	$\mathbf{f}_{region}$	PhrDet		SGDet	
			N3R100	N5R100	N3R100	N5R100
✗	✗	✗	2.33	2.58	0.45	0.6
✓	✓	✗	10.67	11.3	2.81	3.21
✓	✓	✓	**12.74**	**14.37**	**4.77**	**5.48**

Table 1. **Ablation Study on Positional Embeddings in PET.** '$\mathbf{f}_{sub}$', '$\mathbf{f}_{obj}$', and '$\mathbf{f}_{region}$' denotes the learnable positional embeddings for indicating the subject region, the object region, and the complement region in the image tokens.

Filtering Threshold. In Tab. 2, we investigate the effectiveness of setting a filtering threshold θ to filter out the mismatched image region and caption entity pairs. The results show that compared with the region-entity alignment without filtering (row 1), the introduced θ (row 2) is simple yet effective in improving the performance significantly.

Thresh	PhrDet		SGDet	
	N3R100	N5R100	N3R100	N5R100
✗	10.39	10.8	3.09	3.19
✓	**12.74**	**14.37**	**4.77**	**5.48**

Table 2. **Ablation Study on Filtering Threshold.** 'Thresh' denotes the filtering threshold θ for filtering out the mismatched image region and caption entity pairs.

A4.2. More Model Diagnosis.

Here we provide more diagnoses of our framework for a clearer understanding of the efficacy. We answer the following questions. **Q1**: How significantly does the pre-trained GroupViT [18] enhance the learning our framework? **Q2**: How does our framework perform with partial ground truth given? **Q3**: How does our framework perform with BLIP [5] replaced by CLIP [10] for the label prediction?

Pre-trained Weights	PhrDet		SGDet	
	N3R100	N5R100	N3R100	N5R100
✗	0	0	0	0
COCO Caption [2]	1.99	2.51	0.07	0.1
CC12M [1, 12]+YFCC [14]	**12.74**	**14.37**	**4.77**	**5.48**

Table 3. **Examination on Pre-trained GroupViT Weights.**

In Tab. 3, we examine the efficacy of the pre-trained GroupViT [18] in two more training settings: no pre-trained GroupViT weights are used (row 1); initializing weights of GroupViT pre-trained solely on the COCO Caption dataset [2] (row 2). The results show that a pre-trained GroupViT is necessary for the effectiveness of our model. Furthermore, GroupViT pre-trained on a large dataset (row 3) can provide very strong location priors and thus facilitates our model significantly (answering **Q1**).

Method	SGCls		PredCls		SGDet	
	N3R100	N5R100	N3R100	N5R100	N3R100	N5R100
PSGCLIP	7.38	9.11	25.72	26.16	2.83	3.23
Ours	**9.51**	**10.79**	**36.28**	**39.79**	**4.77**	**5.48**

Table 4. **More Evaluation Settings.**

We evaluate the performance of our model on two additional settings with partial ground truth: (i) **SGCls**, where ground truth object masks are known; (ii) **PredCls**, where ground truth object masks and semantics are known. The correctness definition is the same as **SGDet**. The results are shown in Tab. 4 row 2. The results show that both the segmentation and the relation/entity label prediction still have a large space to improve, especially the label prediction. A better method for label prediction in our challenging setting may improve the performance significantly (answering **Q2**).

Substituting BLIP with CLIP in our framework for the label prediction, akin to **PSGCLIP**, results in performance decline across all settings as per Tab. 4. The significant drop in **PredCls** demonstrates CLIP's insensitivity to nuanced relation predicates (answering **Q3**).

A4.3. More Visualization for Qualitative Evaluation

We provide more visualization of the predicted PSGs by TextPSG in Fig. 1 for further qualitative evaluation, comparing with the baseline **SGGNLS-o**.

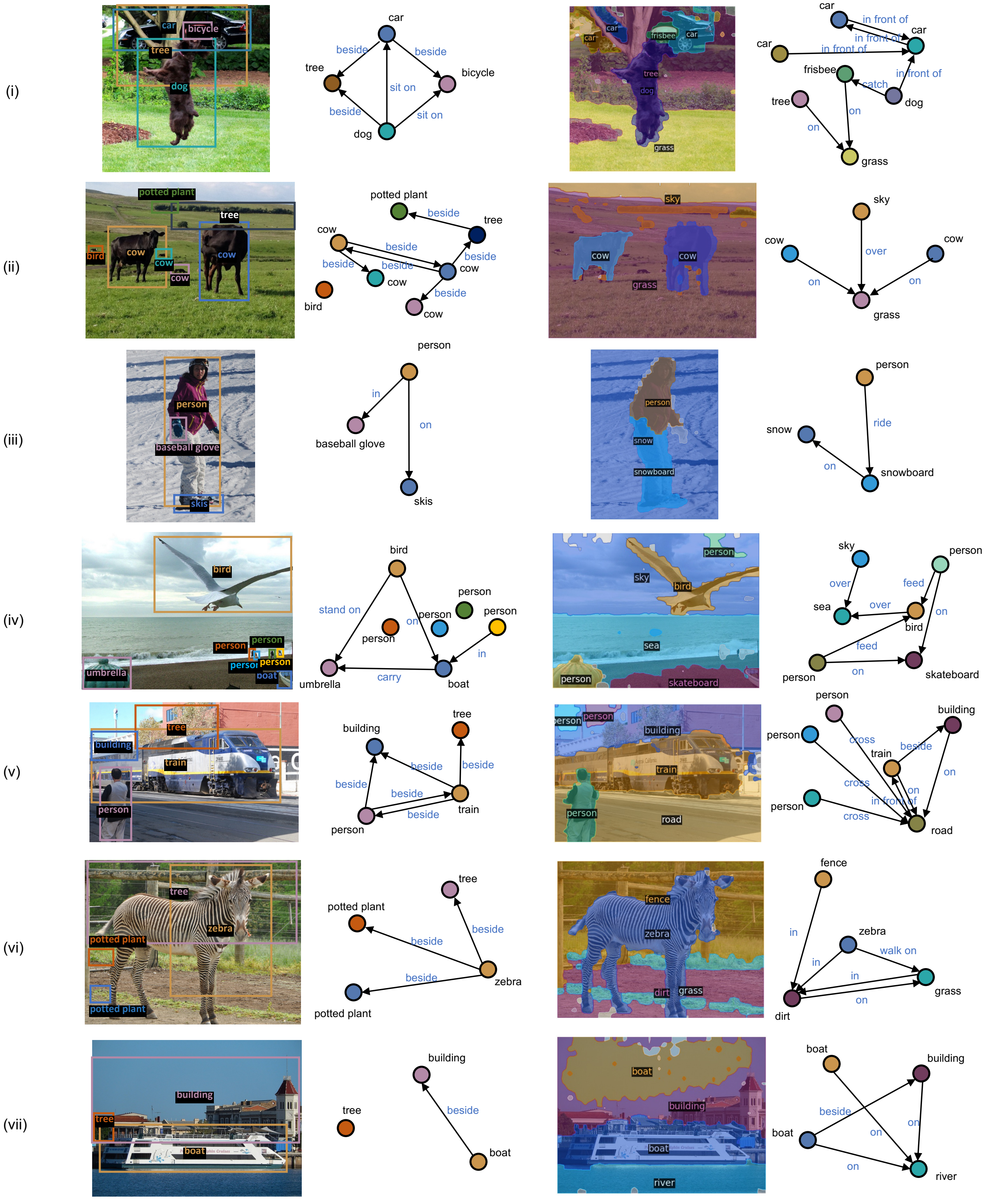
A4.4. Example of Failure Cases

Compared with the baseline **SGGNLS-o**, Fig. 1 shows that our framework is capable of providing more fine-grained labels to each pixel in the image, and is able to reach a panoptic understanding of the scene. However, there are some limitations to our framework that result in some failure cases.

Firstly, the strategy we use to convert the semantic segmentation into instance segmentation is not entirely effective. As shown in Fig. 1, our strategy can successfully separate the two cows in (ii), but mistakenly divides the car behind the tree into three parts in (i).

Secondly, our framework faces difficulty in locating small objects in the scene due to limitations in resolution and the grouping strategy for location. As shown in Fig. 1 (ii) and (iv), our method can identify large objects such as large cows, birds, grass, and sea, but struggles to locate relatively small objects such as small cows in (ii) and people in (iv).

Thirdly, the relation prediction of our framework requires enhancement, as it is not adequately conditioned on the image. As shown in Fig. 1 (i), the relations between the blue mask of the car and the green mask of the car are predicted as both being *in front of*, which is not reasonable. In this case, *beside* may be a more appropriate prediction (in this case, the first limitation about the segmentation conversion also exists).



(a) Results of SGGNLS-o

(b) Results of Ours

Figure 1. **More Qualitative Comparison between SGGNLS-o (a) and Ours (b).** For each method, the results of object location are shown on the left, while the results of scene graph generation are shown on the right. For **SGGNLS-o** and **Ours**, the visualized relations are picked from the top 10 triplets in the scene graph (the predicate score should be greater than 0.6). For **SGGNLS-o**, only proposals matched with ground truth (only requires a correct location, ignores the semantics) are visualized.

References

[1] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 4

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 2

[4] Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956–1981, 2018. 3

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4

[6] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012. 2

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[8] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 3

[9] George A. Miller. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. 3

[10] 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, 2021. 3, 4

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, July 2018. 4

[13] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 2

[14] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 4

[15] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104:154–171, 2013. 3

[16] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[17] 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

[18] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144, June 2022. 1, 4

[19] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 2

[20] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *ICCV*, 2021. 3