DSGG: Dense Relation Transformer for an End-to-end Scene Graph Generation Supplementary Material

Zeeshan Hayder[1], Xuming He[2],
[1]Data61-CSIRO, Australia, [2]ShanghaiTech University,
zeeshan.hayder@data61.csiro.au, hexm@shanghaitech.edu.cn

In this supplementary material, we provide further details on the proposed approach and present supplementary results. Additionally, we showcase the qualitative results obtained from the DSGG method for the scene graph generation and the panoptic scene graph generation tasks.

1. Dense Scene Graph Generation (DSGG)

In this section, we first provide a summary highlighting the key contributions of our method, along with comparisons to [4] and [7]. Following this, we offer an additional comparison with scene graph detection methods using the Visual Genome dataset. Additionally, we present results with no-graph constraints and conduct a per-class performance comparison between our method and [11].

1.1. Comparison to [4] and [7]

The primary factor contributing to achieving state-of-the-art performance while significantly reducing parameters is that DSGG does not depend on a node-based matching approach to acquire queries, which is customary in all transformer-based approaches (including [4] and [7]). In contrast, DSGG employs sub-graph matching as a guiding mechanism to acquire graph-aware queries, thereby enabling direct learning of the scene graph and ultimately leading to a compact model with improved performance.

Specifically, in [4], each entity has K distinct queries for K triplets, and the attention mask is derived from the maximum attention of nK features using softmax. Conversely, in DSGG, each entity is represented by a unique graph-query for all its relations, and the attention is learned via sigmoid to accommodate multiple relations among all the objects. DSGG exhibits improvements of 4.9% & 6.6% in mR@50/100 respectively, as shown in Table T1.

In [7], an additional [rln]-token is learned alongside N [obj]-tokens to train model. Moreover, a major constraint of their approach is its dependence on predicted objects before relation classification among them (Sec 3.4 in [7]), a process distinct from the sub-graph matching that is employed in DSGG. Furthermore, as in Table T1, DSGG shows enhancements of 10.9% and 14.8% in mR@50/100, respectively.

	Scene Graph Detection (SGDet)					
Method	**R@50**	**R@100**	**mR@50**	**mR@100**	**M@50**	**M@100**
Px2Graph [6]	15.5	18.8	-	-	-	-
FCSGG [5]	21.3	25.1	3.6	4.2	12.5	14.7
CoRF+T [1]	18.6	-	3.9	-	-	-
RelTR [2]	27.5	30.7	10.8	12.6	-	-
Relationformer [7]	28.4	31.3	9.3	10.7	18.9	21.0
TraCQ [3]	28.3	35.7	13.8	14.6	-	-
RepSGG [4] †	29.6	34.8	9.3	11.4	19.5	23.1
RepSGG [4]	12.1	14.6	15.3	18.9	13.7	16.8
DSGG (ours) †	**32.9**	**38.5**	13.0	17.3	23.0	28.0
DSGG (ours)	26.5	32.9	**20.2**	**25.5**	**23.4**	**29.2**

Table T1. Additional Evaluation on the Visual Genome test set. The references to the cited works will be added in the final paper.

1.2. Additional comparisons on the VG dataset

In the paper, we included only a representative selection of recent works. Nevertheless, we have included results from [1–7], in Table T1. Additionally, it is worth noting that DSGG outperforms all VG baseline methods in terms of mR@K and M@K metrics.

1.3. Ng-Recall Performance

Table T2 presents DSGG attaining state-of-the-art performance even on the no-graph constraint metric. Note that the DSGG attains a notable improvement on the no-graph constraint mean recall metric.

	ng-Recall			ng-Mean Recall		
Method	**@20**	**@50**	**@100**	**@20**	**@50**	**@100**
MOTIFS [10]	-	30.5	35.8	-	-	-
FCSGG [5]	19.6	26.8	32.1	4.2	6.5	8.6
Relationformer [7]	22.9	31.2	36.8	-	-	-
DSGG (ours)	**23.0**	**32.5**	**39.4**	**9.8**	**15.2**	**18.1**

Table T2. SGDet: No-graph constraint results on the VG test set.

1.4. Per-class Performance

Figure S1 compares the per-class mean-recall (mR@100) performance of our method and HiLo [11] using the Resnet-50 backbone. Note that our approach demonstrated superior mean-recall for 45 categories, yielded similar outcomes for 8 categories, and only exhibited lower performance for 3 categories (*'beside', 'in front of', and 'on back of'*). In addition to that, our method outperformed the baseline method in both rare and non-rare categories.

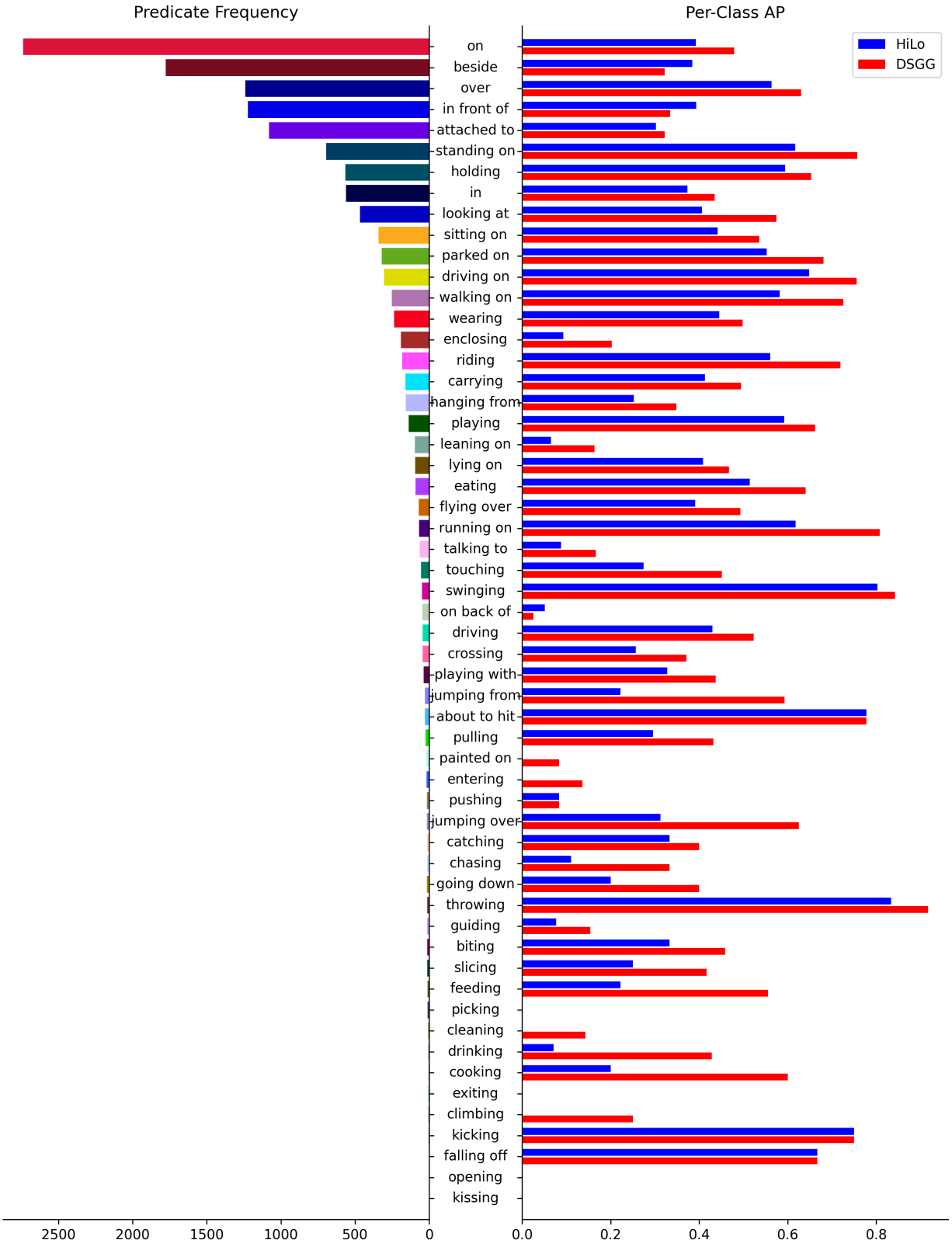


Figure S1. **Per-Class Performance** on the PSG dataset. **Left** column shows the predicate names and their frequency in the test split. **Right** column shows the per-category results of the HiLo method [11] in blue color and our method results in red color.

2. Qualitative Performance

In this section, we provide the qualitative results of the proposed DSGG method on both the scene-graph generation and the panoptic scene-graph generation datasets.

2.1. Panoptic Scene Graph Generation

In this section, we show the qualitative results of our method on the PSG [9] dataset. Because of limited space, we only present the top 9 triplets along with their corresponding subject and object masks. The subjects are highlighted in blue color and the objects are highlighted in red color. The <subject, predicate, object> relationships are shown in the white color. Figure S2 shows that our method can predict multiple simultaneous relationships among the same subject and object. Specifically, our method predicts that in this image a) the sky is over the sea, b) the person is standing on the surfboard, c) the sky is over the person, d) the person is playing the surfboard, e) the sky is over the sea, f) sky over the surfboard, g) person touching the sea, h) person on the sea, and i) person is playing with the surfboard. Figure S3 contains a) the dog is biting the frisbee, b) the dog is playing with the frisbee, c) the dog is running on the grass, d) the dog playing frisbee, e) the dog is walking on the grass, f) the dog is standing on the grass, g) the dog is catching the frisbee, h) the dog is sitting on the grass, and i) the dog is on the grass. In Figure S4, our method predicts a) the sky is over the grass, b) the sky is over the tree, c) the kite is flying over the sky, d) the sky is over the kite, e) the sky is over the person, f) the person is pulling the kite, g) the person is standing on the grass, h) the sky is over the house, and i) the person is holding the kite. Figure S5 shows that a) the person swinging a baseball bat, b) the person wearing a baseball glove, c) the person holding a baseball bat, d) the person standing on the playing field, e) the person standing on the dirt, f) the person wearing the baseball glove, g) the person is about to hit the sports ball, h) the person running on the playing field, and i) the person looking at the person.

2.2. Scene Graph Generation

This section shows our method results on the Visual Genome [8] dataset. Figures S6, S7, S8, and S9 show the ground truth, predicted entities, and their relationships as a bipartite graph for easy visualization. Note that our method can generate more meaningful relationships.

It is worth mentioning that the PSG dataset (as in Figure S2, S3, S4, and S5) has minimal noise, dense relations, and segmentation masks for enhanced model learning, while the VG dataset is sparse, noisy, includes object-part relationships, and lacks complete annotations. The ground truth in the qualitative results demonstrates this effect.

References

[1] George Adaimi, David Mizrahi, and Alexandre Alahi. Composite relationship fields with transformers for scene graph generation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 52–64, 2023. 1

[2] Y. Cong, M. Yang, and B. Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 45(09):11169–11183, 2023. 1

[3] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Single-stage visual relationship learning using conditional queries. In *Advances in Neural Information Processing Systems*, 2022. 1

[4] Hengyue Liu and Bir Bhanu. Repsgg: Novel representations of entities and relationships for scene graph generation, 2023. 1

[5] Hengyue Liu, Ning Yan, Masood S. Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11541–11551, 2021. 1

[6] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *Advances in Neural Information Processing Systems*, 2017-December:2172–2181, 2017. 31st Annual Conference on Neural Information Processing Systems, NIPS 2017 ; Conference date: 04-12-2017 Through 09-12-2017. 1

[7] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, and Bjoern Menze. Relationformer: A unified framework fornbsp;image-to-graph generation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 422–439, Berlin, Heidelberg, 2022. Springer-Verlag. 1

[8] 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 (CVPR)*, 2020. 3

[9] 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

[10] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[11] Zijian Zhou, Miaojing Shi, and Holger Caesar. Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21637–21648, 2023. 1, 2

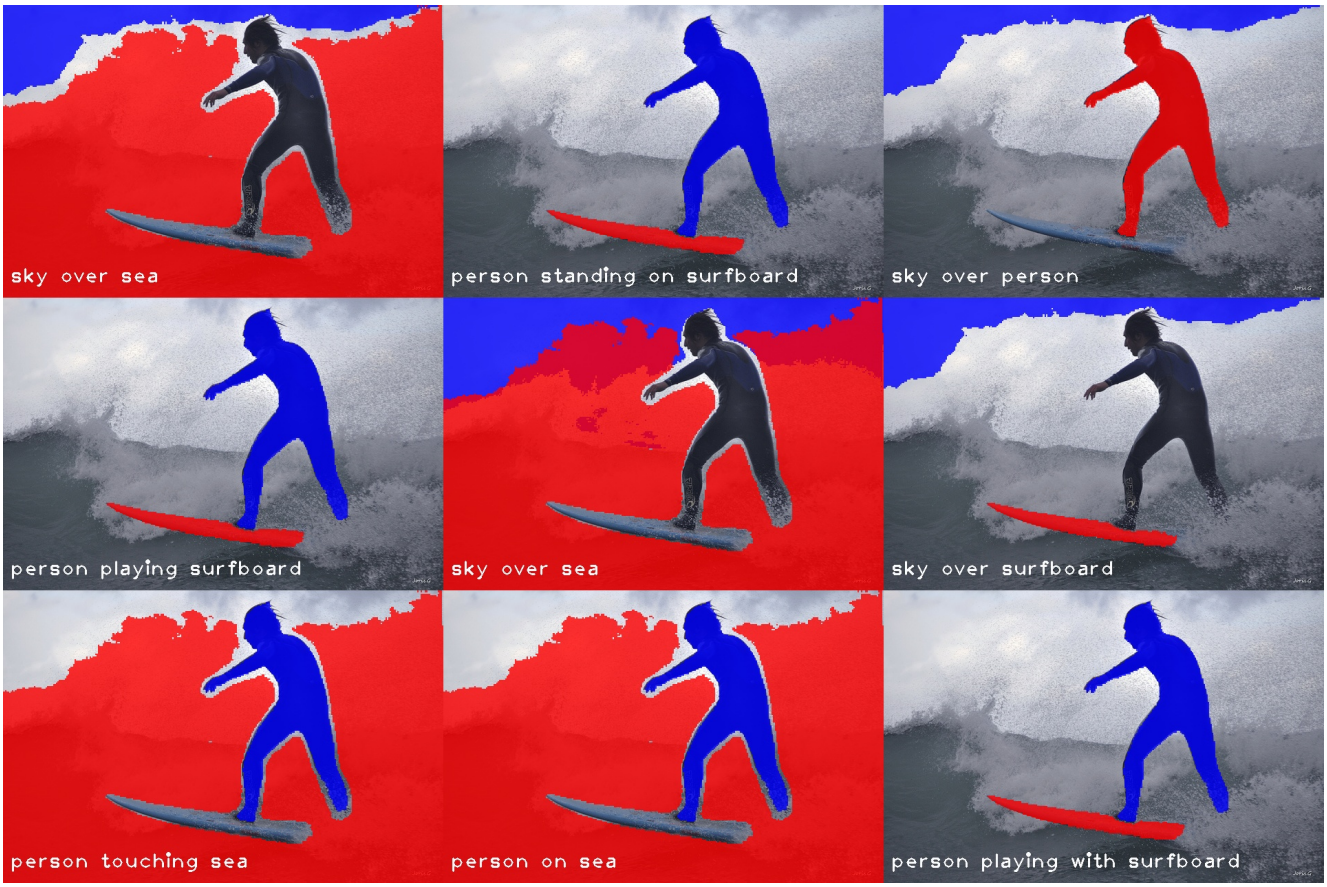


Figure S2. **Qualitative Performance** on the PSG dataset. The subject, object, and relationship triplets are shown in blue, red, and white colors respectively. Specifically, our method predicts that in this image a) the sky is over the sea, b) the person is standing on the surfboard, c) the sky is over the person, d) the person is playing the surfboard, e) the sky is over the sea, f) sky over surfboard, g) person touching the sea, h) person on the sea, and i) person is playing with the surfboard.

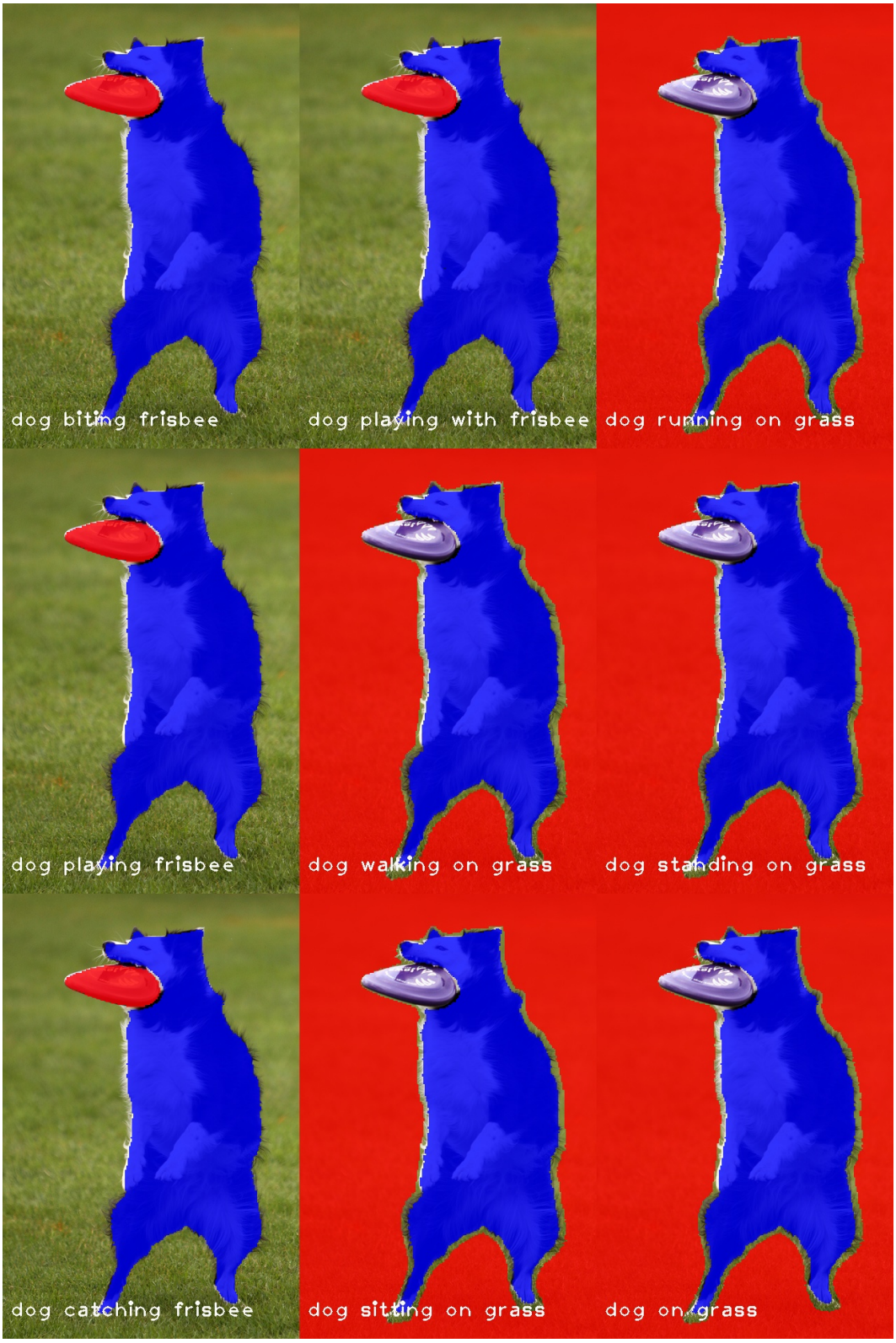


Figure S3. **Qualitative Performance** on the PSG dataset. The subject, object, and relationship triplets are shown in blue, red, and white colors respectively. Specifically, our method predicts that in this image a) the dog is biting the frisbee, b) the dog is playing with the frisbee, c) the dog is running on the grass, d) the dog playing frisbee, e) the dog is walking on the grass, f) the dog is standing on the grass, g) the dog is catching the frisbee, h) the dog is sitting on the grass, and i) the dog is on the grass.

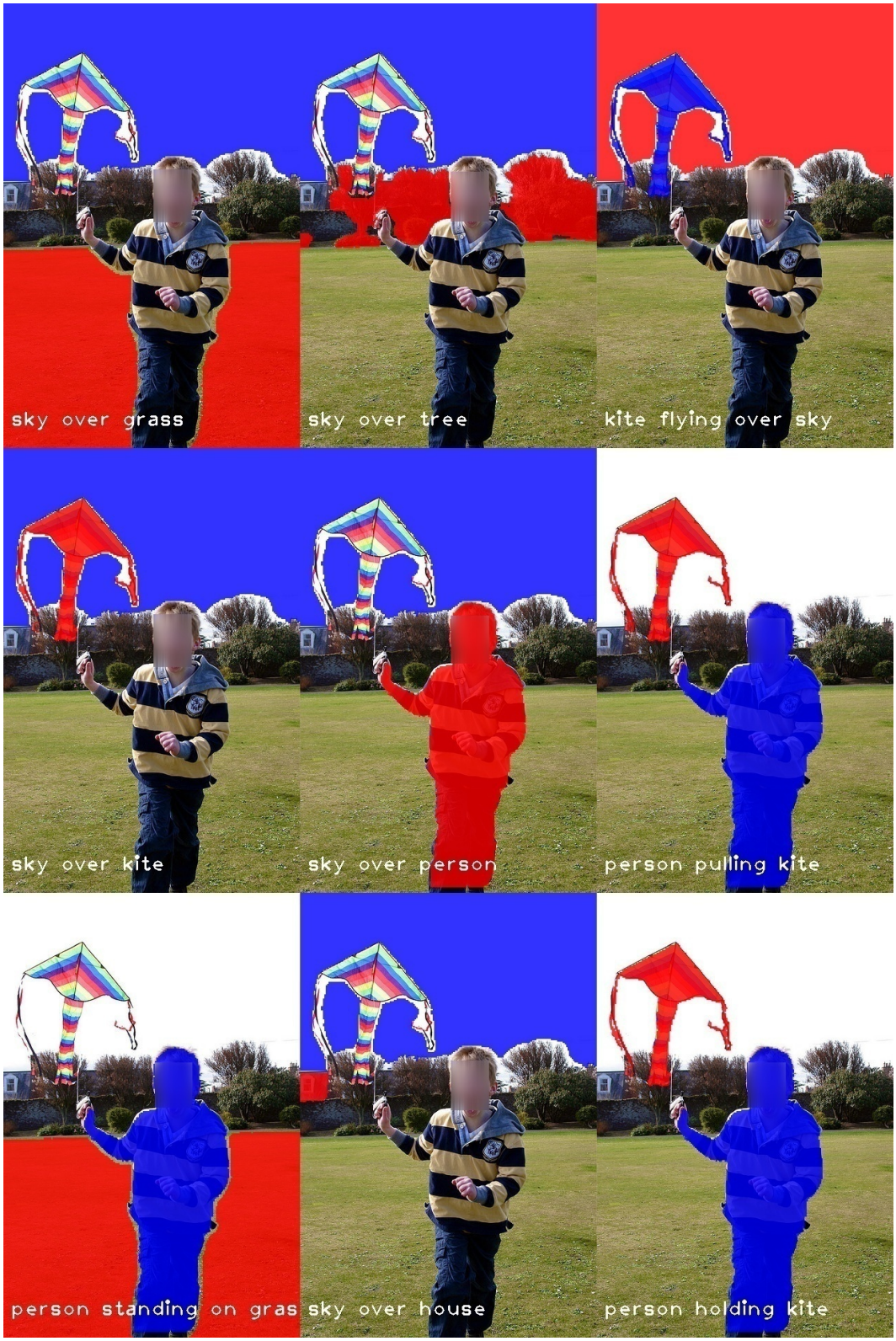
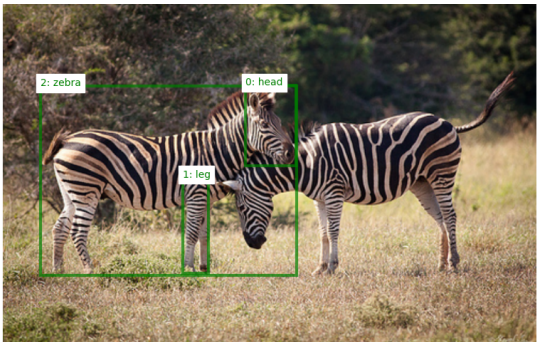
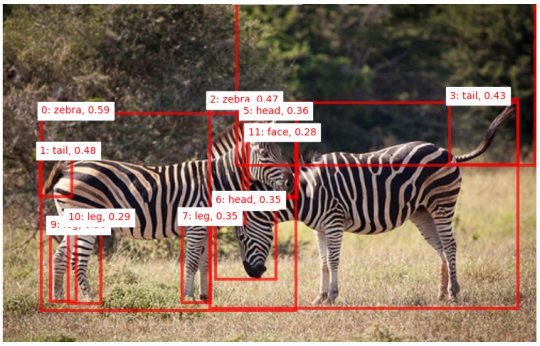


Figure S4. **Qualitative Performance** on the PSG dataset. The subject, object, and relationship triplets are shown in blue, red, and white colors respectively. Specifically, our method predicts that in this image a) the sky is over the grass, b) the sky is over the tree, c) the kite is flying over the sky, d) the sky is over the kite, e) the sky is over the person, f) the person is pulling the kite, g) the person is standing on the grass, h) the sky is over the house, and i) the person is holding the kite.

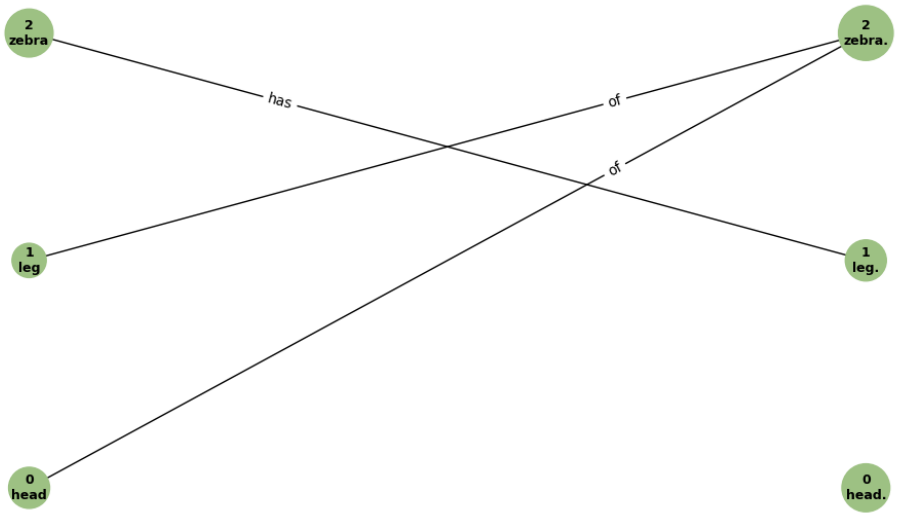
Figure S5. **Qualitative Performance** on the PSG dataset. The subject, object, and relationship triplets are shown in blue, red, and white colors respectively. Specifically, our method predicts that in this image a) the person swinging a baseball bat, b) the person wearing a baseball glove, c) the person holding a baseball bat, d) the person standing on the playing field, e) the person standing on the dirt, f) the person wearing the baseball glove, g) the person is about to hit the sports ball, h) the person running on the playing field, and i) the person looking at the person.



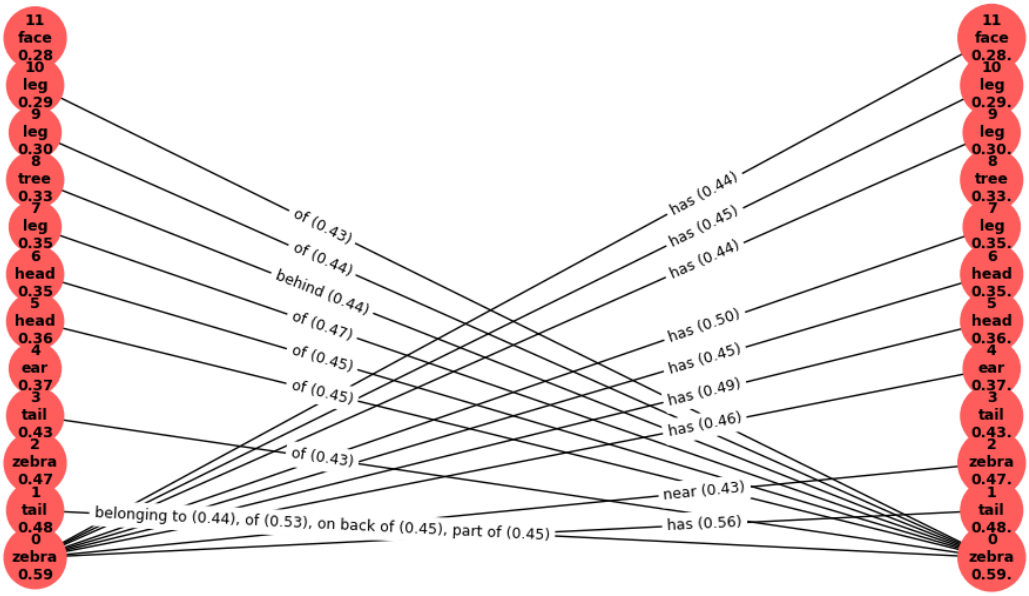
(a) Ground-truth objects with class labels



(b) Detected objects with class labels and score



(c) Ground-truth scene graph

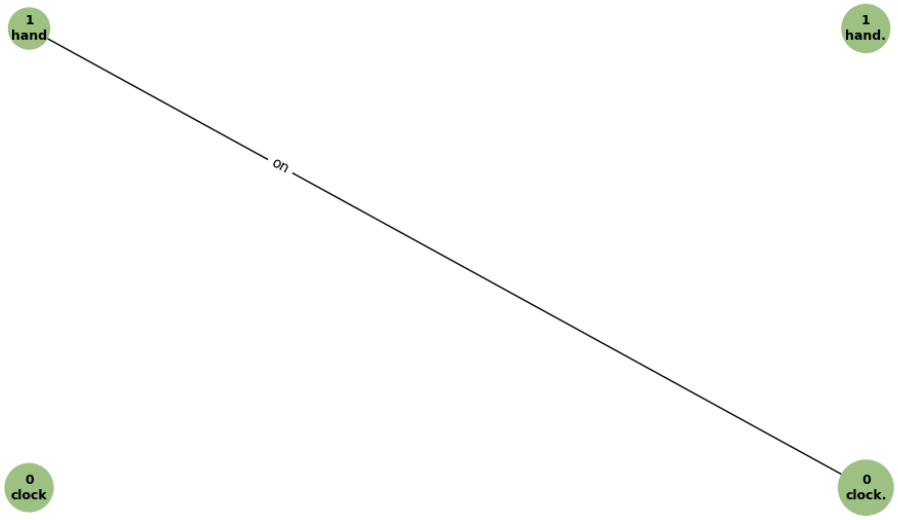


(d) Detected dense scene-graph using DSGG

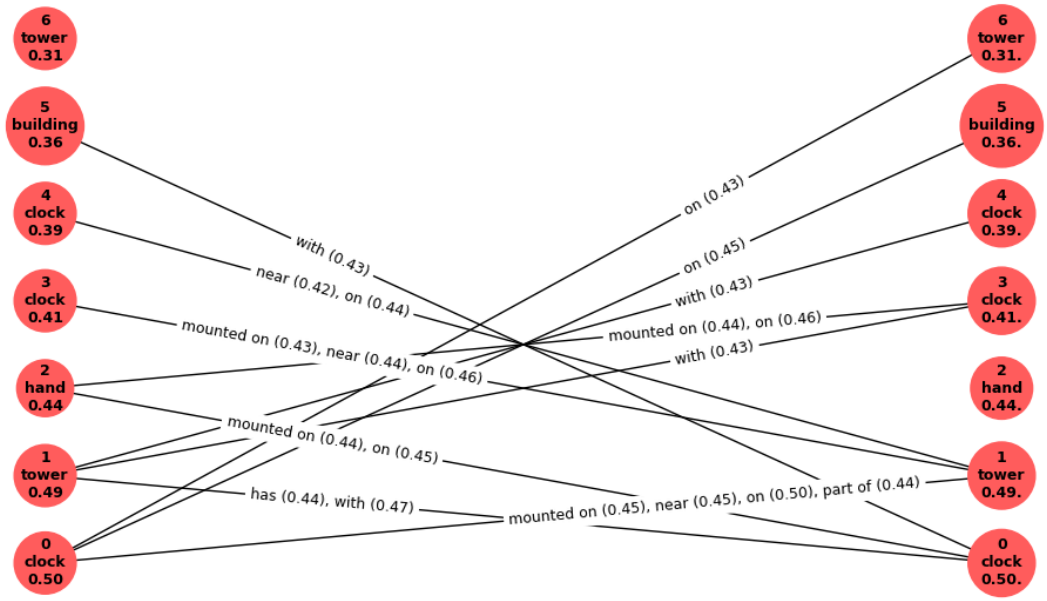
Figure S6. **Qualitative Performance** on the Visual Genome dataset. In the bipartite graph, the left side represents the subjects, while the right side represents the objects. The edges in the graph represent the top-20 detected relations between the subjects and objects.

(a) Ground-truth objects with class labels

(b) Detected objects with class labels and score



(c) Ground-truth scene graph

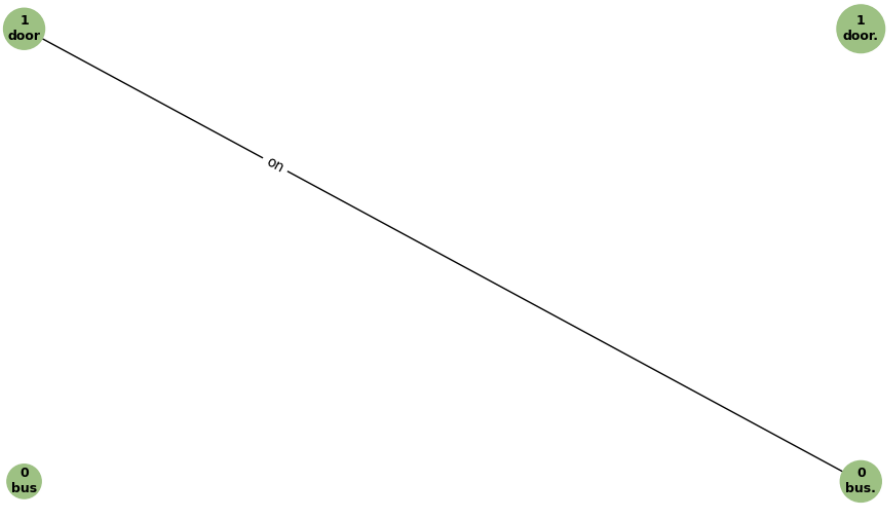


(d) Detected dense scene-graph using DSGG

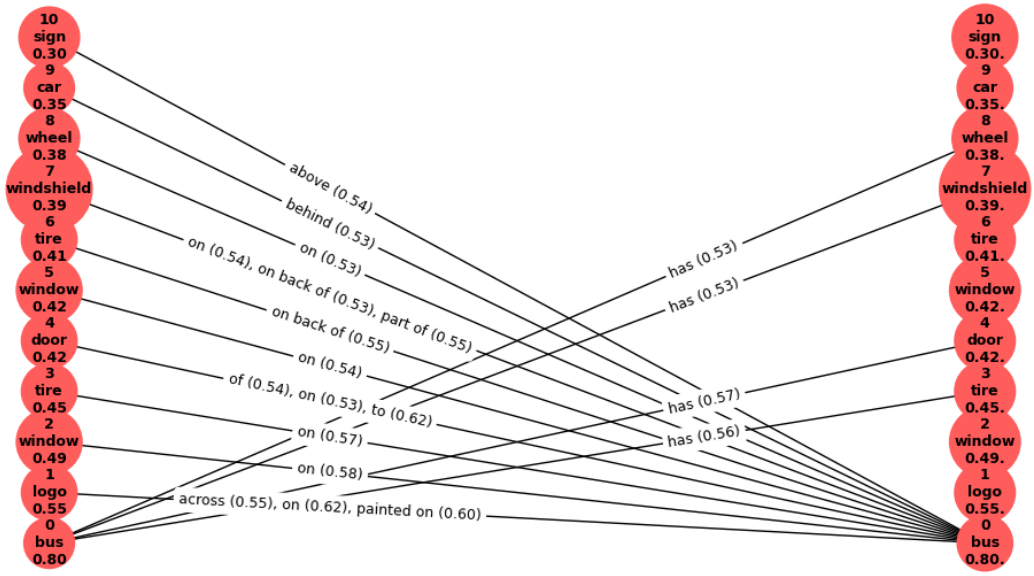
Figure S7. **Qualitative Performance** on the Visual Genome dataset. In the bipartite graph, the left side represents the subjects, while the right side represents the objects. The edges in the graph represent the top-20 detected relations between the subjects and objects.

(a) Ground-truth objects with class labels

(b) Detected objects with class labels and score

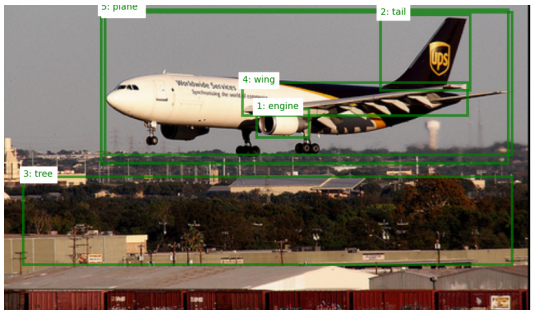


(c) Ground-truth scene graph

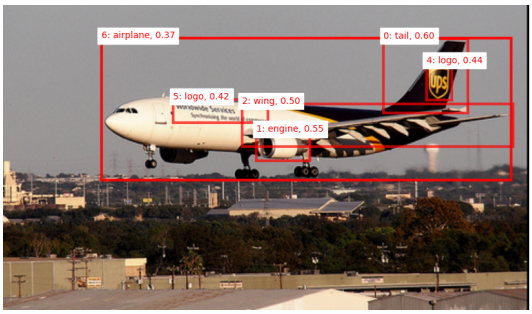


(d) Detected dense scene-graph using DSGG

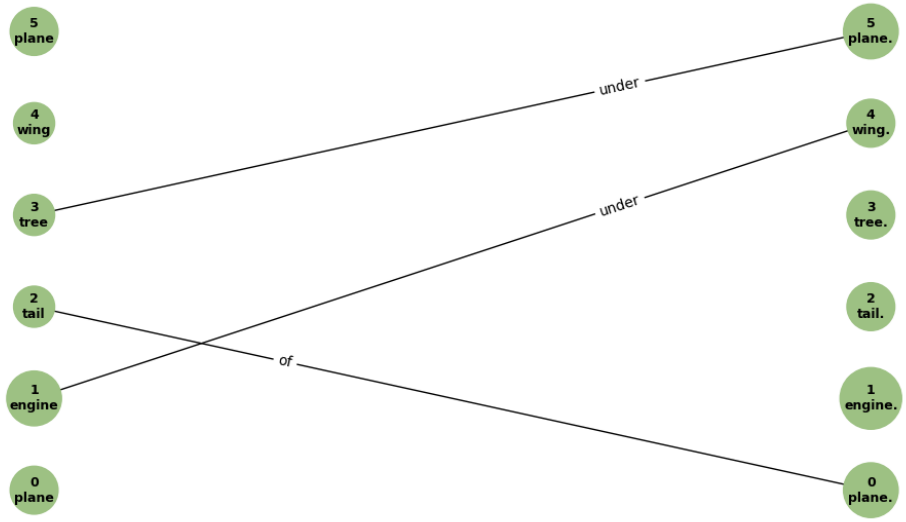
Figure S8. **Qualitative Performance** on the Visual Genome dataset. In the bipartite graph, the left side represents the subjects, while the right side represents the objects. The edges in the graph represent the top-20 detected relations between the subjects and objects.



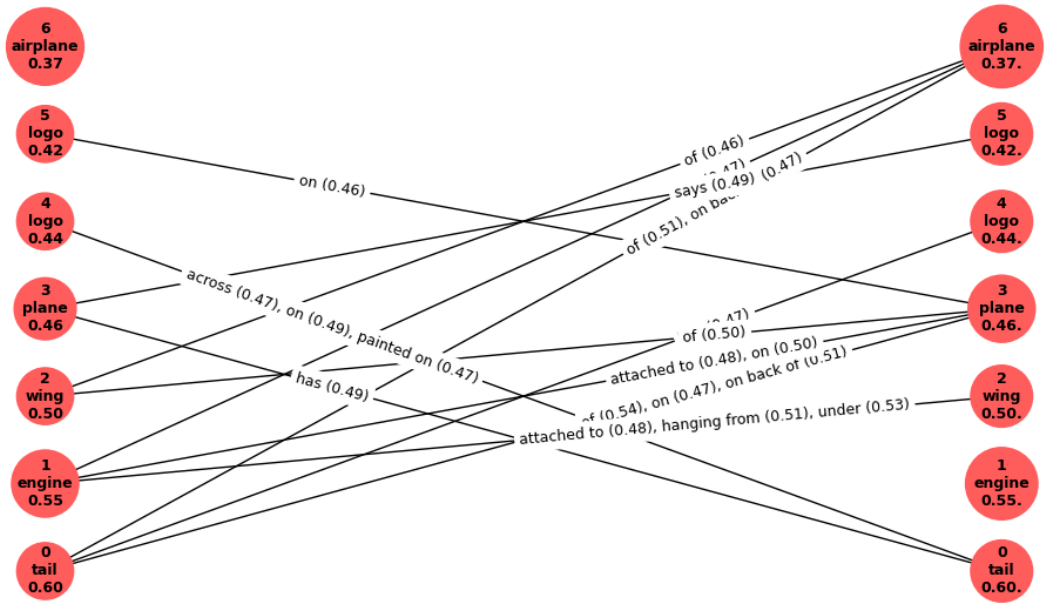
(a) Ground-truth objects with class labels



(b) Detected objects with class labels and score



(c) Ground-truth scene graph



(d) Detected dense scene-graph using DSGG

Figure S9. **Qualitative Performance** on the Visual Genome dataset. In the bipartite graph, the left side represents the subjects, while the right side represents the objects. The edges in the graph represent the top-20 detected relations between the subjects and objects.