Effective Scene Graph Generation by Statistical Relation Distillation
—— Supplementary ——

Thanh-Son Nguyen[1] Hong Yang[1,2] Basura Fernando[1,2]
[1]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research, Singapore
{nguyen_thanh_son, yang_hong,fernando_basura}@ihpc.a-star.edu.sg

In this supplementary, we show additional results that were excluded from the main submission due to length constraint.

1. Impact of Frequent Predicates on Recall

Our hypothesis is that the Recall scores are significantly impacted by a few frequent predicates. To test this, we remove the top K most frequent predicates and then calculate recall for the predicate classification task. Table 1 shows that baseline models have high recall on the entire dataset ($K = 0$), but recall drops sharply when frequent predicates are excluded. For example, the Transformer baseline's recall at 20 falls from 58.7 to 29.74 after removing the top 5 most frequent predicates. Recall scores for baseline methods continue to decrease with higher values of K. Conversely, methods trained on our enhanced dataset show improved trends, supporting our goal of reducing predicate bias in the training set. Our method outperforms the baselines for all three models when frequent predicates are removed ($K = 5$ and higher). Unlike the baselines, our method shows higher recall scores as more frequent predicates are excluded, with substantial differences at larger K values. For instance, with $K = 25$, the Transformer model trained on our enhanced dataset achieves a recall at 20 of 41.47%, compared to 8.99% when trained on the original dataset. This supports our hypothesis and demonstrates the effectiveness of our method in reducing bias in the training dataset.

2. Example for Context-based SRD Motivation

Figure 1 shows the prior pair probability across all the 25 clusters of VG for four pairs of objects. It shows that the likelihood of having a relation between two objects are varied for different clusters. For example, in the context of 'beach', i.e., in cluster 8 and 12 where pair probability between 'man' and 'beach' is high, 'man is more likely to have relation with 'surfboard' than with 'laptop' or 'table'. Similarly, both 'laptop' and 'table' appear to be able to have relation with 'man' in the same context (cluster 22). The visualization of images in each cluster can be found in Figures 10, 11, and 12.

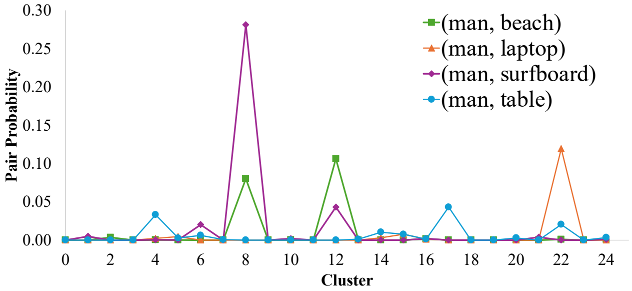


Figure 1. Prior pair probability across all the clusters, showing the dependent of priors on the context. For example, 'man' is more likely to have relation with 'surfboard' than with 'laptop' in the 'beach' context (cluster 8).

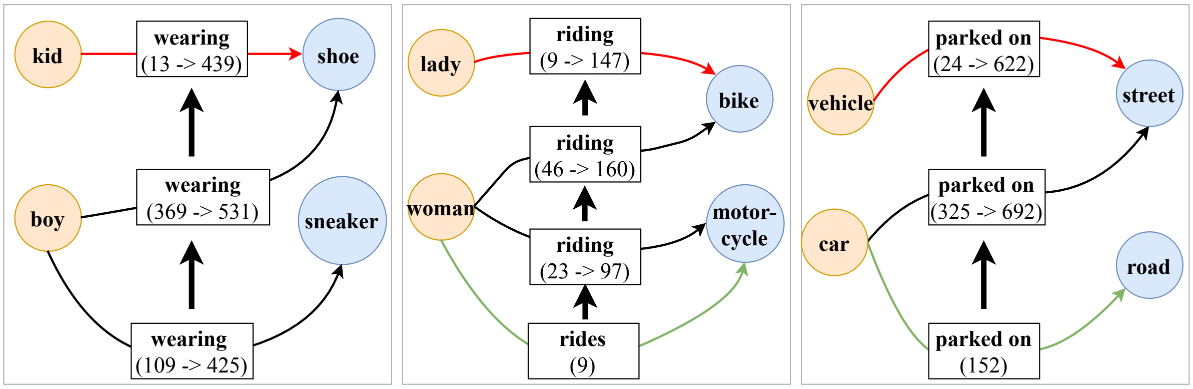
3. Distillation Tracing

SRD updates frequency information for both existing and non-existing relations, and here we analyze discovered relations using original SRD without context for simplicity. Figure 2 illustrates the distillation process for three triplets. In Figure 2a, the frequency of triplet <*kid, wearing, shoe*> is increased from 13 to 439 by getting *enhanced* from triplets <*boy, wearing, sneaker*> and <*boy, wearing, shoe*>, that was done based on the similarity between kid and boy, and shoe and sneaker. Similarly, <*lady, riding, bike*> (Figure 2b) is enhanced based on the similarity between lady and woman, bike and motorcycle, and rides and riding. Lastly, <*vehicle, parked on, street*> is influenced by <*car, parked on, road*> and <*car, parked on, street*> since car and vehicle are similar, the same for street and road. We, however, also observe problems arising when the similarity information is inaccurate.

For example, it is not common for vehicles to park on

Method	$K = 0$	$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 25$
Transformer-Baseline	**58.70 / 65.07 / 66.77**	29.74 / 31.70 / 32.31	19.38 / 19.80 / 19.87	18.25 / 18.51 / 18.55	11.37 / 11.54 / 11.58	8.99 / 9.03 / 9.03
Transformer-Ours	25.46 / 32.13 / 34.25	**31.01 / 36.61 / 38.31**	**42.16 / 47.01 / 48.22**	**45.17 / 48.79 / 49.73**	**40.70 / 44.25 / 45.11**	**41.47 / 43.51 / 44.02**
Motif-Baseline	**57.50 / 64.29 / 66.14**	26.49 / 28.29 / 28.84	15.45 / 15.87 / 16.00	13.54 / 13.83 / 13.94	7.31 / 7.56 / 7.70	8.30 / 8.32 / 8.32
Motif-Ours	26.38 / 33.07 / 35.18	**28.75 / 34.15 / 35.94**	**38.38 / 43.29 / 44.64**	**45.05 / 48.85 / 50.04**	**41.09 / 45 / 46.18**	**41.76 / 44.79 / 45.46**
VCTree-Baseline	**57.84 / 64.30 / 66.04**	27.53 / 29.47 / 30.13	16.60 / 17.03 / 17.15	15.45 / 15.79 / 15.86	7.11 / 7.35 / 7.43	8.38 / 8.40 / 8.40
VCTree-Ours	28.29 / 35.05 / 37.11	**29.19 / 34.25 / 35.77**	**38.6 / 43.13 / 44.32**	**44.71 / 48.65 / 49.49**	**40.19 / 43.94 / 44.71**	**40.02 / 42.56 / 42.99**

Table 1. The recall scores for predicate classification task when not considering the top K frequent predicates. The results are shown in this order: Recall@20 / Recall@50 / Recall@100



a) <kid, wearing, shoe> **b) <lady, riding, bike>** **c) <vehicle, parked on, street>**

Figure 2. Distillation tracing. In the figure, the x→y label below each predicate shows the triplet frequency values before and after distillation. Triplets of non-selected objects or predicates (green arrows) also contribute to the relation statistics.

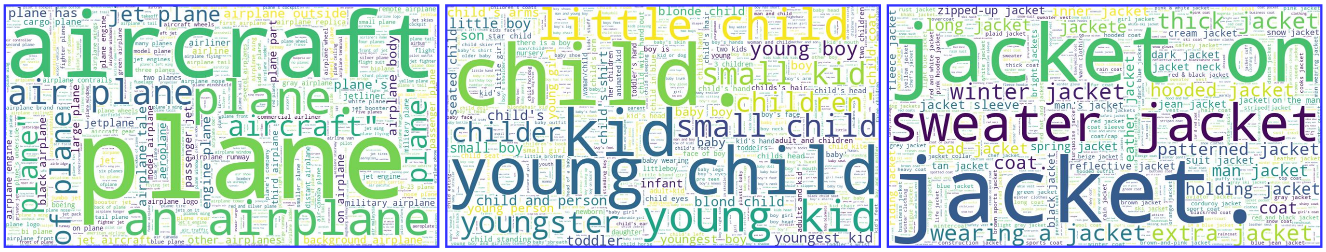
sidewalk, but this triplet obtained more frequency after the distillation because according to BERT, ”sidewalk” and ”street” are similar. Measuring semantic similarities, especially context-dependent ones, is challenging. We ablate text encoders in the supplementary. Leveraging semantic sources like the WordNet [4] and hierarchical ontology such as the VerbNet [2, 5, 9] could offer potential solutions.

Wordcloud Visualizations of Word Similarity To illustrate the quality of similarity between objects and between predicates, we visualize the wordcloud showing all the objects and predicates (including of unselected and selected words) where the weights are based on the similarity with the word of interest. The similarity scores are computed using sentence BERT (Paraphrase-mpnet-base-v2). Figure 3 shows the wordcloud for objects ‘airplane’, ‘child’, and ‘jacket’, and Figure 4 shows the wordclouds for predicates ‘belonging to’, ‘walking in’, and ‘parked on’. The figures show that BERT consistently produced accurate similarity results. Visualizations for all selected objects (150) and predicates (50) are included in the attached HTML files submitted with the paper.

4. Ablation of Word Embedding for SRD

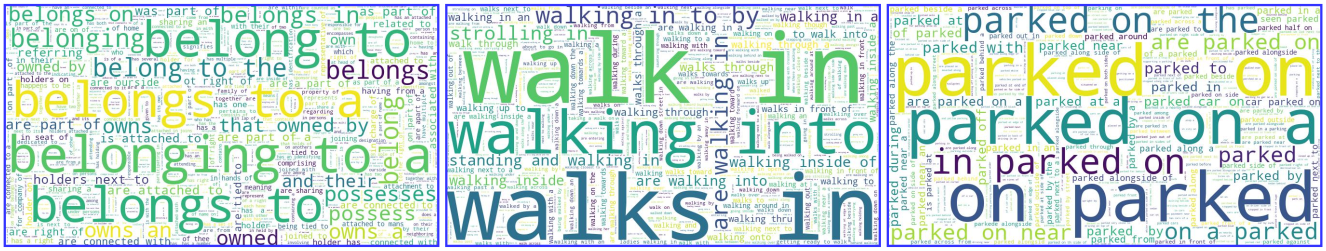
During the distillation process in SRD, the relational frequencies of triplets are passed to one another based on how similar their subjects, objects, and predicates are. To measure the similarity, we use cosine similarity of the corresponding word embeddings generated using pre-trained models such as BERT [3]. The quality of the embeddings is important as it is the main factor to decide where the relational frequency information can be transferred to. In this section we compare the performance of our best setting (i.e., context-based SRD) when using different pre-trained word embedding models including GloVe [6], FastText [1], ConceptNet [10], and sentence-BERT variations [7] (i.e., all-MiniLM-L6-v2, all-mpnet-base-v2, and paraphrase-mpnet-base-v2[1]). We use Transformer as the base model and evaluate using VG dataset. Table 2 shows the comparison results. Among the traditional word embedding models, FastText generally outperforms GloVe and ConceptNet regarding mean recall at 20, 50 and 100. The marginal differences between FastText and the other two models might not be significant but are consistent across all cutoff values. When it comes to transformer-based models (all-MiniLM-L6-v2, all-mpnet-base-v2, and Paraphrase-mpnet-base-v2), Paraphrase-mpnet-base-v2 achieves the best results. Interestingly, using the transformer-based models does not significantly improve the results. In some cases, except

[1] https://www.sbert.net/docs/pretrained_models.html



a) airplane **b) child** **c) jacket**

Figure 3. Wordclouds showing all objects (including unselected and selected words) where the size of words reflect the similarity scores with the corresponding word of interest.



a) belonging to **b) walking in** **c) parked on**

Figure 4. Wordclouds showing all predicates (including unselected and selected words) where the size of words reflect the similarity scores with the corresponding word of interest.

Table 2. Ablation of word embedding models used for SRD. Results are generated using context-based SRD with Transformer on VG. The results reported in the paper is Paraphrase-mpnet-base-v2.

Model	mR@20	mR@50	mR@100
GloVe	32.8	38.5	40.5
FastText	33.5	**39.6**	41.5
ConceptNet	33.0	38.9	41.1
all-MiniLM-L6-v2	33.1	38.8	41.0
all-mpnet-base-v2	33.1	39.0	41.0
Paraphrase-mpnet-base-v2	**34.0**	**39.6**	**41.7**

for Paraphrase-mpnet-base-v2, the performances are even lower than traditional word embedding model like FastText. The reason could be that most of the objects and predicates are single word, so it cannot really make use of the transformer-based models' capability in encoding longer text. Paraphrase-mpnet-base-v2 is also the model we used for the results reported in the main paper.

5. Finding the Number of Clusters for the Context-based Methods

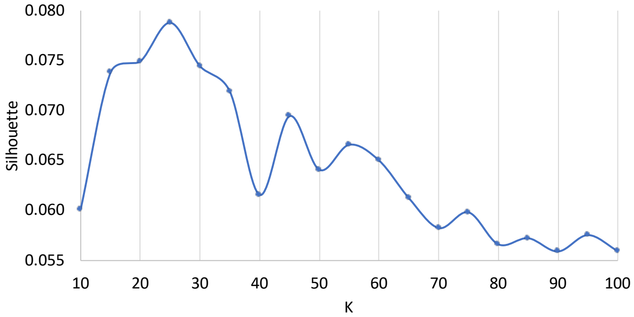


Figure 5. Silhouette coefficient scores for different number of clusters K when running K-Means on Visual Genome train set.

We determine the number of image clusters by the Silhouette coefficient. Figure 5 shows the Silhouette coefficient scores when we vary the number of clusters from 10 to 100 on VG. We choose $K = 25$ clusters as it yields the best Silhouette coefficient [8] of 0.079. Figures 10, 11, and 12 show the 25 clusters with their randomly-chosen images in VG. Qualitatively, most of the clusters contain images of the same topic (cluster). For example, clusters of *snow-skating*, *skateboard*, *wild animals* (top row), *building with clocks*, *baseball*, *traffic with buses* (second row), etc.. The good clusters offer more accurate information when we perform distillation (SRD) and computing statistical priors. That explains the consistent improvements when using context-based SRD. Similar process is applied when finding the number of cluster for GQA-200.

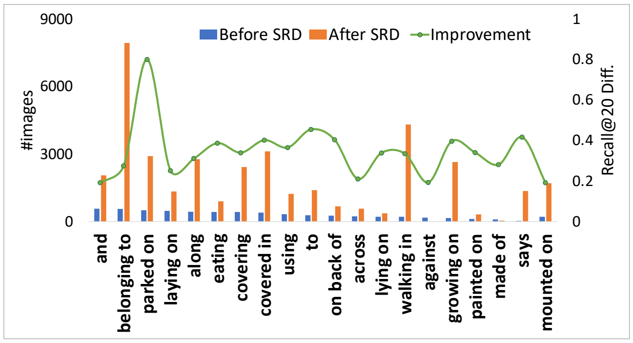


Figure 6. Comparison of original (Before SRD) and augmented data (After SRD) shows increased images with rare predicate relations and improved recall@20 for Visual Genome.

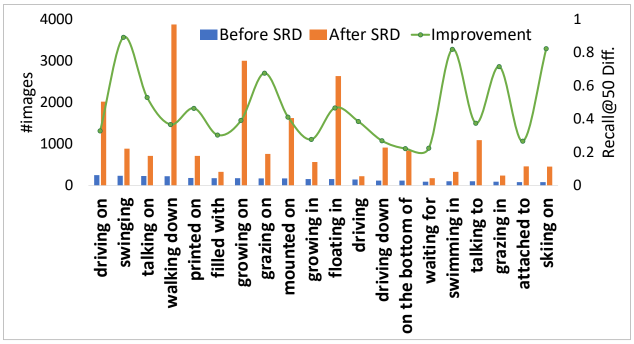


Figure 7. Comparison between the original data (Before SRD) and augmented data (After SRD) in terms of the number of images containing relations of a predicate, and the differences regarding recall@50 for GQA-200. Our proposed method adds more triplets for infrequent predicates that improves their performance. (Results obtained using Transformer)

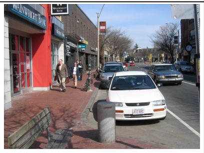
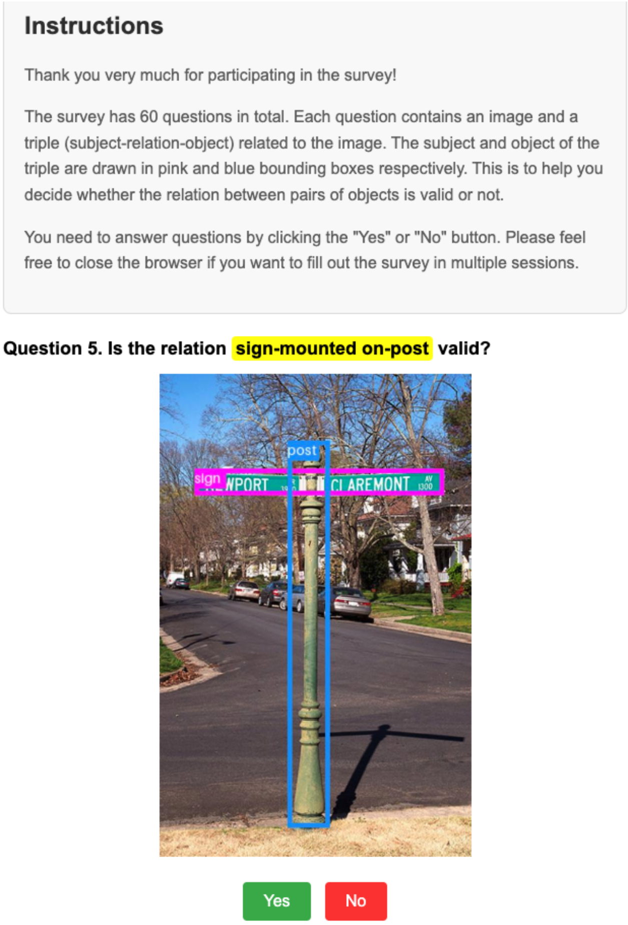
Image	Human-Annotated	Newly-added Relations
	1. (grass, growing on, ground) 2. (leaves, of, tree)	1. (grass, growing on, ground) 2. (leaves, growing on, tree) 3. (zebra, grazing on, grass)
	1. (clouds, in, sky)	1. (boat, floating in, water) 2. (clouds, floating in, sky) 3. (sky, full of, clouds)
	1. (car, on, street) 2. (woman, on, sidewalk)	1. (car, driving down, road) 2. (car, driving down, street) 3. (car, driving on, road) 4. (man, walking down, sidewalk) 5. (man, walking down, street) 6. (people, walking down, sidewalk) 7. (people, walking down, street)

Figure 8. GQA-200– Example of newly-added relations using our method with Transformer.



Instructions

Thank you very much for participating in the survey!

The survey has 60 questions in total. Each question contains an image and a triple (subject-relation-object) related to the image. The subject and object of the triple are drawn in pink and blue bounding boxes respectively. This is to help you decide whether the relation between pairs of objects is valid or not.

You need to answer questions by clicking the "Yes" or "No" button. Please feel free to close the browser if you want to fill out the survey in multiple sessions.

Question 5. Is the relation sign-mounted on-post valid?

Figure 9. Screenshot of the user-study conducted to qualitatively evaluate newly added relations.

6. Qualitative Examples of Actual Relations Added to Train Images

Table 3 shows the examples of the relations that are actually added to the images to enhance the VG training set. For each image, we show the ground-truth annotations (i.e., "Original") and the new relations added by our method with different base SGG models including Motif [13], VCTree [12], and Transformer [11]. Our method enables adding more relations to each image. There are many images with very few annotations, but with the newly added relations, the annotations are more complete. For example, the second image only has two relations, our method can add 9 more relations. Each of the last two images only has 1 ground-truth annotation and we are able to add 5 more relations which enhance the quality of the training images. From these examples, we can see that our method is helpful in adding missing relations to images. Specifically when the bounding boxes are available but the relations were not fully annotated. The same observations can be seen in GQA-200 dataset as shown in Table 4. These newly added relations contribute to the improvement of SGG models as shown in the experiments.

7. Additional Results

Figure 6 shows the impact of SRD on rare predicates for VG dataset.

Figure 7 shows the impact of SRD on rare predicates for GQA-200 dataset.

Figure 8 shows the examples of newly added relations for images in GQA-200 dataset.

Figure 9 shows the screenshot of the user-study conducted to qualitatively evaluate newly added relations.

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 2

[2] Susan Windisch Brown, James Pustejovsky, Annie Zaenen, and Martha Palmer. Integrating generative lexicon event structures into verbnet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[4] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[5] Martha Palmer, Claire Bonial, and Jena D Hwang. 17 verbnet: Capturing english verb behavior, meaning, and usage. *The Oxford handbook of cognitive science*, page 315, 2016. 2

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 2

[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 2

[8] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 3

[9] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005. 2

[10] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017. 2

[11] 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. 4

[12] 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. 4

[13] 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. 4

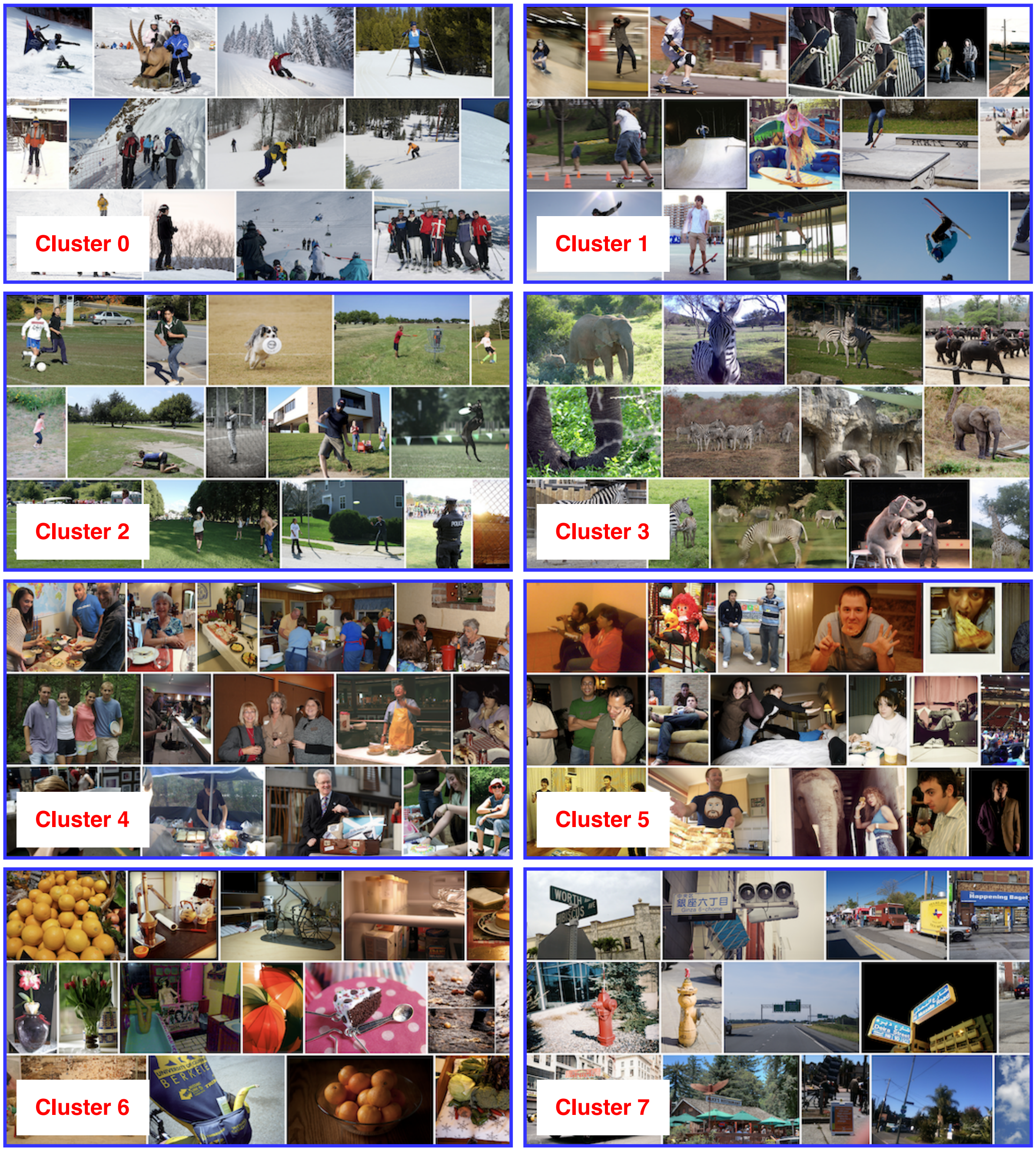


Figure 10. Showing images of the 25 image-clusters obtained by K-Means using CLIP vision features for Visual Genome dataset. Many of the clusters group images of the same context, e.g., snowskating, animals, baseball, train, etc. (Part 1/3)

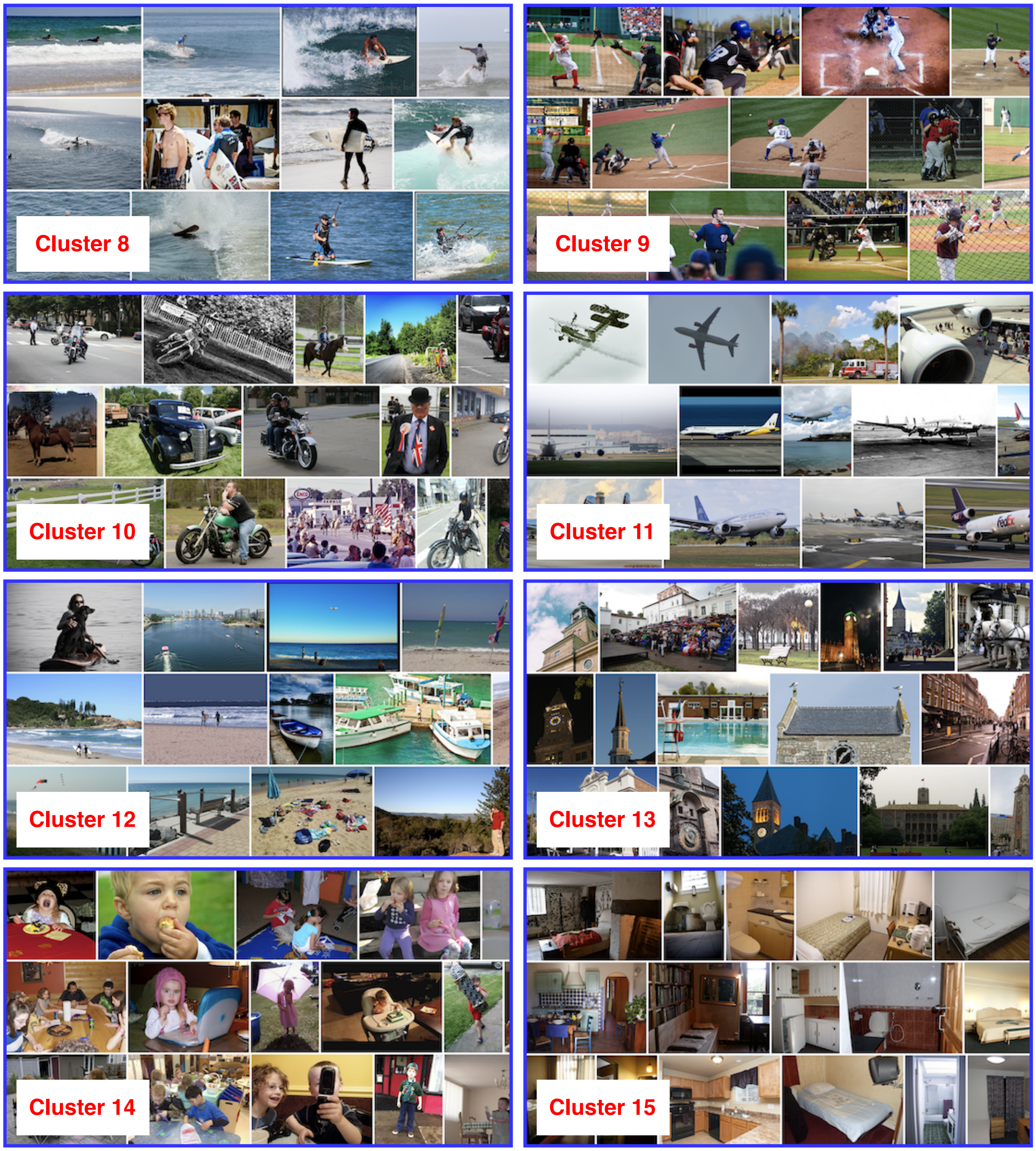


Figure 11. Showing images of the 25 image-clusters obtained by K-Means using CLIP vision features for Visual Genome dataset. Many of the clusters group images of the same context, e.g., snowskating, animals, baseball, train, etc. (Part 2/3)

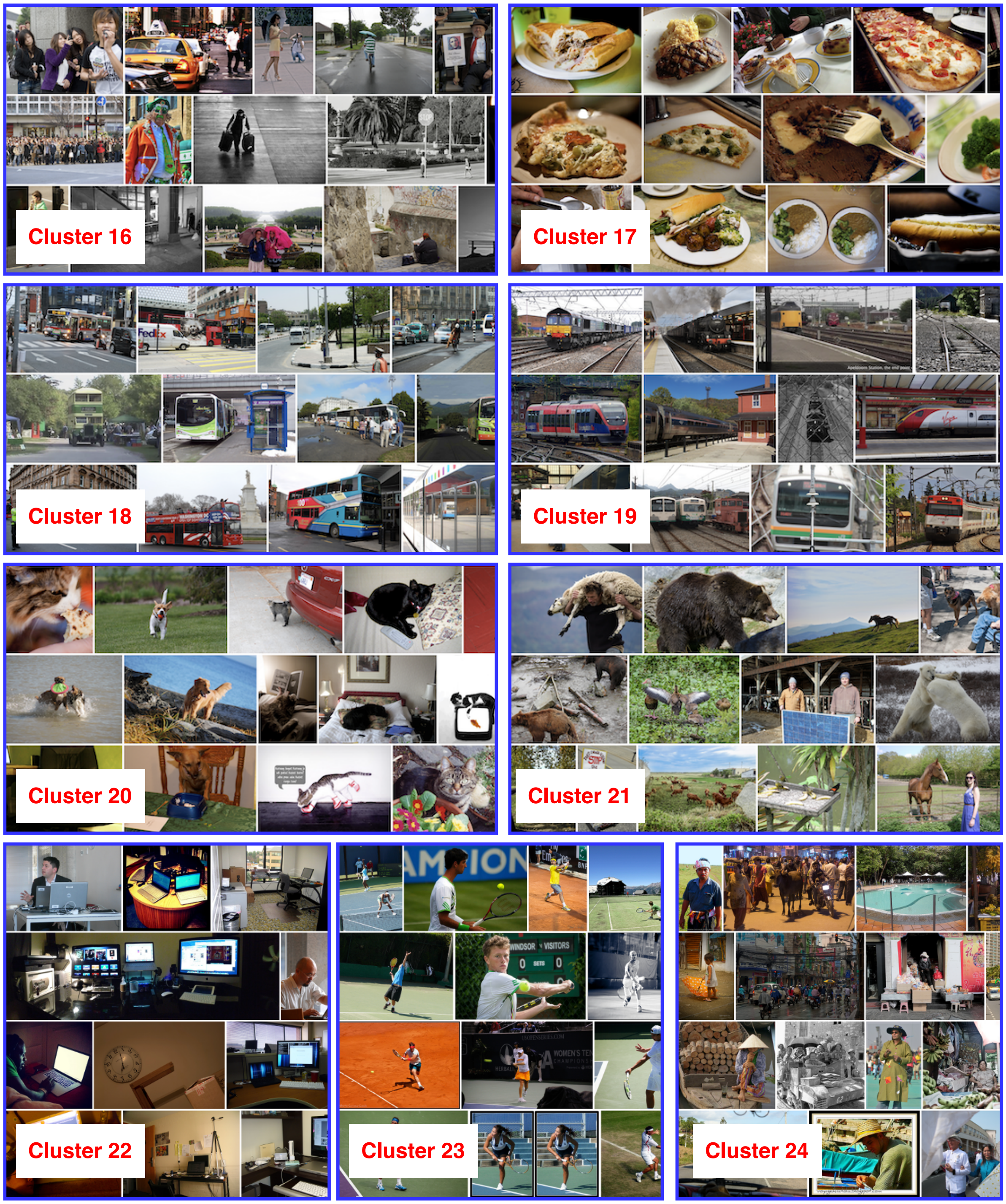


Figure 12. Showing images of the 25 image-clusters obtained by K-Means using CLIP vision features for Visual Genome dataset. Many of the clusters group images of the same context, e.g., snowskating, animals, baseball, train, etc. (Part 3/3)

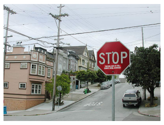
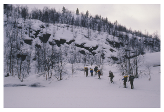
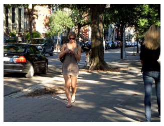
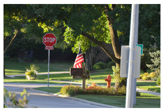
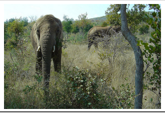
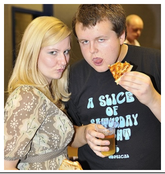
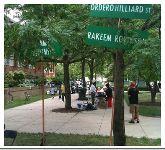
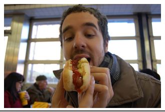
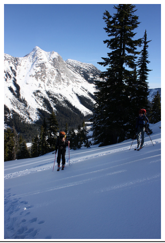
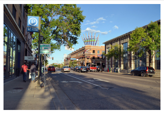
Image	Original	Motif	VCTree	Transformer
	1. (sidewalk, of, street) 2. (sign, on, sidewalk) 3. (sign, on, street) 4. (tree, behind, sign) 5. (tree, on, hill) 6. (tree, on, sidewalk) 7. (window, on, building)	1. (building, along, street) 2. (car, parked on, street) 3. (leaf, growing on, tree) 4. (letter, painted on, sign) 5. (roof, covering, building) 6. (sign, mounted on, pole) 7. (sign, says, letter) 8. (sign, says, sign) 9. (tree, along, sidewalk) 10. (tree, along, street) 11. (tree, covered in, leaf) 12. (window, part of, building)	1. (building, along, street) 2. (car, parked on, street) 3. (leaf, growing on, tree) 4. (letter, painted on, sign) 5. (roof, covering, building) 6. (sign, mounted on, pole) 7. (sign, says, letter) 8. (sign, says, sign) 9. (tree, along, sidewalk) 10. (tree, along, street) 11. (tree, covered in, leaf) 12. (window, part of, building)	1. (building, across, street) 2. (building, along, street) 3. (car, parked on, street) 4. (leaf, growing on, tree) 5. (letter, painted on, sign) 6. (roof, covering, building) 7. (sign, mounted on, pole) 8. (sign, says, letter) 9. (sign, says, sign) 10. (tree, along, sidewalk) 11. (tree, along, street) 12. (tree, covered in, leaf) 13. (window, part of, building)
	1. (people, near, mountain) 2. (tree, on, tree)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (person, walking in, snow) 5. (skier, walking in, snow) 6. (snow, covering, mountain) 7. (snow, covering, tree) 8. (tree, covered in, snow) 9. (tree, growing on, mountain)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (person, walking in, snow) 5. (skier, walking in, snow) 6. (snow, covering, mountain) 7. (snow, covering, tree) 8. (tree, covered in, snow) 9. (tree, growing on, mountain)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (person, walking in, snow) 5. (skier, walking in, snow) 6. (snow, covering, mountain) 7. (snow, covering, tree) 8. (tree, covered in, snow) 9. (tree, growing on, mountain)
	1. (hair, belonging to, woman) 2. (leaf, near, tree) 3. (woman, walking on, sidewalk)	1. (building, across, street) 2. (car, parked on, street) 3. (hair, belonging to, woman) 4. (leaf, growing on, tree) 5. (person, walking in, street) 6. (tree, along, sidewalk) 7. (tree, along, street) 8. (tree, covered in, leaf)	1. (building, across, street) 2. (car, parked on, street) 3. (hair, belonging to, woman) 4. (leaf, growing on, tree) 5. (person, walking in, street) 6. (tree, along, sidewalk) 7. (tree, along, street) 8. (tree, covered in, leaf)	1. (building, across, street) 2. (building, along, street) 3. (car, parked on, street) 4. (hair, belonging to, woman) 5. (leaf, growing on, tree) 6. (person, walking in, street) 7. (tree, along, sidewalk) 8. (tree, along, street) 9. (tree, covered in, leaf)
	1. (flag, on, pole) 2. (plant, near, sidewalk) 3. (pole, with, sign)	1. (branch, growing on, tree) 2. (leaf, growing on, tree) 3. (sign, mounted on, pole) 4. (sign, mounted on, post) 5. (sign, says, sign) 6. (tree, along, street) 7. (tree, covered in, leaf)	1. (branch, growing on, tree) 2. (leaf, growing on, tree) 3. (sign, mounted on, pole) 4. (sign, mounted on, post) 5. (sign, says, sign) 6. (tree, along, street) 7. (tree, covered in, leaf)	1. (branch, growing on, tree) 2. (leaf, growing on, tree) 3. (sign, mounted on, pole) 4. (sign, mounted on, post) 5. (sign, says, sign) 6. (tree, across, street) 7. (tree, covered in, leaf)
	1. (elephant, eating, leaf) 2. (head, of, elephant) 3. (trunk, of, elephant)	1. (branch, from, tree) 2. (ear, belonging to, elephant) 3. (leaf, growing on, tree) 4. (tree, covered in, leaf) 5. (tree, growing on, mountain)	1. (branch, from, tree) 2. (ear, belonging to, elephant) 3. (leaf, growing on, tree) 4. (tree, covered in, leaf) 5. (tree, growing on, mountain)	1. (branch, from, tree) 2. (ear, belonging to, elephant) 3. (leaf, growing on, tree) 4. (tree, covered in, leaf) 5. (tree, growing on, mountain)
	1. (man, has, hand) 2. (woman, has, hand)	1. (hair, belonging to, man) 2. (hair, belonging to, woman) 3. (hand, belonging to, man) 4. (man, and, woman) 5. (man, eating, pizza)	1. (hair, belonging to, man) 2. (hair, belonging to, woman) 3. (hand, belonging to, man) 4. (man, and, woman) 5. (man, eating, pizza)	1. (hair, belonging to, man) 2. (hand, belonging to, man) 3. (man, and, woman) 4. (man, eating, pizza)
	1. (sign, on, post)	1. (leaf, growing on, tree) 2. (sign, mounted on, pole) 3. (sign, mounted on, post) 4. (tree, along, sidewalk) 5. (tree, covered in, leaf)	1. (leaf, growing on, tree) 2. (sign, mounted on, pole) 3. (sign, mounted on, post) 4. (tree, along, sidewalk) 5. (tree, covered in, leaf)	1. (leaf, growing on, tree) 2. (sign, mounted on, pole) 3. (sign, mounted on, post) 4. (tree, along, sidewalk) 5. (tree, covered in, leaf)
	1. (man, with, hat)	1. (hair, belonging to, man) 2. (hair, belonging to, woman) 3. (hand, belonging to, man) 4. (man, and, woman) 5. (man, eating, food)	1. (hair, belonging to, man) 2. (hair, belonging to, woman) 3. (hand, belonging to, man) 4. (man, and, woman) 5. (man, eating, food)	1. (hair, belonging to, man) 2. (hair, belonging to, woman) 3. (man, and, woman) 4. (man, eating, food)
	1. (ski, on, man) 2. (snow, on, mountain)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (snow, covering, mountain) 5. (snow, covering, tree) 6. (tree, covered in, snow) 7. (tree, growing on, mountain)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (snow, covering, mountain) 5. (snow, covering, tree) 6. (tree, covered in, snow) 7. (tree, growing on, mountain)	1. (man, walking in, snow) 2. (mountain, covered in, snow) 3. (people, walking in, snow) 4. (snow, covering, mountain) 5. (snow, covering, tree) 6. (tree, covered in, snow) 7. (tree, growing on, mountain)
	1. (car, on, street) 2. (people, on, sidewalk) 3. (sign, on, building)	1. (building, along, street) 2. (car, parked on, street) 3. (sign, says, sign) 4. (tree, along, street) 5. (truck, parked on, street) 6. (vehicle, parked on, street)	1. (building, along, street) 2. (car, parked on, street) 3. (sign, says, sign) 4. (tree, along, street) 5. (truck, parked on, street) 6. (vehicle, parked on, street)	1. (building, along, street) 2. (car, parked on, street) 3. (sign, says, sign) 4. (tree, along, street) 5. (truck, parked on, street) 6. (vehicle, parked on, street)

Table 3. Example of newly added relations using our proposed method with different base models, i.e., Motif, VCTree, and Transformer (Visual Genome dataset).

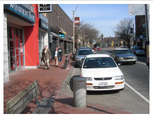
Image	Original	Motif	VCTree	Transformer
	1. (car, on, road) 2. (door, near, window) 3. (tree, on, mountain) 4. (window, near, door) 5. (window, next to, door)	1. (bus, driving on, road) 2. (car, driving down, street) 3. (car, driving on, road) 4. (car, driving on, street) 5. (clouds, floating in, sky) 6. (people, walking down, sidewalk) 7. (people, walking down, street) 8. (sky, full of, clouds)	1. (bus, driving on, road) 2. (bus, driving on, street) 3. (car, driving on, road) 4. (car, driving on, street) 5. (clouds, floating in, sky) 6. (people, walking down, sidewalk) 7. (people, walking down, street) 8. (sky, full of, clouds)	1. (bus, driving on, road) 2. (car, driving down, street) 3. (car, driving on, road) 4. (clouds, floating in, sky) 5. (people, walking down, sidewalk) 6. (people, walking down, street) 7. (sky, full of, clouds)
	1. (clouds, in, sky) 2. (sheep, in, grass) 3. (sheep, lying in, grass) 4. (sheep, lying on, grass) 5. (tree, on, field)	1. (clouds, floating in, sky) 2. (grass, growing in, field) 3. (grass, growing on, ground) 4. (sheep, grazing in, field) 5. (sheep, grazing on, grass) 6. (sheep, lying in, grass) 7. (sky, full of, clouds)	1. (clouds, floating in, sky) 2. (grass, growing in, field) 3. (grass, growing on, ground) 4. (sheep, grazing in, field) 5. (sheep, lying in, grass) 6. (sky, full of, clouds)	1. (clouds, floating in, sky) 2. (grass, growing in, field) 3. (grass, growing on, ground) 4. (sheep, grazing in, field) 5. (sheep, lying in, grass) 6. (sky, full of, clouds)
	1. (car, on, street) 2. (woman, on, sidewalk)	1. (car, driving on, road) 2. (car, driving on, street) 3. (man, walking down, sidewalk) 4. (man, walking down, street) 5. (people, walking down, sidewalk) 6. (people, walking down, street)	1. (car, driving down, street) 2. (car, driving on, road) 3. (car, driving on, street) 4. (man, walking down, sidewalk) 5. (man, walking down, street) 6. (people, walking down, sidewalk) 7. (people, walking down, street)	1. (car, driving down, road) 2. (car, driving down, street) 3. (car, driving on, road) 4. (man, walking down, sidewalk) 5. (man, walking down, street) 6. (people, walking down, sidewalk) 7. (people, walking down, street)
	1. (grass, growing in, field)	1. (grass, growing in, field) 2. (grass, growing on, ground) 3. (leaves, growing on, tree) 4. (zebra, grazing in, field) 5. (zebra, grazing on, grass)	1. (grass, growing in, field) 2. (grass, growing on, ground) 3. (leaves, growing on, tree) 4. (zebra, grazing in, field) 5. (zebra, grazing on, grass)	1. (grass, growing in, field) 2. (grass, growing on, ground) 3. (leaves, growing on, tree) 4. (zebra, grazing in, field) 5. (zebra, grazing on, grass)
	1. (leg, of, man) 2. (man, wearing, glass) 3. (man, wearing, shirt) 4. (man, wearing, shorts)	1. (man, hitting, ball) 2. (man, swinging, racket) 3. (player, hitting, ball) 4. (player, swinging, racket) 5. (racket, hitting, ball)	1. (man, hitting, ball) 2. (man, swinging, racket) 3. (player, hitting, ball) 4. (player, swinging, racket) 5. (racket, hitting, ball)	1. (man, hitting, ball) 2. (man, swinging, racket) 3. (player, hitting, ball) 4. (player, swinging, racket) 5. (racket, hitting, ball)
	1. (horse, wearing, collar)	1. (bird, swimming in, water) 2. (clouds, floating in, sky) 3. (grass, growing on, ground) 4. (leaves, growing on, tree)	1. (bird, swimming in, water) 2. (clouds, floating in, sky) 3. (grass, growing on, ground) 4. (leaves, growing on, tree) 5. (sky, full of, clouds)	1. (clouds, floating in, sky) 2. (grass, growing on, ground) 3. (leaves, growing on, tree)
	1. (letter, of, sign) 2. (tree, around, building)	1. (clouds, floating in, sky) 2. (leaves, growing on, tree) 3. (letter, printed on, sign) 4. (sign, mounted on, pole)	1. (clouds, floating in, sky) 2. (leaves, growing on, tree) 3. (letter, printed on, sign) 4. (sign, mounted on, pole) 5. (sky, full of, clouds)	1. (clouds, floating in, sky) 2. (leaves, growing on, tree) 3. (letter, printed on, sign) 4. (sign, mounted on, pole) 5. (sky, full of, clouds)
	1. (woman, carrying, bag)	1. (bus, driving on, road) 2. (bus, driving on, street) 3. (car, driving on, road) 4. (car, driving on, street) 5. (woman, walking down, sidewalk)	1. (bus, driving on, road) 2. (bus, driving on, street) 3. (car, driving on, road) 4. (car, driving on, street) 5. (woman, walking down, sidewalk)	1. (bus, driving on, road) 2. (bus, driving on, street) 3. (car, driving down, road) 4. (car, driving down, street) 5. (car, driving on, road) 6. (car, driving on, street) 7. (woman, walking down, sidewalk)
	1. (man, wearing, shirt) 2. (person, walking on, sidewalk)	1. (man, talking on, cell phone) 2. (man, talking on, phone) 3. (man, walking down, sidewalk) 4. (people, walking down, sidewalk) 5. (woman, talking on, cell phone) 6. (woman, talking on, phone) 7. (woman, walking down, sidewalk)	1. (man, talking on, cell phone) 2. (man, talking on, phone) 3. (man, walking down, sidewalk) 4. (people, walking down, sidewalk) 5. (woman, talking on, cell phone) 6. (woman, talking on, phone) 7. (woman, walking down, sidewalk)	1. (man, talking on, cell phone) 2. (man, talking on, phone) 3. (man, walking down, sidewalk) 4. (people, walking down, sidewalk) 5. (woman, talking on, cell phone) 6. (woman, talking on, phone) 7. (woman, walking down, sidewalk)
	1. (bear, wearing, shoe) 2. (man, wearing, glass) 3. (man, wearing, shirt) 4. (man, wearing, sweater) 5. (pants, on, bear)	1. (man, walking down, sidewalk) 2. (man, walking down, street) 3. (people, walking down, sidewalk) 4. (people, walking down, street) 5. (person, walking down, sidewalk) 6. (woman, walking down, sidewalk)	1. (man, walking down, sidewalk) 2. (man, walking down, street) 3. (people, walking down, sidewalk) 4. (people, walking down, street) 5. (person, walking down, sidewalk) 6. (woman, walking down, sidewalk)	1. (man, walking down, sidewalk) 2. (man, walking down, street) 3. (people, walking down, sidewalk) 4. (people, walking down, street) 5. (person, walking down, sidewalk) 6. (woman, walking down, sidewalk)
	1. (boat, in, water) 2. (man, on, bike) 3. (person, on, bike)	1. (boat, floating in, water) 2. (man, walking down, sidewalk) 3. (person, walking down, sidewalk) 4. (sign, mounted on, pole)	1. (boat, floating in, water) 2. (man, walking down, sidewalk) 3. (person, walking down, sidewalk) 4. (sign, mounted on, pole)	1. (boat, floating in, water) 2. (man, walking down, sidewalk) 3. (person, walking down, sidewalk) 4. (sign, mounted on, pole)
	1. (paw, of, dog) 2. (tail, of, dog)	1. (dog, catching, frisbee) 2. (man, catching, frisbee) 3. (man, throwing, frisbee)	1. (dog, catching, frisbee) 2. (man, catching, frisbee)	1. (dog, catching, frisbee) 2. (man, catching, frisbee)

Table 4. Example of newly added relations using our proposed method with different base models, i.e., Motif, VCTree, and Transformer (GQA-200 dataset).