# Supplementary Material for "Energy-Based Learning for Scene Graph Generation"

Mohammed Suhail[1,2] Abhay Mittal[4] Behjat Siddiquie[4] Chris Broaddus[4]
Jayan Eledath[4] Gerard Medioni[4] Leonid Sigal[1,2,3]
[1]University of British Columbia [2]Vector Institute for AI [3]Canada CIFAR AI Chair [4] Amazon
suhail33@cs.ubc.ca mrmittal@amazon.com behjats@amazon.com chrispb@amazon.com
eledathj@amazon.com medioni@amazon.com lsigal@cs.ubc.ca

## 1. Quantitative Studies

We present the complete results of our experiments using both the mR@K[2] and the R@K[3] metrics in Table 1.

**Visual Genome.** VCTree, Motif and IMP models, suffer a small drop in regular R@K [3] performance when trained using the energy-based loss. This can be attributed to the heavy-trailed distribution of Visual Genome dataset [1]. Our model predicts granular and informative relations which get penalized on the R@K metric due to biased ground truth annotations. When comparing the mR@K metric [2] our model fares significantly better. For an unbiased scene graph generation framework such as VCTree-TDE, our model is able to improve on both the mR@K and the R@K metrics. We plot the difference in predicate level R@100 performance (sorted in descending order of sampling fraction) of a Motif model trained using the energy-based loss and the cross-entropy loss in Figure 1, along with the sampling fractions of the relations in the visual genome dataset in Figure 2. We observe that for generic relations such as on with a large number of training samples, the baseline model has slightly higher recall. However, granular relations with a smaller number training samples have significantly higher recall rates.

**GQA.** We observe an almost consistent improvement in both R@K and mR@K. This is primarily due to removing the frequency bias component which leads to relatively unbiased predictions. We plot difference in predicate level R@100 performance of the energy-loss based model and the cross entropy based model in Figure 3 and the corresponding sampling fractions of the relations in Figure 4. Note that the sampling fractions in Figure 4 were plotted using a log-scale on the y-axis for clarity. To keep the visualization simple, we do not plot relations where both the energy model and the baseline model have zero recall. We observe a similar trend where our model performance is comparable to baseline on relations with more labels and significantly better on relations with lesser annotations.

| | | | PredCls | | | | | | SGCls | | | | | | SGDet | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R@K | | | mR@K | | | R@K | | | mR@K | | | R@K | | | mR@K | | |
| Dataset | Model | Method | @20 | @50 | @100 | @20 | @50 | @100 | @20 | @50 | @100 | @20 | @50 | @100 | @20 | @50 | @100 | @20 | @50 | @100 |
| Visual Genome | VCTree | CE | **59.82** | **65.93** | **67.57** | 13.07 | 16.53 | 17.77 | **41.49** | **45.16** | **46.1** | 8.5 | 10.53 | 11.24 | **24.9** | **32.02** | **36.3** | 5.31 | 7.16 | 8.35 |
| | | EBM | 57.31 | 63.99 | 65.84 | **14.2** | **18.19** | **19.72** | 40.31 | 44.72 | 45.84 | **10.04** | **12.54** | **13.45** | 24.21 | 31.36 | 35.87 | **5.67** | **7.71** | **9.1** |
| | Motif | CE | **58.56** | **64.38** | **67.13** | 12.45 | 15.71 | 16.98 | **35.95** | **39.18** | **39.96** | 6.95 | 8.58 | 9.05 | **25.62** | **32.97** | **37.41** | 5.07 | 6.91 | 8.12 |
| | | EBM | 58.39 | 65.19 | 67.33 | **14.17** | **18.02** | **19.53** | 35.65 | 39.16 | 40.04 | **8.18** | **10.22** | **10.98** | 24.29 | 31.74 | 36.29 | **5.66** | **7.72** | **9.27** |
| | IMP | CE | **54.34** | **61.05** | **63.06** | 8.85 | 10.97 | 11.77 | **34.02** | **37.39** | **38.26** | 5.4 | 6.4 | 6.74 | **16.34** | **23.64** | **28.71** | 2.2 | 3.29 | 4.14 |
| | | EBM | 54.61 | 61.49 | 63.49 | **9.43** | **11.83** | **12.77** | 34.03 | 37.24 | 38.09 | **5.66** | **6.81** | **7.17** | 18.14 | 25.86 | 31.13 | **2.78** | **4.23** | **5.44** |
| | VCTree-TDE | CE | 40.12 | 50.83 | 54.91 | 16.3 | 22.85 | 26.26 | 26 | 33.03 | 35.97 | 11.85 | 15.81 | 17.99 | 13.97 | 19.43 | 23.34 | 6.59 | 8.99 | 10.78 |
| | | EBM | **41.62** | **51.22** | **54.29** | **19.87** | **26.66** | **29.97** | **29.53** | **36.49** | **38.92** | **13.86** | **18.2** | **20.45** | **14.66** | **20.55** | **24.74** | **7.1** | **9.69** | **11.6** |
| GQA | Transformer | CE | 34.68 | 50.86 | 58.46 | 1.17 | 2.48 | 3.69 | 11.05 | 14.86 | 16.42 | 0.54 | 0.97 | 1.29 | - | - | - | - | - | - |
| | | EBM | **35.61** | **51.88** | **59.5** | **1.28** | **2.94** | **4.71** | **12.14** | **16.12** | **17.66** | **0.68** | **1.32** | **1.77** | - | - | - | - | - | - |
| | Motif | CE | 32.73 | 47.51 | 54.32 | 0.89 | 1.83 | 2.75 | 11.34 | 15.31 | 16.93 | 0.49 | 0.87 | 1.18 | - | - | - | - | - | - |
| | | EBM | **34.9** | **50.66** | **57.98** | **1.04** | **2.29** | **3.49** | **11.64** | **15.74** | **17.39** | **0.57** | **0.9** | **1.26** | - | - | - | - | - | - |
| | IMP | CE | 29.4 | 42.44 | 48.49 | 0.5 | 0.95 | 1.34 | **11.87** | **15.82** | **17.44** | 0.28 | 0.5 | 0.65 | - | - | - | - | - | - |
| | | EBM | **29.85** | **43.3** | **49.13** | **0.57** | **1.07** | **1.5** | 11.64 | 15.47 | 17.02 | **0.34** | **0.58** | **0.76** | - | - | - | - | - | - |

Table 1: **Quantitative Results.** Table shows the **R@K** and **mR@K** comparison between models trained using the proposed framework and energy based formulation

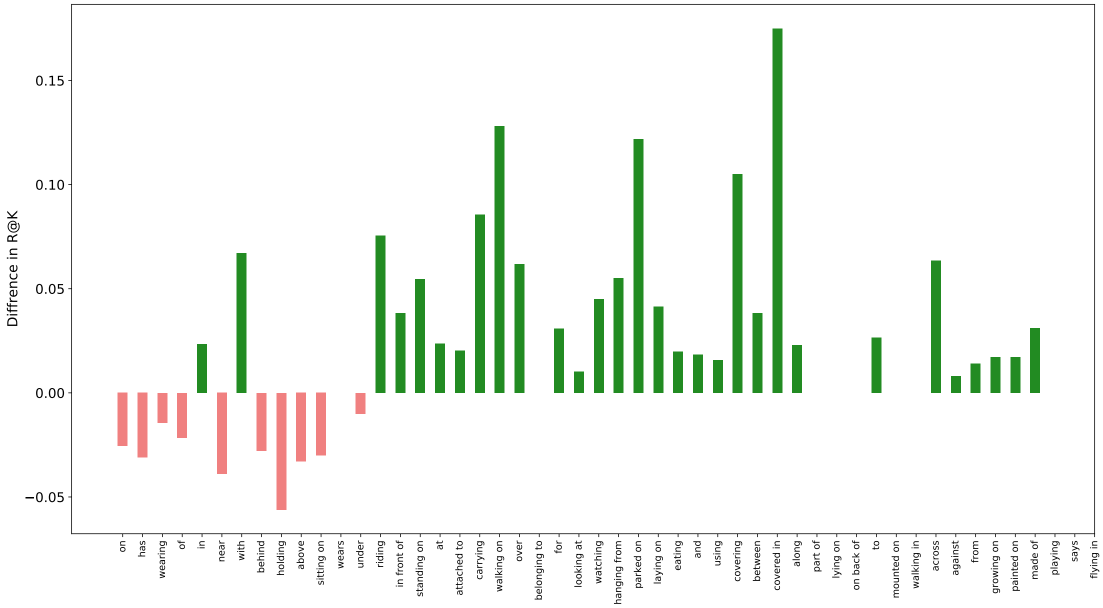

Figure 1: This figure plots the relation wise difference in Recall@100, sorted by descending order of sampling fractions, between a Motif model trained using the proposed energy-based framework and a similar model trained using the standard cross-entropy based framework. The green (red) bars correspond to a relative improvement (regression) in the performance of the energy-based model. We note that using our proposed methodology, we obtain large improvements in the performance of relations with relatively less data. The slight degradation in performance on the commonly occurring relations such a `on` is due to relatively unbiased/granular predictions from energy based models.

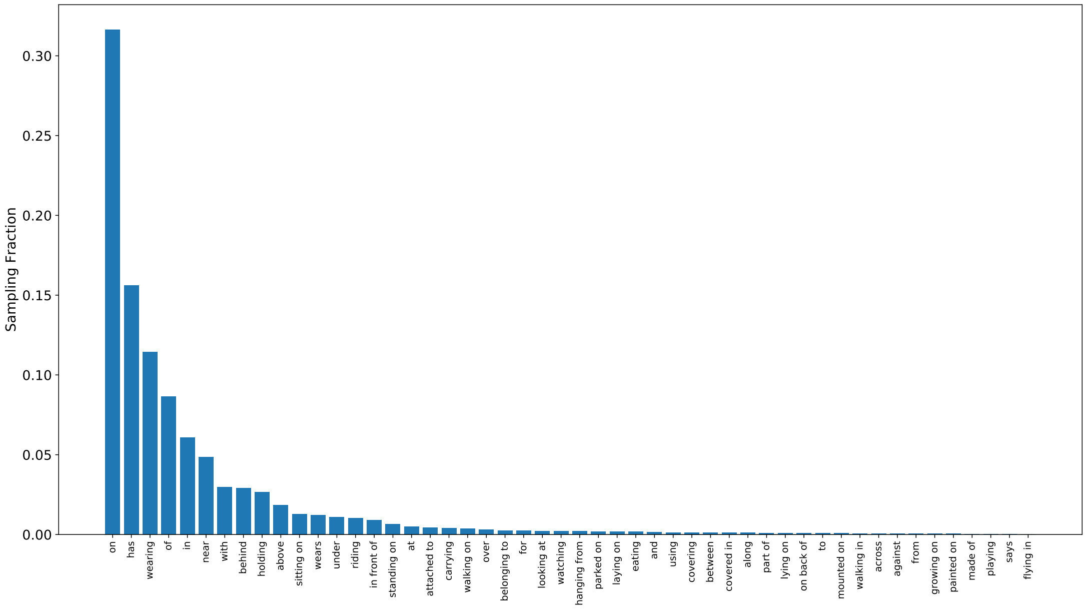

Figure 2: Figure shows the sampling fraction of the different relations in the Visual Genome dataset.

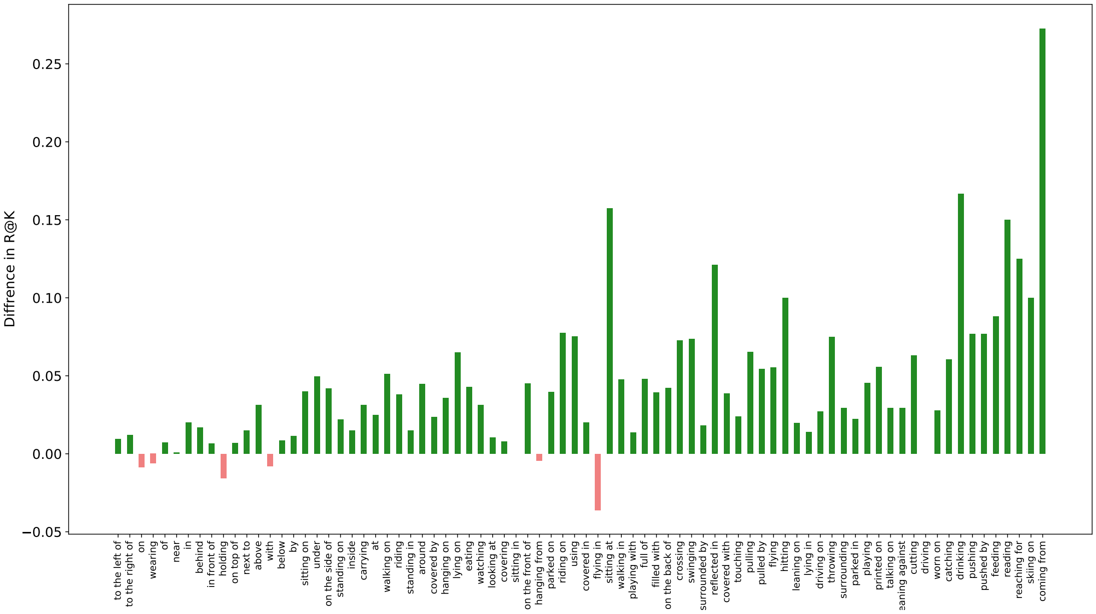

Figure 3: This figure shows the performance difference in relation wise recall on a Motif model trained using an energy loss and using a cross-entropyloss on the GQA dataset sorted in descending order of sampling fractions. We observe that for commonly occurring relations in the dataset, the performance of the baseline and proposed framework is comparable. As we move to the left, we observe larger improvements in performance from energy-based training.

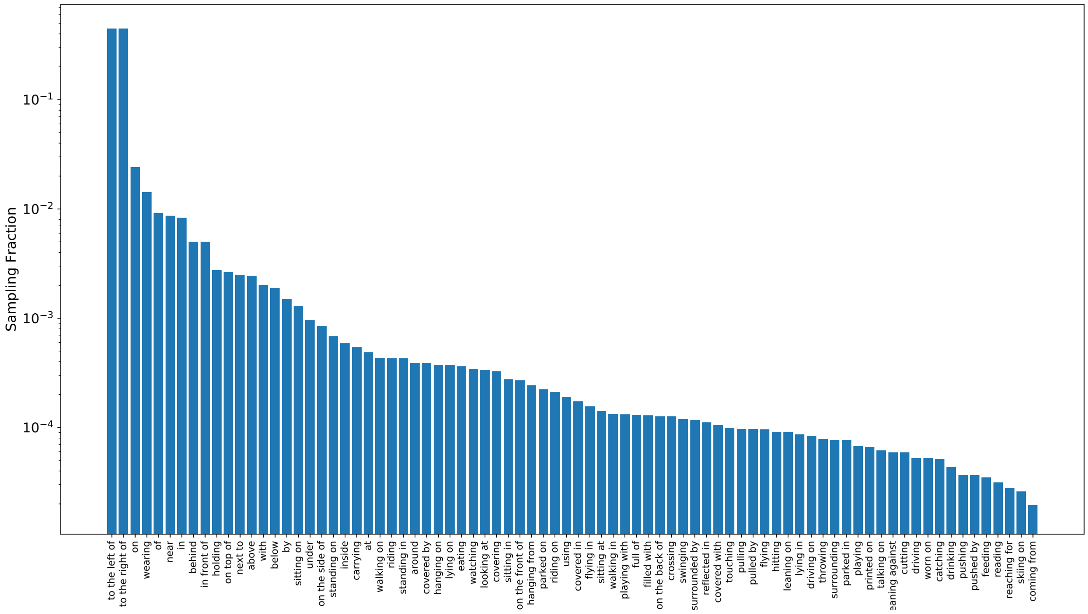

Figure 4: This figure shows the sampling fraction for relation labels in the GQA dataset. The graph is plotted using a log-scale on the y-axis due to a large disparity in the sample fractions. The linear downward trend depicted in the plot corresponds to an exponential reduction in the occurrence of the less frequently occurring relations.

## References

[1] 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. 1

[2] 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

[3] 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*, pages 5831–5840, 2018. 1