# Learning to Generate Language-supervised and Open-vocabulary Scene Graph using Pre-trained Visual-Semantic Space — CVPR2023 Supplementary Material

Yong Zhang[†], Yingwei Pan[‡], Ting Yao[‡], Rui Huang[†*], Tao Mei[‡], and Chang-Wen Chen[§]

[†] The Chinese University of Hong Kong, Shenzhen [‡] HiDream.ai Inc. [§] The Hong Kong Polytechnic University

yongzhang@link.cuhk.edu.cn, {panyw.ustc, tingyao.ustc}@gmail.com, ruihuang@cuhk.edu.cn, tmei@hidream.ai, changwen.chen@polyu.edu.hk

The supplementary material contains: (1) more details about the evaluation setups on the VG150 dataset in our experiments; (2) more evaluation results.

## A. Evaluation Setups

As mentioned in the main paper [8] (Section 4.2 – Evaluation protocols and metrics), we compute all recall metrics over test images of the VG150 dataset. However, considering that the adopted GLIP pre-trained visual-semantic space (VSS) has seen part of images in the original VG150 test split (∼26k) during pre-training [1], we exclude these overlapped images and achieve a new split of ∼15k test images. Our approach adopts the same VG150 train split and computes evaluation metrics over the new test split.

Since the new test split is a subset of the original test split, one concern is whether or not the performances over these two test splits would exhibit significant variations. In an effort to delve into this concern, we compare the performances of several classic SGG methods on the original and the new test split. These methods are IMP [5], VTransE [7], VCTREE [4] and MOTIFS [6], with stable implementations in codebase [2]. Table 1 summarizes the results. We observe that the computed recalls show only $< 0.15\%$ variations (relatively $< 0.8\%$) between the two different test splits. Such results basically validate that the new VG150 test split can lead to stable recall metrics with mostly same performance trends as in the original split. The performance variations between these two test splits are trivial in comparison with the performance differences between different SGG methods. For example, in Table 1, our VS$^3$ *(Swin-L)* achieves $> 2.65\%$ performance boosts than the mentioned baselines. Accordingly, we directly compare the performances obtained on the new VG150 test split by our method against the performances reported in previous works on the original VG150 test split.

*Corresponding author

| Methods | Original test split (∼26k) | New test split (∼15k) | Difference | Relative |
|---|---|---|---|---|
| IMP [5] | 18.59 / 26.36 / 31.62 | 18.45 / 26.35 / 31.57 | -0.14 / -0.01 / -0.05 | 0.75% |
| VTransE [7] | 23.06 / 29.99 / 34.69 | 23.06 / 29.91 / 34.59 | -0.00 / -0.08 / -0.10 | 0.29% |
| VCTREE [4] | 24.51 / 31.29 / 35.98 | 24.45 / 31.19 / 35.87 | -0.06 / -0.10 / -0.11 | 0.32% |
| MOTIFS [6] | 25.29 / 32.30 / 37.08 | 25.16 / 32.21 / 36.94 | -0.13 / -0.09 / -0.14 | 0.51% |
| VS$^3$ *(Swin-T)* | - | 26.10 / 34.53 / 39.18 | - | - |
| VS$^3$ *(Swin-L)* | - | **27.81 / 36.63 / 41.50** | - | - |

Table 1. Performance comparisons in the evaluation metrics (R@20/50/100) between the original VG150 test split and the new test split (removing invalid images that have already been seen during GLIP [1] pre-training). We observe the evaluation differences between these two splits ($< 0.15\%$ variations) are somewhat trivial in comparison with the performance differences between different SGG methods (e.g., VTransE improves over IMP by $> 3\%$). Difference = New test split metrics - Original test split metrics; Relative = $max$(|Difference| / Original test split metrics).

| SGG model | Detector | Backbone | mR@20 | mR@50 | mR@100 |
|---|---|---|---|---|---|
| IMP [5] | Faster-RCNN | RX-101 | 2.8 | 4.2 | 5.3 |
| VTransE [7] | Faster-RCNN | RX-101 | 3.7 | 5.0 | 6.0 |
| VCTREE [4] | Faster-RCNN | RX-101 | 4.2 | 5.7 | 6.9 |
| MOTIFS [6] | Faster-RCNN | RX-101 | 4.1 | 5.5 | 6.8 |
| VS$^3$ | - | Swin-T | **4.3** | **6.6** | **8.1** |

Table 2. Experimental results of fully supervised SGG. All metrics are computed under the SGDET protocol on VG150 test images. Results of previous models come from Tang et al. [3].

## B. More Evaluation Results

We report additional results on the unbiased metric under the same fully supervised SGG in Table 2. Concretely, the adopted unbiased metric is mean Recall@K (mR@K), which averages Recall@K across all predicate categories. The results show that the pre-trained VSS does not mitigate the bias issue very significantly, since the bias issue is naturally rooted in the pre-training image-text corpus. Note that we can easily apply debias techniques (e.g., reweight, TDE [3]) in our VS$^3$ framework to further mitigate the bias issue.

## References


[1] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, et al. Grounded language-image pre-training. In *CVPR*, 2022. 1

[2] Kaihua Tang. A scene graph generation codebase in pytorch, 2020. https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch. 1

[3] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 1

[4] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 1

[5] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 1

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1

[7] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 1

[8] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *CVPR*, 2023. 1