Appendix: Fine-Grained Scene Graph Generation with Data Transfer

Ao Zhang[1*], Yuan Yao[2*], Qianyu Chen[2], Wei Ji[1†], Zhiyuan Liu[2†], Maosong Sun[2], and Tat-Seng Chua[1]

[1] Sea-NExT Joint Lab, Singapore
School of Computing, National University of Singapore, Singapore
[2] Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
aozhang@u.nus.edu, yaoyuanthu@163.com

1 VG-1800 Dataset

The VG-1800 dataset aims to provide reliable evaluation for the large-scale scene graph generation.

1.1 Dataset Construction

We construct the dataset based on original Visual Genome dataset [2] by the following steps: (1) **Filtration.** Instead of simply auto-filtering [7] and choosing the top frequent predicate categories [1], we manually filter out unreasonable predicate categories, including misspelling predicates (e.g., `i frot of`), adjectives (e.g., `white`), nouns (e.g., `car`), and relative clauses (e.g., `who has`). To provide enough relation instances for robust evaluation, we retain all object categories and predicate categories with over 5 samples. (2) **Split.** We split the VG dataset into 70% training and 30% test. Following VG-50 split [5], we further split out $5,000$ images from the training set as the validation set, and ensure at least 5 samples on the test set and at least 1 samples on the training set for each predicate category.

1.2 Dataset Statistics

Finally, the dataset contains $70,098$ object categories, $1,807$ predicate categories and $272,084$ distinct relation triplets. It consists of $66,289$, $4,995$, and $32,893$ images for training set, validation set and test set respectively. There are on average 19.5 objects and 16.0 relations for each image.

Comparison with Other VG Splits. We also compare our VG-1800 split with other splits based on Visual Genome [2] dataset, including a conventional VG-50

∗ indicates equal contribution.

† Corresponding author: jiwei@nus.edu.sg, liuzy@tsinghua.edu.cn

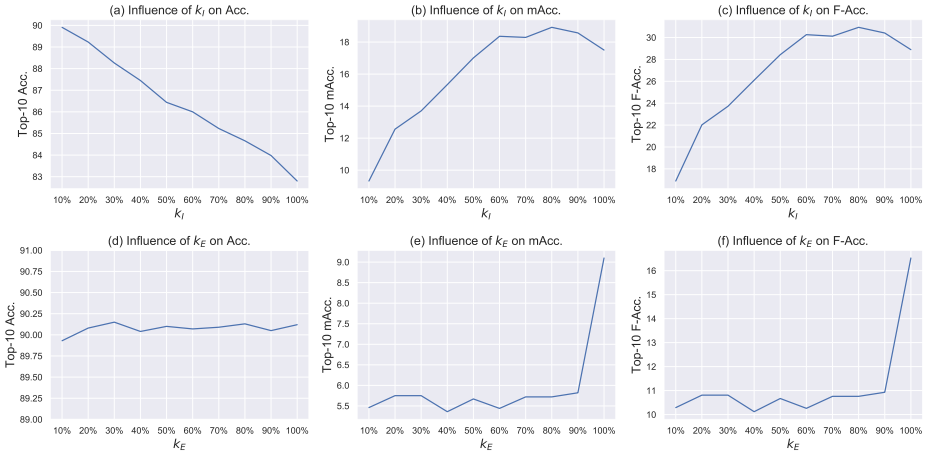

Fig. 1: Influence of k_I and k_E in different metrics.

and the other two large-scale SGG splits VG8K and VG8K-LT. Our VG-1800 can provide a more reliable evaluation for large-scale SGG. (1) When compared with VG8K, we provide a much cleaner dataset by manually cleaning the noise. For example, VG8K does not filter out nouns and adjectives, which will lead to an unreliable evaluation. (2) When compared with VG8K-LT, a cleaner version of VG8K, we provide a much stable evaluation for large amount of tail classes. More specifically, as shown in Table 1, our VG-1800 contains more test images. Meanwhile, VG-1800 also contains more samples of tail classes. As shown in Figure 2, our VG-1800 has 1,807 predicate classes with no less than 5 samples, while VG8K-LT has only 526 classes that have no less than 5 samples.

Table 1: Comparison between different datasets' predicate filtration and split ratio. Filtration denotes the method to remove noisy predicates. The Train, Val, and Test denote number of images in training set, validation set and test set.

Dataset	Filtration	Train	Val	Test
VG-50 [5]	-	57,723	5,000	26,446
VG8K [7]	Auto	97,961	2,000	4,871
VG8K-LT [1]	Auto	97,623	1,999	4,860
VG-1800	Manual	66,289	4,995	32,893

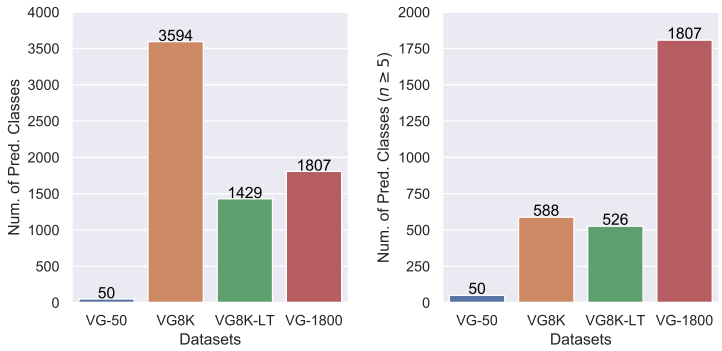

Fig. 2: Comparison of number of predicate classes between different splits on test set. The $n \geq 5$ denotes predicate classes having no less than 5 samples.

2 Implementation Details

2.1 VG-50

All normally trained baseline models including Motif, Transformer, VCTree, and GPS-Net are reproduced by us. For fair comparison, we equally remove all resampling and reweighting strategies. Morever, to encourage informative SGG, we remove the frequency bias in the training and inference of base models, which may make the result tend to have higher mR@K, F@K and lower R@K than results in their original papers.

For Transformer, some implementation details are different. The batch size can be enlarged to 16 on 2 GPUs. The learning rate is reduced to 0.08 for PREDCLS and SGDET for training stability. For Transformer on SGCLS, where the training is even more unstable, we further lower the learning rate to 0.016. All experiments are done on RTX-2080ti GPUs.

2.2 VG-1800

Compared with VG-50, the same backbone, parameter fixation, learning rate, optimizer, and learning schedule are used on VG-1800 dataset. Specially, due to the significant increase of (*subject*, `predicate`, *object*) combinations, we equally remove all frequency bias items on VG-1800 to reduce machines' memory usage. For internal and external transfer, the k_I is set to 90% and k_E is set to 100%.

For baselines, we find that due to the significant difference between the number of head predicates and tail predicates, the bi-level resampling in BGNN [3] will make the model pay most of the attention on tail classes while ignoring head classes. The drop out rate of images that do not contain rare predicates are set to almost 100%. This lead to a bad convergence of BGNN. Thus, we remove bi-level resampling for BGNN results. For RelMix [1], we equip the proposed VilHub loss and predicate feature mixup to the Neural Motif model. Similar

with bi-level sampling, we find that the reweighting strategy also lead to worse results. Thus, we do not include a reweighting version like VG-50.

3 Supplementary Experiments

3.1 Influence of k_I abd k_E

Influence of k_I. To provide a more detailed analysis on the influence of k_I, we report the performance on Acc and mAcc with different k_I. As shown in Figure 1, with the increase of internal transfer percentage k_I, Acc decreases linearly, while mAcc first increases when $k_I \leq 80\%$ and then decreases. The phenomenon shows that transferring more in internal transfer does not necessarily mean higher mAcc. For VG-1800, The first 80% internal data transfer is helpful to improve mAcc, while the last confident 20% will harm the overall performance. We guess the last 20% data may contain too noise, which will lower the data quality for model training.

Influence of k_E. As for k_E, external transfer shows almost no influence on Acc and mAcc when $k_E \leq 90\%$, while significantly boost mAcc when $k_E = 100\%$. Contrary to our observations for k_I, the last 10% samples which are believed to be unuseful by models, seem to bring the most profitable boost for mAcc. We guess the reason is that model can not distinguish well between tail classes and `NA` samples, while this part of the data is essential to provide more training samples for tail classes.

3.2 Adaptive Threshold.

In our IETrans, when determining how much data to transfer, we equally use a fixed percentage number for all relational triplets, which seems to be sub-optimal. Thus, we also tried an adaptive threshold by considering the prediction score of concrete relational triplet instances.

For internal transfer, given a general relational triplet instance (o_s, p_G, o_o), we are required to decide whether to transfer to its corresponding informative type p_I. We denote the model's prediction score of object pair (o_s, o_o) on p_I as $s^*_{p_I}$. We denote the average and standard error of all (c_{o_s}, p_I, c_{o_o}) relational triplet instances' prediction score on p_I as μ_I and σ_I. We conduct the transfer when $s^*_{p_I}$ satisfies:

$$s^*_{p_I} > \mu_I + k\sigma_I, \tag{1}$$

where k is a hyperparameter. The intuition is that if a general instance (o_s, p_G, o_o)'s prediction score is over the average of all real (c_{o_s}, p_I, c_{o_o})'s prediction scores on p_I, (o_s, p_G, o_o) can probably be relabeled as an informative one. Meanwhile, the standard error is considered to further control the adaptive threshold.

By choosing different k including $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$, we can get a curve with different Acc and mAcc trade-offs. As shown in Figure 3, the model with adaptive thresholds is overall worse than our fixed percentage.

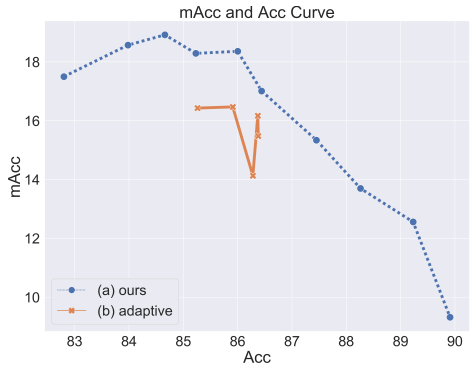

Fig. 3: Comparison of adaptive thresholds and fixed percentages for Internal Transfer.

A possible explanation is that the prediction score of an instance is non-linearly dependent on the number of its own instances and its similarity with different general classes, which results in inconsistency among different relational triplets. Especially when the number of an informative relational triplet is very small, the average of its prediction score is often near to zero, which will easily lead to an over-transfer problem. Thus, we leave the design of a more intelligent adaptive threshold for future work.

For external transfer, as shown in the paper, the prediction scores on `NA` of missed annotated samples are unreliable, i.e. the samples with the highest `NA` score bring maximum benefits for the model's performance.

SGCLS Results on VG-1800. We also evaluate our method on SGCLS task

Table 2: SGCLS triplet-level evaluation results on VG-1800 dataset.

Models	Top-1				Top-5				Top-10			
	Acc	mAcc	F-Acc	Non-Zero	Acc	mAcc	F-Acc	Non-Zero	Acc	mAcc	F-Acc	Non-Zero
BGNN [3]	16.29	0.18	0.36	22	22.99	0.86	1.65	159	24.15	1.48	2.78	221
Motif [6]	**18.93**	0.18	0.36	37	**26.08**	0.74	1.43	90	**27.28**	1.15	2.21	121
-Focal Loss	18.55	0.14	0.28	29	25.91	0.51	1.00	52	27.14	0.76	1.48	80
-TDE [4]	18.02	0.11	0.22	15	24.85	0.38	0.75	38	26.14	0.56	1.10	53
-RelMix [1]	18.27	0.22	0.43	47	25.57	0.83	1.60	100	26.71	1.26	2.42	130
-IETrans ($k_I = 10\%$) (ours)	18.24	1.68	3.16	212	25.80	1.68	3.16	212	27.25	2.54	4.65	264
-IETrans ($k_I = 90\%$) (ours)	4.91	**1.78**	**2.62**	**298**	20.66	**4.72**	**7.68**	**538**	24.85	**6.54**	**10.36**	**637**

on VG-1800 dataset. As shown in table 2, the comparison with other baselines is similar to the results on PREDCLS task. When compared with Motif, our IETrans ($k_I = 10\%$) can achieve significant improvement on top-1 mAcc and Non-Zero metrics (over 5 times of Motif) with negligible degeneration (less than

1 point) on top-1 Acc metric. Our IETrans ($k_I = 90\%$) can further boost the mAcc and Non-Zero metrics, which shows the ability of our IETrans to generate informative scene graphs. When compared with other baselines, our IEtrans can achieve the best F-Acc metrics across top-1, top-5, and top-10 evaluations. However, there is a large gap between SGCLS results and PREDCLS results (*e.g.*, 2.62% vs. 4.70% for top-1 F-Acc of IETrans ($k_I = 90\%$)), which indicates that further effort should be made to explore the joint optimization of both objects and predicates.

4 Discovered Visual Hierarchy Analysis

Visual Hierarchy Evaluation. A key element of conducting correct internal transfer is to find reasonable general-informative relation pairs. To evaluate the precision, we randomly choose 50 pairs with over 3 samples being transferred, so as to avoid involving too many noise-to-noise pairs. Then, human evaluation is conducted. The ratio of reasonable general-informative pairs is **76%** for VG-50, and **74%** for VG-1800.

Visualization. In the following, we show 100 discovered general-informative pairs for both VG-50 and VG-1800. The pairs are ranked by the number of samples which are transferred.

Table 3: Examples of discovered visual hierarchy in VG-50

(*window*, `on`, *building*) → (*window*, `part of`, *building*)
(*man*, `wearing`, *arm*) → (*man*, `wears`, *arm*)
(*boy*, `wearing`, *boy*) → (*boy*, `wears`, *boy*)
(*pillow*, `on`, *bed*) → (*pillow*, `lying on`, *bed*)
(*building*, `has`, *building*) → (*building*, `made of`, *building*)
(*sign*, `on`, *building*) → (*sign*, `mounted on`, *building*)
(*arm*, `of`, *arm*) → (*arm*, `belonging to`, *arm*)
(*sign*, `on`, *building*) → (*sign*, `hanging from`, *building*)
(*man*, `wearing`, *bag*) → (*man*, `wears`, *bag*)
(*window*, `on`, *building*) → (*window*, `belonging to`, *building*)
(*car*, `on`, *building*) → (*car*, `parked on`, *building*)
(*clock*, `on`, *building*) → (*clock*, `mounted on`, *building*)
(*man*, `wearing`, *building*) → (*man*, `wears`, *building*)
(*man*, `has`, *arm*) → (*man*, `wears`, *arm*)
(*man*, `wearing`, *boot*) → (*man*, `wears`, *boot*)
(*bottle*, `on`, *bottle*) → (*bottle*, `sitting on`, *bottle*)
(*window*, `on`, *bus*) → (*window*, `belonging to`, *bus*)
(*book*, `on`, *book*) → (*book*, `above`, *book*)
(*man*, `has`, *arm*) → (*man*, `with`, *arm*)
(*ear*, `of`, *ear*) → (*ear*, `belonging to`, *ear*)
(*hand*, `of`, *arm*) → (*hand*, `belonging to`, *arm*)

(*bottle*, `on`, *bottle*) → (*bottle*, `above`, *bottle*)
(*window*, `in`, *building*) → (*window*, `part of`, *building*)
(*light*, `on`, *building*) → (*light*, `mounted on`, *building*)
(*door*, `on`, *building*) → (*door*, `to`, *building*)
(*food*, `on`, *food*) → (*food*, `lying on`, *food*)
(*bowl*, `on`, *bowl*) → (*bowl*, `above`, *bowl*)
(*man*, `wearing`, *man*) → (*man*, `wears`, *man*)
(*flower*, `in`, *flower*) → (*flower*, `painted on`, *flower*)
(*woman*, `wearing`, *bag*) → (*woman*, `wears`, *bag*)
(*tree*, `near`, *building*) → (*tree*, `in front of`, *building*)
(*woman*, `wearing`, *arm*) → (*woman*, `wears`, *arm*)
(*window*, `of`, *building*) → (*window*, `part of`, *building*)
(*clock*, `on`, *building*) → (*clock*, `part of`, *building*)
(*bus*, `on`, *building*) → (*bus*, `parked on`, *building*)
(*man*, `wearing`, *coat*) → (*man*, `wears`, *coat*)
(*building*, `has`, *building*) → (*building*, `with`, *building*)
(*boy*, `wearing`, *arm*) → (*boy*, `wears`, *arm*)
(*man*, `on`, *arm*) → (*man*, `riding`, *arm*)
(*tree*, `has`, *branch*) → (*tree*, `with`, *branch*)
(*woman*, `wearing`, *boot*) → (*woman*, `wears`, *boot*)
(*pillow*, `on`, *bed*) → (*pillow*, `above`, *bed*)
(*woman*, `has`, *arm*) → (*woman*, `with`, *arm*)
(*window*, `on`, *building*) → (*window*, `to`, *building*)
(*glass*, `on`, *bottle*) → (*glass*, `sitting on`, *bottle*)
(*sign*, `on`, *building*) → (*sign*, `says`, *building*)
(*man*, `wearing`, *bike*) → (*man*, `wears`, *bike*)
(*tire*, `on`, *building*) → (*tire*, `on back of`, *building*)
(*branch*, `on`, *branch*) → (*branch*, `growing on`, *branch*)
(*book*, `on`, *book*) → (*book*, `laying on`, *book*)
(*car*, `on`, *car*) → (*car*, `parked on`, *car*)
(*man*, `wearing`, *bench*) → (*man*, `wears`, *bench*)
(*elephant*, `has`, *ear*) → (*elephant*, `using`, *ear*)
(*person*, `on`, *beach*) → (*person*, `standing on`, *beach*)
(*wing*, `on`, *plane*) → (*wing*, `attached to`, *plane*)
(*windshield*, `on`, *building*) → (*windshield*, `of`, *building*)
(*arm*, `on`, *arm*) → (*arm*, `belonging to`, *arm*)
(*woman*, `holding`, *bag*) → (*woman*, `carrying`, *bag*)
(*window*, `on`, *bike*) → (*window*, `part of`, *bike*)
(*ear*, `on`, *ear*) → (*ear*, `belonging to`, *ear*)
(*man*, `on`, *beach*) → (*man*, `walking on`, *beach*)
(*boy*, `has`, *boy*) → (*boy*, `wears`, *boy*)
(*roof*, `on`, *building*) → (*roof*, `covering`, *building*)
(*leaf*, `on`, *branch*) → (*leaf*, `growing on`, *branch*)

(*head*, `of`, *arm*) → (*head*, `belonging to`, *arm*)
(*wheel*, `on`, *building*) → (*wheel*, `on back of`, *building*)
(*tree*, `near`, *building*) → (*tree*, `along`, *building*)
(*bird*, `on`, *bird*) → (*bird*, `sitting on`, *bird*)
(*door*, `on`, *door*) → (*door*, `to`, *door*)
(*woman*, `has`, *bag*) → (*woman*, `with`, *bag*)
(*man*, `wearing`, *hat*) → (*man*, `wears`, *hat*)
(*man*, `on`, *arm*) → (*man*, `standing on`, *arm*)
(*sign*, `on`, *building*) → (*sign*, `attached to`, *building*)
(*letter*, `on`, *building*) → (*letter*, `painted on`, *building*)
(*bird*, `on`, *bird*) → (*bird*, `standing on`, *bird*)
(*ear*, `of`, *cat*) → (*ear*, `belonging to`, *cat*)
(*window*, `on`, *bench*) → (*window*, `part of`, *bench*)
(*window*, `near`, *building*) → (*window*, `part of`, *building*)
(*wheel*, `on`, *bike*) → (*wheel*, `on back of`, *bike*)
(*building*, `near`, *building*) → (*building*, `across`, *building*)
(*elephant*, `has`, *ear*) → (*elephant*, `between`, *ear*)
(*man*, `has`, *bag*) → (*man*, `with`, *bag*)
(*engine*, `on`, *plane*) → (*engine*, `mounted on`, *plane*)
(*man*, `wearing`, *chair*) → (*man*, `wears`, *chair*)
(*woman*, `wearing`, *building*) → (*woman*, `wears`, *building*)
(*sign*, `on`, *sign*) → (*sign*, `mounted on`, *sign*)
(*plate*, `on`, *bowl*) → (*plate*, `above`, *bowl*)
(*man*, `wearing`, *face*) → (*man*, `wears`, *face*)
(*leg*, `of`, *giraffe*) → (*leg*, `part of`, *giraffe*)
(*pillow*, `above`, *bed*) → (*pillow*, `lying on`, *bed*)
(*tire*, `on`, *bike*) → (*tire*, `on back of`, *bike*)
(*leg*, `of`, *arm*) → (*leg*, `belonging to`, *arm*)
(*man*, `wearing`, *cap*) → (*man*, `wears`, *cap*)
(*bird*, `has`, *bird*) → (*bird*, `with`, *bird*)
(*trunk*, `of`, *ear*) → (*trunk*, `belonging to`, *ear*)
(*roof*, `of`, *building*) → (*roof*, `covering`, *building*)
(*plate*, `on`, *bottle*) → (*plate*, `above`, *bottle*)
(*man*, `wearing`, *ear*) → (*man*, `wears`, *ear*)
(*man*, `has`, *building*) → (*man*, `wears`, *building*)
(*window*, `on`, *arm*) → (*window*, `part of`, *arm*)

Table 4: Examples of discovered visual hierarchy in VG-1800

(*window*, `on`, *building*) → (*window*, `on exterior of`, *building*)
(*man*, `wearing`, *arm*) → (*man*, `wearing striped`, *arm*)
(*arm*, `of`, *arm*) → (*arm*, `stretched out on`, *arm*)

(*man*, has, *arm*) → (*man*, stretching out, *arm*)
(*cloud*, in, *cloud*) → (*cloud*, floating through, *cloud*)
(*pillow*, on, *bed*) → (*pillow*, propped up on, *bed*)
(*tree*, in, *background*) → (*tree*, visible in, *background*)
(*leg*, of, *arm*) → (*leg*, belonging to, *arm*)
(*boat*, in, *boat*) → (*boat*, sailing on, *boat*)
(*building*, has, *building*) → (*building*, seen outside, *building*)
(*cloud*, in, *building*) → (*cloud*, floating through, *building*)
(*hand*, of, *arm*) → (*hand*, hand of, *arm*)
(*boat*, on, *boat*) → (*boat*, sailing on, *boat*)
(*cloud*, in, *sky*) → (*cloud*, floating through, *sky*)
(*building*, in, *background*) → (*building*, visible in, *background*)
(*man*, has, *arm*) → (*man*, sheltering, *arm*)
(*man*, wearing, *arm*) → (*man*, dressed in, *arm*)
(*head*, of, *arm*) → (*head*, turning, *arm*)
(*man*, has, *arm*) → (*man*, losing, *arm*)
(*cloud*, in, *airplane*) → (*cloud*, floating through, *airplane*)
(*cloud*, in, *background*) → (*cloud*, floating through, *background*)
(*window*, on, *awning*) → (*window*, on exterior of, *awning*)
(*man*, has, *arm*) → (*man*, pointing with, *arm*)
(*man*, has, *arm*) → (*man*, spreading, *arm*)
(*window*, on, *bus*) → (*window*, lining side of, *bus*)
(*window*, on, *balcony*) → (*window*, on exterior of, *balcony*)
(*cloud*, in, *beach*) → (*cloud*, floating through, *beach*)
(*car*, on, *building*) → (*car*, driving alongside, *building*)
(*window*, on, *building*) → (*window*, lining side of, *building*)
(*man*, wearing, *arm*) → (*man*, lifting up, *arm*)
(*shirt*, on, *arm*) → (*shirt*, worn by, *arm*)
(*man*, wearing, *bag*) → (*man*, wearing striped, *bag*)
(*bottle*, on, *bottle*) → (*bottle*, kept on, *bottle*)
(*man*, wearing, *background*) → (*man*, wearing striped, *background*)
(*boy*, wearing, *boy*) → (*boy*, striped, *boy*)
(*cloud*, in, *air*) → (*cloud*, floating through, *air*)
(*woman*, has, *arm*) → (*woman*, raising, *arm*)
(*car*, on, *building*) → (*car*, moving down, *building*)
(*airplane*, in, *airplane*) → (*airplane*, flying under, *airplane*)
(*tile*, on, *bathroom*) → (*tile*, fixed to, *bathroom*)
(*arm*, on, *arm*) → (*arm*, stretched out on, *arm*)
(*cloud*, in, *arm*) → (*cloud*, floating through, *arm*)
(*head*, of, *arm*) → (*head*, belonging to, *arm*)
(*man*, wearing, *arm*) → (*man*, kicking up, *arm*)
(*airplane*, on, *airplane*) → (*airplane*, taking off from, *airplane*)
(*sign*, on, *building*) → (*sign*, strapped, *building*)

(*man*, `wearing`, *air*) → (*man*, `wearing striped`, *air*)
(*sign*, `on`, *arrow*) → (*sign*, `strapped`, *arrow*)
(*window*, `on`, *building*) → (*window*, `adorning`, *building*)
(*cat*, `has`, *cat*) → (*cat*, `possesses`, *cat*)
(*cloud*, `in`, *boat*) → (*cloud*, `floating through`, *boat*)
(*person*, `has`, *arm*) → (*person*, `stretching out`, *arm*)
(*clock*, `on`, *building*) → (*clock*, `attached to side of`, *building*)
(*train*, `on`, *building*) → (*train*, `switching`, *building*)
(*woman*, `has`, *arm*) → (*woman*, `combing`, *arm*)
(*boy*, `wearing`, *arm*) → (*boy*, `striped`, *arm*)
(*window*, `on`, *building*) → (*window*, `on the side of`, *building*)
(*window*, `of`, *building*) → (*window*, `on exterior of`, *building*)
(*arrow*, `on`, *arrow*) → (*arrow*, `printed`, *arrow*)
(*woman*, `wearing`, *arm*) → (*woman*, `wearing striped`, *arm*)
(*head*, `of`, *arm*) → (*head*, `turned to`, *arm*)
(*branch*, `on`, *branch*) → (*branch*, `sticking up on`, *branch*)
(*man*, `on`, *arm*) → (*man*, `swimming with`, *arm*)
(*wing*, `on`, *airplane*) → (*wing*, `on left side of`, *airplane*)
(*mountain*, `in`, *background*) → (*mountain*, `visible in`, *background*)
(*bowl*, `on`, *bowl*) → (*bowl*, `placed on`, *bowl*)
(*cloud*, `in`, *blue sky*) → (*cloud*, `floating through`, *blue sky*)
(*window*, `on`, *arrow*) → (*window*, `on exterior of`, *arrow*)
(*foot*, `of`, *arm*) → (*foot*, `belonging to`, *arm*)
(*toilet*, `in`, *bathroom*) → (*toilet*, `installed in`, *bathroom*)
(*sign*, `on`, *building*) → (*sign*, `anchored to`, *building*)
(*wall*, `on`, *building*) → (*wall*, `making up`, *building*)
(*kite*, `in`, *air*) → (*kite*, `flying through`, *air*)
(*leaf*, `on`, *building*) → (*leaf*, `growing on`, *building*)
(*man*, `wearing`, *arm*) → (*man*, `adjusting`, *arm*)
(*person*, `wearing`, *arm*) → (*person*, `striped`, *arm*)
(*tile*, `on`, *bathroom*) → (*tile*, `installed on`, *bathroom*)
(*bag*, `on`, *bag*) → (*bag*, `kept in`, *bag*)
(*sky*, `in`, *sky*) → (*sky*, `stretched across`, *sky*)
(*bear*, `has`, *bear*) → (*bear*, `scratching`, *bear*)
(*line*, `on`, *building*) → (*line*, `painted in`, *building*)
(*man*, `wearing`, *building*) → (*man*, `wearing striped`, *building*)
(*book*, `on`, *book*) → (*book*, `arranged on`, *book*)
(*wall*, `near`, *building*) → (*wall*, `making up`, *building*)
(*window*, `on`, *advertisement*) → (*window*, `lining side of`, *advertisement*)
(*sink*, `in`, *bathroom*) → (*sink*, `mounted in`, *bathroom*)
(*wave*, `in`, *arm*) → (*wave*, `cresting in`, *arm*)
(*window*, `in`, *building*) → (*window*, `on exterior of`, *building*)
(*blanket*, `on`, *bed*) → (*blanket*, `laying over`, *bed*)

(*man*, `wearing`, *backpack*) → (*man*, `carrying`, *backpack*)
(*hair*, `of`, *arm*) → (*hair*, `on head of`, *arm*)
(*kite*, `in`, *beach*) → (*kite*, `kite in`, *beach*)
(*man*, `wearing`, *arm*) → (*man*, `dressed`, *arm*)
(*boy*, `has`, *arm*) → (*boy*, `outstretched`, *arm*)
(*picture*, `on`, *bed*) → (*picture*, `framed on`, *bed*)
(*window*, `on`, *banner*) → (*window*, `on exterior of`, *banner*)
(*arm*, `of`, *arm*) → (*arm*, `belonging to`, *arm*)
(*man*, `on`, *board*) → (*man*, `going off`, *board*)
(*shadow*, `on`, *arm*) → (*shadow*, `cast over`, *arm*)
(*arm*, `of`, *arm*) → (*arm*, `around neck of`, *arm*)

References

1. Abdelkarim, S., Agarwal, A., Achlioptas, P., Chen, J., Huang, J., Li, B., Church, K., Elhoseiny, M.: Exploring long tail visual relationship recognition with large vocabulary. In: Proceedings of ICCV. pp. 15921–15930 (2021)
2. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV pp. 32–73 (2017)
3. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of CVPR. pp. 11109–11119 (2021)
4. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of CVPR. pp. 3716–3725 (2020)
5. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of CVPR. pp. 5410–5419 (2017)
6. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural Motifs: Scene graph parsing with global context. In: Proceedings of CVPR. pp. 5831–5840 (2018)
7. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Proceedings of the AAAI. pp. 9185–9194 (2019)