

A. Overview

In this supplementary material, we present:

- Detailed dataset statistics in experiments (Section B).
- More detailed analysis of CaCao (Section C).
- Human evaluation of CaCao (Section D).
- Additional experimental results (Section E).
- Additional examples (Section F).

B. Dataset Statistics

Visual Genome. Table 1 and 3 show the coarse-grained predicates and fine-grained predicates with the number of training instances for each predicate in the Visual Genome dataset [9]. Table 2 and 4 show the coarse-grained predicates and fine-grained predicates with the number of training instances for each predicate after cross-modal boosting by our CaCao. We can observe that CaCao increases dataset scale, especially the tail predicates, which significantly alleviates the long-tail distribution problem in SGG.

GQA. For large-scale benchmark SGG, GQA [8] contains 113K images and over 3.8M relation annotations. In order to ensure the quality of the dataset, we perform a manual cleaning process to remove annotations that had poor quality or ambiguous meanings following prior works [5]. We finally select the top 200 object classes and top 100 predicate classes as the GQA-200 split like VG-50 to explore the generalization ability of CaCao in large-scale SGG.

VG-1800. VG-1800 [20] is another large-scale benchmark dataset, which filters out spelling errors and unreasonable relations, ultimately preserving 70,098 object classes and 1,807 predicate classes for more challenging scenarios. For each predicate category in VG-1800, there exist over 5 samples on the test set to provide a reliable evaluation.

C. Cross-Modal Predicate Boosting

C.1. Data Preprocessing

We first collect as many detailed pictures as possible from the Internet (*i.e.* CC3M, COCO caption) as the original data for training and get nearly 80k images and 2k predicate categories with corresponding descriptions. Then we conduct semantic analysis of the corresponding description statement of each image through **Stanford CoreNLP** and preserve those informative chunks (*i.e.* V, P, N, NP, and VP) to extract fine-grained triplets contained in captions. and Since the raw data contains much noise, we further design heuristic rules (*i.e.* corpus co-occurrence frequency, layer depth of lexical analysis) to filter out predicates that are not informative or misspelling automatically instead of handling them manually. We finally eliminate

those coarse-grained predicates and preserve 585 categories of diverse predicates to obtain informative <subject, predicate, object> relationships, which nearly cover most of the common situations in the real world, as shown in Table 7. Since the VG dataset also contains some fine-grained predicates, there are 27 categories of informative predicates we obtained have overlap with them.

C.2. Adaptive Semantic Cluster Loss

Importance of semantic co-reference. We list more semantic co-reference words and some clustering results as shown in the table 5, such as he “walks through” / “is passing through” / “passed by” a street may correspond to the same predicate “walking on. To address the semantic co-reference challenge, we proceed to train CaCao using the ASCL based on predicate semantic clusters. Since there are strong dependencies between triples in complex scenarios, for each predicate class, we represent and average the embeddings of all triples corresponding to it. To achieve this, we use the feature map of the last BERT layer as the representation of each entire triplet. We initialize the target predicate according to different similarity thresholds, and then confirm the number of initial centroids.

Importance of semantic ambiguity. Although semantic clustering is static to contexts, CaCao dynamically adjusts the predicted results based on context-aware labels, which are sensitive to various contexts. Then semantic clustering promotes diverse expressions for the adjusted synonyms, which are also context-sensitive. Besides, we find only a few semantic ambiguities caused by contexts (6% for ‘wearing’ to ‘has’) in the current dataset and analyze that the influence of contexts on synonyms in SGG is small during training. For a few failure cases caused by complex semantic ambiguities, we provide several candidates to correct the mapping and obtain more accurate prediction results.

C.3. Fine-Grained Predicate Boosting

In Figure 1a and 1b, we show the predicate distributions of the standard SGG dataset and open-world boosted data from CaCao. To enhance predicates into the target scene graphs, we need to establish the mapping from diversity predicates to target predicates, as shown in Table 10.

Moreover, we notice that there exists ambiguity and overlap between coarse-grained predicates and fine-grained predicates in fact. We further create the mapping between fine-grained predicates and coarse-grained predicates based on the semantic association between predicates [2]. We then figure out those low-confidence fine-grained predicates and map them into general predicates as final predicted results to achieve better trade-offs on long-tail recognition.

Coarse-grained Predicates	above	across	against	along	and	at	behind	between	for	from
Number of Predicates	47341	1996	3092	3624	3477	9903	41356	3411	9145	2945
Coarse-grained Predicates	has	in	in front of	near	of	on	over	to	under	with
Number of Predicates	277936	251756	13715	96589	146339	712409	9317	2517	22596	66425

Table 1. Statistics of **coarse-grained predicates** for the VG-50.

Coarse-grained Predicates	above	across	against	along	and	at	behind	between	for	from
Number of Predicates	47829	60320	88810	3722	10254	38305	43345	94138	10643	17149
Coarse-grained Predicates	has	in	in front of	near	of	on	over	to	under	with
Number of Predicates	300695	296474	24950	141494	197294	787048	12820	8672	43535	93078

Table 2. Statistics of **coarse-grained predicates** for the boosted VG-50 from CaCao.

Fine-grained Predicates	attached to	belonging to	carrying	covered in	covering	eating	flying in	growing on	hanging from	holding
Number of Predicates	10190	3288	5213	2312	3806	4688	1973	1853	9894	42722
Fine-grained Predicates	laying on	looking at	lying on	made of	mounted on	on back of	painted on	parked on	part of	playing
Number of Predicates	3739	3083	1869	2380	2253	1914	3095	2721	2065	3810
Fine-grained Predicates	riding	says	sitting on	standing on	using	walking in	walking on	watching	wearing	wears
Number of Predicates	8856	2241	18643	14185	1925	1740	4613	3490	136099	15457

Table 3. Statistics of **fine-grained predicates** for the VG-50.

Fine-grained Predicates	attached to	belonging to	carrying	covered in	covering	eating	flying in	growing on	hanging from	holding
Number of Predicates	80066	20858	79148	54015	17879	100241	6752	20290	90025	68378
Fine-grained Predicates	laying on	looking at	lying on	made of	mounted on	on back of	painted on	parked on	part of	playing
Number of Predicates	31783	150817	21944	27189	62583	20628	36882	68218	14727	20789
Fine-grained Predicates	riding	says	sitting on	standing on	using	walking in	walking on	watching	wearing	wears
Number of Predicates	62625	22273	68474	70311	63777	32956	38853	235425	258332	60328

Table 4. Statistics of **fine-grained predicates** for the VG-50.

Predicted Predicates	Semantic Co-reference Predicates
'wearing'	['wearing', 'worn on', 'carrying']
'holding'	['holding', 'carrying', 'pulling']
'next to'	['next to', 'sitting next to', 'standing next to']
'standing in'	['standing in', 'standing on', 'standing by']
'below'	['below', 'beneath', 'standing behind']
'flying in'	['flying', 'flying in', 'floating in']
'sitting on'	['sitting at', 'sitting in', 'is seated on']
'hang on'	['hang on', 'hanging on', 'hanging from']
'covered in'	['covered in', 'covered with', 'covered by']
'surrounded by'	['surrounded by', 'covered by', 'pulled by']
'walks through'	['walks through', 'is passing through', 'passed by']

Table 5. The examples of top clustering results for semantic co-reference predicates

D. Human Evaluation

A key element of effective SGG boosting is to obtain high-quality data. Thus, we conduct a human evaluation for automatically obtained labels from CaCao to verify the quality. We randomly select 100 images containing 545 base relationships and 3543 novel relationships to validate the accuracy and informativeness of the predicates associated with these augmented relationships, ensuring their utility in facilitating open-world predicate scene graph generation. We show the result in Table 8. We observe the ratio of reasonable fine-grained predicates in CaCao is **73.4%** and the proportion of coarse-grained predicates is greatly reduced by CaCao’s enhanced predicates. Consequently, the

Models	<i>PredCls</i>	<i>SGCls</i>	<i>SGDet</i>
	zsR@50/100 ↑	zsR@50/100 ↑	zsR@50/100 ↑
MOTIFS [19]	10.9 / 14.5	2.2 / 3.0	0.1 / 0.2
+Resample [1]	11.1 / 14.3	2.3 / 3.1	0.1 / 0.3
+TDE-GATE [13]	5.9 / 8.1	3.0 / 3.7	2.2 / 2.8
+Label Refine [6]	14.4 / -	3.0 / -	3.1 / -
+QuatRE [17]	11.9 / 15.2	2.8 / 3.6	0.2 / 0.4
+CaCao	12.0 / 13.1	5.1 / 5.8	3.6 / 3.9
VCTree [14]	10.8 / 14.3	1.9 / 2.6	0.2 / 0.7
+TDE-GATE [13]	7.7 / 11.0	1.9 / 2.6	1.9 / 2.5
+Label Refine [6]	13.5 / -	6.2 / -	3.3 / -
+QuatRE [17]	11.3 / 14.4	3.5 / 4.4	0.5 / 0.9
+CaCao	13.6 / 14.9	6.5 / 7.2	3.3 / 5.2
Transformer [13]	11.3 / 14.7	2.5 / 3.3	0.9 / 1.1
+CaCao	14.5 / 15.9	4.8 / 5.7	4.4 / 5.7

Table 6. Comparisons of the VG-50 SGG results on zero-shot combinational generalization performance (zsR@K) among various approaches.

results indicate that the predicates enhanced by CaCao can effectively provide fine-grained information.

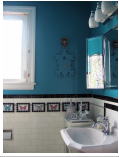


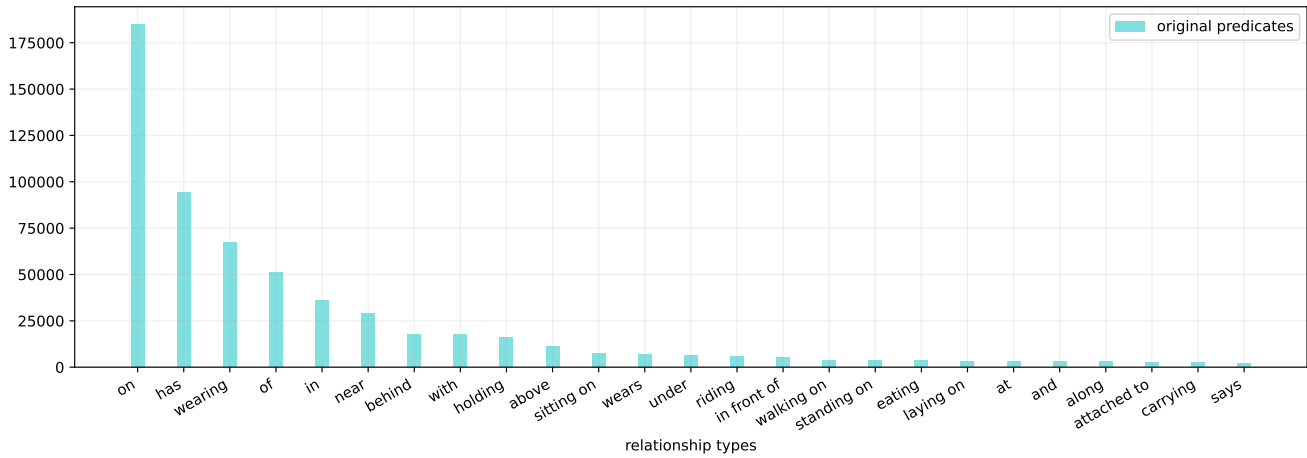
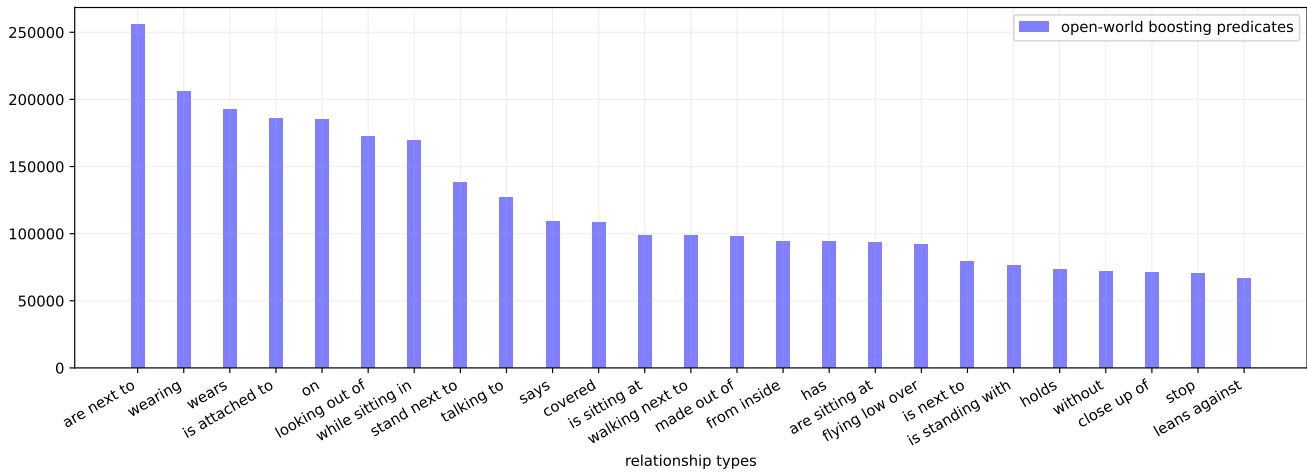
Image	Description	Extracted Relationships
	A clock that blends in with the wall hangs in a bathroom.	('clock', 'blends in with', 'wall') ('clock', 'in with', 'wall') ('clock', 'with', 'wall') ('clock', 'hangs in', 'bathroom') ('clock', 'in', 'bathroom')
	A couple at the beach walking with their surfboards.	('couple', 'at', 'beach') ('couple', 'walking with', 'their-surf') ('couple', 'with', 'their-surf')
	A yellow and black bird standing on and hanging with a bike rack.	('black-bird', 'on', 'bike-rack') ('yellow-bird', 'on', 'bike-rack') ('black-bird', 'standing on', 'bike-rack') ('black-bird', 'hanging with', 'bike-rack')

Table 7. The examples of <subject, predicate, object> extraction from raw data for prompt tuning.



(a) Top-25 predicates of original distribution



(b) Top-25 predicates of open-world distribution

Figure 1. Qualitative predicate distributions of the standard SGG dataset and the open-world enhanced data from CaCao.

E. Additional Experiment Analyses

Compositional Generalization. Thanks to the remarkable performance of our CaCao in the open-world scenario, it

	Total Predicate	True Predicate	Fine-Grained Predicate (%) \uparrow	Coarse-Grained Predicate (%) \downarrow
Original	545	545	119 (21.8%)	426 (78.2%)
CaCao	3543	2427	1781 (73.4%)	646 (26.6%)
Overall	4088	2972	1900 (63.9%)	1072 (36.1%)

Table 8. Human evaluation for the accuracy and variety of enhanced predicates from CaCao.

Model Type	Methods	Scene Graph Detection			
		R@50/100 \uparrow	mR@50/100 \uparrow	F@50/100 \uparrow	
Specific	BGNN [10]	31.0 / 35.8	10.7 / 12.6	15.9 / 18.6	
	SVRP [7]	31.8 / 35.8	10.5 / 12.8	15.8 / 18.9	
	DT2-ACBS [4]	15.0 / 16.3	22.0 / 24.0	17.8 / 19.4	
One-stage	SSRCNN [15]	23.7 / 27.3	18.6 / 22.5	20.8 / 24.7	
	+CaCao (ours)	25.4 / 30.0	18.7 / 23.1	21.5 / 26.1	
Model-Agnostic strategy	Motif [19]	31.0 / 35.1	6.7 / 7.7	11.0 / 12.6	
	Resample	+Resample [1]	30.5 / 35.4	8.2 / 9.7	12.9 / 15.2
		+Reweight [16]	24.4 / 29.3	10.5 / 13.2	14.7 / 18.2
	Reweight	+CogTree [18]	20.0 / 22.1	10.4 / 11.8	13.7 / 15.4
		+FGPL [12]	21.3 / 24.3	15.4 / 18.2	17.9 / 20.8
		+GCL [5]	18.4 / 22.0	16.8 / 19.3	17.6 / 20.6
	Causal Rule	+TDE [13]	16.9 / 20.3	8.2 / 9.8	11.0 / 13.2
		+Only Caption Relations	20.3 / 25.0	8.2 / 10.0	11.7 / 14.3
	Data Enhancement	+DLFE [3]	25.4 / 29.4	11.7 / 13.8	16.0 / 18.8
		+IETrans [20]	23.5 / 27.2	15.5 / 18.0	18.7 / 21.7
+CaCao (ours)		24.4 / 29.1	17.1 / 20.0	20.5 / 23.7	

Table 9. Performance (%) of our method **CaCao** and other baselines with different model types for both **head** and **tail** categories on VG-50 dataset.

demonstrates the potential to improve the model compositional generalization ability in traditional zero-shot scene graph generation tasks [11, 6, 17]. Table 6 presents the zero-shot Recall@K metrics in each task (*i.e.*, *PredCls*, *SGCls*, and *SGDet*), providing a comprehensive evaluation of the compositional generalization performance. We compare our proposed CaCao with other state-of-the-art approaches. Our proposed method achieves improvements in most of the settings with different SGG backbones, except for MOTIFS in *PredCls*. MOTIFS being a textual-only model fails to effectively utilize the enhanced data to learn implicit features for discerning the combination of relations and hence performs poorly when given the ground truth contexts. Conversely, the multi-modal VCTree and Transformer models effectively utilize extra triplet-level data due to their ability to align more visual information, facilitating generalization to unseen triplets during testing.

Further Evaluation on Head and Tail Predicates. Since CaCao brings much extensive visual relation knowledge on various visual predicates from powerful VL-models, the CaCao may achieve a better trade-off on long-tail distribution

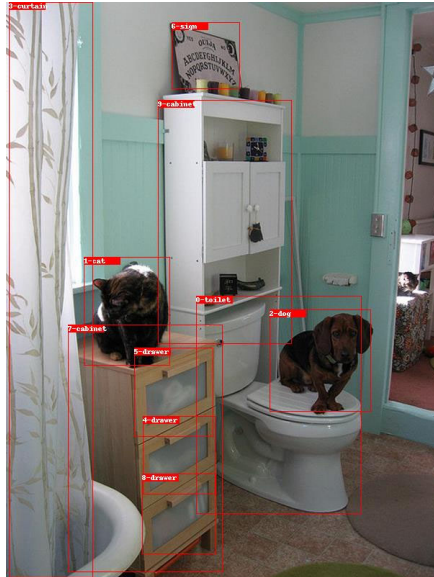
SGG. Our results on the whole category set partly give evidence that CaCao can achieve a better balance in the long-tail distribution. Additionally, we inspect the performance of CaCao across non-rare head predicates to further verify its better balance between head and tail predicate categories in Table 9 **R@K**. Following prior works [20], we further use the harmonic average of R@K and mR@K to jointly evaluate R@K and mR@K, which is denoted as **F@K**. From Table 9, we observe that **CaCao** outperforms other SOTA model-agnostic methods and specific string baseline according to the joint metric **F@K** (**20.5 / 23.7** of F@50/100 on SGGDet), showing the effectiveness of CaCao on both head and tail categories.

F. Additional Examples

Figure 2 shows some more examples for qualitative visualizations of enhanced SGG based on our CaCao.

Open-world predicate relationships → Target predicate relationships
['sidewalk', 'in between', 'car'] → ['sidewalk', 'between', 'car']
['sidewalk', 'walking across', 'street'] → ['sidewalk', 'across', 'street']
['tree', 'hanging in', 'building'] → ['tree', 'hanging from', 'building']
['tree', 'uses', 'phone'] → ['tree', 'using', 'phone']
['car', 'are parked on', 'street'] → ['car', 'parked on', 'street']
['street', 'parked at', 'sidewalk'] → ['street', 'parked on', 'sidewalk']
['street', 'among', 'car'] → ['street', 'between', 'car']
['phone', 'hanging on', 'tree'] → ['phone', 'hanging from', 'tree']
['motorcycle', 'displaying', 'person'] → ['motorcycle', 'carrying', 'person']
['building', 'connected to', 'pole'] → ['building', 'attached to', 'pole']
['street', 'parked at', 'sidewalk'] → ['street', 'parked on', 'sidewalk']
['shirt', 'leans against', 'woman'] → ['shirt', 'against', 'woman']
['glass', 'hanging on', 'head'] → ['glass', 'hanging from', 'head']
['chair', 'to make', 'leg'] → ['chair', 'made of', 'leg']
['man', 'watch', 'woman'] → ['man', 'watching', 'woman']
['man', 'leaning up against', 'table'] → ['man', 'against', 'table']
['screen', 'laying on', 'paper'] → ['screen', 'lying on', 'paper']
['paper', 'looking up at', 'screen'] → ['paper', 'looking at', 'screen']
['tree', 'hanging over', 'trunk'] → ['tree', 'hanging from', 'trunk']
['car', 'hooked up to', 'pole'] → ['car', 'attached to', 'pole']
['tree', 'across from', 'fence'] → ['tree', 'between', 'fence']
['sidewalk', 'hanging in', 'trunk'] → ['sidewalk', 'hanging from', 'trunk']
['sidewalk', 'traveling on', 'leaf'] → ['sidewalk', 'growing on', 'leaf']
['boy', 'looking down at', 'car'] → ['boy', 'looking at', 'car']
['woman', 'is using', 'pant'] → ['woman', 'using', 'pant']
['woman', 'towing', 'shirt'] → ['woman', 'carrying', 'shirt']
['head', 'connected to', 'nose'] → ['head', 'attached to', 'nose']
['hair', 'is looking at', 'child'] → ['hair', 'looking at', 'child']
['nose', 'tied to', 'head'] → ['nose', 'attached to', 'head']
['finger', 'is parked on', 'hand'] → ['finger', 'painted on', 'hand']
['man', 'eaten', 'pizza'] → ['man', 'eating', 'pizza']
['windshield', 'towing', 'umbrella'] → ['windshield', 'carrying', 'umbrella']
['airplane', 'hanging on', 'wing'] → ['airplane', 'hanging from', 'wing']
['airplane', 'hanging on', 'wing'] → ['airplane', 'hanging from', 'wing']
['airplane', 'flying high in', 'sky'] → ['airplane', 'flying in', 'sky']
['sign', 'strapped', 'arrow'] → ['sign', 'on', 'arrow']
['face', 'connected to', 'neck'] → ['face', 'above', 'neck']
['tree', 'across from', 'building'] → ['tree', 'across', 'building']
['roof', 'across from', 'building'] → ['roof', 'along', 'building']
['jacket', 'is cluttered with', 'man'] → ['jacket', 'with', 'man']
['sign', 'are showing on', 'building'] → ['sign', 'says', 'building']
['short', 'in between', 'man'] → ['short', 'with', 'man']
['jean', 'stacked on', 'man'] → ['jean', 'painted on', 'man']
['person', 'is walking on', 'sidewalk'] → ['person', 'walking on', 'sidewalk']
['chair', 'are looking at', 'boy'] → ['chair', 'in front of', 'boy']

Table 10. Examples of open-world predicates to target predicates mapping.

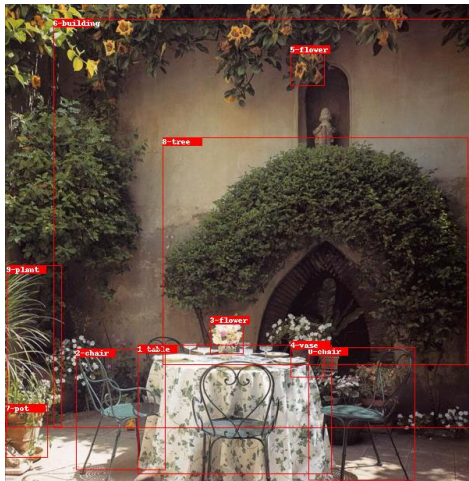


Ground Truth Triplets

6_sign --- **on** --- 9_cabinet
 2_dog --- **laying on** --- 11_seat
 6_sign --- **above** --- 0_toilet
 1_cat --- **laying on** --- 5_drawer
 3_curtain --- **near** --- 1_cat
 7_cabinet --- **near** --- 3_curtain
 8_drawer --- **on** --- 7_cabinet

Predicted Triplets

2_dog --- **laying on** --- 11_seat
 0_toilet --- **covered in** --- 11_seat
 7_cabinet --- **between** --- 5_drawer
 1_cat --- **laying on** --- 7_cabinet
 7_cabinet --- **made of** --- 3_curtain
 6_sign --- **on** --- 9_cabinet
 7_cabinet --- **near** --- 0_toilet
 8_drawer --- **on** --- 7_cabinet
 6_sign --- **attached to** --- 9_cabinet
 1_cat --- **laying on** --- 5_drawer
 6_sign --- **above** --- 0_toilet
 0_toilet --- **under** --- 2_dog



Ground Truth Triplets

2_chair --- **in front of** --- 6_building
 2_chair --- **near** --- 7_pot
 3_flower --- **on** --- 1_table
 8_tree --- **near** --- 6_building
 0_chair --- **near** --- 1_table

Predicted Triplets

8_tree --- **in front of** --- 6_building
 0_chair --- **near** --- 1_table
 8_tree --- **against** --- 11_tree
 2_chair --- **in front of** --- 6_building
 3_flower --- **in front of** --- 6_building
 2_chair --- **near** --- 7_pot
 9_plant --- **above** --- 7_pot
 5_flower --- **hanging from** --- 6_building
 3_flower --- **attached to** --- 1_table
 6_building --- **playing** --- 8_tree
 4_vase --- **on** --- 1_table

Figure 2. Additional Qualitative Results for Transformer equipped with our CaCao framework for predicate enhancement with the ground truth relationships. The predicted triplets are from the SGDet setting.

References

- [1] Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *Eighth international conference on machine vision (ICMV 2015)*, volume 9875, pages 423–427. SPIE, 2015. 2, 4
- [2] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2580–2590, 2019. 1
- [3] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 4
- [4] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021. 4
- [5] Xingning Dong, Tian Gan, Xueming Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 1, 4
- [6] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 2, 4
- [7] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. *arXiv preprint arXiv:2208.08165*, 2022. 4
- [8] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1
- [9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1
- [10] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021. 4
- [11] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 4
- [12] Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, and Jingkuan Song. Fine-grained predicates learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19467–19475, 2022. 4
- [13] 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. 2, 4
- [14] 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. 2
- [15] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19437–19446, 2022. 4
- [16] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021. 4
- [17] Zheng Wang, Xing Xu, Guoqing Wang, Yang Yang, and Heng Tao Shen. Quaternion relation embedding for scene graph generation. *IEEE Transactions on Multimedia*, 2023. 2, 4
- [18] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020. 4
- [19] 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. 2, 4
- [20] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. *arXiv preprint arXiv:2203.11654*, 2022. 1, 4