Supplementary Material for Fine-Grained Scene Graph Generation via Sample-Level Bias Prediction

Yansheng Li[1], Tingzhu Wang[1†], Kang Wu[1], Linlin Wang[1], Xin Guo[2], and Wenbin Wang[3]

[1] School of Remote Sensing and Information Engineering, Wuhan University
[2] Ant Group
[3] Hubei Key Laboratory of Intelligent Vision Monitoring for Hydroelectric Engineering, College of Computer and Information Technology, China Three Gorges University
tingzhu.wang@whu.edu.cn

A Dataset Details

GQA [1] is a vision-and-language dataset, consisting of a total of 113k images. We retain images only with the most frequent 160 object and 60 relationship categories for experiments. Then it contains 59,588 images, of which 41,773 (70%) images are used for the training, and 17,815 (30%) images are used for the testing. We follow [4] to sample a 5k validation set from the training set for parameter tuning. The detailed list of the most frequent 160 object and 60 relationship categories is shown in Tab. 1.

We visualize the quantity distribution for each relationship as shown in Fig. 1, GQA exhibits a severe long-tailed effect, with a highly imbalanced distribution between head categories (e.g., "*on*", "*wearing*", "*of*") and tail categories (e.g., "*contain*", "*pulling*", "*pulled by*").

B Ablation Studies

iii) The Effect of Weight Factor α: To assess the impact of α for SBG, we conduct the PredCls task on Transformer model. We validate a range of values (0.050, 0.075, 0.100) for α. The performance is presented in Tab. 2. From the results, it can be observed that the A@50/100 metric achieves the highest performance when α is set to 0.075, indicating the optimal performance of SBG.

iV) The Effect of Training Mode: In Section 3.2, we employ a gradual training mode, where the parameters of the classic SGG model are frozen after the training, and subsequently, the training of BGAN is conducted. The comparison between the gradual training and integrated training of SBG on the PredCls task of Transformer model is presented in Tab. 3. The results indicate

† indicates the corresponding author.

Table 1: List of object and relationship categories in GQA.

	categories
object	'window', 'man', 'shirt', 'tree', 'wall', 'person', 'building', 'ground', 'sky', 'sign', 'head', 'pole', 'hand', 'grass', 'hair', 'leg', 'car', 'woman', 'leaves', 'trees', 'table', 'ear', 'pants', 'people', 'eye', 'water', 'door', 'fence', 'nose', 'wheel', 'chair', 'floor', 'arm', 'jacket', 'hat', 'shoe', 'tail', 'clouds', 'leaf', 'face',' letter', 'plate', 'number', 'windows', 'shorts', 'road', 'flower', 'sidewalk', 'bag', 'helmet', 'snow', 'rock', 'boy', 'cloud', 'tire', 'logo', 'roof', 'glass', 'street', 'foot', 'umbrella', 'legs', 'post', 'jeans', 'mouth', 'boat', 'cap', 'bottle', 'bush', 'girl', 'flowers', 'shoes', 'picture', 'glasses', 'field', 'mirror', 'bench', 'box', 'dirt', 'bird', 'clock', 'neck', 'bowl', 'food', 'bus', 'letters', 'pillow', 'shelf', 'train', 'trunk', 'horse', 'airplane', 'plant', 'coat', 'lamp', 'kite', 'wing', 'elephant', 'house', 'cup', 'paper', 'dog', 'seat', 'sheep', 'street light', 'counter', 'branch', 'glove', 'banana', 'giraffe', 'book', 'rocks', 'cow', 'truck', 'racket', 'ceiling', 'flag', 'skateboard', 'cabinet', 'zebra', 'eyes', 'ball', 'bike', 'wheels', 'sand', 'surfboard', 'frame', 'hands', 'motorcycle', 'feet', 'windshield', 'finger', 'bushes', 'player', 'child', 'hill', 'sink', 'bed', 'cat', 'container', 'traffic light', 'sock', 'tie', 'towel', 'pizza', 'paw', 'backpack', 'collar', 'basket', 'mountain', 'vase', 'lid', 'phone', 'branches', 'animal', 'donut', 'fur', 'license plate', 'laptop', 'lady'
relationship	'on', 'wearing', 'of', 'near', 'in', 'behind', 'in front of', 'holding', 'on top of', 'next to', 'above', 'with', 'below', 'by', 'sitting on', 'under', 'on the side of', 'beside', 'standing on', 'inside', 'carrying', 'at', 'walking on', 'riding', 'standing in', 'around', 'covered by', 'hanging on', 'lying on', 'eating', 'watching', 'looking at', 'covering', 'sitting in', 'on the front of', 'hanging from', 'parked on', 'riding on', 'using', 'covered in', 'flying in', 'sitting at', 'walking in', 'playing with', 'full of', 'filled with', 'on the back of', 'crossing', 'swinging', 'surrounded by', 'standing next to', 'reflected in', 'covered with', 'contain', 'touching', 'pulling', 'pulled by', 'flying', 'leaning on', 'hitting'

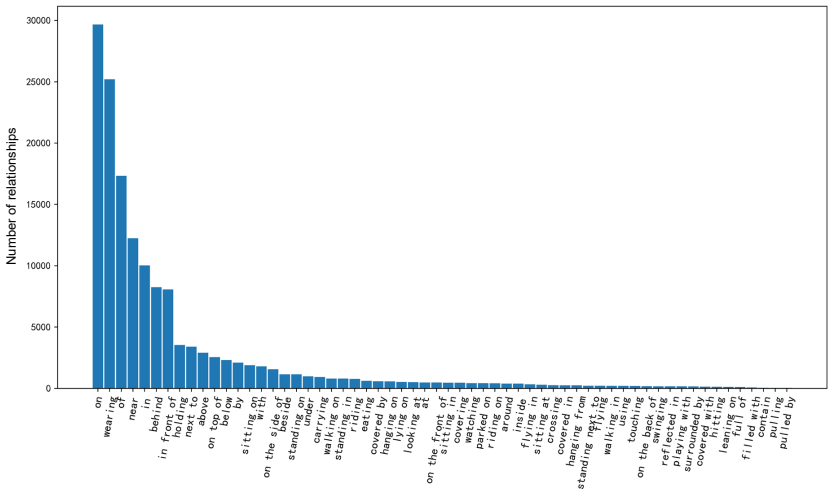

Fig. 1: Quantity distribution for each relationship varies from many to few.

Table 2: The effect of weight factor α.

	PredCls		
Weight Factor α	R@50/100	mR@50/100	A@50/100
0.050	55.9 / 57.7	33.0 / 35.3	44.5 / 46.5
0.075	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**
0.100	57.2 / 59.0	32.0 / 34.1	**44.6** / 46.6

that the gradual training outperforms the integrated training. This is because SBG is trained based on the output of the classic SGG model. However, the output of the classic SGG model using the integrated training is continuously varied, thus leading to the unstable training for SBG.

Table 3: The effect of training mode.

	PredCls		
Training Mode	R@50/100	mR@50/100	A@50/100
integrally	56.4 / 58.1	32.6 / 34.8	44.5 / 46.5
gradually	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**

V) The Superiority of BGAN for Sample-Level Bias Prediction: To demonstrate the sample-level bias's prediction capability of BGAN, which employs the one-dimensional convolution network, we conduct a comparison involving three networks: a conventional 5-layer fully connected network (denoted as FC_5), a 5-layer one-dimensional convolutional network (denoted as $1D_5$), and a fully connected BGAN (denoted as BGAN_{FC}), on the Predcls task of Transformer model. The results are presented in Tab. 4. It can be observed that in the case of FC_5 and $1D_5$ networks, the $1D_5$ network outperforms the FC_5 network slightly, as the $1D_5$ network benefits from the translation invariance and strong local receptive field provided by one-dimensional convolutions. Similarly, the performance of the BGAN based on one-dimensional convolutions is slightly better than that of BGAN_{FC} which uses the fully connected networks. Furthermore, by comparing the first and last two rows, BGAN exhibits stronger capabilities for the sample-level bias prediction than conventional neural networks.

Table 4: The superiority of BGAN for sample-level bias prediction.

	PredCls		
Network	R@50/100	mR@50/100	M@50/100
FC_5	42.3 / 44.0	37.4 / 39.7	39.9 / 41.9
$1D_5$	41.9 / 43.7	38.1 / 40.2	40.0 / 42.0
BGAN_{FC}	60.1 / 62.9	28.4 / 30.0	44.3 / 46.5
BGAN	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**

Vi) The Analysis for Feature Mapping ϕ: In Section 3.2, when constructing the correction bias set, we utilize an encoder that includes a single layer of transformer (denoted as $Trans_1$) to map high-dimensional features to one-dimensional features. We compare this approach with a conventional fully connected mapping (denoted as FC) and an encoder containing two layers of transformer (denoted as $Trans_2$), based on the Predcls task of Transformer model. The results are presented in Table Tab. 5. It is evident that using $Trans_1$ for feature mapping yields the best performance. Compared to FC, $Trans_1$ demonstrates superior performance by leveraging the strong interaction capabilities of the transformer. Moreover, $Trans_2$ is relatively complex and results in a performance decline.

Table 5: The analysis for feature mapping ϕ.

	PredCls		
Mapping Method	R@50/100	mR@50/100	M@50/100
FC	55.8 / 57.8	32.1 / 34.7	44.0 / 46.3
$Trans_1$	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**
$Trans_2$	55.8 / 57.6	32.9 / 35.3	44.4 / 46.5

Vii) The Structure Analysis of BGAN: The generator G and discriminator D in BGAN consist of multiple layers of one-dimensional convolution networks. The performance of G and D directly impacts the performance of BGAN. To assess their impact, we conduct experiments using various combinations of one-dimensional convolution layers for G and D based on the Predcls task of Transformer model. The results are presented in Tab. 6. Based on the combination (5, 3) of G and D (last row in the table), we individually keep the number of layers fixed for G and D while modifying the number of layers for the other. It is evident that among these combinations, the combination (5, 3) yields the best performance for both G and D.

Table 6: The structure analysis towards G and D in BGAN.

BGAN		PredCls		
G(layers)	D(layers)	R@50/100	mR@50/100	M@50/100
5	2	55.0 / 56.2	33.9 / 36.3	44.5 / 46.3
5	4	60.4 / 62.1	28.7 / 31.0	44.6 / **46.7**
4	3	57.0 / 59.1	32.2 / 34.1	**44.6** / 46.6
6	3	56.9 / 58.7	32.1 / 34.3	44.5 / 46.5
5	3	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**

Viii) The Effect of Small Non-Zero Value ε: In constructing the correction bias set (Section 3.2), we utilize the ε which is set to 0.0001. In order to

assess the impact of ε for SBG, we conduct the Predcls task using the Transformer model. We test a range of values (0.001, 0.0001, 0.00001) for ε, and the performance of our SBG is presented in Tab. 7. It can be observed that the M@50/100 metric achieves the highest performance when ε is set to 0.0001, indicating optimal comprehensive performance.

Table 7: The effect of small non-zero value ε.

	PredCls		
ε value	R@50/100	mR@50/100	M@50/100
0.001	56.6 / 58.4	32.5 / 34.8	**44.6** / 46.6
0.0001	55.8 / 57.6	33.3 / 35.7	**44.6** / **46.7**
0.00001	56.5 / 58.3	32.6 / 34.9	**44.6** / 46.6

iX) The Improvements of Long-Tailed Classes: In Fig. 2, we present the R@100 of each relationship for the PredCls task, comparing Transformer and our SBG. It shows that all tail classes are improved significantly.

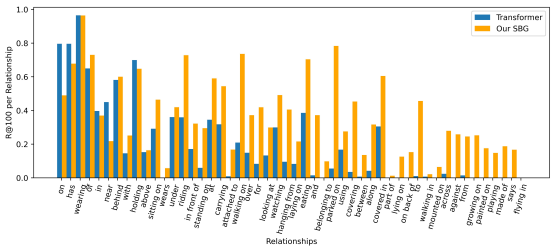

Fig. 2: Comparison of R@100 on Transformer and our SBG. The relationships are listed by the long-tailed order. Only "flying in" is not improved, whose training samples are only 5, affecting the correction effect of our method.

X) The Rationale for Generative Model. The bias in our SBG is non-linear and its continuity is very important for correction, so we compare generative models with non-generative models for bias prediction in Fig. 3. GAN has the dual optimisation that helps to predict the more non-linear bias, and that G and D of GAN supervise each other and promote each other making the $\mathbf{b}^{pre}$ predicted by GAN more closely approximate to the $\mathbf{b}^{tru}$ and capture the continuity of the $\mathbf{b}^{tru}$ better. These are also reflected in HiFi-GAN [3] and VCA-GAN [2].

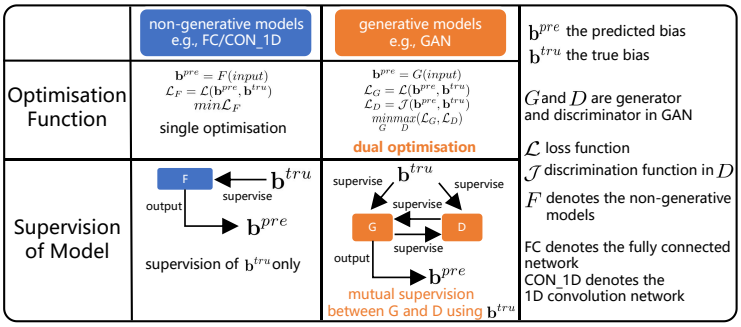

Fig. 3: Comparison of generative and non-generative models.

C Visualization for Bias Correction

In order to specifically demonstrate the process of sample-level bias correction, we illustrate the corrections of relationships for object pairs <man, boat> and <man, pole> as depicted in Fig. 4 (a) and Fig. 4 (b). The original predictions are the coarse-grained relationships of "*on*" and "*holding*". Utilizing the contextual information (from union region) of <man, boat> and <man, pole>, the relationships' global bias, and the original predictions, the generator in BGAN predicts the sample-specific biases to refine the coarse-grained "*on*" and "*holding*" to the fine-grained "*sitting on*" and "*using*".

References

1. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
2. Kim, M., Hong, J., Ro, Y.M.: Lip to speech synthesis with visual context attentional gan. Advances in Neural Information Processing Systems **34**, 2758–2770 (2021)
3. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems **33**, 17022–17033 (2020)
4. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020)

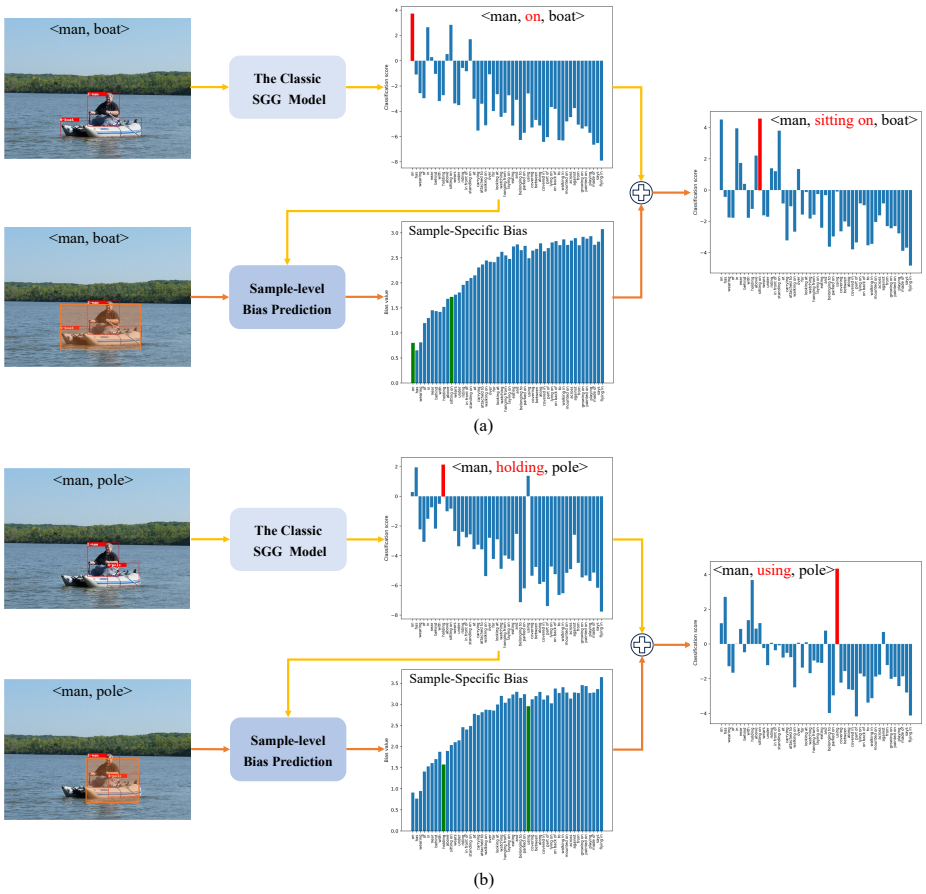

Fig. 4: The bias corrections of relationships for object pairs <man, boat> and <man, pole>.