Supplementary Material
Paper ID 499: Is An Image Worth Five Sentences?
A New Look into Semantics for Image-Text Matching

October 15, 2021

1 Implementation Details

In this section, we describe the hyper-parameters and the training procedure used to obtain the models shown in the main paper for the reduced data scenario (Table 2 - main paper) and the state-of-the-art comparison (Table 3 - main paper). Specifically, for each model we employ the training procedure described in the original paper. Each of the models use the 36 most confident regions obtained by a Faster R-CNN [4] from the object detector proposed by [1]. The visual features used by each model are the same in all cases. The best model is selected according to the sum of NCS with CIDEr as a similarity metric on the validation set. It is important to note that NCS with SPICE is the one we employ for evaluation.

For the model VSRN [7] + SAM, we start with the public pretrained weights. We add our SAM loss function and re-train the model for 30 epochs. For Flickr30K we employ a soft negative sampling, with a temperature parameter $\tau = 3$ and a weight on the SAM triplet loss of 20. When training on MSCOCO, we employ a soft negative sampling, with a temperature parameter $\tau = 10$ and a weight on the SAM triplet loss of 5. In both datasets, the original triplet loss is kept alongside with our SAM formulation. As in the original paper, the word embedding size is 300-d and the dimension of the final joint embedding space is 2048-d. The mini-batch size employed is 128. At training, the Adam optimizer [5] is used. The original triplet loss margin is 0.2. The model is trained for 30 epochs, with a learning rate of 0.0002 for the initial 15 epochs and is divided by 10 for the remaining 15 epochs.

In the CVSE [9] + SAM model, we train the model from scratch alongside with the proposed SAM loss function as follows. In Flickr30K, we employ a random sampling strategy with a temperature parameter $\tau = 7$ and a weight on the SAM triplet loss of 5 alongside with the original triplet. In MSCOCO, we employ a soft negative sampling strategy, with a temperature parameter $\tau = 5$ and a weight on the SAM triplet loss set to 5, and only our SAM triplet loss is used. Following the original CVSE model, the joint space dimension is 1024-d. The consensus exploitation is performed with a 300-d GloVe [8] representation. The loss formulation contains the following weights for each term are kept, $\lambda_1 = 3, \lambda_2 = 5, \lambda_3 = 1, \lambda_4 = 2$. The mini-batch size employed is 128. At training, the Adam optimizer [5] is used. The original triplet loss margin is 0.2. The model is trained for 30 epochs, with a learning rate of 0.0002 for the initial 15 epochs and is divided by 10 for the remaining 15 epochs.

Finally, for the SGR [2] + SAM model, we train the model from scratch in both datasets. In Flickr30K, we employ a random sampling strategy, with a temperature parameter $\tau = 10$ and a weight on the SAM triplet loss of 10. In MSCOCO, we use a random sampling strategy, with a temperature parameter $\tau = 5$ and a weight on the SAM triplet loss of 5. In both datasets, the original triplet is kept alongside with our SAM triplet. Training of the original model is performed as described by the authors. The word embedding size is 300-d and the number of hidden states is 1024-d. The dimension of the similarity representation is 256. The original triplet loss margin is 0.2. The number of reasoning steps is 3. The initial learning rate is set to 0.0002 for 10 epochs and is decreased by a 10 on the final 10 epochs on MSCOCO. For Flickr30K, the initial learning rate is kept for 30 epochs and it decays by 0.1 for the next 10 epochs.

2 Effect of Temperature and Sampling

In this section, we extend the results of the effect of setting different temperature values τ and sampling strategies. In all the experiments in this section, CVSE [9] is used. We evaluate the impact of these parameters on Flickr30K on Table 1 and removing GT items (non-GT) on Table 2. When calculating the results on MSCOCO 1K, we keep the soft negative (SN) sampling strategy while the impact of different values of τ and the usage of the original triplet is measured. Table 3 shows the results with the inclusion of GT and Table 4 depicts the results obtained with non-GT relevant items.

τ	S	T	Recall							Normalized Cumulative Semantic Recall						
			I2T			T2I				I2T			T2I			
			R@1	R@5	R@10	R@1	R@5	R@10	Rsum	N@1	N@5	N@10	N@1	N@5	N@10	Nsum
3	SN	✓	69.8	87.8	93.1	55.5	83.2	89.7	479.1	59.6	65.4	64.6	65.9	60.1	55.8	371.3
3	SN	✗	70.0	88.1	92.6	54.1	82.4	89.3	476.5	60.0	64.7	64.5	64.8	60.0	55.8	369.7
5	SN	✓	69.7	88.4	93.1	54.5	82.4	89.4	477.5	59.9	65.2	64.4	65.2	59.7	55.5	370.0
5	SN	✗	70.4	88.2	92.5	55.1	82.6	89.4	478.2	60.3	65.6	64.3	65.5	59.8	55.9	371.3
10	SN	✓	69.6	89.1	92.3	54.9	82.2	89.1	477.2	60.0	65.4	64.2	65.3	59.6	55.3	369.7
10	SN	✗	70.6	86.5	92.9	54.1	82.3	89.4	475.8	60.9	65.1	64.3	64.7	59.6	55.7	370.2
3	RS	✓	65.4	85.7	91.6	50.9	79.1	87.4	460.1	56.3	61.1	61.5	62.1	59.1	55.9	355.9
3	RS	✗	60.9	84.3	89.4	46.9	76.0	84.2	441.7	53.8	58.9	59.7	59.0	58.0	55.1	344.4
5	RS	✓	68.3	88.1	92.7	54.0	81.5	88.8	473.4	59.1	63.9	63.8	64.7	59.6	56.0	367.0
5	RS	✗	67.5	87.4	92.1	52.9	80.3	88.1	468.3	58.1	63.0	63.0	63.8	59.4	56.0	363.3
10	RS	✓	70.7	87.8	93.7	54.7	82.4	89.4	478.7	60.9	64.8	64.0	65.2	59.8	55.4	370.1
10	RS	✗	68.0	87.9	93.0	52.8	81.2	88.3	471.2	59.5	63.9	63.5	63.8	59.5	55.5	365.7
3	HN	✓	59.6	81.4	89.3	44.9	75.6	84.3	435.1	54.1	57.4	58.9	56.9	57.2	54.4	338.9
3	HN	✗	60.5	85.8	91.5	44.6	73.9	83.2	439.5	55.5	58.7	59.6	57.8	57.5	55.0	344.2
5	HN	✓	71.1	88.7	92.7	54.4	82.6	89.3	478.8	60.6	64.6	64.1	65.1	59.8	55.7	369.9
5	HN	✗	62.9	86.3	92.7	47.4	76.7	84.4	450.4	57.6	60.1	60.9	59.6	58.2	55.3	351.7
10	HN	✓	70.0	88.3	92.6	54.7	82.4	89.2	477.2	60.2	64.5	64.0	65.1	59.8	55.5	369.0
10	HN	✗	63.5	85.1	92.0	47.2	76.3	84.5	448.6	57.2	60.6	60.8	59.5	58.1	55.0	351.2

Table 1: Experiments on Flickr30K regarding the effect of (τ), soft (SN), random (RS) and hard negative (HN) sampling . The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[9] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

τ	S	T	Normalized Cumulative Semantic Recall (Non-GT)						
			I2T			T2I			
			N@1	N@5	N@10	N@1	N@5	N@10	Nsum
3	SN	✓	40.9	42.4	43.1	42.7	44.1	44.4	257.5
3	SN	✗	41.2	42.3	43.1	42.8	44.2	44.5	258.1
5	SN	✓	41.4	42.3	42.8	42.8	43.9	44.1	257.2
5	SN	✗	41.5	42.6	43.4	42.7	44.0	44.6	258.8
10	SN	✓	42.3	42.3	42.8	42.6	43.7	43.9	257.6
10	SN	✗	41.3	42.5	43.0	42.7	44.0	44.4	257.8
3	RS	✓	40.4	42.1	43.1	42.6	44.6	45.1	257.9
3	RS	✗	40.7	42.2	43.0	42.4	44.4	45.0	257.7
5	RS	✓	40.9	42.6	43.1	42.5	44.2	45.0	258.3
5	RS	✗	41.0	43.1	43.5	42.6	44.5	45.1	259.8
10	RS	✓	41.7	42.6	42.8	42.6	43.8	43.9	257.4
10	RS	✗	41.5	43.0	43.3	42.5	44.1	44.4	258.8
3	HN	✓	38.4	40.5	41.4	41.1	43.5	44.3	249.0
3	HN	✗	41.2	41.9	42.7	42.8	44.4	45.3	258.3
5	HN	✓	41.2	42.4	43.0	42.9	44.0	44.4	258.0
5	HN	✗	40.1	42.3	43.0	42.6	44.4	45.2	257.7
10	HN	✓	41.0	42.4	43.0	42.3	43.9	44.1	256.8
10	HN	✗	41.2	42.6	43.0	42.1	44.3	44.6	257.8

Table 2: Experiments on Flickr30K (Non-GT) regarding the effect of (τ), soft (SN), random (RS) and hard negative (HN) sampling . The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[9] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

τ	S	T	Recall							Normalized Cumulative Semantic Recall						
			I2T			T2I				I2T			T2I			
			R@1	R@5	R@10	R@1	R@5	R@10	Rsum	N@1	N@5	N@10	N@1	N@5	N@10	Nsum
3	SN	✓	76.2	93.7	96.3	63.6	91.7	96.4	517.9	69.2	73.7	69.7	75.8	72.5	68.6	429.4
3	SN	✗	74.8	93.5	96.8	62.2	91.5	96.2	515.0	68.3	73.2	69.4	74.8	72.7	69.0	427.3
5	SN	✓	76.4	94.0	97.3	64.2	92.2	96.4	520.5	68.8	74.1	70.0	76.1	72.5	68.4	429.7
5	SN	✗	76.1	93.8	96.9	63.2	91.4	96.6	518.0	68.2	73.6	69.6	75.5	72.4	68.9	428.2
10	SN	✓	76.9	94.2	97.7	64.4	91.8	96.3	521.3	69.9	73.8	69.6	76.3	71.9	67.7	429.1
10	SN	✗	76.8	94.2	97.4	64.2	91.8	96.2	520.6	70.1	73.7	69.7	76.3	72.1	68.0	429.9

Table 3: Experiments on MSCOCO 1K regarding the effect of (τ), employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all our experiments, we employ CVSE[9]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

τ	S	T	Normalized Cumulative Semantic Recall (Non-GT)						
			I2T			T2I			
			N@1	N@5	N@10	N@1	N@5	N@10	Nsum
3	SN	✓	45.4	46.0	46.0	56.9	58.9	59.0	312.2
3	SN	✗	45.2	46.2	46.2	56.9	59.1	59.5	313.1
5	SN	✓	45.4	46.2	46.2	56.6	58.6	58.6	311.5
5	SN	✗	45.3	46.2	46.2	56.9	59.1	59.3	312.9
10	SN	✓	44.1	45.2	45.1	56.6	57.8	57.7	306.5
10	SN	✗	44.4	45.7	45.5	57.1	58.3	58.2	309.3

Table 4: Experiments on MSCOCO 1K (Non-GT) regarding the effect of (τ), employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all our experiments, we employ CVSE[9]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

3 Qualitative Samples for Reduced Data Scenario

In this section, we provide qualitative samples on image-to-text and text-to-image in Flickr30K and MSCOCO 1K coming from the reduced data scenario (please refer to Section 5.2 in the main paper) by only using 10% of the training set in Flickr30K and 5% in MSCOCO. To offer additional insights, we provide the Recall (R^v) and NCS per each sample.

3.1 Image-to-Text

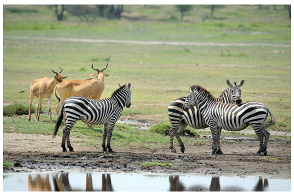	**Ground Truth**: Three zebras and other wild animals out in a semi-green field. Three zebras and two other animals grazing. Wildlife standing near water area in natural setting. Three zebras near the shore line of a body of water. A group of animals stand next to a watering hole.
CVSE + SAM	**Three zebras near the shore line of a body of water**. R_i : 1, S_i : 0.46 A heard of zebra on the plains at a watering hole. R_i : 0, S_i : 0.21 **A group of animals stand next to a watering hole**. R_i : 1, S_i : 0.33 A group of zebras and birds are gathered around water. R_i : 0, S_i : 0.17 There is a herd of zebras standing around. R_i : 0, S_i : 0.06 $R_i@5 = 1, N_i@5 = 0.61$
CVSE	**Three zebras near the shore line of a body of water**. R_i : 1, S_i : 0.46 There is a herd of zebras standing around. R_i : 0, S_i : 0.06 A heard of zebra on the plains at a watering hole. R_i : 0, S_i : 0.21 There are several zebras grazing near the water as a bird flies over them. R_i : 0, S_i : 0.05 A group of zebras and birds are gathered around water. R_i : 0, S_i : 0.17 $R_i@5 = 1, N_i@5 = 0.47$
SGR + SAM	**Three zebras near the shore line of a body of water**. R_i : 1,S_i : 0.46 Two zebras fighting in a cloud of dust. R_i : 0, S_i : 0.05 Three zebra in the middle of a field with a body of water in the distance. R_i : 0, S_i : 0.24 Three zebras standing in a sandy desert area. R_i : 0, S_i : 0.17 **Three zebras and other wild animals out in a semi-green field**. R_i : 1, S_i : 0.42 $R_i@5 = 1, N_i@5 = 0.67$
SGR	A zebra grazing on long dry grass in a field. R_i : 0,S_i : 0.05 Four zebras are grazing at a nature reserve. R_i : 0, S_i : 0.06 A group of animals stand next to a watering hole. R_i : 0, S_i : 0.06 Three zebras standing in a sandy desert area. R_i : 0, S_i : 0.17 The small herd of sheep are grazing on the grassy field. R_i : 0, S_i : 0.05 $R_i@5 = 0, N_i@5 = 0.19$

Table 5: Image-to-Text qualitative results in MSCOCO 1K. The initial row depicts the queried image and the associated ground truth captions. Each retrieved caption shows the Recall (R_i) and SPICE (S_i) score when compared with the GT captions. Each sample showcases the final per sample Recall ($R_i@5$) and NCS ($N_i@5$) score obtained. Bolded captions represent the correctly retrieved ground truth items.

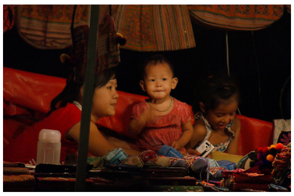	**Ground Truth**: Two children and a woman are sitting on a sofa, one of the children has a camera. Three Asian children sitting on a couch with tapestries hanging in the background. An Asian woman and her two children sit at a table doing crafts. A woman in a red shirt sitting with two young girls in dresses. Three young girls in arts and crafts room.
CVSE + SAM	**Two children and a woman are sitting on a sofa, one of the children has a camera.** R_i : 1, S_i : 0.39 **An Asian woman and her two children sit at a table doing crafts.** R_i : 1, S_i : 0.43 **Three Asian children sitting on a couch with tapestries hanging in the background.** R_i : 1, S_i : 0.39 A woman and three children are in a room full of toys. R_i : 0, S_i : 0.18 A group of children sitting on the floor, eating snacks at school R_i : 0, S_i : 0.05 $R_i@5 = 1, N_i@5 = 0.72$
CVSE	Three college-age women sit in upholstered chairs. R_i : 0, S_i : 0.05 Three young women face each other while sitting on red plush chairs. R_i : 0, S_i : 0.04 A plat is sitting on the floor next to a blond girl. R_i : 0, S_i : 0.05 Three girls talking in a lobby. R_i : 0, S_i : 0.10 Two kids sitting at a table eating. R_i : 0, S_i : 0.17 $R_i@5 = 0, N_i@5 = 0.16$
SGR + SAM	**Three Asian children sitting on a couch with tapestries hanging in the background.** R_i : 1,S_i : 0.39 **An Asian woman and her two children sit at a table doing crafts.** R_i : 1, S_i : 0.43 Six children are sitting around taking notes together. R_i : 0, S_i : 0.05 Woman on four way seesaw with 2 kids. R_i : 0, S_i : 0.14 A group of mostly asian children sitting at cubicles in blue chairs. R_i : 0, S_i : 0.09 $R_i@5 = 1, N_i@5 = 0.54$
SGR	A child playing in the ocean. R_i : 0, S_i : 0.05 Construction workers deal with removing railroad tracks. R_i : 0, S_i : 0.00 A mural of children on a brick wall. R_i : 0, S_i : 0.05 Four people in the subway are having fun. R_i : 0, S_i : 0.00 Several elderly men are grouped around a table. R_i : 0, S_i : 0.05 $R_i@5 = 0, N_i@5 = 0.07$

Table 6: Image-to-Text qualitative results in Flickr30K. The initial row depicts the queried image and the associated ground truth captions. Each retrieved caption shows the Recall (R_i) and SPICE (S_i) score when compared with the GT captions. Each sample showcases the final per sample Recall (R_i@5) and NCS (N_i@5) score obtained. Bold captions represent the correctly retrieved ground truth items.

3.2 Text-to-Image

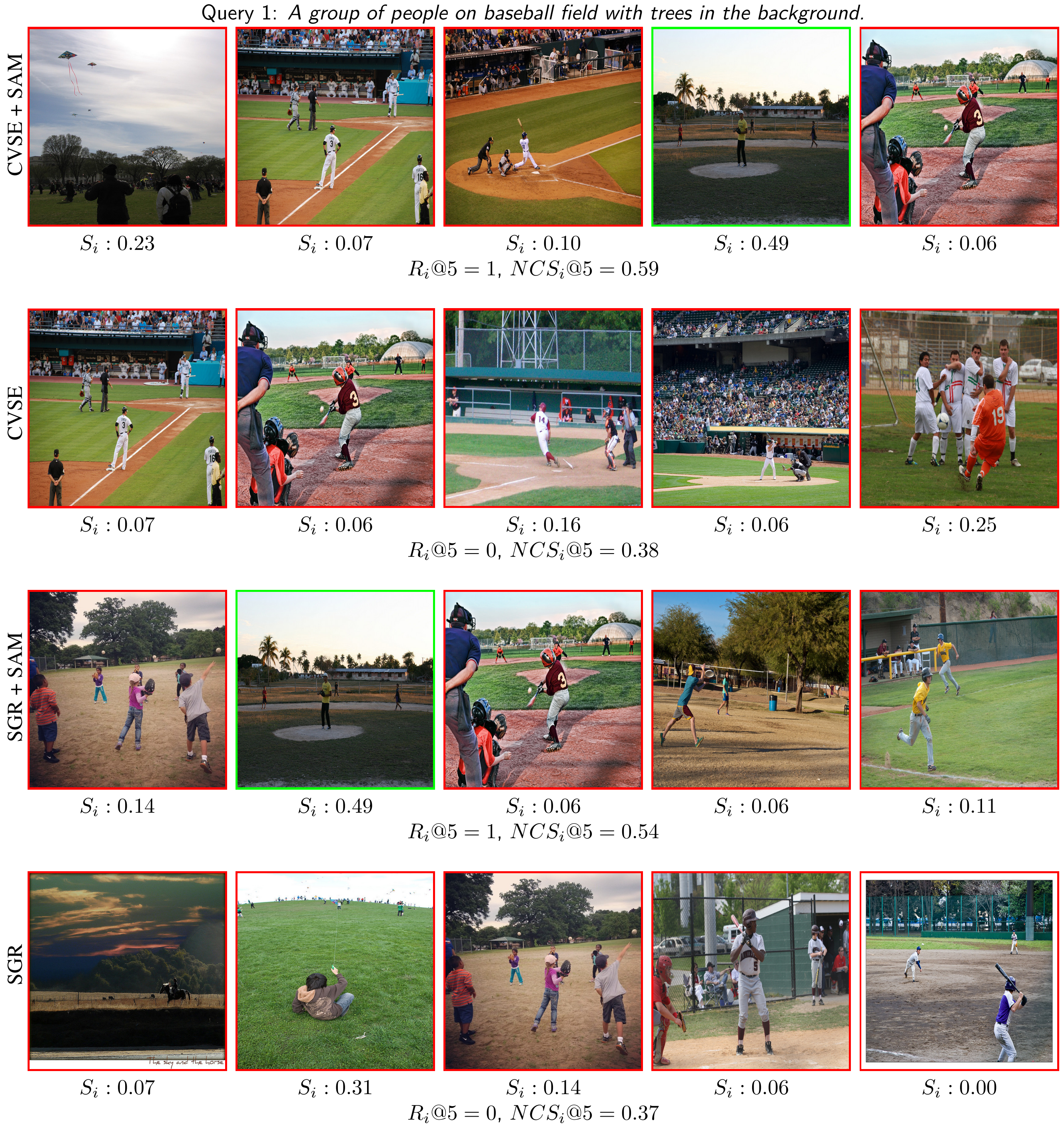

Figure 1: MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).

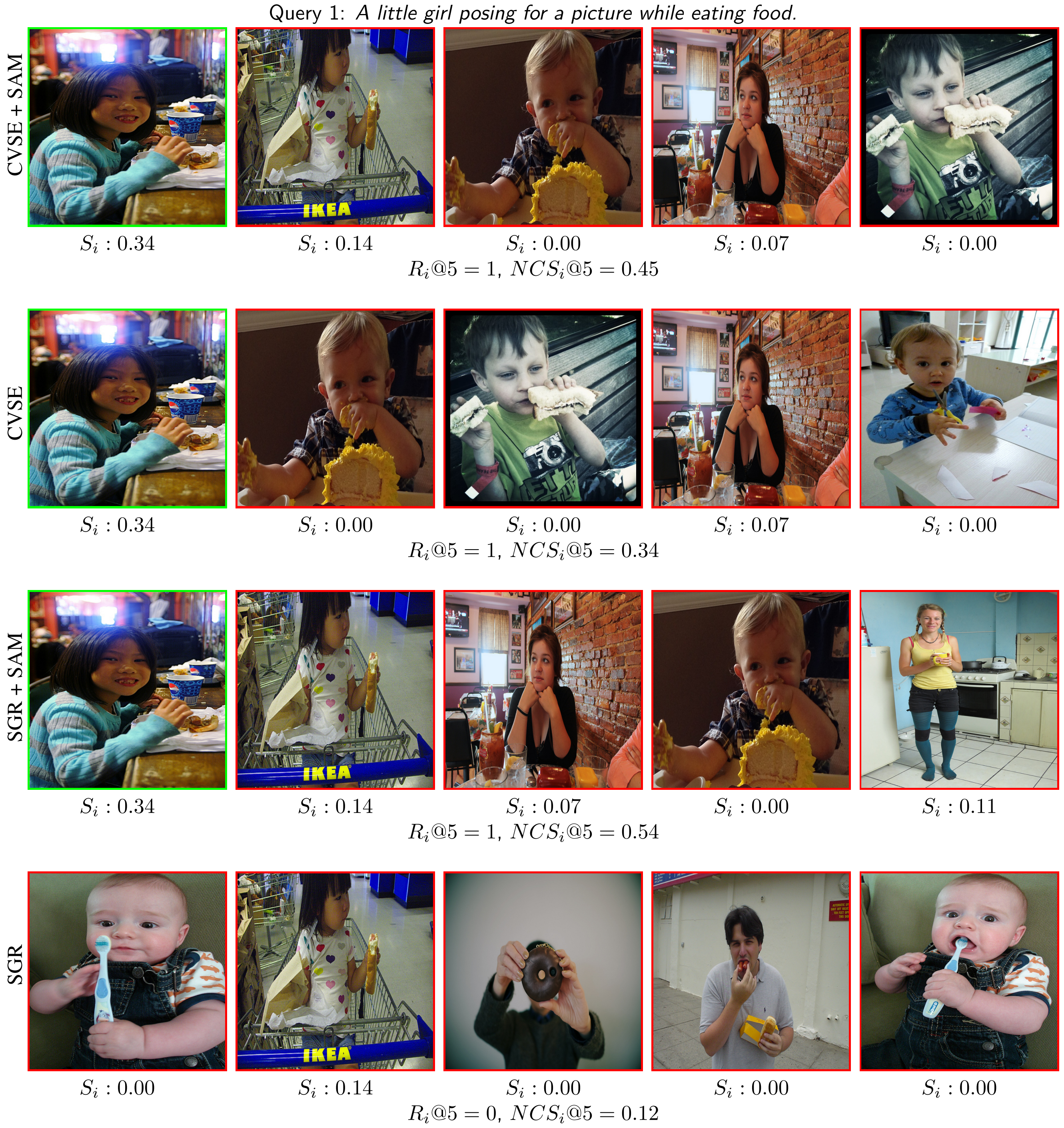

Figure 2: MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).

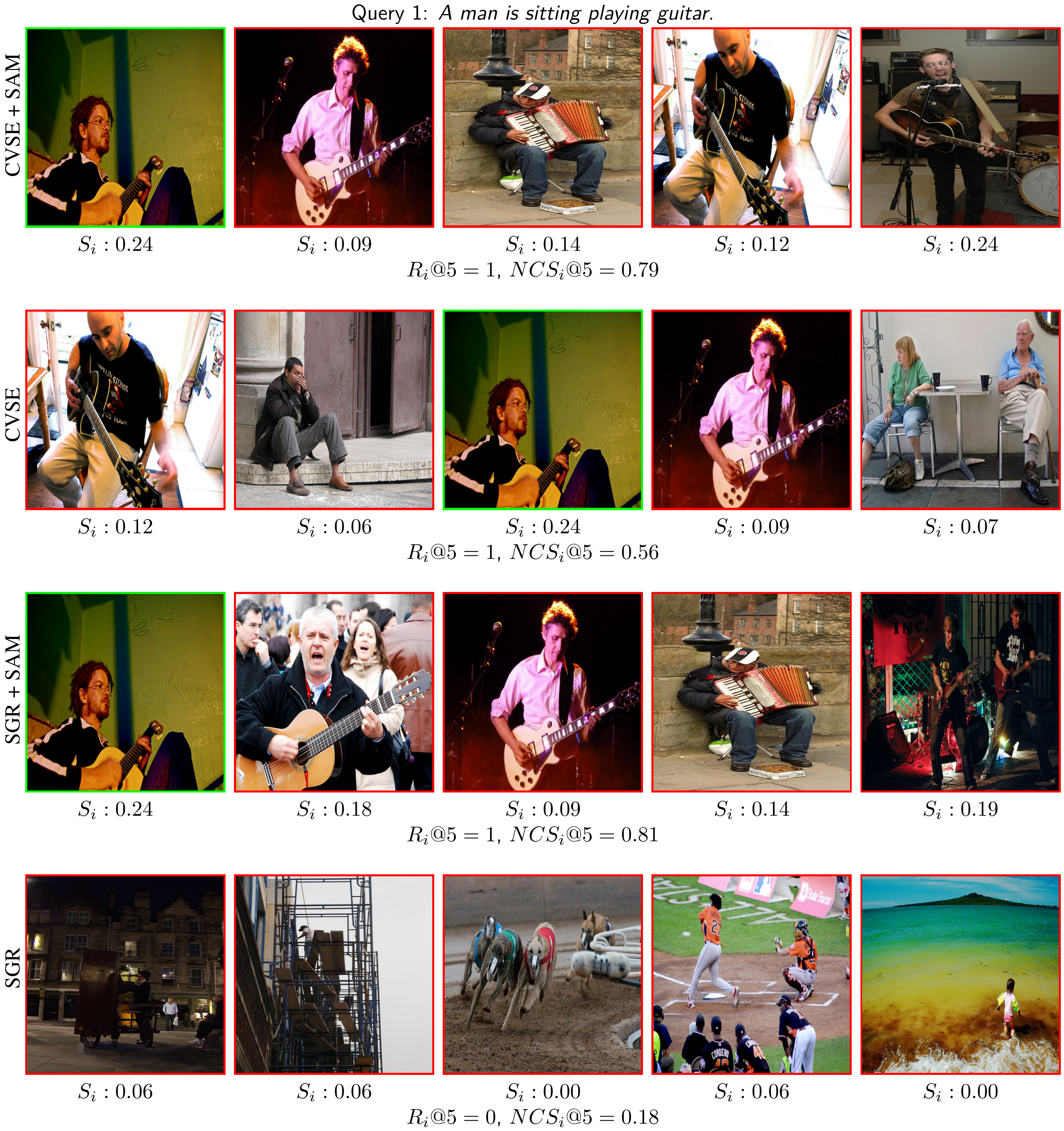

Figure 3: Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).

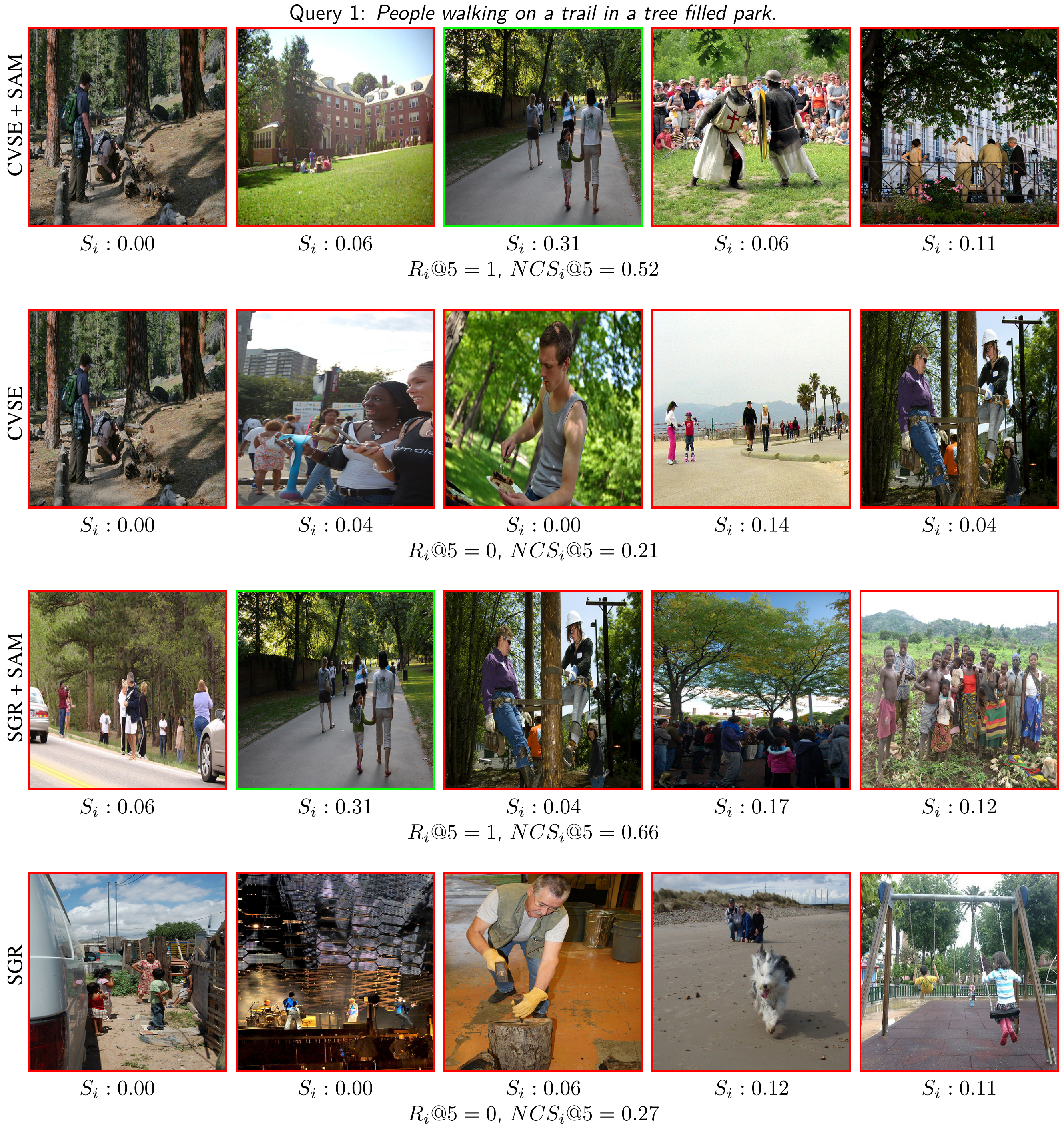

Figure 4: Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).

4 Insights on State-of-the-Art Retrieval

In this section, we extend the performance results of state-of-the-art ITM retrieval models. Similarly as in the main paper, the following models are evaluated: VSE++ [3], SCAN [6], VSRN [7], CVSE [9], SGR and SAF [2]. The figures comparing the models at different cut-off points (@k), evaluated with and without GT samples are shown in Table 7. The same behaviour described in the main document can be observed. As the number of relevant images increase, the Recall(R) depicts a constant score increase, while the Semantic Recall (SR) shows the opposite.

On one hand, state-of-the-art models tend to miss semantic relevant samples that are not labelled in the dataset as ground truth. On another hand, even though there is a significant performance gap between recently proposed (CVSE, SAF, SGR) and previous (VSE++, SCAN) models when they are evaluated in terms of Recall, this statement does not necessarily hold. The aforementioned effect is even more significant when models are evaluated on the non-GT annotations. It is important to mention that aside missing 4 annotated GT samples in the image-to-text scenario, the limitations of evaluating with Recall R^v are present at the moment of considering relevant samples. As an overview, these results suggest that there is still a considerable room for improvement in ITM pipelines.

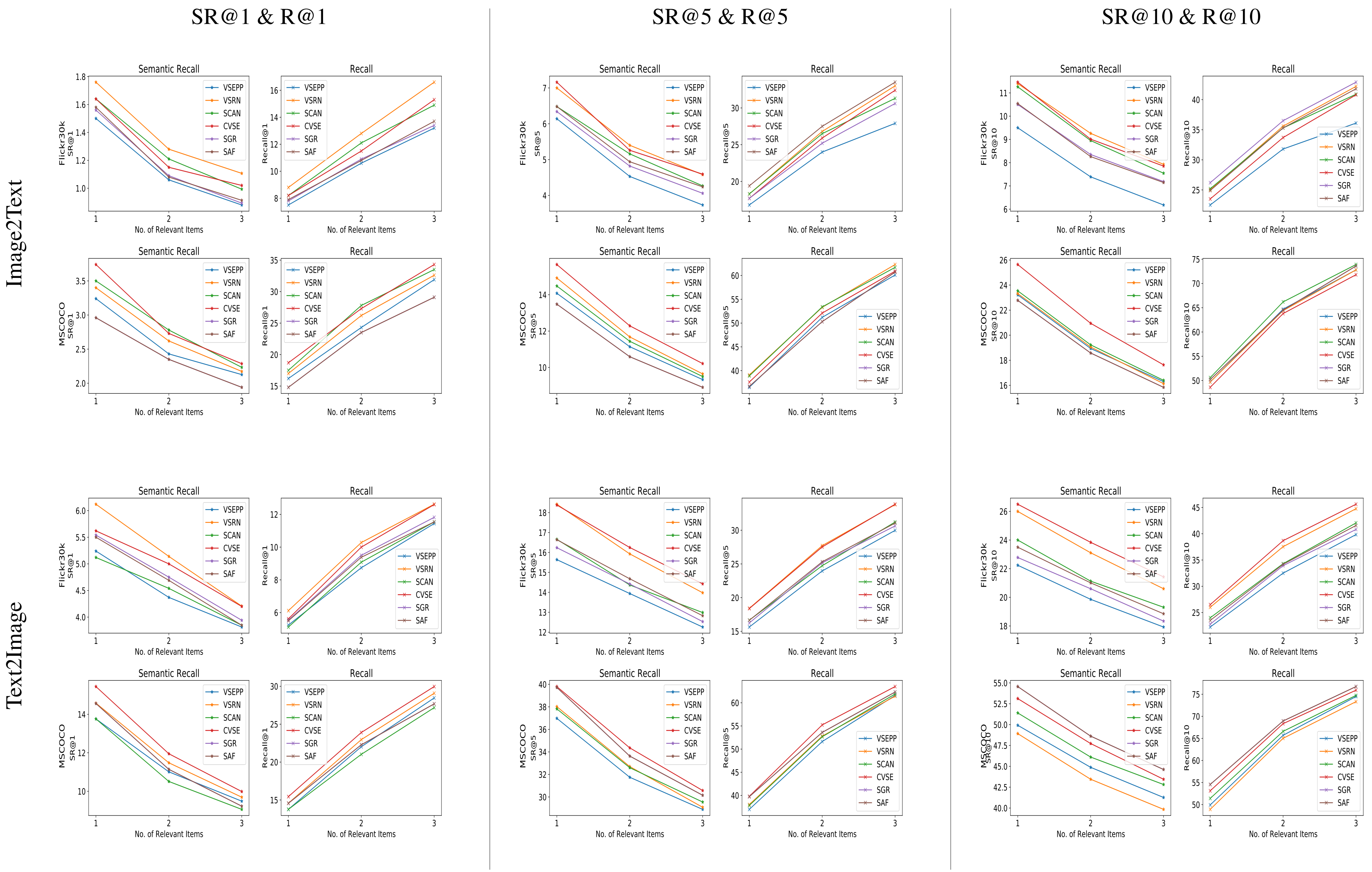

Table 7: Results evaluated with Recall(R) and the the proposed Semantic Recall(SR) for GT and Non-GT items. Each column represents the figures obtained at a cut-off point @1, @5 and @10 respectively. Best viewed in color.

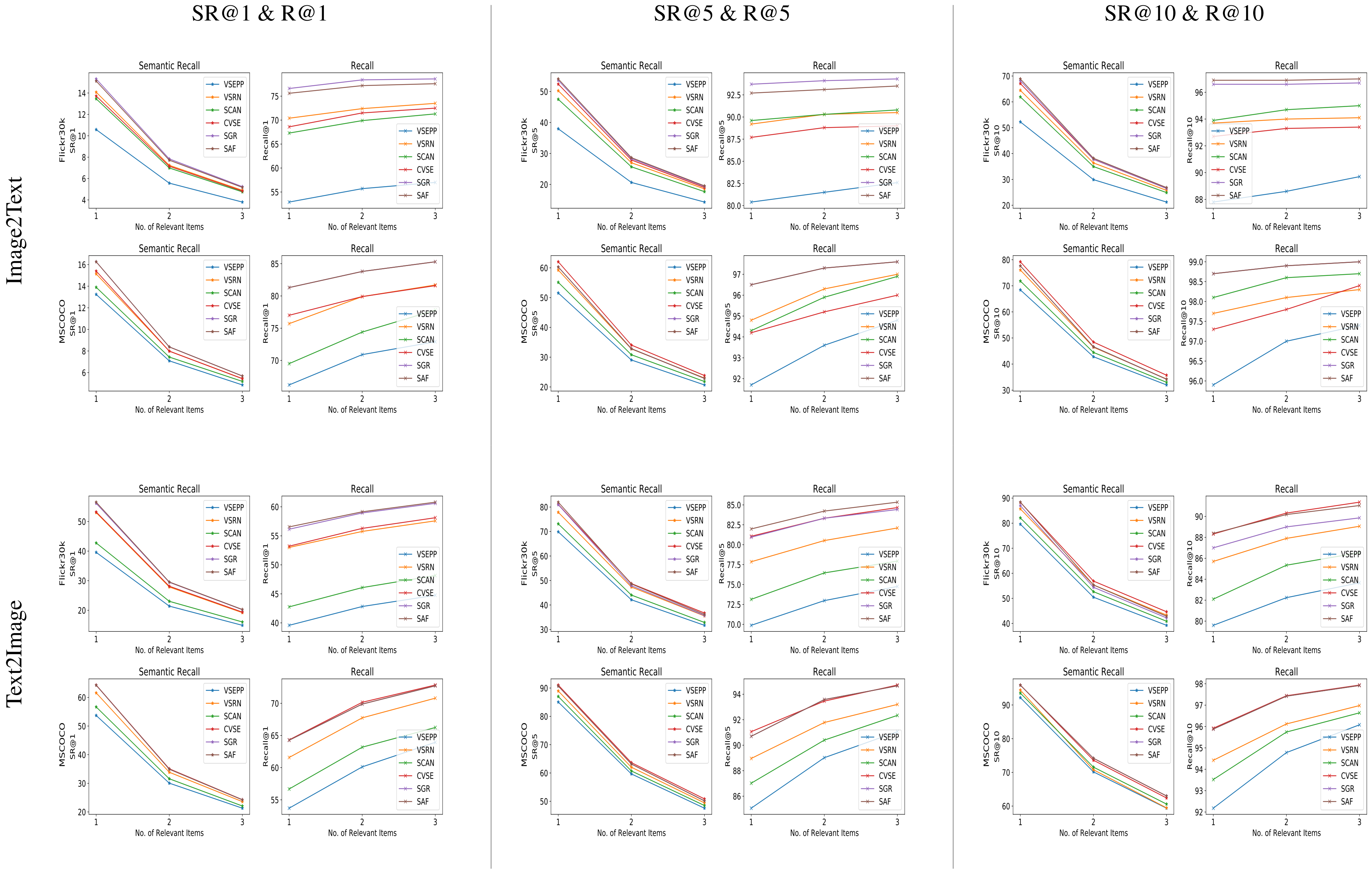

Table 8: Results evaluated with Recall(R) and the the proposed Semantic Recall(SR) for GT and Non-GT items. Each column represents the figures obtained at a cut-off point @1, @5 and @10 respectively. Best viewed in color.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017.

[2] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*, 2021.

[3] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. 2018.

[7] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

[9] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. *arXiv preprint arXiv:2007.08883*, 2020.