Compare and Reweight: Distinctive Image Captioning Using Similar Images Sets Supplementary material

The organization of the supplementary material is as follows. We include the ablation study of the CIDErBtw training strategies and the hyperparameter α_w in Section A. Then, the vocabulary frequency statistic is shown in Section B. We evaluate our model on the official online testing server in Section C. More details of user study, as well as the comparison with SOTA methods are introduced in Section D. In Section E, we present additional qualitative results.

A Ablation study

In this section, two ablation studies are performed. One is for the influence of training with different CIDErBtw strategies. The other is for the influence of different α_w settings.

A.1 Analysis of CIDErBtw training strategies

Here we use Transformer+SCST as the baseline model to do an ablation study in Table A.1. The effects of our three CIDErBtw training strategies are shown in each row after the baseline performance. The second row "+ CIDErBtw weighted XE loss" means that CIDErBtw weighted XE loss is used in the first training step, and the original SCST is used in the second training step. The third row "+ CIDErBtw weighted reward" indicates that the CIDErBtw weighted XE loss is used in the first training step, and the CIDErBtw weighted reward is used in the second training step. The fourth row "+ CIDErBtwReward" means combining three strategies of CIDErBtw weighted XE loss, CIDErBtw weighted reward, and CIDErBtw reward.

The CIDErBtw score of the models in these four conditions decreases sequentially from 86.68 to 82.74, and the retrieval metrics increase sequentially. This demonstrates that the three CIDErBtw training strategies can effectively improve the distinctiveness of the captioning model. In particular, weighted XE loss effectively increases the scores of other metrics, leading to an improvement of 2.26 in CIDEr and 1.0 in BLEU4. Analyzing the performance change after CIDErBtw reward, it can be seen that the results on accuracy related metrics are slightly worse, while distinctiveness related metrics are significantly better, *e.g.*, R@10 surges from 68.44 to 71.28. This shows that combining CIDErBtw in the reward can guide the model to generate words with better CIDErBtw score, thus enhancing the distinctness of captions within similar images set. The CIDEr score slightly decreases since the CIDErBtw reward encourages generating distinctive words.

Method Condition	CIDEr ↑	CIDErBtw ↓	BLEU3↑	BLEU4↑	METEOR↑	ROUGE-L↑	SPICE↑	R@1 ↑	R@5 ↑	R@10 ↑
Transformer+SCST	125.13	86.68	50.26	38.04	27.96	58.60	22.30	23.38	54.34	68.44
+ CIDErBtw weighted XE loss	127.39	84.38	**51.31**	**39.04**	29.02	59.16	22.93	24.60	54.62	68.72
+ CIDErBtw weighted reward	**128.11**	84.70	51.29	39.00	**29.12**	**59.24**	22.92	24.46	55.22	69.02
+ CIDErBtwReward	127.78	**82.74**	50.97	38.52	29.09	58.82	**22.96**	**26.46**	**57.98**	**71.28**

Table A.1: Ablation study of our three CIDErBtw training strategies on MSCOCO test split.

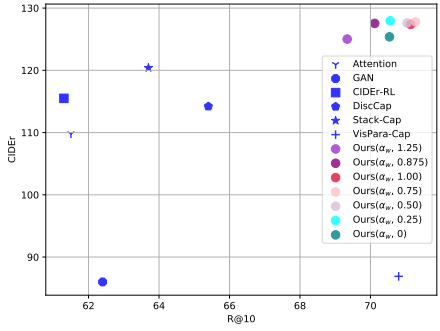


Figure A.1: The performance of different models considering CIDEr and R@10. Here we compare our method with other previous works, *i.e.*, Attention [9], GAN [2], CIDEr-RL [9], DiscCap [8], Stack-Cap [4], and VisPara-Cap [7].

A.2 Analysis of α_w

We use Transformer+SCST as baseline model, and use fixed α_r as 0.4 for CIDErBtw reward. So it means "Transformer+SCST+CIDErBtwReward" when α_w equals to 0, otherwise the models are trained with all our three strategies. Compared with "Transformer+SCST" in Table 1, **Our(α_w,0)** has good performance, especially for CIDErBtw and retrieval metrics. The training weight of the data on the parameters is refined due to reweighting loss and reward when α_w is not equal to 0, thus we get better performance. Figure A.1 shows that our method can obtain high CIDEr and R@10 at the same time.

A larger α_w means more effect of CIDErBtw weight to the training process. As can be seen from Table A.2, the model performs best in terms of distinctness when α_w equals to 0.75, while it performs best in terms of accuracy when α_w equals to 0.25. It reflects the model's trade-off in accuracy and distinctness among different α_w. With the increase of α_w, the training weight for low-frequency words gradually increases, which tends to decreases the CIDEr score. The distinctiveness related metrics are first improved by increasing α_w; however, too large α_w (greater than 0.875) is also not conducive to the learning of general language usage, which degrades the model performance. Generally speaking, our model performs well when α_w between 0.25 and 0.875, reflecting that our strategies are robust.

Method Condition	CIDEr ↑	CIDErBtw ↓	BLEU3↑	BLEU4↑	METEOR↑	ROUGE-L↑	SPICE↑	R@1 ↑	R@5 ↑	R@10 ↑
Ours (α_w, 1.25)	125.02	83.29	49.50	37.11	28.72	58.37	22.56	24.54	54.76	69.34
Ours (α_w, 1.00)	127.54	83.35	50.70	38.36	29.09	59.05	23.04	25.74	55.90	70.12
Ours (α_w, 0.875)	127.38	**82.48**	50.77	38.34	29.04	58.77	22.85	26.34	57.16	71.14
Ours (α_w, 0.75)	127.78	82.74	50.97	38.52	29.09	58.82	22.96	**26.46**	**57.98**	**71.28**
Ours (α_w, 0.50)	127.61	83.54	51.22	38.82	29.1	59.08	**23.09**	25.94	57.16	71.04
Ours (α_w, 0.25)	**127.96**	83.85	**51.33**	**38.94**	**29.12**	**59.13**	22.9	25.72	56.04	70.56
Ours (α_w, 0)	125.38	85.73	50.39	38.28	28.42	58.93	22.61	25.3	56.74	70.54

Table A.2: The performance of our model under different CIDErBtw weight parameter α_w. Our baseline model here is "Transformer+SCST+CIDErBtwReward", and "Our (α_w, x)" means set α_w as x.

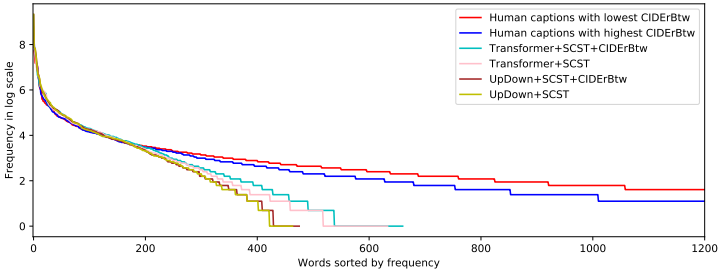


Figure B.1: Statistic of words frequency on test split. In ground truth captions, we pick the caption with lowest CIDErBtw score for each image as "human captions with lowest CIDErBtw", and the caption with highest CIDErBtw score as "human captions with highest CIDErBtw". We also compute the vocabulary frequency of generated captions for different models (*i.e.*, Transformer+SCST+CIDErBtw, Transformer+SCST, UpDown+SCST+CIDErBtw and UpDown+SCST).

B Vocabulary frequency statistics

We show the vocabulary frequency plots of different models in Figure B.1. Each curve counts word frequencies from 5,000 captions of test split (one caption for each image). We choose the ground truth caption with the lowest CIDErBtw for "human caption with lowest CIDErBtw" and vice versa. If a model uses diverse words, the plot should have a longer tail. We find that human captions with the lowest CIDErBtw contain 6,506 unique words, while those with the highest CIDErBtw only contain 3,681 unique words, indicating that ground truth captions with lower CIDErBtw are also more diverse. Captions generated by models in this figure use less than 700 unique words, indicating that there is still an obvious gap between machine generated captions and human ground truth captions. We can observe the impact of CIDErBtw from the perspective of vocabulary frequency. The models trained with CIDErBtw weight have a longer tail and higher frequency than models without CIDErBtw weight. It indicates that our training strategies also guides the model to generate more diverse captions.

Model	BLEU1		BLEU2		BLEU3		BLEU4		METEOR		ROUGE-L		CIDEr-D	
Metric	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [9]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116
LSTM-A [12]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
UpDown [1]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [6]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SGAE [10]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet [5]	81.0	95.0	65.8	89.6	51.4	81.3	**39.4**	**71.2**	**29.1**	**38.5**	58.9	**74.5**	126.9	129.6
GCN-LSTM+HIP [11]	**81.6**	**95.9**	**66.2**	**90.4**	**51.5**	**81.6**	39.3	71.0	28.8	38.1	**59.0**	74.1	**127.9**	**130.2**
Ours	80.2	94.6	64.8	88.6	50.2	79.8	38.3	69.4	28.6	37.8	58.3	73.4	123.8	126.0

Table C.1: Leaderboard of the top ranking published state-of-the-art image captioning models on the online MS-COCO test server with 5 reference captions (c5) and 40 reference captions (c40).

C Online Evaluation

We also evaluate our model on the official online testing server in Table C.1. We compare our best model in Table 1 (*i.e.*, Transformer+SCST+CIDErBtw) with other latest published captioning models. Note that we only use the single model on the online server, while other SOTA methods prefer to ensemble several models to obtain better results. Our model is designed to generate distinctive captions, and its accuracy is competitive.

D User Study

D.1 More details about User Study

We conduct two user studies to fairly evaluate the quality of the generated captions. The first user study, the image retrieval experiment, can assess the distinctiveness of captions. The task involves displaying the target image, a similar image with the same semantic meaning, and a generated caption describing the target image. The users are asked to choose the image that more closely matches the caption. The interface can be found in Figure D.1. In the second experiment, we compare two captions generated from a baseline model trained with and without CIDErBtw. In each trial, an image and two captions are displayed, and the user is asked to choose the better caption with respect to two criteria: distinctiveness and accuracy. The interface is shown in Figure D.2.

In each experiment, we randomly sample 50 similar images pair from the test split. Twenty users participated in the experiments. They are graduate students without prior experience on the image captioning task, in order to avoid biases.

D.2 Compare with SOTA Methods

We compare our method and other state-of-the-art distinctive methods [3,4,8] by user study. Two experiments are performed. Firstly, we performed image retrieval experiments with captions generated by these methods and report the accuracy in Table D.1. The image retrieval accuracy is consistent with other

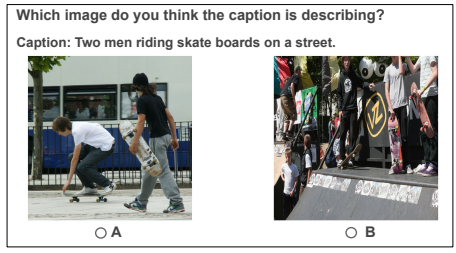


Figure D.1: User study interface for image retrieval experiment.

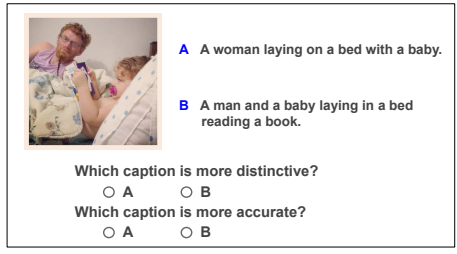


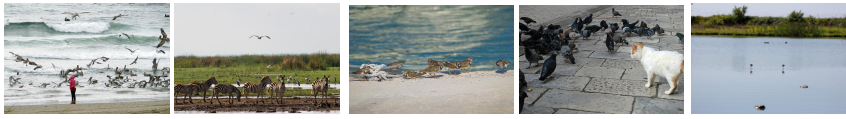
Figure D.2: User study interface for selecting caption experiment.

metrics reported in Table 1, where our method achieves the most distinctive result, higher than the second-best, CL-Cap, by a large margin (3.8% in retrieval accuracy).

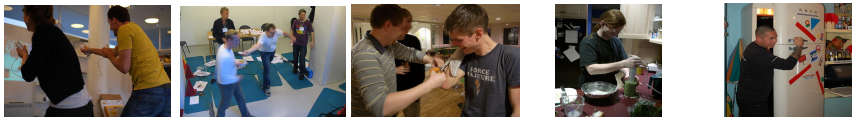
In the second experiment, we compare captions generated by four methods, Stack-Cap [4], DiscCap [8], CL-Cap [3], and our model Transformer trained with SCST and CIDErBtw weight (denoted as TF+CIDErBtw). In each trial, an image and four captions are displayed, and the user is asked to rate these captions from the 1-5 scale (higher is better) with respect to two criteria, distinctiveness and accuracy. In this experiment, we randomly sampled 50 images from the test split and eight users are invited to rate the captions. The results are shown in Table D.1. Our method achieves higher scores in both distinctiveness and accuracy, which is consistent with the automatic evaluation reported in Table 1.

Method	image retrieval	distinctiveness	accuracy
Stack-Cap [4]	72.6%	3.34	3.33
DiscCap [8]	73.9%	3.37	3.41
CL-Cap [3]	75.8%	3.41	3.12
TF+CIDErBtw(ours)	**79.6%** *	**3.57****	**3.45**

Table D.1: User study results on image retrieval and caption rating. Compared with SOTA methods, our models achieves higher image retrieval accuracy and rating scores (in statistical significant test with all baselines, * indicates 2-sample z-test on proportions with $p<0.05$, ** indicates paired t-test with $p<0.05$).



Baseline:	(116.0) A group of people flying kites on the beach.	(139.6) A group of zebras and zebras standing in a field.	(85.1) A group of birds standing on the beach.	(99.6) A flock of birds walking on a city street.	(69.8) A group of birds standing in the water.
Ours:	(65.5) A woman standing on the beach with a flock of birds.	(88.1) A herd of zebras and birds standing in a field.	(69.0) A group of birds standing on the beach in front of water.	(89.4) A cat and a flock of birds on a sidewalk.	(60.1) A group of ducks swimming in the water.



Baseline:	(119.2) Two people standing in a room.	(78.8) A group of people playing video game in a room.	(78.0) A man holding a pair of scissors in his hand.	(24.9) A woman is preparing food in a kitchen	(155.4) A man holding a refrigerator in front of a.
Ours:	(95.6) Two people playing a video game in a living room.	(76.7) Four men playing a video game in a room.	(41.8) Four men cutting paper with scissors in a room	(13.5) A man standing in a kitchen with a blender	(140.9) A man standing in front of a white refrigerator.

Baseline:	(56.6) A plate of food on a table.	(80.5) A pizza with cheese and sitting on a table.	(73.1) A plate of food on a table.	(69.8) A plate of food with and on a table.	(73.8) A plate of food on a table.
Ours:	(41.9) A plate of food with a fork and a glass on a table.	(96.7) A pizza sitting on a table with a glass of wine.	(50.8) A plate of food with meat and vegetables on a table.	(43.0) Plates of food with meat and bread on a table	(58.7) A table with plates of food and drinks on it

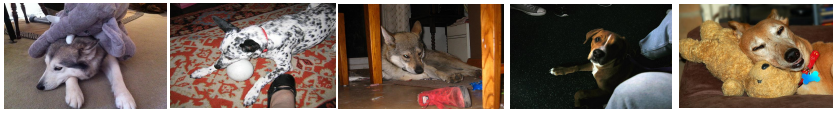
Baseline:	(38.4) A man in a suit and sitting.	(53.3) A man sitting on the beach with a frisbee.	(88.2) A man in a suit.	(78.3) A man in a hat and wearing a hat.	(84.2) A man holding a cell phone in his hand.
Ours:	(15.2) A man sitting in front of a table with a clock.	(39.2) A man in a hat holding a frisbee in a parking lot.	(60.2) A black and white photo of a man in a suit.	(50.5) A man wearing a black hat and a tie.	(52.2) A man wearing a hat and tie holding a cell phone.

Baseline:	(134.5) A boy holding a teddy bear.	(1.1804) A black dog laying on a couch.	(66.7) A teddy bear sitting in the back of a chair.	(61.5) A woman sitting on a couch with a dog in the.	(92.4) A dog laying on a couch.
Ours:	(95.5) A young boy holding a teddy bear in front of a wall.	(1.012) A black dog laying on a bed in front of a book shelf.	(59.1) A teddy bear sitting in a basket in the snow.	(39.4) A girl sitting in a chair with a bunch of dogs.	(79.6) A dog laying on a bed with a teddy bear.

Figure D.3: More qualitative results.

Baseline:	(25.7) A plate of food with bananas and on a table.	(67.9) A piece of cake on a plate with a fork.	(21.0) A pizza with cheese and sitting on a plate.	(66.8) A bunch of bananas and sitting on a table.	(114.8) A pizza sitting on a plate.
Ours:	(21.9) A piece of food with bananas and strawberries on a white plate.	(56.4) A piece of cake on a plate with an apple and a fork.	(30.7) A piece of pizza on a plate on a table.	(37.8) A banana and pineaple sitting on a counter.	(16.1) A pizza with cheese and toppings sitting on a white plate.

Baseline:	(77.1) A group of people with umbrellas.	(103.8) A man walking in the rain with an umbrella.	(108.6) A group of people walking in the rain with umbrellas.	(106.4) A group of people walking in the rain with umbrellas.	(180.9) A man walking in the rain with an umbrella.
Ours:	(54.3) A group of people walking down a street with an umbrella.	(24.9) A person holding a colorful umbrella standing in a park.	(113.1) Four people walking in the rain with umbrellas.	(52.0) A black and white photo of a group of people walking with umbrellas.	(131.3) A man walking in the rain with an umbrella in front of a building.



Baseline:	(142.6) A dog laying on the floor with a teddy bear.	(1.1146) A black and white dog laying on a couch.	(127.1) A dog laying on the floor.	(150.4) A dog sitting on the floor.	(85.8) Two brown teddy bears sitting on a couch.
Ours:	(127.7) A white dog laying on the floor with a stuffed elephant.	(0.8343) A dog playing with a ball on the floor.	(81.6) A dog laying on the floor next to a pair of shoes.	(92.9) A dog laying on the floor next to a person.	(40.8) A brown dog laying on a couch with a teddy bear.

Baseline:	(52.1) A person holding a hot dog in their hand.	(90.9) A woman standing next to a giraffe.	(99.6) A woman standing next to a giraffe.	(95.0) A man and a woman standing next to a giraffe.	(84.0) A man standing next to a giraffe.
Ours:	(7.5) A person holding an apple in front of a cow.	(47.2) A woman taking a picture with a giraffe at a zoo.	(0.9950) A woman feeding a giraffe at a zoo.	(78.9) A man feeding a giraffe on a grass field.	(10.3) A person working in front of a giraffe and a tree.

Baseline:	(47.3) A red bus parked in a parking lot.	(68.7) A wooden bench sitting in front of a park.	(93.0) A red and white boat in the water.	(65.1) A group of boats on a body of water.	(37.9) A boat in the water on a body of water.
Ours:	(30.3) A red bus is parked in a park with trees.	(41.4) A white bench sitting in front of a garden.	(85.1) A red and white ship sailing on the sea.	(58.8) A goup of boat is in the water near a hill.	(25.6) A boat is on a river in front of trees.

Figure D.4: More qualitative results.

E More Qualitative Results

We show more qualitative results in Figures D.3 and D.4. Each row is a similar images set.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. pp. 6077–6086 (2018) 4
2. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a Conditional GAN. In: ICCV (2017) 2
3. Dai, B., Lin, D.: Contrastive learning for image captioning. In: NeurIPS (2017) 4, 5
4. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. In: AAAI (2018) 2, 4, 5
5. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: ICCV. pp. 4634–4643 (2019) 4
6. Jiang, W., Ma, L., Jiang, Y.G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: ECCV. pp. 499–515 (2018) 4
7. Liu, L., Tang, J., Wan, X., Guo, Z.: Generating diverse and descriptive image captions using visual paraphrases. In: CVPR. pp. 4240–4249 (2019) 2
8. Luo, R., Price, B., Cohen, S., Shakhnarovich, G.: Discriminability objective for training descriptive captions. In: CVPR. pp. 6964–6974 (2018) 2, 4, 5
9. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR. pp. 7008–7024 (2017) 2, 4
10. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR. pp. 10685–10694 (2019) 4
11. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. In: ICCV. pp. 2621–2629 (2019) 4
12. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: ICCV. pp. 4894–4902 (2017) 4