FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation Supplementary Material

Sijin Wang[1,2], Ziwei Yao[1,2], Ruiping Wang[1,2], Zhongqin Wu[3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Tomorrow Advancing Life Education Group, Beijing, 100080, China
{sijin.wang, ziwei.yao}@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn, wuzhongqin@tal.com

In this supplementary document, we introduce in detail the datasets (HM-MSCOCO and MSCOCO Entities [5]), the score distribution of our model, evaluation cases by different automatic metrics, visualizing cases of grounding analysis, extended quantitative experiments and extra discussion on some concerned issues.

1. Datasets

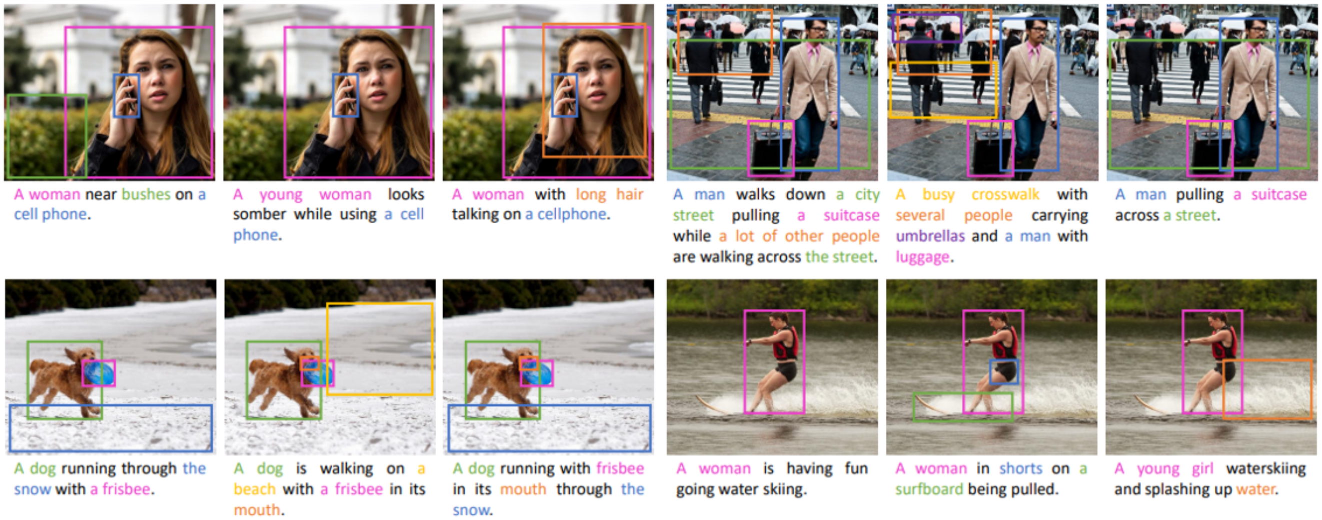
HM-MSCOCO. To exam the consistency of different automatic metrics, we collect a new dataset named HM-MSCOCO. HM-MSCOCO contains $5,000$ images coming from MS COCO test split [9]. Each image has 3 human-machine candidate caption pairs and 4 reference captions, in which the human candidate caption is randomly picked from one of the 5 human-labeled ground-truth captions of this image, and the reference captions are the remaining 4 ground-truth captions. Therefore, human candidate captions can be considered to have high fidelity and adequacy. Three machine candidate captions are derived from [6], which are generated by three remarkable image captioning models (NeuralTalk [9], Show and tell [15], and Show attend and tell [16]). Though PASCAL-50S [14] also contains $1,000$ HM candidate pairs, which is of a small scale and easy to identify the human caption in each pair, we use the more difficult HM-MSCOCO to investigate the consistency of different automatic metrics.

Table 2 in our main paper shows the results of different metrics on HM-MSCOCO. The accuracy is defined as the percentage of pairs whose human caption gets a higher score than the machine-generated one. The average score is computed among all the $15,000$ human/machine captions, respectively, using the given metric (note that the scores of different metrics are incomparable). Rule-based metrics show a low accuracy; even some of them undesirably give the machine higher scores than human (e.g. BLEU-1, BLEU-4, and ROUGE-L). Though the human written candidate for each image is randomly picked from its five ground-truth captions in MS COCO, it cannot obtain a higher score by rule-based metrics, which from another view again reveals the deficiencies of the reference-based evaluation mode. In contrast, the results also show the high consistency of FAIEr with the human judgment, which demonstrates the evaluation strategy of our metric can more accurately reveal the human evaluation intentions.

MSCOCO Entities[1]**.** It [5] associates the noun chunks in the caption with the visual object regions in the target image on the MS COCO dataset (examples are shown in Fig.1 of this supplementary document). Each visual region has a category label, such as “people” and “dog”. We collect the images belonging to our test split and their visual regions and corresponding noun chunks from MSCOCO Entities. Totally, there are $5,000$ images with $23,940$ captions, $51,992$ noun chunks, and $50,166$ visual regions (each visual region can correspond to several noun chunks in a caption).

To evaluate our FAIEr, we keep the annotations where the category label is exactly equal to the last word in the noun chunk. (*e.g*. the region “people” with noun chunk “a group of” will be eliminated and the region “ men” with noun chunk “four men” will be preserved.). We set the GT (ground-truth) word of each visual region as the last word of its corresponding noun chunk, and words not in this noun chunk are non-GT words. Next, we select the images with 5 captions, and each caption contains at least one noun chunk labeled with visual regions. After filtering, $3,098$ images (each with 5 captions), $28,933$ visual regions and $29,207$ noun chunks constitute our MSCOCO Entities test split. For each image, we randomly select one caption as the candidate caption, and the remaining four are reference captions. We first compute the matching scores between the image region and its GT/non-GT words in the reference captions,

[1] https://github.com/aimagelab/show-control-and-tell


A woman near bushes on a cell phone.

A young woman looks somber while using a cell phone.

A woman with long hair talking on a cellphone.

A man walks down a city street pulling a suitcase while a lot of other people are walking across the street.

A busy crosswalk with several people carrying umbrellas and a man with luggage.

A man pulling a suitcase across a street.

A dog running through the snow with a frisbee.

A dog is walking on a beach with a frisbee in its mouth.

A dog running with frisbee in its mouth through the snow.

A woman is having fun going water skiing.

A woman in shorts on a surfboard being pulled.

A young girl waterskiing and splashing up water.

Figure 1. Examples of MSCOCO Entities [5].

and then the scores are averaged by the number of GT/non-GT words, respectively. For the candidate captions, we similarly calculate the average matching scores of union objects and the GT/non-GT words. The results are illustrated in Table 6 in our main paper, which can reveal the effectiveness of our matching module and attention fusion module.

2. Range of scores

We test the full model **FAIEr-4 ref** on more than 20 million samples from MS COCO, Flickr8K, Composite, PASCAL-50S and Nocaps, and find that overwhelming majority of the scores are distributed between -2 and 6, which can be seen as the empirical output range of our model. Fig.2 shows the score distribution over all samples. Besides, the score distribution is related to the margin in the loss function. Specifically, the range of results expands as the chosen margin parameter m increases. This parameter here is set to 0.2, following the same settings in our main paper.

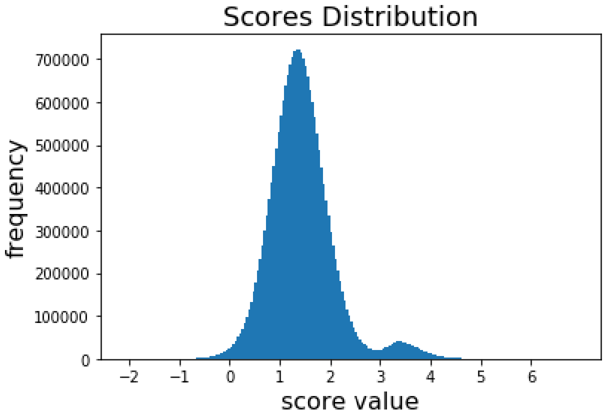


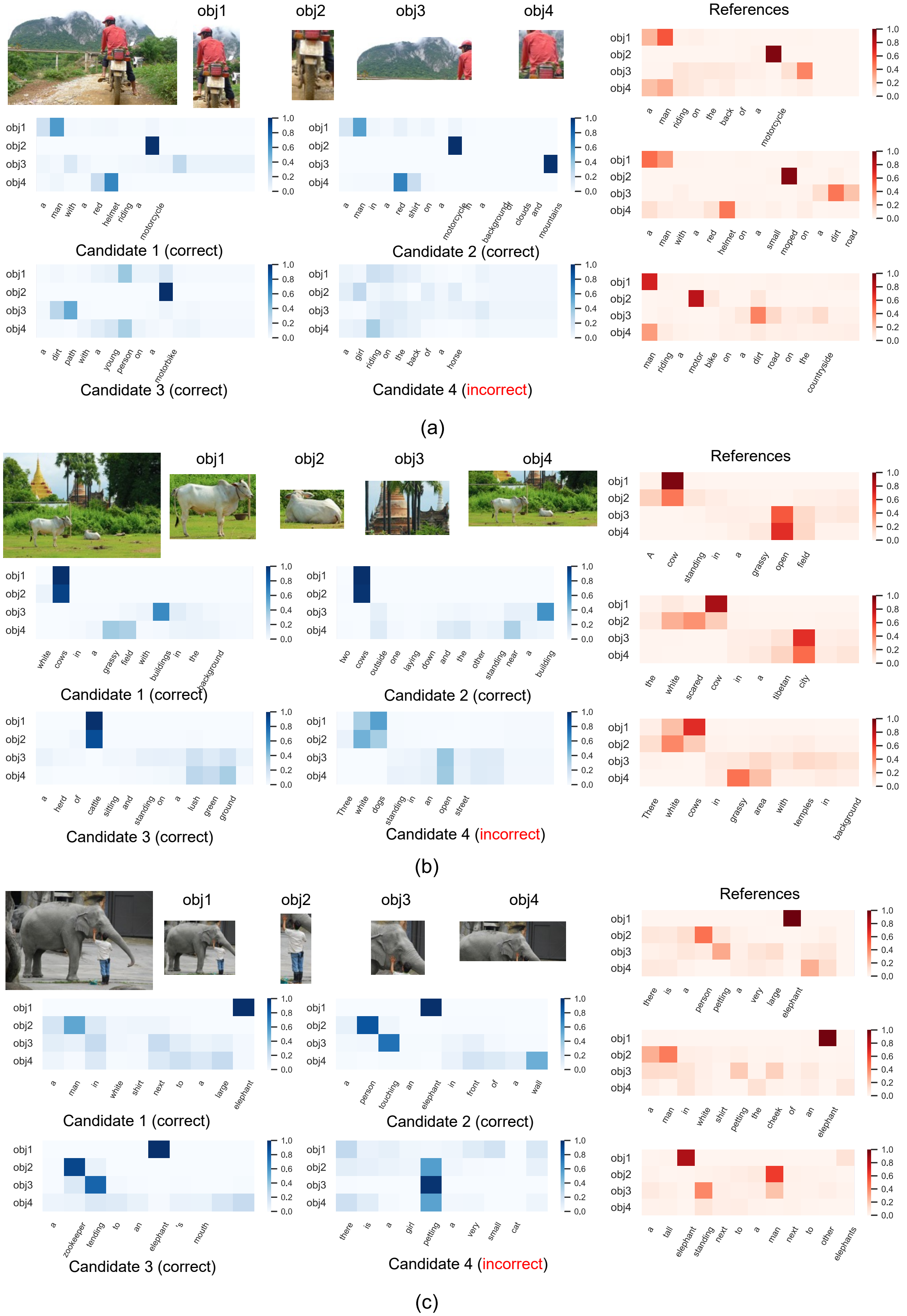
Figure 2. Score distribution over 20,862,680 different samples. The margin parameter m of loss funtion is 0.2.

3. Evaluation examples

In this section, we show more evaluation examples to expand Fig.5 and Fig.6 in our main paper.

Fig.4 in this supplementary document shows several evaluation cases, in which the images and reference captions come from MS COCO [11] test split. Each case has five candidate captions: the first three are correct captions from the original dataset with high fidelity and adequacy, while the latter two are captions we build to measure evaluation metrics from different aspects. The 4th one is incorrect and contains words similar to reference captions, and the 5th one is correct while with low adequacy. From the results in this figure, it shows that FAIEr can give the faithful and adequate candidate high scores. It also can give higher scores to the correct candidates of low adequacy than the incorrect ones, which is more fair.

Fig.3 in this supplementary document illustrates the grounding analysis of more cases. The matching heatmaps between the union/visual objects and the candidate/reference words demonstrate the effectiveness of our FAIEr model. Note that in this figure, to allow equal display size of the heatmaps for both long and short candidates, the heatmaps for the longest sentence do not display the "." and ending characters "<end>", and for the short sentences, we have hidden the "." and "<end>" on the axis after the last word but display their values.



(a)

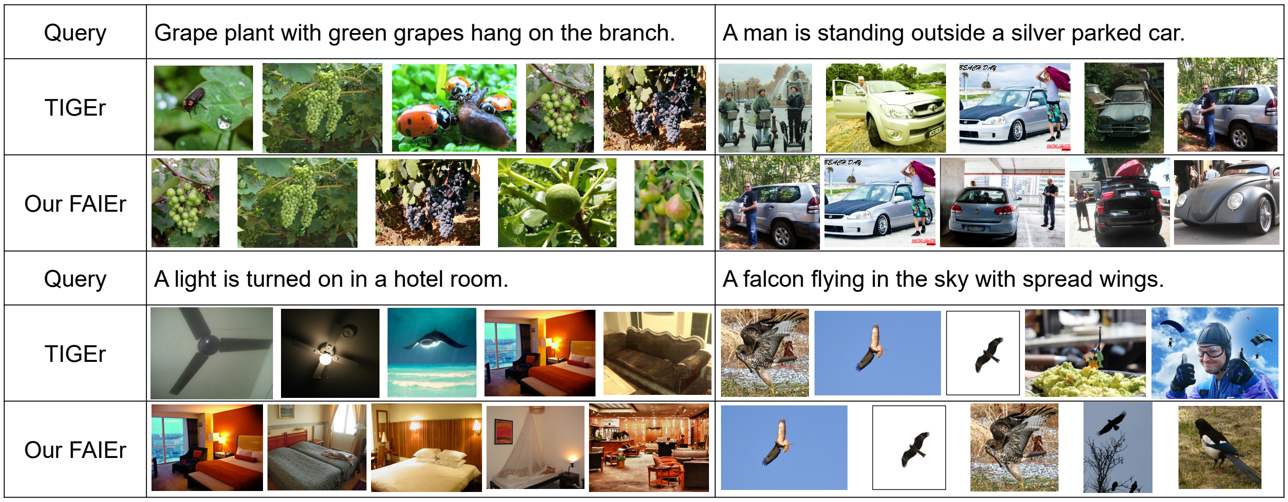
(b)

(c)

Figure 3. Visualizing object-level evaluation of FAIEr\rel.

Image	References	Candidates	w/o image				with image	
			BLEU-4	METEOR	ROUGE-L	SPICE	FAIEr\rel	FAIEr
	• A man riding on the back of a motorcycle. • A man with a red helmet on a small moped on a dirt road. • Man riding a motor bike on a dirt road on the countryside.	A man with a red helmet riding a motorcycle.	0.707107	0.2914	0.6292	0.4545	3.95	3.58
		A man in a red shirt on a motorcycle in a background of clouds and mountains.	5.36E-09	0.2357	0.5002	0.1429	3.60	3.61
		A dirt path with a young person on a motorbike.	7.73E-09	0.2333	0.4045	0.0870	3.66	3.32
		A girl riding on the back of a horse.	0.610474	0.3449	0.7778	0.1000	1.98	1.22
		A verdant area with a bridge and a background of clouds and mountains.	4.13E-09	0.0659	0.2943	0.0000	2.26	2.23
	• A cow standing in a grassy open field. • The white scared cow in a Tibetan city. • There white cows in grassy area with temples in background.	White cows in a grassy field with buildings in the background.	5.64E-05	0.3537	0.6724	0.4800	3.91	3.70
		Two cows outside one laying down and the other standing near a building.	6.95E-13	0.0827	0.1990	0.1600	2.77	2.73
		A herd of cattle sitting and standing on a lush green ground.	7.09E-13	0.1417	0.3112	0.0870	2.75	2.72
		Three white dogs standing in an open street.	6.99E-09	0.1709	0.3750	0.0870	1.97	1.88
		Lush trees in front of the buildings.	1.08E-12	0.0535	0.1317	0.0000	2.85	2.53
	• A person with a basketball stands in front of a goal. • A basket ball player is posing in front of a basket. • A basketball player holds a basketball for a picture.	A basketball player stands in front of a basket holding a ball.	0.558395	0.3507	0.6135	0.3636	4.45	4.07
		A young man in a green jersey is holding a ball in front of the wall.	2.46E-05	0.1983	0.3832	0.0800	3.60	3.77
		A uniformed boy is holding a basketball with his back to the hoop.	4.13E-09	0.1623	0.2820	0.1667	4.02	3.15
		A dog with a ball stands in front of a door.	0.429694	0.2896	0.7273	0.0952	2.39	2.15
		There is a backboard on the wall.	9.33E-13	0.0703	0.2137	0.0000	2.75	2.52
	• There is a person petting a very large elephant. • A man in white shirt petting the cheek of an elephant. • A tall elephant standing next to a man next to other elephants.	A man in white shirt next to a large elephant.	0.57735	0.3176	0.5666	0.5217	3.99	3.69
		A person touching an elephant in front of a wall.	7.26E-09	0.1811	0.3188	0.0870	3.73	3.27
		A zookeeper tending to an elephant's mouth.	6.16E-09	0.1011	0.3070	0.0952	3.47	3.22
		There is a girl petting a very small cat.	6.31E-05	0.2728	0.6667	0.0000	2.00	2.21
		A man wearing blue pants and with shirt.	5.74E-09	0.1405	0.3070	0.1905	2.12	2.24
	• A little girl in the grass wearing sunglasses holding a frisbee. • A young girl is holding a frisbee in the grass. • Young girl in sunglasses standing in a lawn, holding a frisbee.	A little girl standing in the grass holding a frisbee.	8.03E-05	0.4012	0.8498	0.0952	4.26	4.05
		A girl in blue shirt and shorts holding a frisbee in front of a fence.	0.203334	0.2749	0.4980	0.1600	3.72	3.32
		A little girl with a red frisbee on a lush green field.	3.03E-05	0.2004	0.4382	0.2609	3.94	3.89
		A little dog in the pool wearing collar holding a frisbee.	4.48E-05	0.3057	0.7273	0.0952	2.84	2.06
		Wildflowers grow in the grass in front of the fence.	3.55E-05	0.1296	0.3000	0.1000	2.55	2.26

Figure 4. Evaluation examples of different metrics.



Grape plant with green grapes hang on the branch.

A man is standing outside a silver parked car.

A light is turned on in a hotel room.

A falcon flying in the sky with spread wings.

Figure 5. Image retrieval examples on Nocaps.

4. Extended quantitative experiments

4.1. Model-level correlation

To validate the model-level human correlation of FAIEr, we conduct experiments that give scores to captions generated by different image captioning models on the same dataset and then compare them with human judgements. We calculate FAIEr 4-ref scores for three typical models, NeuralTalk [9], Show&Tell [15] and Up-Down [4]. As shown in Table 1, the scores increase with the development of image captioning methods, which is consistent with human's evaluation.

Table 1. FAIEr scores of three typical image captioning models' results on MS COCO.

NeuralTalk	Show&Tell	Up-Down
2.926	2.964	3.268

4.2. Extended experiments on Composite Dataset

In this section, we extend the human correlation experiments on Composite Dataset for comparison with VIFIDEL. We tried but failed to reproduce the results reported in their original paper, because some critical data, e.g., word vector, is missing in the released code. In our main paper, the utilizations of Composite Dataset of VIFIDEL and our FAIEr are different. Following SPICE and TIGEr, we use the full set containing captions from three datasets, MS COCO, Flickr8k and Flickr30k, while VIFIDEL only uses system-generated captions in MS COCO part. VIFIDEL evaluates the Spearman's correlation between the automated metrics and human judgments regarding both relevance and thoroughness on Composite Dataset, while FAIEr assesses three different correlation coefficients regarding relevance in the main paper. Therefore, for further comparison, we test FAIEr and other rule-based methods under VIFIDEL's experimental settings. As displayed in Table 2, "*VIFIDEL" means results are copied from their original paper [12], and FAIEr performs best among all metrics.

Table 2. Comparisons of Spearman's correlation on Composite Dataset under VIFIDEL's experimental settings.

Method	5Refs		1Ref	
	Relevance	Thoroughness	Relevance	Thoroughness
BLEU-1	0.28	0.25	0.24	0.20
BLEU-4	0.26	0.24	0.24	0.20
METEOR	0.31	0.27	0.26	0.22
ROUGE-L	0.28	0.26	0.25	0.21
CIDEr	0.33	0.29	0.28	0.23
SPICE	0.36	0.33	0.28	0.26
*VIFIDEL[12]	0.30	0.27	0.29	0.27
FAIEr-4 ref	**0.49**	**0.43**	**0.45**	**0.40**

4.3. Extended experiments of the reference number

In this section, we extend the experiments in Fig.4 of the main paper. Some candidates in Composite Dataset [1] come from one of the five human-labeled reference captions of the target image. The test setting of Fig.4 in the main paper does not remove the reference that is the same as the candidate, so the test cases where the candidate is highly similar to the reference often occur, especially when testing with increasing number of references. As shown in the Fig.4 of the main paper, these cases will favor the rule-based metrics (BLEU [13], ROUGE-L [10], METEOR [7], CIDEr [14], and SPICE [3]) and lead them to exceed TIGEr [8], when testing with more references.

Here we provide further investigation of another setting that was similarly adopted in TIGEr [8] to alleviate the effect of the above situation. Specifically, we remove the repetitive references from Composite Dataset and test different automatic metrics using different numbers of reference captions, which is displayed in Fig.6 here. In the new setting, FAIEr also outperforms other metics in almost all cases and other similar conclusions to Fig.4 of the main paper still hold. Comparing Fig.6 in the supplementary document and Fig.4 in the main paper, we can find that no matter whether the candidate is highly similar to the references or not, learning based methods FAIEr and TIGEr both perform very consistently; while the rule-based metrics perform obviously better in the cases where the candidate is highly similar to the references, which again demonstrates the advantages of the learning based metrics.

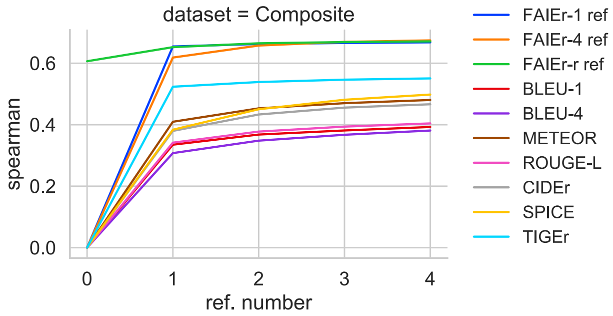


Figure 6. Testing metrics using different numbers of reference captions on Composite (the setting is different from Fig.4 in the main paper).

5. Extra discussion

In this section, we provide more discussion about our method as follows.

1) Generalization on novel visual concepts.

Generalization on new categories is a common challenge for learning-based methods using class-limited object detectors, which is a key problem closely related to zero-shot recognition, basically considering the relevance between seen and unseen categories to solve this issue. As for FAIEr, if the novel concept is semantically/visually similar to a known category, the detector can probably ground it to a similar region.

As shown in Section 4.5 of the main paper, we conduct experiments on a subset of the validation set of Nocaps Dataset to validate the generalization ability of FAIEr. Nocaps divides testing images into three splits, in-domain, near-domain and out-of-domain, according to whether objects in the image are seen during training. Except for testing on the full set as in the main paper, we additionally test on its out-of-domain split(R@1/5/10): **Image-to-Text: [0.965/1.000/1.000] ; Text-to-Image: [0.795/0.932/0.958]**. Compared to that in Table 6 of the main paper, results of Image-to-Text are nearly the same, while performance of Text-to-Image shows slight, acceptable decline. These results further prove FAIEr's generalization ability. In addition, we show image retrieval examples on Nocaps [2] in detail, as a visualizing extension of results displayed in Table 6 in our main paper. Taking a caption from Nocaps as a query, we score it on all 1,000 images from the validation set examples in the explore page of Nocaps website[2] and display the top 5 results in Fig.5. The results show that our FAIEr can find out the most relevant images, which means that it is better at understanding both image and text modalities, on account of its comprehensive scene graph representations and delicate attention fusion mechanism.

2) About our three evaluating orientations.

Back to our motivation, we tried to decompose the complex and subjective human evaluation intentions as fidelity, adequacy, and fluency for image captioning. We think it is impossible to capture all information in an image by a caption because of word limitation. If evaluation metrics only take references into account, some correct details not in reference might be considered wrong and do harm to evaluation. We note that the SPICE and CIDEr papers show that adding more than 5 reference captions seems not helpful. This is probably because more captions do not necessarily bring more information, since humans usually have similar content preferences when describing pictures with limited words. Fidelity measures whether the caption is related to the target image, and adequacy measures how much humans' common attention it conveys. To evaluate these two aspects, FAIEr **mainly** uses images to check captions' correctness and uses references to highlight the image gist. We did not mean that fidelity couldn't be assessed from references, but since images contain whole information and references come from images, mainly using images to measure fidelity is enough. Fidelity and adequacy are not separately assessed, but are fused to train in our model.

As for fluency, though not measuring it explicitly, the RNN encoder in FAIEr utilizes context information and encodes word order information, thus taking fluency into account to some extent. To further display FAIEr's potentiality for evaluating fluency, we followed VIFIDEL to average FAIEr and CIDEr scores on Flickr8K and test human correlation: **P-ρ(0.711), S-ρ(0.742), K-τ(0.589)**. Compared with FAIEr 4-ref in Table 1 of the main paper, performance gains 2%-3% improvement. Such results show that fluency is worth further exploration, probably with more advanced NLP techniques, and the merge of learning-based and rule-based models is a feasible solution.

References

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermuller. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45, 2017. 5

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019. 6

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398. Springer, 2016. 5

[4] 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 visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 5

[5] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812, 2018. 1

[7] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[8] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019. 5

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 1, 5

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 5

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

[2] http://nocaps.org/explore

Proceedings of the European Conference on Computer Vision, pages 740–755. Springer, 2014. 2

[12] Pranava Madhyastha, Josiah Wang, and Lucia Specia. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy, July 2019. Association for Computational Linguistics. 5

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 5

[14] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 1, 5

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 1, 5

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 1