

Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization

Supplementary Material

Nicholas Moratelli*¹
nicholas.moratelli@unimore.it

Davide Caffagni*¹
davide.caffagni@unimore.it

Marcella Cornia¹
marcella.cornia@unimore.it

Lorenzo Baraldi¹
lorenzo.baraldi@unimore.it

Rita Cucchiara^{1,2}
rita.cucchiara@unimore.it

¹ University of Modena and Reggio Emilia
Modena, Italy

² IIT-CNR
Pisa, Italy

In the following, we present additional materials about DiCO. In particular, we provide additional analyses and ablation studies, comparing DiCO with the standard SCST training paradigm. Moreover, we report further implementation details and qualitative results on all considered datasets and settings.

A Preliminaries

In this section, we first recap the definition of the SCST and Reinforcement Learning from Human Feedback (RLHF) training protocols [13, 15]. Then, we introduce captioning metrics based on contrastive embedding spaces [14].

Self-critical sequence training. SCST [15] is a two-step training methodology which (1) pre-trains a captioner f_θ using a time-wise cross-entropy loss with respect to ground-truth sequences, and (2) fine-tunes the same network by maximizing the CIDEr score [20] using a reinforcement learning (RL) approach. We assume that the captioner takes as input an image I described with a sequence of visual features (v_1, v_2, \dots, v_R) , and a ground-truth sequence $s = (w_1, w_2, \dots, w_T)$, where w_i is a token belonging to a pre-defined vocabulary. Noticeably, depending on the dataset there might be multiple ground-truth sequences associated with each image. During the first training stage, the network is conditioned on visual features and all ground-truth tokens up to the current prediction step t , and f_θ is optimized using the cross-entropy loss (teacher forcing). In the second training stage, instead, the network

*These authors contributed equally to this work.

© 2024. The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

is only conditioned on the input image and generates an entire caption $s' = (w'_1, w'_2, \dots, w'_{T'})$ by sampling input tokens from the output probability distribution generated at the previous time step. For instance, w'_t might be chosen as $w'_t = \operatorname{argmax}_f f_\theta(w_t | w'_{t-1}, \dots, w'_1, v_1, \dots, v_R)$, or multiple sentences can be sampled via beam search. The generated sentences are then employed to compute the CIDEr metric, which is later used as a reward to guide a policy-gradient RL update step (see [13] for details).

Reinforcement learning from human feedback. Recent NLP literature has employed techniques based on RLHF [13] to align the behavior of a large language model to human preferences. This approach is usually based on the collection of large-scale datasets of human preferences: the language model f_θ^1 is prompted with a prompt x to produce pairs of answers $(s'_1, s'_2) \sim f_\theta$, which are then presented to human labelers who express preference for one answer, *i.e.* $s'_w \succ s'_l$, where s'_w and s'_l indicate, respectively, the preferred and dis-preferred completion. The resulting dataset of human preferences $\mathcal{D} = \{x_i, s'_{w,i}, s'_{l,i}\}_{i=1}^N$ is then employed to train a reward model on top of it [9], for subsequent optimization with reinforcement learning. In image captioning, due to the lack in size of existing human preference datasets [11, 9, 20], *training a learnable reward model to follow the RLHF approach is impracticable* (see also Sec. C).

Learnable contrastive captioning metrics. As pointed out by recent literature on captioning evaluation, a model learned with language-image pre-training [14] can be straightforwardly employed as a captioning metric. Given a caption s' generated from I , indeed, its correctness score can be defined as a function of the similarity predicted by the image-text model, *i.e.* $\operatorname{sim}(I, s')$. A popular choice [8] is to define the score to be proportional to the ReLU of the predicted similarity and to employ a scalar multiplier w to stretch the resulting score within the range of $[0, 1]$:

$$\text{CLIP-S}(I, s') = w \cdot \operatorname{ReLU}(\operatorname{sim}(I, s')). \quad (8)$$

In the original formulation of [8] (termed CLIP-S), the backbone employed for computing similarities was pre-trained on 400M noisy (image, text) pairs collected from the internet. While CLIP-S shows a significantly higher alignment with human judgments compared to traditional metrics (*e.g.* BLEU, METEOR, CIDEr), the noisy nature of the training data limits the CLIP-S capability to distinguish fluent human-generated captions. To overcome this issue, a recent choice [16] is that of fine-tuning the backbone on cleaned data, which further boosts the correlation with human judgments. Specifically, the PAC-S score [16] trains on the basis of a similarity matrix built with human-collected captions and machine-generated ones, where the latter are obtained from a captioner trained to mimic the same distribution of human captions. In case a set of reference captions $R = \{r_i\}_{i=1}^N$ is given, there exists a version of the CLIP-based metrics accounting for them [8], which is defined as follows:

$$\operatorname{RefCLIP-S}(I, s', R) = \operatorname{H-Mean}(\operatorname{CLIP-S}(I, s'), \operatorname{ReLU}(\max_{r \in R} \cos(s', r))). \quad (9)$$

Following [16], the same formula can be applied to compute the reference-based version of PAC-S (RefPAC-S).

¹With a slight abuse of notation, in this paragraph we use f_θ to refer to a single-modality language model.

Training	Reward	Reference-based					Reference-free			
		B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	R@1	MRR
DiCO (w/o quality distances)	CLIP-S	19.3	25.6	79.5	0.820	0.858	0.836	0.851	45.7	57.2
DiCO	CLIP-S	21.4	27.1	82.6	0.824	0.863	0.837	0.856	46.5	58.4
DiCO (w/o quality distances)	PAC-S	24.9	27.6	91.7	0.812	0.873	0.809	0.875	50.6	62.4
DiCO	PAC-S	25.2	28.4	89.1	0.815	0.875	0.812	0.877	50.9	62.9

Table 5: Effectiveness analysis of using quality distances to weight rewards. Results are reported on the COCO test set using ViT-L/14 as backbone.

B Ablation Studies

Early stopping condition. When comparing multiple training strategies, we always employ an early stopping condition based on the validation value of the reference-based version of the metric used as a reward. In practice, when optimizing for CLIP-S, we early stop the training according to the validation RefCLIP-S, while when optimizing for PAC-S we early stop based on the validation RefPAC-S. We then take the model state corresponding to the epoch with the highest validation score and report its evaluation metrics. While this provides a reasonable evaluation strategy that equally promotes all compared approaches, evaluating a single model state does not capture the full training behavior of different fine-tuning strategies.

To complement Fig. 1 of the main paper, in Fig. 4 we report the test curves of CIDEr, RefCLIP-S, and CLIP-S obtained when optimizing the CLIP-S score and again those of CIDEr, RefPAC-S, and PAC-S obtained when optimizing the PAC-S score. For both cases, we compare the results using DiCO and SCST. With a red marker, we indicate the model state chosen by the early stopping condition, while a star marker indicates the model state after XE pre-training. As it can be seen, SCST hacks the reward metric immediately after the start of the fine-tuning phase, at the expense of CIDEr, RefCLIP-S, and RefPAC-S. Correspondingly, when optimizing using PAC-S as reward, the early stopping condition is forced to select the model state corresponding to the first fine-tuning epoch, which indeed showcases the highest RefPAC-S. Continuing the fine-tuning, though, would let SCST hack the reward metric even further and provide lower-quality captions.

On the contrary, DiCO showcases a more robust training behavior. While CLIP-S and PAC-S values increase during fine-tuning as a result of the optimization process, the decrease in CIDEr is well restrained, while RefCLIP-S and RefPAC-S even increase with respect to the XE state. This highlights that DiCO can optimize modern captioning metrics without incurring reward hacking and without deviating from a fluent and high-quality generation. Finally, in Fig. 6 we also report sample captions from the COCO Karpathy test split when optimizing the PAC-S score with SCST at different training stages, in comparison with DiCO. While SCST optimization tends to produce degraded and repetitive captions over time, DiCO maintains fluency and generation quality.

Effectiveness of using quality distances. We also evaluate the effectiveness of weighting rewards with quality distances (cf. Eq. 4) and train a different version of our DiCO approach setting $\gamma_i = \frac{1}{k}$. Table 5 reports the results of this analysis, using both CLIP-S and PAC-S as rewards. Notably, using quality distances to weight rewards improves the performance on both reference-based and reference-free metrics, thus demonstrating the usefulness of our strategy.

Effect of varying the β parameter. Fig. 5 shows how evaluation metrics vary when chang-

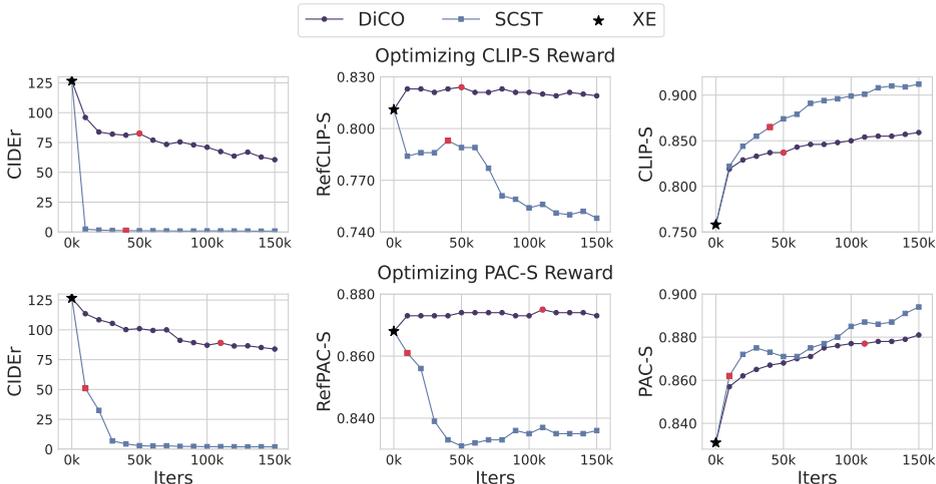


Figure 4: Metric curves when optimizing CLIP-S (top) and PAC-S (bottom) scores with DiCO and SCST. The red dot indicates the early stopping point we employ.

ing the β parameter, which regularizes the deviation from the pre-trained model. In particular, we report CIDEr, CLIP-S, and PAC-S scores using six different β values (*i.e.* from 0.05 to 0.3). As it can be seen, a higher β value prevents the model from deviating from the original pre-trained captioner (trained with XE loss), with CIDEr scores greater than 100 and, as a consequence, lower CLIP-S and PAC-S. On the contrary, when using a lower β value, reference-based metrics like CIDEr are penalized as the model is more inclined to deviate from the original version, thus boosting CLIP-S and PAC-S metrics. Overall, we find that $\beta = 0.2$ represents a good compromise between reference-based and reference-free metrics, and therefore we employ this value for all experiments.

Number of loser captions. DiCO requires generating $k + 1$ captions at each training step, of which the k worst are selected as losers according to the metric employed as reward. Table 6 shows the results as we vary the parameter k . In our experiments, we select $k = 4$ as it achieves the highest scores on reference-free metrics while keeping competitive performance on reference-based metrics.

C Additional Experimental Results

Comparison with SCST and RLHF. To complement the analyses reported in the main paper, we compare our fine-tuning strategy with SCST [15] and RLHF [6]. As we focus on the optimization of modern captioning metrics, for SCST experiments we directly apply a

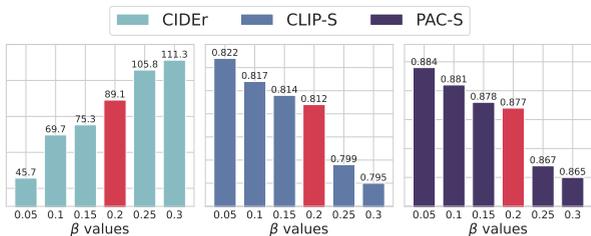


Figure 5: CIDEr, CLIP-S, and PAC-S scores when changing the β parameter using ViT-L/14 as backbone. Higher β values prevent the model from deviating from the pre-trained captioner, while penalizing reference-free metrics. The best trade-off is given by $\beta = 0.2$.

Training	Reward	k	Reference-based				Reference-free				
			B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	R@1	MRR
SCST	PAC-S	-	22.3	28.4	51.1	0.801	0.861	0.805	0.862	46.7	58.8
DiCO (Ours)	PAC-S	1	30.4	28.2	109.5	0.819	0.876	0.790	0.861	41.4	54.0
DiCO (Ours)	PAC-S	2	27.1	28.3	99.9	0.818	0.876	0.802	0.871	46.2	58.9
DiCO (Ours)	PAC-S	4	25.2	28.4	89.1	0.815	0.875	0.812	0.877	50.9	62.9
DiCO (Ours)	PAC-S	6	24.9	28.6	86.9	0.813	0.874	0.811	0.876	50.9	62.5
DiCO (Ours)	PAC-S	7	24.8	28.5	86.5	0.812	0.873	0.811	0.876	50.5	62.4

Table 6: Performance varying the number of “loser” captions k . Results are reported on the COCO test set using ViT-L/14 as backbone.

CLIP-based reward using either CLIP-S or PAC-S. Further, we adapt the RLHF paradigm to a captioning setting by first training a reward model based on human feedback and then optimizing the captioning model via reinforcement learning based on the PPO objective [17], using the score from the reward model as a reward. To train the reward model, we employ a combination of datasets typically used to evaluate the correlation of captioning metrics with human judgments, namely Flickr8k-Experts, Flickr8k-CF, and Composite [4, 9]. All datasets contain multiple candidate captions, either human-annotated or generated by a captioning model, associated with a given image and corresponding human ratings that evaluate whether each caption correctly describes the image. Overall, we obtain around 3.5k unique images and 50k captions each associated with a normalized rating between 0 and 1. At training time, we sample a pair of candidate captions for each image and use the associated human ratings to train the reward model, using maximum likelihood estimation. The reward model is built by modifying the captioner pre-trained with XE so to have a single final output and is trained with a negative log-likelihood loss following [13]. In addition to this adaptation of the RLHF training strategy, we also design a variant in which the human preferences-based reward model is replaced with a CLIP-based evaluator, directly employing CLIP-S or PAC-S as reward. For completeness, we also include the results of the model trained with XE loss only, which is the starting point for all other fine-tuning strategies.

Results are reported in Table 7 in terms of reference-based and reference-free evaluation metrics. As it can be seen, the proposed optimization strategy generally leads to better results across all metrics, surpassing both SCST and RLHF by a significant margin. Specifically, we can notice that optimizing the captioner with human feedback does not improve the final results. This is probably due to the limited size of available captioning datasets with human ratings, that prevent the effective application of standard RLHF fine-tuning to a cap-

Training	Reward	Reference-based				Reference-free				
		B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	R@1	MRR
XE	-	37.3	30.4	126.6	0.811	0.868	0.758	0.831	27.7	38.5
RLHF	HF	21.4	27.8	57.9	0.776	0.843	0.745	0.819	24.7	34.9
RLHF	CLIP-S	12.9	24.2	2.3	0.714	0.800	0.732	0.794	19.5	29.0
SCST	CLIP-S	10.2	23.0	1.1	0.793	0.827	0.865	0.834	43.3	55.0
DiCO	CLIP-S	21.4	27.1	82.6	0.824	0.863	0.837	0.856	46.5	58.4
RLHF	PAC-S	12.4	23.7	2.0	0.712	0.798	0.726	0.790	18.1	27.5
SCST	PAC-S	22.3	28.4	51.1	0.801	0.861	0.805	0.862	46.7	58.8
DiCO	PAC-S	25.2	28.4	89.1	0.815	0.875	0.812	0.877	50.9	62.9

Table 7: Comparison with different fine-tuning strategies. Results are reported on the COCO test set using ViT-L/14 as backbone.

Model	Backbone	Reward	Reference-based					Reference-free	
			B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
Cho <i>et al.</i> (SCST) [10]	RN50	CLIP-S	5.9	14.2	13.9	0.689	0.769	0.721	0.803
Cho <i>et al.</i> (SCST) [10]	RN50	CLIP-S+Gr.	11.1	16.3	19.1	0.683	0.784	0.684	0.808
DiCO (Ours)	RN50	CLIP-S	14.2	16.1	17.2	0.688	0.782	0.696	0.805
DiCO (Ours)	RN50	PAC-S	13.6	16.4	19.2	0.695	0.800	0.704	0.835
DiCO (Ours)	ViT-B/32	PAC-S	13.8	16.7	19.3	0.708	0.811	0.726	0.855
DiCO (Ours)	ViT-L/14	PAC-S	15.0	17.4	23.2	0.722	0.822	0.731	0.855

Table 8: Fine-grained image captioning results on the FineCapEval dataset.

Model	Reward	Backbone	nocaps			VizWiz			TextCaps			CC3M		
			C	CLIP-S	PAC-S									
Cho <i>et al.</i> (SCST) [10]	CLIP-S	RN50	10.9	0.765	0.819	4.7	0.693	0.784	7.6	0.731	0.813	3.6	0.717	0.784
Cho <i>et al.</i> (SCST) [10]	CLIP-S+Gr.	RN50	54.0	0.712	0.822	20.4	0.648	0.774	26.8	0.680	0.814	18.0	0.671	0.790
SCST	PAC-S	RN50	20.9	0.741	0.850	13.0	0.668	0.795	22.0	0.683	0.822	5.8	0.699	0.797
DiCO (Ours)	PAC-S	RN50	64.6	0.733	0.851	29.3	0.680	0.813	30.8	0.696	0.838	21.4	0.690	0.815
SCST	PAC-S	ViT-B/32	35.7	0.750	0.854	20.1	0.715	0.837	21.9	0.699	0.835	9.8	0.698	0.809
DiCO (Ours)	PAC-S	ViT-B/32	66.5	0.754	0.869	32.7	0.710	0.842	31.8	0.712	0.853	23.4	0.697	0.821
SCST	PAC-S	ViT-L/14	44.8	0.746	0.850	26.8	0.701	0.820	23.6	0.705	0.836	13.2	0.701	0.811
DiCO (Ours)	PAC-S	ViT-L/14	74.3	0.755	0.865	40.6	0.706	0.832	33.7	0.717	0.852	26.7	0.704	0.824

Table 9: Image captioning results on out-of-domain datasets like nocaps, VizWiz, TextCaps, and CC3M.

tioning model. When instead using CLIP-S and PAC-S as rewards, both SCST and RLHF experience a significant drop in standard image captioning metrics. In terms of CLIP-based metrics, SCST obtains quite good results which however are not supported with robustness on all other metrics. Overall, our DiCO strategy exhibits good performance in all evaluation directions, obtaining the best results in terms of CLIP-based and retrieval-based scores while maintaining competitive performance on standard metrics.

Fine-grained image captioning evaluation. As an additional analysis, we report in Table 8 fine-grained image captioning results on the FineCapEval dataset [10], which contains 1,000 images from COCO and CC3M [18] annotated with 5 detailed and fine-grained captions, describing the background of the scene, the objects and their attributes, and the relations between them. Also in this setting, DiCO confirms its superior performance compared to other CLIP-based optimized captioners [10], thus further demonstrating the effectiveness of directly optimizing a captioning model with the proposed solution. Specifically, when considering the same backbone used in [10] (*i.e.* RN50), DiCO achieves the best results in terms of both standard captioning metrics and CLIP-based scores with the sole exception of CLIP-S.

Out-of-domain evaluation. To evaluate generalization capabilities to out-of-domain images, we extend our analysis by considering diverse image captioning datasets, including nocaps [10], which has been introduced for novel object captioning and contains object classes that are not present in the COCO dataset, VizWiz [10], composed of images taken by blind people, TextCaps [19], which is instead focused on text-rich images, and CC3M [18], composed of image-caption pairs collected from the web. In Table 9 we report the results of our approach using PAC-S as reward, compared to the CLIP-based training strategy proposed in [10] and the standard SCST with the same reward as in our approach. Even in this challenging context, DiCO achieves the best results across all datasets and backbones, demonstrating better descriptive capabilities than competitors.

Additional results on Flickr30k. Finally, we also benchmark our method on images from the Flickr30k dataset [20]. We report the results in Table 10, using both PAC-S and CLIP-S

Model	Backbone	Reward	Reference-based					Reference-free	
			B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
Cho <i>et al.</i> (SCST) [10]	RN50	CLIP-S	4.0	16.5	10.0	0.751	0.806	0.818	0.839
Cho <i>et al.</i> (SCST) [10]	RN50	CLIP-S+Gr.	11.0	20.9	36.8	0.750	0.826	0.755	0.839
DiCO (Ours)	RN50	CLIP-S	16.8	22.0	44.9	0.762	0.829	0.774	0.839
DiCO (Ours)	RN50	PAC-S	17.2	22.6	46.8	0.769	0.846	0.786	0.871
DiCO (Ours)	ViT-B/32	PAC-S	17.8	22.8	48.6	0.780	0.855	0.810	0.890
DiCO (Ours)	ViT-L/14	PAC-S	19.0	24.5	55.8	0.790	0.862	0.804	0.883

Table 10: Image captioning results on the Flickr30k dataset.

as reward and comparing with the approach proposed in [10]. As it can be noticed, DiCO demonstrates strong generalization capabilities, achieving the best results on almost all evaluation metrics further confirming the effectiveness of our training strategy.

D Additional Details

Additional implementation and training details. During cross-entropy pre-training, we accumulate gradients for 8 training steps over 2 GPUs, resulting in 1,024 samples per batch. For this training stage, the learning rate is linearly increased up to $2.5 \cdot 10^{-4}$. Each fine-tuning experiment starts from the XE checkpoint with the highest CIDEr, leveraging 2 GPUs and a global batch size of 16. Training the reward models for RLHF follows the same settings as the fine-tuning phase.

CIDEr-based optimization. In computing quality distances with CIDEr metric as reward (see Table 4 of the main paper), we set the softmax temperature τ to 1, a higher value than the one used for CLIP-S and PAC-S optimization (equal to $1/(3 \cdot 10^2)$). We argue that the CIDEr score is discriminant enough to discern the goodness of similar captions sampled from a beam search. On the contrary, CLIP-based metrics are less sensible to small changes, thus needing a lower temperature to amplify the score differences.

Human-based evaluation. As shown in the main paper, we conducted a user study to evaluate the quality of generated captions. To

this end, we developed a custom web interface that presents the users with an image and two captions, one generated by DiCO, and one drawn from a different model (cf. Table 2), and asks them to select the best caption based on correctness and helpfulness. We show a screenshot of the developed interface is shown in Fig. 7. Overall, the evaluation involved more than 50 different users, collecting approximately 3,000 evaluations for both criteria.

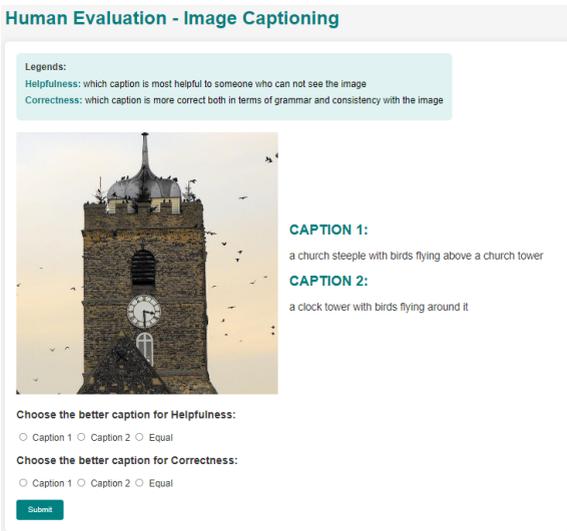


Figure 7: User study interface to evaluate helpfulness and correctness of given captions.

LLM-based evaluation. GPT-3.5 Turbo proves itself very compliant with our requests. However, we find about a hundred failure cases (*e.g.* wrong JSON format, more scores than the number of candidate captions, etc.) out of 7,000 requests. We opt for simply discarding them in the winner rate computations. For fair evaluation, we randomly swap the order in which we insert the two candidate captions in the prompt. This ensures that the descriptions generated by our competitors have on average the same probability as ours to be processed first by the LLM causal attention, which may influence the final score. Following [4], the prompt we used is:

```
You are trying to tell if each sentence in a candidate set of captions is
describing the same image as a reference set of captions.
Candidate set: {candidate captions}
Reference set: {target captions}
You have to determine how likely is that each of the sentences in the
candidate set is describing the same image as the reference set, on a scale
from 0 to 100. Please output exclusively a JSON list, with an item for each
candidate. Each item should be a dictionary with a key "score", containing
a value between 0 and 100, and a key "reason" with a string value containing
the justification of the rating. Start directly with the json.
```

E Additional Qualitative Results

Finally, we report additional qualitative results to qualitatively validate the effectiveness of our training strategy. In particular, Fig. 8 and Fig. 9 show sample images from the COCO dataset and captions predicted by DiCO in comparison to those generated by SCST, the model proposed in [4], and the large-scale model BLIP-2 [10]. As it can be seen, DiCO generates significantly more detailed captions than BLIP-2, while reducing repetitions typically present in SCST-generated sentences. To qualitatively validate the generalization capabilities to out-of-domain images, we report sample captions predicted by DiCO and SCST using PAC-S as reward on nocaps [0] (Fig. 10), VizWiz [0] (Fig. 11), TextCaps [19] (Fig. 12), and CC3M [18] (Fig. 13).

In Fig. 14, we instead show some qualitative results when using CIDEr as reward. In this case, we compare DiCO with standard image captioning models, including a vanilla Transformers trained with the same visual features used in our approach, COS-Net [10], and \mathcal{M}^2 Transformer [6]. All competitors have been trained with a standard XE+SCST training protocol. Also in this setting, DiCO is able to generate high-quality captions compared to competitors, confirming that it can also be employed as a valid alternative to SCST for training standard image captioning models.

F Limitations

As with all image captioning models, we acknowledge that our method might fail to provide informative captions in some rare contexts. To qualitatively evaluate the limitations of our approach, we report some failure cases in Fig. 15. As it can be seen, DiCO may produce factual errors, *e.g.* mistaking balloons for *kites* (first sample, first row) or a *stuffed animal* for a seal (first sample, second row). Additionally, DiCO may fail to recognize known entities, thus providing only a broad description of the scene (*e.g.* a *white monument* rather than the Taj Mahal mausoleum, or a *black silver car* rather than an Aston Martin). This can be conducted to the image-caption pairs contained in the COCO dataset, which lack open-world knowledge. Finally, when the main subject of the image is uncertain (second sample,

third row), DiCO may overlook the picture and generate captions based on its learned priors, resulting in hallucinations.

References

- [1] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. *arXiv preprint arXiv:1511.03292*, 2015.
- [2] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019.
- [3] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating Image Captions with Large Language Models. In *EMNLP*, 2023.
- [4] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained Image Captioning with CLIP Reward. In *NAACL*, 2022.
- [5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.
- [6] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020.
- [7] Danna Gurari, Yanan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In *ECCV*, 2020.
- [8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIP-Score: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021.
- [9] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013.
- [10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023.
- [11] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *CVPR*, 2022.
- [12] 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 *ECCV*, 2014.
- [13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

- [15] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017.
- [16] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, 2023.
- [17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018.
- [19] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, 2020.
- [20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015.
- [21] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.



SCST (after 10k iters): A group of people with umbrellas walking down a street with people walking down a wet sidewalk holding pink umbrellas in rain.

SCST (after 50k iters): Many people crossing wet wet alley with people walking with colorful umbrellas outside a building with wet alley with people walking under umbrellas outside rainy surface.

SCST (after 100k iters): Pedestrians walking down wet wet road with pedestrians carrying pink umbrellas outside a building on wet sidewalk outside rainy wall with buildings outside surface poles surface poles.

DiCO (Ours): A group of people walking down a wet sidewalk with umbrellas in the rain.



SCST (after 10k iters): A group of young boys kicking around a soccer ball on a soccer field with other young boys running around with net in background.

SCST (after 50k iters): Young boys kicking soccer ball around soccer goal kicking grass underneath a goal on grass behind background behind surface surface with trees in background behind surface surface.

SCST (after 100k iters): Young boys soccer teams chasing after after soccer soccer goalie in background with green leaves on grass behind background surface court with young boy.

DiCO (Ours): A group of young children kicking a soccer ball in a field.



SCST (after 10k iters): A group of people playing frisbee with a man laying on ground with a person laying on ground with other people in background.

SCST (after 50k iters): Group of kids playing ultimate frisbee with man laying on ground with people on sand floor with frisbees while people gather around background behind surface surface.

SCST (after 100k iters): A group of kids playground with man laying on cement floor playing frisbee game with man laying outside a crowd in background surface outside surface poles leg.

DiCO (Ours): A group of people playing with a frisbee on a beach with other people in the background.



SCST (after 10k iters): A black motorcycle parked on a sidewalk next to a parked motorcycle on a sidewalk next to a rack with bicycles in background.

SCST (after 50k iters): An old motorcycle parked on sidewalk with parked bicycle outside a brick background with other bikes on sidewalk outside clear background behind surface background behind surface surface.

SCST (after 100k iters): Antique motorcycle parked outside a brick building with a silver seat outside a bike on a sidewalk with other bikes outside background surface outside surface poles top.

DiCO (Ours): A small black motorcycle parked on a sidewalk next to other bikes.



SCST (after 10k iters): A small pizza with vegetables on a wooden picnic table with a pizza on a picnic table with silverware and wine in background.

SCST (after 50k iters): Small vegetable pizza with vegetable vegetable on wooden picnic table with serving dish with other foods on grass outside clear surface behind background behind surface outside surface.

SCST (after 100k iters): Cooked vegetable vegetable vegetable vegetable pizza served outside outside table with fork on picnic table outside a wine holder on sun surface outside background surface poles hand.

DiCO (Ours): A small pizza on a wooden picnic table with silverware and a wine glass in the background.

Figure 6: Qualitative results on sample images from the COCO Karpathy test split [12] using SCST optimization with PAC-S reward at different fine-tuning states, in comparison with DiCO.



BLIP-2 [11]: A group of colorful umbrellas under a covered area.
Cho et al. [9]: A large blue vase sitting on the dirt ground with colorful decorations next to a market.
SCST: Several colorful colorful umbrellas hanging from a wooden structure under a tree tree with statues on display in outdoor market under palm trees on clear background.
DiCO (Ours): A display of colorful umbrellas in a shop with decorations.



BLIP-2 [11]: A green and yellow train pulling into a station.
Cho et al. [9]: A green commuter train parked near a platform area with a green trees area motion stance ear stance.
SCST: A green and yellow passenger train traveling down train tracks next to a loading platform with a green passenger on a platform with trees in background.
DiCO (Ours): A green and yellow passenger train traveling down train tracks next to a platform.



BLIP-2 [11]: A black and white photo of a train.
Cho et al. [9]: A large metal train driving next to a lot of tanks on the tracks.
SCST: Black and white photograph of freight freight freight freight cars on railroad tracks with tanker cars on track with wires in background on background.
DiCO (Ours): A black and white photo of a freight train traveling down railroad tracks next to wires.



BLIP-2 [11]: A woman in a boat selling food on the water.
Cho et al. [9]: A couple of women preparing a tray of food in the river with bananas.
SCST: Two women in canoes with baskets full of bananas and other asian asian workers carrying baskets on shelves with baskets on clear surface in background.
DiCO (Ours): Two asian women in a small boat filled with food and bananas.



BLIP-2 [11]: A birthday cake with dora the explorer on it.
Cho et al. [9]: A large blue birthday cake with toys and toys on the table.
SCST: A colorful birthday cake decorated with purple and green flowers on top of purple birthday cake with decorations on table in background on background.
DiCO (Ours): A birthday cake with purple and green decorations on it.



BLIP-2 [11]: A bunch of carrots next to a plate of food.
Cho et al. [9]: A bunch of carrots and other carrots on a white plate with a knife behind them.
SCST: A white plate topped with carrots and other vegetables on a clear surface with other vegetables on display in background on background on background.
DiCO (Ours): A bunch of carrots and other vegetables on a white plate.

Figure 8: Qualitative results on sample images from the COCO Karpathy test split using DiCO with PAC-S reward. We compare our approach with SCST using PAC-S as reward, the model proposed in [9] with CLIP+S+Grammar as reward, and the BLIP-2 model [11] which has been trained on large-scale vision-and-language datasets.



BLIP-2 []: A group of teddy bears on a boat.

Cho et al. []: A couple of teddy bears wearing hats sitting on a boat with a plant behind them.

SCST: Teddy bears dressed in green costumes riding a miniature boat decorated with green hats on a blue wall in military uniform on display in background.

DiCO (Ours): Stuffed animals dressed in green costumes riding in a boat.



BLIP-2 []: A herd of zebras walking through a grassy field.

Cho et al. []: A large herd of zebra and other animals grazing in the prairie.

SCST: A herd of zebras running through tall brown grass in savanna with distance in distance in background on clear surface in background on background.

DiCO (Ours): A large herd of zebras walking through tall brown grass in a large field.



BLIP-2 []: A table topped with food and a bottle of wine.

Cho et al. []: Two plates of food and a bottle of wine on the table with a bottle.

SCST: A white plate topped with meat cheese and vegetables next to a bottle of wine and bread with cheese and tomatoes on wooden surface in background.

DiCO (Ours): A table topped with two bowls of food next to a bottle of wine and cheese.



BLIP-2 []: A small train that is on display in a mall.

Cho et al. []: A large red train driving on a busy street with people near it.

SCST: A miniature miniature train with a miniature train on a sidewalk with people walking around a mall with a mall on a mall platform in background.

DiCO (Ours): People walking around a miniature train on a sidewalk in a shopping mall.



BLIP-2 []: A woman with a bunch of bananas on her head.

Cho et al. []: A smiling woman wearing a colorful costume holding a bunch of bananas on the background.

SCST: A woman dressed in colorful costume with yellow bananas on her head with a man's head dressed in colorful costume in background on background.

DiCO (Ours): A smiling woman wearing a large banana costume on her head with people in the background.



BLIP-2 []: Two horses walking in the desert with mountains in the background.

Cho et al. []: A group of three brown horses walking together in the desert.

SCST: Two brown horses walking through dry desert desert with sand on clear surface in distance with clear background on clear surface in background.

DiCO (Ours): Two brown horses walking through a desert plain with sand and bushes in background.

Figure 9: Qualitative results on sample images from the COCO Karpathy test split using DiCO with PAC-S reward. We compare our approach with SCST using PAC-S as reward, the model proposed in [] with CLIP+S+Grammar as reward, and the BLIP-2 model [] which has been trained on large-scale vision-and-language datasets.



SCST: A woman dressed in colorful costume holding a colorful umbrella in a rain-outfit with chinese writing on her face.

DiCO (Ours): A asian woman wearing a colorful costume holding a parasol.



SCST: A little girl walking through water with a backpack walking through a flooded street with water in a girl's hand.

DiCO (Ours): A girl walking through a flooded street with a backpack.



SCST: A person sitting on a red motorcycle with a helmet on a red motorcycle at a show.

DiCO (Ours): A person in a red and white outfit sitting on a red motorcycle at a show.



SCST: A military military vehicle driving down a rain soaked road with people in a military military vehicle on a rainy day.

DiCO (Ours): A military vehicle driving down a wet road with people standing on it.



SCST: A police car driving down a road with lights on driving down a busy road with other vehicles on a sunny day.

DiCO (Ours): A line of emergency vehicles driving down a road with trees in the background.



SCST: A chocolate cake topped with strawberries and strawberries on a plate with ice cream with strawberries on a white surface.

DiCO (Ours): A cake topped with strawberries and whipped cream.

Figure 10: Qualitative results on sample images from nocaps.



SCST: A black and white dog laying in front of a Christmas tree with a dog laying on the floor next to it.

DiCO (Ours): A black and white dog laying in front of a christmas tree.



SCST: A plastic statue of a person wearing aluminum foil on a wooden board with aluminum foil on a counter.

DiCO (Ours): A sculpture of a person wearing a dress standing on a wooden board.



SCST: A person sitting on a tiled floor wearing pink pants sitting on the floor with his feet on the floor.

DiCO (Ours): A man sitting on the floor with his legs crossed.



SCST: A plastic plastic container filled with meat on a wooden table next to a water bottle on a wood surface.

DiCO (Ours): A plastic container of chicken on a wooden table next to a water bottle.



SCST: An electric guitar on carpet with yarn holder behind background bottom a hawk on a couch cushion behind background.

DiCO (Ours): An old fashioned guitar sitting on a colorful rug.



SCST: A blue vase filled with purple and white flowers on a blue table with other flowers on a blue background.

DiCO (Ours): A picture of purple and white flowers on a blue table.

Figure 11: Qualitative results on sample images from VizWiz.



SCST: A person holding a cell phone with a beer in front of a hand holding a cell phone with a beer in it.

DiCO (Ours): A person holding a smart phone next to a beer.



SCST: Three young children standing around an orange orange statue with three young girls in an orange building.

DiCO (Ours): Three young children reaching up on an orange statue.



SCST: Two chefs working in an industrial industrial kitchen with a grill in a stainless steel oven with lots of meat on the counter.

DiCO (Ours): Two chefs working in a commercial kitchen with a metal grill.



SCST: A group of young women dressed in yellow school uniforms posing for a picture in a school uniform.

DiCO (Ours): A group of young women wearing yellow and blue school uniforms posing for a picture together.



SCST: A yellow taxi cab driving down a busy city street with cars on a busy city street with buildings in the background.

DiCO (Ours): A yellow taxi cab driving down a busy city street with other cars and buildings.



SCST: A group of young women running on a race race track with two girls running around a race track with a crowd in the background.

DiCO (Ours): A group of young women running across a track at a competition.

Figure 12: Qualitative results on sample images from TextCaps.



SCST: A person with a backpack walking through deep snow with a backpack walking through a forest with mountains in the background.

DiCO (Ours): A person standing in the snow with a backpack near a large hill with a mountain in the background.



SCST: A group of people walking down a sidewalk with lots of people with backpacks walking around a path with trees in the background.

DiCO (Ours): A crowd of people standing on a sidewalk next to a tree covered with colorful leaves.



SCST: A beautiful young woman wearing a yellow costume posing for a picture wearing a yellow dress with her hand on her face.

DiCO (Ours): A beautiful young woman wearing a yellow dress posing with her arm around her neck.



SCST: A professional baseball player running with a blue helmet in a blue uniform in a stadium with a crowd crowd in the background.

DiCO (Ours): A professional baseball player in blue and white uniform running through a stadium.



SCST: Black and white photograph of a woman dressed dressed in black and white dress with long black and white clothing on a dark surface.

DiCO (Ours): A black and white photo of a woman dressed in black and white clothing.



SCST: A row of basketball balls sitting next to a row of orange balls in front of a body of water with city in background.

DiCO (Ours): A row of basketball balls in front of a view of a city skyline in the background.

Figure 13: Qualitative results on sample images from CC3M.

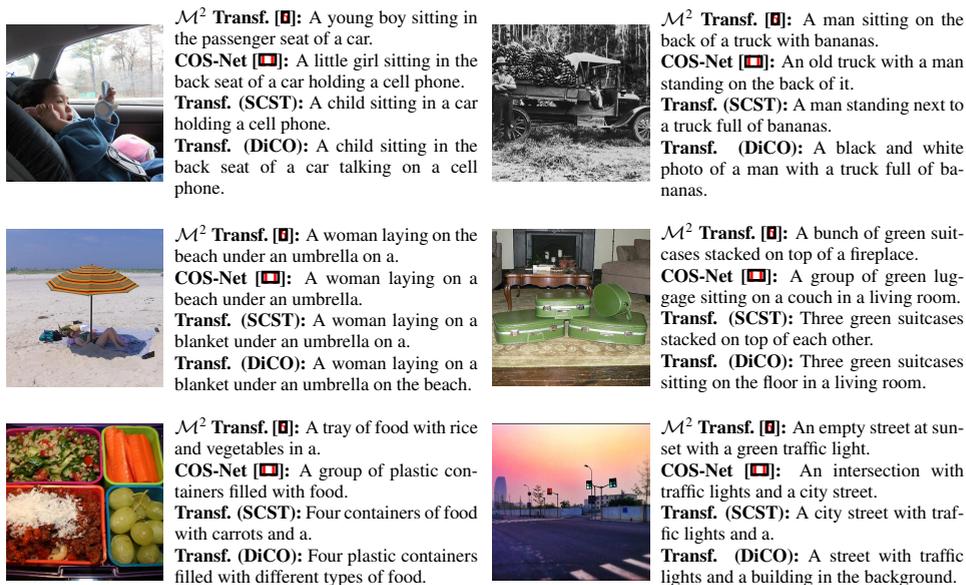


Figure 14: Qualitative results on sample images from the COCO Karpathy test split using DiCO with CIDEr reward. We compare our approach with a standard Transformer trained with SCST and CIDEr as reward, \mathcal{M}^2 Transformer [B] and COS-Net [C].

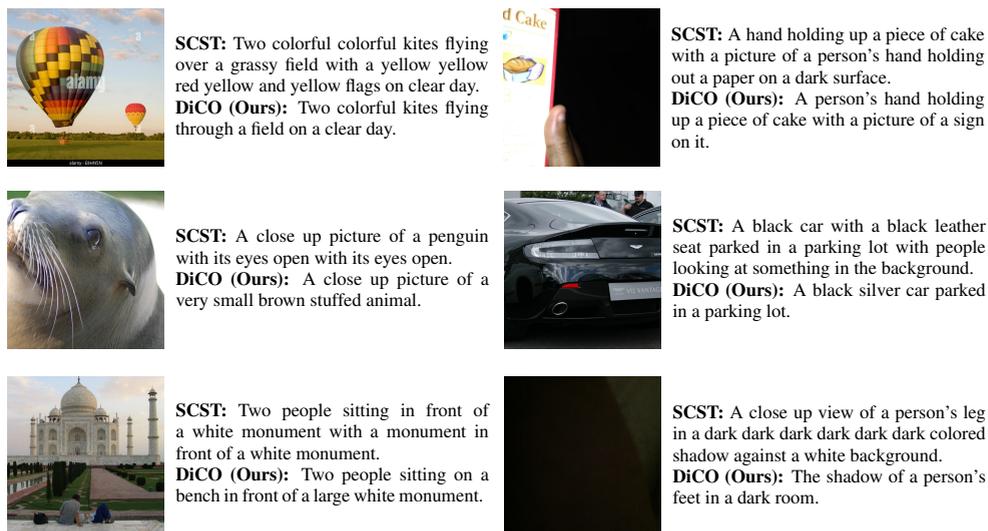


Figure 15: Qualitative results showcasing samples where DiCO fails in comparison to the SCST training methodology.