Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization

Nicholas Moratelli*[1]
nicholas.moratelli@unimore.it
Davide Caffagni*[1]
davide.caffagni@unimore.it
Marcella Cornia[1]
marcella.cornia@unimore.it
Lorenzo Baraldi[1]
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

Abstract

The conventional training approach for image captioning involves pre-training a network using teacher forcing and subsequent fine-tuning with Self-Critical Sequence Training to maximize hand-crafted captioning metrics. However, when attempting to optimize modern and higher-quality metrics like CLIP-Score and PAC-Score, this training method often encounters instability and fails to acquire the genuine descriptive capabilities needed to produce fluent and informative captions. In this paper, we propose a new training paradigm termed *Direct CLIP-Based Optimization* (DiCO). Our approach jointly learns and optimizes a reward model that is distilled from a learnable captioning evaluator with high human correlation. This is done by solving a weighted classification problem directly inside the captioner. At the same time, DiCO prevents divergence from the original model, ensuring that fluency is maintained. DiCO not only exhibits improved stability and enhanced quality in the generated captions but also aligns more closely with human preferences compared to existing methods, especially in modern metrics. Additionally, it maintains competitive performance in traditional metrics. Our source code and trained models are publicly available at https://github.com/aimagelab/DiCO.

1 Introduction

The task of image captioning [21, 53, 58, 64] requires an algorithm to describe a visual input in natural language. As a captioner should ideally match the level of detail and precision desired by the user, over time there has been an increasing interest in developing training strategies for aligning the behavior of a captioner to mimic a desired style and quality level.

*These authors contributed equally to this work.

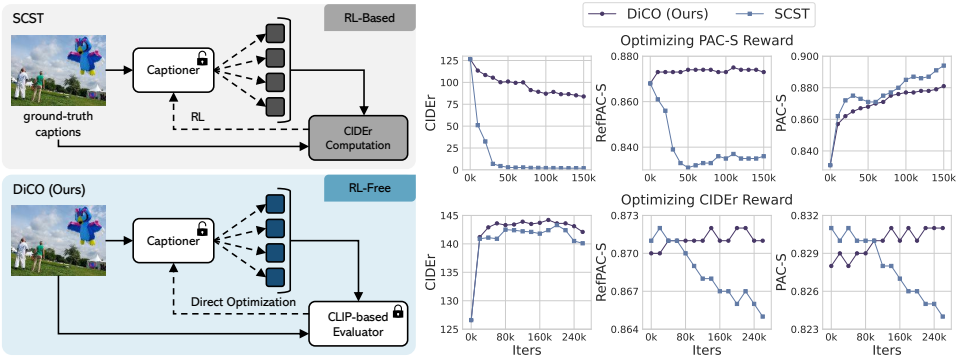

Figure 1: Comparison between SCST [45] and our *Direct CLIP-Based Optimization* (DiCO). DiCO distills a reward model from a learnable CLIP-based captioning evaluator, without requiring reinforcement learning and preventing reward hacking and divergence.

Traditionally, the quality of captions has been measured with textual similarity metrics, so captioners have been trained to maximize a non-differentiable metric like CIDEr [57] during a fine-tuning stage based on reinforcement learning, *i.e.* Self-Critical Sequence Training (SCST) [34, 44, 45]. As this strategy requires the availability of multiple reference captions and tends to produce less distinctive descriptions that ignore the fine detailed aspects of an image, recently there have been preliminary attempts to optimize higher-quality image captioning metrics based on embedding spaces that do not require human references [13, 24, 59], like CLIP-Score [17] and PAC-Score [47]. Besides, these metrics also consider the actual multi-modal alignment between the generated text and the visual content of the input image rather than just comparing texts. Most importantly, they also showcase a superior alignment with human judgment, making them ideal candidates for tuning the behavior of captioners towards a higher quality of generation.

Unfortunately, optimizing modern metrics with pre-existing strategies like SCST results in instability and model collapse [13]. We showcase this in Fig. 1, where we employ SCST for optimizing either PAC-S or CIDEr (light blue lines). When we try to optimize PAC-S, the fine-tuned captioner hacks the metric and deviates from a fluent and high-quality generation, resulting in a rapid decrease according to all other metrics and leading to repetitions and grammatical errors. To solve these issues, we propose DiCO, a novel training methodology that can align a captioner towards better quality captions by distilling from an external contrastive-based evaluator like CLIP-S or PAC-S, without incurring model collapse and without employing a reinforcement learning objective. Our approach achieves this goal by learning a reward model directly into the captioner and mimicking pairwise quality relations expressed by the external evaluator. This ensures a high degree of alignment with human preferences while avoiding reward hacking. This is visually represented in Fig. 1 (dark blue lines): DiCO can optimize both a modern metric like PAC-S and a traditional one like CIDEr by maintaining good scores across all metrics.

We assess the quality of the proposed training methodology by conducting extensive experiments on the COCO dataset [31]. Furthermore, in the supplementary materials, we prove the generalization capabilities of DiCO over other six image captioning benchmarks. Our experimental results demonstrate that DiCO features state-of-the-art quality in the generated captions and improved training stability. This also results in a better performance in terms of modern captioning metrics, while also balancing with competitive performances

on traditional handcrafted metrics. On the other hand, when adopted to maximize standard captioning metrics like CIDEr [57], DiCO achieves state-of-the-art results also in this setting. Going beyond automatic image captioning metrics, we confirm the effectiveness of our approach by also employing human-based evaluation.

To sum up, our proposal markedly differs from all fine-tuning strategies in the current image captioning literature. Presently, this field remains closely tied to traditional techniques, employing the classic SCST algorithm with rewards based on ground-truth captions, while overlooking a concerted emphasis on semantic and syntactic richness, as well as alignment with human cognition. Extensive experiments on standard image captioning datasets demonstrate the effectiveness of the proposal.

2 Related Work

Standard image captioning. Early attempts in the field of image captioning were based on an encoder-decoder architecture, wherein the visual input content is encoded through a CNN, while the textual output is aptly generated by an RNN conditioned on the visual encoding [6, 21, 45, 58]. Subsequently, this approach witnessed refinement through the integration of different attention-based strategies [64], eventually applied to image regions [3] and enhanced with spatial and semantic graphs [65, 66]. More recently, an alternative trend encompasses Transformer-based architectures, where numerous works have been developed exploring varied directions [5, 14, 19, 29]. While the aforementioned approaches exploited the same fine-tuning strategy usually composed of a pre-training with cross-entropy loss followed by reinforcement learning, we explore a different perspective. Along this line, Cho *et al.* [13] stands out as the method that is closely related to our proposal, as it defines a CLIP-based fine-tuning scheme that, however, relies on reinforcement learning. Concurrently, large-scale vision-and-language pre-training has been used to perform several tasks requiring multimodal capabilities, such as image captioning. These models [18, 28, 59, 60, 68] are pre-trained on millions or even billions of image-text pairs, usually collected from the web, and fine-tuned for a target task.

LLM-based image captioning. To leverage the power of LLMs demonstrated in different contexts, many attempts have emerged to bestow vision capabilities to a pre-trained LLM [15, 22, 23, 46, 70], resulting in impressive performance over various vision-and-language tasks like image captioning. In this context, ZeroCap [53] runs a few optimization steps for each new token, to align the text produced by GPT-2 [40] to the input image, using CLIP [41] as guidance. Other works [37, 43], instead, start from a pre-trained LLM and only learn cross-attention layers to mimic the interaction between textual and visual domains. Research efforts have also been dedicated to developing large-scale multimodal models [7], usually based on LLMs and trained on huge amounts of multimodal data [8, 11, 26, 27]. In this context, image captioning is employed as a pre-training task to help vision-and-language alignment, and eventually in the instruction-tuning stage [32, 33]. Thanks to the underlying LLM, all these solutions usually lead to image captioners with greater descriptive capabilities. In this work, we show how to increase the quality and descriptiveness of generated captions without relying on any pre-trained LLM.

Training strategies for LLMs. Aligning models with human judgment constitutes a well-known issue in both NLP and captioning literature. In this context, several strategies for fine-tuning LLMs have been explored. For example, a common research direction is to guide the model through a combination of input-output pairs and explicit instructions [12, 20, 61, 62].

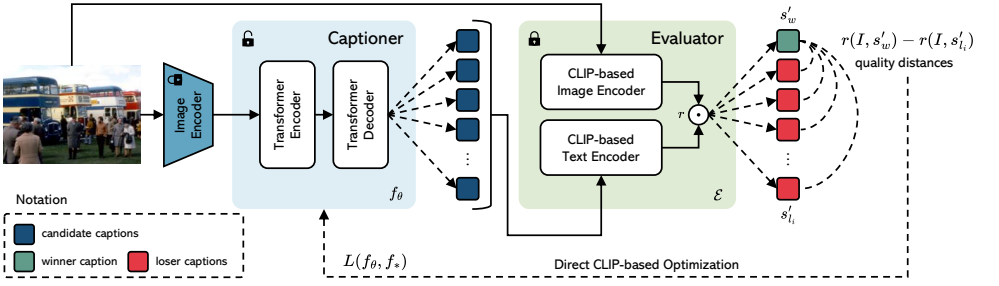

Figure 2: Overview of our approach. Given an image and candidate generations, the figure shows the process for captioner fine-tuning by distilling from a CLIP-based evaluator.

However, LLMs often exhibit a tendency to generate biased and potentially harmful text. To solve this issue, some works have attempted to align models with human judgment through reinforcement learning [38, 54, 56, 71], also designing methodologies for efficient fine-tuning to tackle the substantial memory requirements inherent in training LLMs [25, 35].

3 Proposed Method

Preliminaries. Self-critical sequence training (SCST [45]) is a traditional training paradigm for image captioning. It consists of a two-step training methodology which pre-trains a captioner using a time-wise cross-entropy loss with respect to ground-truth sequences, and fine-tunes the same network by maximizing the CIDEr score [57] using an reinforcement learning approach. Recently, it has been applied also with learnable metrics [17, 47, 48] such as CLIP-S [17], which employs a CLIP [41] embedding space trained to align the embeddings of 400M images and caption pairs. Consequently, a high similarity between a pair of visual-textual CLIP embeddings means that the image-caption pair is highly correlated as well. On the other hand, reinforcement learning from human feedback (RLHF [38]) has been shown to be effective in making LLMs behave more like humans. It starts with a self-supervised pre-trained LLM, then goes through a supervised training phase, and finally, a fine-tuning stage using reinforcement learning. This last step is focused on improving the quality of generated responses by maximizing the score given by a reward model trained to imitate human judgment when comparing two candidate answers. We refer to Appendix A for more details about SCST, RLHF, and image captioning metrics based on contrastive embedding spaces, *i.e.* CLIP-S [17] and PAC-S [47].

Motivation. While adopting significantly different technical choices, there are striking conceptual similarities between the modern RLHF paradigm employed in LLMs and the traditional SCST approach employed in image captioning. Both approaches, indeed, employ reinforcement learning to optimize a reward function, which nevertheless in SCST is a hand-crafted metric, while in RLHF is a learned function from human data. While using RLHF in captioning is impracticable due to the insufficient amount of human preference data to train the reward model (see the comparison with RLHF in Appendix C), contrastive-based learnable metrics offer a compelling alternative to it, as they show a significant alignment with human judgment [47]. Our proposal solves this issue by distilling a reward model from a pre-trained captioning evaluator, considering pairwise relationships from candidate captions. In addition, it also avoids model collapse which is frequent in SCST (cf. Fig. 1).

Deriving the fine-tuning objective. Following recent works on LLM alignment [38], we aim at fine-tuning a captioner f_θ with a Proximal Policy Optimization (PPO) objective [49], where given an image I and a caption s' sampled from the model, the environment produces a reward $r(s',I)$ through a reward model. In addition, we add a per-token KL penalty with the output of the pre-trained model to mitigate overoptimization of the fine-tuned captioner to the reward model. Our objective is therefore defined as

$$\max_{f_\theta} \mathbb{E}_{I\sim\mathcal{D},s'\sim f_\theta(\cdot|I)}\left[r(s',I)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[f_\theta(s'|I)||f_*(s'|I)\right], \tag{1}$$

where β controls the deviation from the pre-trained model, termed as f_*. As it can be seen, the second term has a crucial role, as it prevents the fine-tuned model f_θ from deviating from the distribution on which the reward model is accurate, and prevents the captioner from hacking it, *i.e.* collapsing to high-rewarded answers.

Under this objective, it can be shown [42] that the optimal solution to the fine-tuning problem is given by a model f_r defined as

$$f_r(s'|I) = \frac{1}{Z(I)} f_*(s'|I)\exp\left(\frac{1}{\beta}r(s',I)\right), \tag{2}$$

where $Z(I) = \sum_s f_*(s|I)\exp\left(\frac{1}{\beta}r(s,I)\right)$ is the partition function over possible captions. Although the partition function is difficult to estimate, we can still manipulate Eq. 2 to express the reward function in terms of the optimal captioner, the pre-trained captioner, and the partition function, as follows:

$$r(s',I) = \beta\log\frac{f_r(s'|I)}{f_*(s'|I)} + \beta\log Z(I). \tag{3}$$

Defining a distilled reward model. Since we do not have access to sufficiently large human preference data, defining the reward model in a purely data-driven way would be cumbersome. Instead, we learn our reward model by *distilling* it from a contrastive-based captioning evaluator $\mathcal{E}$. We assume that, given an image and a candidate sentence (I,s'), the evaluator returns a matching score $\mathcal{E}(s',I)$ proportional to the similarity between s' and I.

Given a dataset $\mathcal{D}$ comprising images, we let the captioner generate $k+1$ candidate captions (*e.g.* through beam search). Then, for each image, we select the caption with the highest score according to $\mathcal{E}$ and denote it as s'_w (*i.e.* "winner"). The others, instead, are denoted as $\{s'_{l_i}\}_{i=1}^k$ (*i.e.* "losers"). Based on the evaluator, we define a reward model which distinguishes between the winner caption s'_w and the loser captions $\{s'_{l_i}\}_i$. To make the reward model more robust and accurate, we also impose that it can predict the *relative quality distances* between the winner and the loser captions. Formally, we define our reward model through the following objective:

$$\mathcal{L}_R(r) = -\mathbb{E}\left[\log\sigma\left(\sum_{i=1}^{k}\gamma_i(r(I,s'_w) - r(I,s'_{l_i}))\right)\right], \tag{4}$$

where the expectation is taken over images in the dataset and winner and loser captions. Also, γ_i weights the relative distance between the winner caption s'_w and the i-th loser caption s'_{l_i} according to the evaluator $\mathcal{E}$. Specifically, it is computed as a normalized probability distribution between score distances, as follows:

$$\gamma_i = \mathrm{softmax}_{s'_{l_1},\dots,s'_{l_k}}\left(\frac{\mathcal{E}(I,s'_w) - \mathcal{E}(I,s'_{l_i})}{\tau}\right), \tag{5}$$

where τ is a temperature parameter. Clearly, considering that γ_i sum up to 1, the reward model objective can be rewritten as

$$\mathcal{L}_R(r) = -\mathbb{E}\left[\log\sigma\left(r(I,s'_w) - \sum_{i=1}^{k}\gamma_i r(I,s'_{l_i})\right)\right]. \tag{6}$$

Overall loss function. Following [42], we learn the reward model directly into the captioner. Recalling that the Bradley-Terry model depends only on the difference in rewards between two completions and that γ_i are a valid probability distribution, we replace the definition of $r(s',I)$ as a function of the optimal fine-tuned and pre-trained captioner (Eq. 3) into the reward model objective (Eq. 6), and obtain the final fine-tuning loss of DiCO as

$$L(f_\theta, f_*) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{f_\theta(s'_w|I)}{f_*(s'_w|I)} - \beta\sum_{i=1}^{k}\gamma_i\log\frac{f_\theta(s'_{l_i}|I)}{f_*(s'_{l_i}|I)}\right)\right], \tag{7}$$

where, noticeably, the unknown partition function $Z(I)$ has been cancelled out. Furthermore, the obtained fine-tuning loss, while being derived from the optimal solution to a PPO objective, can be directly optimized through gradient descent, without the need of employing reinforcement learning techniques.

Comparing DiCO with SCST and RLHF. DiCO fine-tunes a captioning model by aligning it to a contrastive-based evaluator while avoiding over-parametrization and model collapse. In comparison with SCST and RLHF, its unique feature is that of *distilling a reward model from an external evaluator* by *learning it directly inside of the captioner*. Further, this is done by avoiding the usage of reinforcement learning at fine-tuning time, which is common to both SCST and RLHF. Differently from RLHF, also, caption candidates are directly sampled from the model, so that a dataset of human-annotated preferences can be avoided. Finally, differently from SCST, DiCO embeds a regularizer to prevent the fine-tuned model from deviating too much from the pre-trained captioner.

4 Experiments

4.1 Experimental Setting

Datasets. All experiments are performed on the COCO dataset [31], using the standard splits defined in [21] with 5,000 images for both test and validation and the rest for training. We report our experimental results on the test set of COCO. Further, we refer the reader to Appendix C for results on six additional datasets, namely nocaps [1], VizWiz [16], TextCaps [52], Conceptual Captions 3M (CC3M) [50], FineCapEval [13], and Flickr30k [67].

Evaluation metrics. In addition to the standard image captioning metrics like BLEU [39], METEOR [9], and CIDEr [57], we employ two CLIP-based scores, namely CLIP-S [17] and PAC-S [47], in both their reference-free and reference-based versions, using the ViT-B/32 backbone for both metrics (also see Appendix A). Moreover, following recent works [10, 24], we measure the quality of generated captions in distinguishing images in a dataset and compute the percentage of the times the image corresponding to each generated caption is retrieved among the first K retrieved items. This is done by ranking the images in terms of CLIP similarity between visual and textual embeddings, using the CLIP ViT-B/32 model, and computing recall at K with $K = 1, 5, 10$. We also compute the mean reciprocal rank

Model			Reference-based Metrics					Reference-free Metrics					
			B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	R@1	R@5	R@10	MRR
Standard Captioners		**Backbone**											
CLIP-VL [6]		RN50×4	40.2	29.7	134.2	0.820	0.862	0.770	0.826	24.0	48.9	61.5	34.8
COS-Net [29]		RN101	42.0	30.6	141.1	0.814	0.870	0.758	0.832	25.8	52.3	64.9	37.1
PMA-Net [8]		ViT-L/14	43.0	30.6	144.1	0.814	0.869	0.755	0.821	-	-	-	-
LLM-based Captioners		**Backbone**											
ZeroCap [53]		ViT-B/32	2.3	10.1	15.1	0.771	0.800	0.810	0.816	-	-	-	-
ClipCap [30]		ViT-B/32	32.3	28.1	108.5	0.809	0.862	0.766	0.833	27.1	53.3	65.5	38.3
SmallCap [43]		ViT-B/32	37.0	27.9	119.7	0.804	0.863	0.748	0.826	23.1	48.2	60.0	33.7
MiniGPT-v2 [10]		ViT-g/14	18.8	24.6	80.4	0.795	0.848	0.752	0.818	27.4	52.0	63.0	37.9
BLIP-2 [24]		ViT-g/14	43.7	32.0	145.8	0.823	0.877	0.767	0.837	31.4	57.5	69.1	42.7
CLIP-based Optimization	**Reward**	**Backbone**											
Cho *et al.* (SCST) [43]	CLIP-S	RN50	6.3	19.7	11.2	0.786	0.823	**0.843**	0.837	43.2	71.9	82.3	55.5
Cho *et al.* (SCST) [43]	CLIP-S+Gr.	RN50	16.9	24.9	71.0	0.792	0.849	0.779	0.839	35.3	63.4	75.2	47.4
SCST	CLIP-S	RN50	14.3	24.7	3.1	0.765	0.830	0.804	0.837	36.9	64.9	75.9	48.7
SCST	PAC-S	RN50	18.5	26.5	32.2	0.785	0.849	0.799	0.860	**44.3**	73.2	83.4	56.5
DiCO (Ours)	CLIP-S	RN50	20.7	25.7	78.9	**0.811**	0.852	0.815	0.842	37.5	66.6	78.1	49.8
DiCO (Ours)	PAC-S	RN50	**22.7**	**27.0**	**79.8**	0.801	**0.865**	0.797	**0.869**	**44.3**	**73.9**	**84.2**	**56.8**
SCST	CLIP-S	ViT-B/32	11.4	23.1	1.1	0.778	0.830	**0.851**	0.846	43.4	70.8	81.1	55.1
SCST	PAC-S	ViT-B/32	20.3	27.1	40.7	0.796	0.858	0.810	0.870	50.0	77.6	87.0	61.8
DiCO (Ours)	CLIP-S	ViT-B/32	22.6	27.0	81.7	**0.817**	0.861	0.825	0.858	46.3	74.0	83.7	58.0
DiCO (Ours)	PAC-S	ViT-B/32	**23.7**	**27.3**	**84.8**	0.810	**0.872**	0.814	**0.882**	**52.9**	**80.8**	**89.5**	**64.8**
SCST	CLIP-S	ViT-L/14	10.2	23.0	1.1	0.793	0.827	**0.865**	0.834	43.3	70.7	80.5	55.0
SCST	PAC-S	ViT-L/14	22.3	**28.4**	51.1	0.801	0.861	0.805	0.862	46.7	74.7	84.8	58.8
DiCO (Ours)	CLIP-S	ViT-L/14	21.4	27.1	82.6	**0.824**	0.863	0.837	0.856	46.5	74.7	84.7	58.4
DiCO (Ours)	PAC-S	ViT-L/14	**25.2**	**28.4**	**89.1**	0.815	**0.875**	0.812	**0.877**	**50.9**	**78.7**	**87.6**	**62.9**

Table 1: Comparison with state-of-the-art models on the COCO test set. Bold font indicates the best results among captioners optimized via CLIP-based rewards with comparable backbones, while underlined indicates the overall best results.

(MRR) for each generated caption: higher MRR scores indicate that captions are more discriminative and therefore usually more detailed.

Implementation and training details. Our baseline architecture is a standard Transformer model with 3 layers in both encoder and decoder, a hidden dimensionality equal to 512, and 8 attention heads. To extract visual features, we use either RN50, ViT-B/32, or ViT-L/14 pre-trained with a CLIP-based objective [41]. Our code is based on the popular Hugging Face Transformers[1] library. All experiments are performed using the Adam optimizer, initially pre-training all the models with cross-entropy. During fine-tuning, we use a batch size of 16, a fixed learning rate equal to $1 \cdot 10^{-6}$, and a beam size of 5 (*i.e.* the number k of looser captions is set to 4). For efficiency, we train with ZeRo memory offloading and mixed-precision [36]. Unless otherwise specified, the β parameter is set to 0.2 and the ViT-L/14 backbone is used to extract visual features. The temperature parameter τ defined in Eq. 5 is set to $1/(3 \cdot 10^2)$. Early stopping is performed according to the reference-based version of the CLIP metrics used as reward. Ablation studies and analyses with different hyperparameters are reported in Appendix B.

4.2 Comparison with the State of the Art

Results on COCO test set. We compare our model trained with the proposed DiCO strategy with other state-of-the-art solutions. We restrain the comparison by only considering captioning models that use CLIP-based visual features to encode images, which have proven to be the most widely employed choice in recent works. In particular, we include some

[1] https://huggingface.co/docs/transformers

SCST: A rusted rusty rusty rusty rusty rusty scissors in a garden with plants in the background.
DiCO (Ours): A rusted scissors sticking out of a metal fence with plants in the background.

SCST: Several cows grazing in a field on a mountain range with a mountain in distance under mountain range in background.
DiCO (Ours): A herd of cattle grazing on a lush green hill next to a large mountain range.

SCST: A little boy with headphones sitting in front of a computer with headphones on a desk in corner with headphones on background.
DiCO (Ours): A little boy with headphones sitting at a desk using a computer.

SCST: A person holding up a clear plastic plastic container filled with sugar covered donuts with people in background in background on background.
DiCO (Ours): A hand holding a plastic container filled with sugar covered donuts.

Figure 3: Qualitative results on COCO sample images, using PAC-S as reward.

recent standard image captioning models exclusively trained on the COCO dataset with a standard XE+SCST training paradigm like CLIP-VL [51], COS-Net [29], and PMA-Net [5]. Moreover, we compare with LLM-based captioning models focused on zero-shot generation capabilities such as ZeroCap [55], lightweight architectures like ClipCap [37] and SmallCap [43], or large-scale training paradigms such as the recently proposed MiniGPT-v2 [11] and BLIP-2 [27] models. To directly compare our solution with other CLIP-based optimization strategies, we also report the results of our base model trained with SCST using CLIP-S or PAC-S as reward and those of the model proposed in [13] in which a standard Transformer is optimized via SCST with a CLIP-based reward, eventually regularizing the training with an additional score that considers the grammatical correctness of generated sentences.

Results are shown in Table 1, including our model trained with DiCO using RN50, ViT-B/32, and ViT-L/14 as visual backbones. Notably, all versions of our model achieve better results than other methods optimized with CLIP-based rewards on almost all evaluation metrics. For example, when comparing our solution optimized via PAC-S reward with SCST and the model proposed in [13], we can notice how not only DiCO improves the performance in terms of standard metrics (*e.g.* 79.8 CIDEr points using RN50 features vs. 32.2 and 71.0 respectively obtained by SCST and [13]), but also obtains increased retrieval-based scores indicating that captions generated by our model are more discriminative and detailed than those generated by competitors. Additionally, DiCO leads to the overall best results on reference-free metrics also surpassing huge models trained on millions or even billions of data like MiniGPT-v2 and BLIP-2, further confirming the effectiveness of our training strategy. To validate the quality of generated captions, we report in Fig. 3 some qualitative results on sample images from the COCO dataset. DiCO can generate more descriptive and detailed captions while reducing repetitions and grammatical errors typically generated using SCST. We refer to Appendix E for additional qualitative results.

Human-based and LLM-based evaluations. As a complement of standard metrics, we also perform a user study and an evaluation based on a widely used LLM (*i.e.* GPT-3.5). To perform the user study, we present the users with an image and a pair of captions, one generated by our model and the other generated by a competitor, and ask them to select the preferred caption judging in terms of (1) *helpfulness* (*i.e.* which caption is most helpful to someone who can not see the image), and (2) *correctness* (*i.e.* which caption is more correct both in terms of grammar and consistency with the image). Users could also state that captions are equivalent on one or both evaluation axes. In this case, 0.5 points are given to both captions. To perform LLM-based evaluation, instead, we leverage the Turbo version of GPT-3.5 and directly ask it to evaluate a pair of captions taking into account the

corresponding reference sentences. In particular, we ask the LLM to return a score between 0 and 100 for each caption between the two in the prompt, where one is generated by our model and the other by a competitor, and use this score to compute the number of times GPT-3.5 prefers our solution against a competitor and vice versa. If the score is the same for both captions, we give 0.5 points to both of them. To force the model to produce a more accurate evaluation, we also ask it to produce a reason for each score, which has been shown to lead to ratings that correlate well with human judgment [9].

	Humans		
	Helpfulness	Correctness	GPT-3.5
ZeroCap [53]	20.3	27.8	20.8
SmallCap [43]	27.8	36.1	50.0
MiniGPT-v2 [11]	33.3	42.9	44.8
BLIP-2 [27]	49.1	48.6	51.2
Cho *et al.* (CLIP-S Reward) [13]	11.2	17.9	21.5
Cho *et al.* (CLIP-S+Gr. Reward) [13]	41.3	36.7	43.0
SCST (PAC-S Reward)	44.6	40.6	48.5

Table 2: Percentage of times a caption from a competitor is preferred against that generated by our proposal, using either human-based evaluations or GPT-3.5. Our solution is preferred more than 50% of the time in almost all cases.

Table 2 shows one-to-one comparisons between our model and one of the considered competitors in terms of both human-based and LLM-based evaluations. Results are reported on a subset of 1,000 images randomly taken from the COCO test set. For each comparison, we report the percentage of times a caption generated by one of the competitors is preferred against the one generated by our solution with PAC-S reward. As it can be seen, DiCO is almost always preferred more than 50% of the time, having a comparable number of preferences only when compared with BLIP-2. When instead considering other CLIP-based optimized models, captions generated by our solution are selected in a considerable number of cases from both human evaluators and GPT-3.5 (*e.g.* more than 55-60% compared to [13] with CLIP-S+Grammar reward). Additional details are reported in Appendix D.

Additional results on grammar metrics. Besides the semantic coherence between images and their descriptions, we compare our method against SCST from the point of view of the fluency and grammatical correctness of the generated captions. To this end, in Table 3 we report the average number of n-gram repetitions per caption (*i.e.* n_i with $i = 1,2,3,4$), com-

			Semantic			Grammar					
Model	**Reward**	**Backbone**	C	CLIP-S	PAC-S	n_1	n_2	n_3	n_4	RE	%Correct
SCST	CLIP-S	RN50	3.1	0.804	0.837	11.762	5.168	2.809	1.518	6.0	24.7
SCST	PAC-S	RN50	32.2	0.799	0.860	5.453	1.588	0.645	0.288	1.6	71.6
DiCO	CLIP-S	RN50	78.9	**0.815**	0.842	**1.583**	**0.143**	**0.039**	**0.015**	**0.1**	**96.1**
DiCO	PAC-S	RN50	**79.8**	0.797	**0.869**	2.051	0.219	0.055	0.017	**0.1**	94.4
SCST	CLIP-S	ViT-B/32	1.1	**0.851**	0.846	11.166	3.566	1.232	0.395	1.5	3.9
SCST	PAC-S	ViT-B/32	40.7	0.810	0.870	5.078	1.443	0.584	0.260	1.6	73.3
DiCO	CLIP-S	ViT-B/32	81.7	0.825	0.858	**1.938**	0.230	0.071	0.026	0.2	94.8
DiCO	PAC-S	ViT-B/32	**84.8**	0.814	**0.882**	1.939	**0.190**	**0.048**	**0.014**	**0.1**	**96.4**
SCST	CLIP-S	ViT-L/14	1.1	**0.865**	0.834	8.788	2.114	0.716	0.248	1.0	2.6
SCST	PAC-S	ViT-L/14	51.1	0.805	0.862	8.788	4.611	1.200	0.479	1.3	72.6
DiCO	CLIP-S	ViT-L/14	82.6	0.837	0.856	**1.710**	**0.142**	**0.039**	**0.014**	**0.1**	**95.4**
DiCO	PAC-S	ViT-L/14	**89.1**	0.812	**0.877**	2.107	0.218	0.056	0.017	**0.1**	94.3

Table 3: Comparison on semantic and grammar metrics. n_i means i-gram repetitions. Results are reported on the COCO test set.

Model	Reference-based							Reference-free	
	B-4	M	R	C	S	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
Up-Down [3]	36.3	27.7	56.9	120.1	21.4	0.787	0.848	0.723	0.803
SGAE [53]	39.0	28.4	58.9	129.1	22.2	0.796	0.855	0.734	0.812
AoANet [19]	38.9	29.2	58.8	129.8	22.4	0.797	0.857	0.737	0.815
$\mathcal{M}^2$ Transformer [14]	39.1	29.8	58.3	131.3	22.6	0.793	0.852	0.734	0.813
COS-Net [45]	42.0	30.6	60.6	141.1	**24.6**	0.814	0.870	**0.758**	**0.832**
PMA-Net [5]	43.0	30.6	61.1	144.1	24.0	0.814	0.869	0.755	0.821
Transformer (SCST)	43.6	30.8	61.0	143.3	23.2	0.809	0.866	0.750	0.826
Transformer (DiCO w/ $\beta = 0.05$)	43.2	31.2	61.1	**144.2**	24.4	0.815	0.871	0.756	0.831
Transformer (DiCO w/ $\beta = 0.1$)	**43.7**	31.2	61.2	143.8	24.5	**0.817**	**0.872**	0.757	**0.832**
Transformer (DiCO w/ $\beta = 0.2$)	**43.7**	**31.3**	**61.3**	143.5	24.4	0.816	**0.872**	0.756	0.831

Table 4: Comparison with standard captioners using CIDEr-based optimization.

puted using the `nltk` language toolkit[2]. We also include the Repetition Evaluation (RE) proposed in [63], which measures the redundancy of n-grams inside a caption (where $n = 4$ as in the original paper). Additionally, we employ the text encoder from [13] and present the percentage of captions classified as grammatically correct (*i.e.* %Correct). Experiments across different backbones confirm that SCST reaches high scores on the optimized metrics, but collapses to predictions that exhibit many repetitions, undermining the fluency of the generated text. DiCO does not suffer from the same problem, keeping low values for repetitions while showcasing state-of-the-art performance over the reward metrics.

CIDEr-based optimization. Finally, we assess whether our training paradigm can also be applied using the CIDEr score as reward, as usually done in standard image captioning approaches. Results are reported in Table 4, showing the performance of a standard Transformer model fine-tuned with the classical SCST procedure and that of other captioners. For completeness, we also include the results in terms of ROUGE [30] and SPICE [2] which are typically used in standard image captioning evaluation. In this case, we apply DiCO with different β values on the same baseline architecture used in previous experiments (*i.e.* a vanilla Transformer with 3 layers in both encoder and decoder). Interestingly, our solution achieves better results than SCST also in this setting, with 144.2 CIDEr points vs. 143.3 obtained by SCST. As an additional result, DiCO reaches better or comparable performance to that obtained by recent captioning models based on more complex architectures and optimized via SCST, thus proving to be a valid alternative also in a standard CIDEr-based setting.

5 Conclusion

We presented DiCO, a novel fine-tuning strategy for image captioning which aligns a model to a learnable evaluator with high human correlation. Our approach optimizes a distilled reward model by solving a weighted classification problem directly inside the captioner, which allows it to capture fine-grained differences between multiple candidate captions. Experimental results on several datasets, conducted through automatic metrics and human evaluations, validate the effectiveness of our approach, which can generate more descriptive and detailed captions than competitors. At the same time, it achieves state-of-the-art results when trained to optimize traditional reference-based metrics.

[2] https://www.nltk.org/

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. This work has been conducted under a research grant co-funded by Altilia s.r.l. and supported by the PNRR-M4C2 (PE00000013) project "FAIR - Future Artificial Intelligence Research", by the PNRR project "ITSERR - Italian Strengthening of Esfri RI Resilience" (CUP B53C22001770006), and by the PRIN project "MUSMA" (CUP G53D23002930006 - M4C2 I1.1), all funded by EU - Next-Generation EU.

References

[1] 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.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016.

[3] 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 *CVPR*, 2018.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005.

[5] Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In *ICCV*, 2023.

[6] Federico Bolelli, Lorenzo Baraldi, and Costantino Grana. A Hierarchical Quasi-Recurrent approach to Video Captioning. In *IPAS*, 2018.

[7] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The Revolution of Multimodal Large Language Models: A Survey. In *ACL Findings*, 2024.

[8] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *CVPR Workshops*, 2024.

[9] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating Image Captions with Large Language Models. In *EMNLP*, 2023.

[10] David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. IC^3: Image Captioning by Committee Consensus. In *EMNLP*, 2023.

[11] Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. *arXiv preprint arXiv:2310.09478*, 2023.

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023.

[13] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained Image Captioning with CLIP Reward. In *NAACL*, 2022.

[14] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020.

[15] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*, 2023.

[16] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In *ECCV*, 2020.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021.

[18] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling Up Vision-Language Pre-training for Image Captioning. In *CVPR*, 2022.

[19] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In *ICCV*, 2019.

[20] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, et al. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[22] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating Images with Multimodal Language Models. In *NeurIPS*, 2023.

[23] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding Language Models to Images for Multimodal Inputs and Outputs. In *ICML*, 2023.

[24] Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. Guiding Image Captioning Models Toward More Specific Captions. In *ICCV*, 2023.

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022.

[27] 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.

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, 2020.

[29] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *CVPR*, 2022.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshops*, 2004.

[31] 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.

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023.

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[34] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved Image Captioning via Policy Gradient Optimization of SPIDEr. In *ICCV*, 2017.

[35] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *AI Open*, 2023.

[36] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training. In *ICLR*, 2018.

[37] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[38] 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.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] 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.

[42] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024.

[43] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation. In *CVPR*, 2023.

[44] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. In *ICLR*, 2015.

[45] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017.

[46] Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. FuseCap: Leveraging Large Language Models to Fuse Visual Data into Enriched Image Captions. In *WACV*, 2024.

[47] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, 2023.

[48] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues. In *ECCV*, 2024.

[49] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[50] 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.

[51] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How Much Can CLIP Benefit Vision-and-Language Tasks? In *ICLR*, 2022.

[52] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, 2020.

[53] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Trans. PAMI*, 45(1):539–559, 2022.

[54] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *NeurIPS*, 2023.

[55] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *CVPR*, 2022.

[56] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[57] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015.

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[59] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*, 2022.

[60] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*, 2022.

[61] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *EMNLP*, 2022.

[62] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *ACL*, 2023.

[63] Yilei Xiong, Bo Dai, and Dahua Lin. Move Forward and Tell: A Progressive Generator of Video Descriptions. In *ECCV*, 2018.

[64] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015.

[65] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 2019.

[66] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018.

[67] 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.

[68] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[69] Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distinctive Image Captioning via CLIP Guided Group Optimization. In *ECCV*, 2022.

[70] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *arXiv preprint arXiv:2303.06594*, 2023.

[71] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*, 2019.