Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis

Davide Bucciarelli[1], Nicholas Moratelli[1], Marcella Cornia[1], Lorenzo Baraldi[1], and Rita Cucchiara[1,2]

[1] University of Modena and Reggio Emilia, Italy
[2] IIT-CNR, Italy
{name.surname}@unimore.it

Abstract. The task of image captioning demands an algorithm to generate natural language descriptions of visual inputs. Recent advancements have seen a convergence between image captioning research and the development of Large Language Models (LLMs) and Multimodal LLMs – like GPT-4V and Gemini – which extend the capabilities of text-only LLMs to multiple modalities. This paper investigates whether Multimodal LLMs can supplant traditional image captioning networks by evaluating their performance on various image description benchmarks. We explore both the zero-shot capabilities of these models and their adaptability to different semantic domains through fine-tuning methods, including prompt learning, prefix tuning, and low-rank adaptation. Our results demonstrate that while Multimodal LLMs achieve impressive zero-shot performance, fine-tuning for specific domains while maintaining their generalization capabilities intact remains challenging. We discuss the implications of these findings for future research in image captioning and the development of more adaptable Multimodal LLMs.

Keywords: Image Captioning · Multimodal LLMs · Parameter Efficient Fine-tuning.

1 Introduction

The task of image captioning requires an algorithm to describe a visual input in natural language. Over the last years, researchers have made remarkable progress in developing approaches specifically devoted to image description, with the aim of increasing visual encoding capabilities [5,63], finding proper architectures and multimodal connectors [11,36], and improving linguistic fluency, relevance, and adherence to a desired description style [21,31,48]. These advancements have not only enhanced the ability of models to generate accurate and contextually appropriate captions but have also contributed to bridging the gap between visual understanding and language generation.

Given the inherent multimodal nature of the task, the evolution of image captioning research has many times intersected that of Large Language Models (LLMs) [31,46,52] and, more recently, it is crossing that of Multimodal LLMs

(MLLMs) [3,14,34,65], which are their natural multimodal extension. The surge of sophisticated text-only LLMs [12, 17, 61], and particularly their capacity for in-context learning, has indeed encouraged researchers to broaden the scope of these models to encompass multiple modalities, both as inputs and outputs. This expansion has led to the development of cutting-edge models such as GPT-4V [1] and Gemini [6], which showcase state-of-the-art performance in various multimodal tasks and applications.

A significant example of the interplay between image captioning research and large multimodal models can be found by analyzing the evolution of their training methodologies. Starting from 2015, captioners have been fine-tuned with reinforcement learning objectives to maximize non-differentiable metrics like CIDEr [62], *i.e.* through self-critical sequence training (SCST) [53]. This approach, although conducted on a smaller scale, closely resembles that of the reinforcement learning from human feedback (RLHF) paradigm [49], which has been a fundamental tool to develop instruction-aligned LLMs and, ultimately, increase their utility in real-world scenarios. In RLHF, indeed, the LLM is fine-tuned to align itself to a trained reward model – replace the trained reward with a non-differentiable metric, and you immediately get a fine-tuning strategy that is conceptually equivalent to SCST. Coming to MLLMs, the similarities between the two tasks are evident, with research on MLLMs questioning the best way to fuse visual features into a Transformer decoder [13, 59, 60] – a question that image captioning literature has been tackling several times in the past.

Considering this overlap in technical goals, the recent surge of MLLMs and the variety of multimodal tasks that they can perform, a natural question arises: are MLLMs the definite replacement for image captioning networks? In this paper, we contribute to finding an answer to this question, by analyzing the performance of different MLLMs on multiple image description benchmarks. In addition to investigating the zero-shot performance of pre-trained models in comparison with that of a state-of-the-art captioner, we also move a step forward and test the adaptation capabilities of MLLMs when it comes to adhering to the classical description style of captioners, which is very concise, grammatically correct, and focuses on everyday objects. Also, we test whether this adaptation can still maintain the generalization capabilities of the MLLM and work well on other semantic domains.

To test this adaptation – or, better to say, *personalization* – capabilities of MLLMs, we employ different fine-tuning strategies, ranging from full fine-tuning to a wide range of parameter-efficient fine-tuning (PEFT) techniques, including prompt learning [33], prefix tuning [35], low-rank adaptation [27], and weight-decomposition in low-rank adaptation [44]. By assessing the performance of current MLLMs for image description, and their adaptation capabilities to different semantic and description domains, we aim to provide a comprehensive evaluation of whether these models can truly replace specialized image captioning networks. Our findings reveal that, while MLLMs exhibit strong zero-shot performance across various benchmarks, their adaptability to specific description styles through fine-tuning is still an open challenge. We conclude by discussing

the implications of our results for future research directions in both image captioning and the development of more versatile and adaptive MLLMs.

2 Related Work

Standard Image Captioning. Early efforts in image captioning primarily focused on detecting key objects within a scene to populate predefined templates [58, 67]. Subsequent research evolved to employ RNN-based encoder-decoder architectures, where visual input was encoded using a CNN and then exploited to condition the generation process through an RNN [30, 63]. These methodologies were further refined with the introduction of attention-based strategies, which applied attention mechanisms to either spatial regions [5] or semantic graphs [66]. Recently, Transformer-based architectures have emerged as the standard in image captioning [19,20,28], often in combination with CLIP-based [51] visual features which demonstrate increased semantics leading to better performance [10,11,36]. Despite being a well-established task in literature, it has historically struggled with generalization and tends to produce very literal captions. In this regard, recent approaches have proposed fine-tuning strategies guided by open-vocabulary metrics [26,54,55] to enhance the descriptive capacity of the models [18,31,48].

Image Captioning with Multimodal LLMs. In the last year, MLLMs have become predominant in performing a wide range of vision-and-language tasks including visual dialogue, image description, and visual question answering [13]. Almost all existing MLLMs, indeed, adopt large-scale architectures to tackle the challenge of bridging visual and language modalities, connecting a pre-trained LLM with a large-scale visual encoder (*i.e.* typically CLIP or its variants).

MLLMs can be categorized considering the type of multimodal connections they employ. Following the widely-used LLaVA model family [41–43], the prevalent strategy in this domain involves using an MLP [65, 70] or a single linear layer [16, 39] to establish multimodal connections. Several variations have been introduced, such as LLaMA-Adapter [23] that proposes an alternative attention mechanism with zero gating, and the approach introduced by Cha *et al.* [15] that replaces linear layers with convolutions. Another significant category of models is built upon Q-Former architecture introduced in [34]. On this line, mPLUG-Owl [68] streamlines Q-Former by incorporating a visual abstractor component that condenses visual information into distinct trainable tokens. Similarly, Qwen-VL [8] employs a single-layer cross-attention module with learnable queries to compress visual features. Other approaches integrate dense cross-attention blocks within the existing pre-trained layers of the LLM [3, 7]. This method is often used in conjunction with a Perceiver model [29], reducing the number of visual tokens before their integration into the language model.

Despite their rapid evolution, the performance analysis of MLLMs in image captioning remains significantly under-explored. Only a few MLLMs are directly trained and evaluated on this task using standard benchmarks, while others treat image description as an inherent capability. On a different line, some recent

studies [40,64,69] have started to estimate the hallucination degree of MLLMs, a crucial aspect in this domain given the level of detail they can generate even when describing input images. Unlike existing literature, this paper aims to analyze standard MLLMs when generating image descriptions and explore how they can be better adapted to the task by comparing different fine-tuning techniques.

Parameter Efficient Fine-tuning Techniques. Adapting LLMs to a specific task may prove impractical due to the substantial computational resources required for complete fine-tuning. In such scenarios, the adoption of PEFT techniques represents a feasible alternative. The principal strategies include (i) prompt-tuning that entails learning a small set of vectors, used as soft prompts fed into the model before the input text [25,33,35,45,47]; (ii) LoRA [27], where pre-trained model weights remain frozen while introducing trainable rank decomposition matrices into each layer; (iii) QLoRA [22], designed to reduce the memory footprint of LLMs while preserving full 16-bit fine-tuning task performance and (iv) DoRA [44] that decomposes a pre-trained model into magnitude and directional components, utilizing LoRA for directional adjustments, thereby efficiently reducing the count of trainable parameters. Despite the availability of a diverse range of techniques, to the best of our knowledge, there has been no experimental analysis conducted to compare them. This paper investigates the impact of PEFT optimization on model performance when customizing the MLLM for a specific task (*i.e.* that of image captioning).

3 Proposed Method

3.1 Preliminaries

An MLLM usually takes as input a multimodal input, comprising both image and text, and generates a textual output in an autoregressive manner. Formally, the architecture is trained to model a probability distribution $p(w_t|I, w_0, w_1, ..., w_{t-1}, \theta)$, where θ denotes the parameters of the model, I represents an input image, and $w_0, .., w_{t-1}$ denotes the textual prompt. The textual prompt usually includes a pre-defined system-level prompt and a question related to the input image, given by the user. Clearly, a standard MLLM can only rely on the user prompt, the input image, and the knowledge stored in its internal parameters (*i.e.* θ) to accommodate requests.

In the rest of the paper, we employ LLaVA [43] as our reference MLLM. LLaVA exploits the capabilities of a pre-trained LLM (*i.e.* Vicuna [17]) and a pre-trained visual model (*i.e.* a CLIP-based visual encoder [51]), which are interconnected through an MLP adapter, in charge of converting CLIP features to dense input tokens. For an input image I, therefore, LLaVA utilizes a pre-trained CLIP visual encoder E_v, extracts a dense grid of visual features $Z_v = E_v(I)$, which is then projected via a learnable MLP to produce a sequence of dense embedding tokens $v_o, v_1, ..., v_N$. Finally, these are prepended to the system prompt, and the full sequence of visual and textual tokens is then given as input to the LLM component of the model.

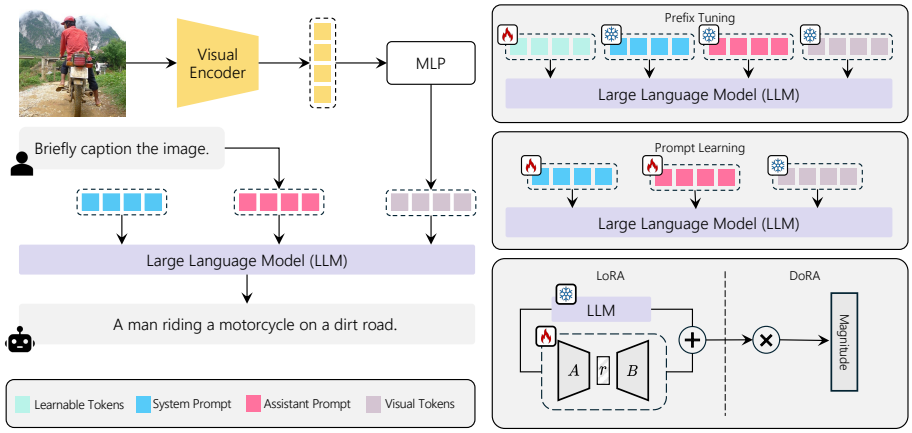


Fig. 1: Overview of our approach. We investigate whether Multimodal LLMs can supplant traditional captioners by assessing their adaptability to different semantic domains and through the usage of different adaptation techniques.

3.2 Personalization Strategies

To adapt MLLMs to specific description styles and semantic domains, we investigate several parameter-efficient fine-tuning (PEFT) techniques.

Prompt Learning. To adapt the MLLM to perform classical image captioning, the most straightforward option is to enrich the input context by injecting learnable vectors into its embedding. This is usually done by adding *new* embedding vectors to an existing prompt, which are initialized from scratch and trained through stochastic gradient descent. In our preliminary experiments, however, we found it beneficial to fine-tune the user prompt and the system prompt embeddings, rather than injecting new embeddings which might be more complex to initialize. Formally, the distribution of the MLLM is conditioned on visual tokens, a system prompt, and a trainable user prompt, leading to

$$p(w_t|\overbrace{v_o, v_1, ..., v_N}^{\text{Visual tokens}}, \underbrace{w_0, w_1, ..., w_{t-1}}_{\text{Learnable System prompt}}, \overbrace{e_0, e_1, ..., e_\tau}^{\text{Learnable User prompt}}), \tag{1}$$

where $e_0, ..., e_\tau$ represents the trainable embeddings of the user prompt. The set of trainable parameters θ^*, in this case, is simply $\theta^* = \{e_0, ..., e_\tau\}$. Differently from the standard formulation of MLLMs, by fine-tuning a portion of the input context, we allow the model to generate more specific answers.

Prefix Tuning. Differently from the previous case, in this case we add a sequence of learnable embeddings to every layer of the Transformer decoder of the MLLM. While this formulation does not allow a straightforward meaningful initialization of the embeddings, like in the case of prompt learning, it comes with the advantage of injecting trainable knowledge at different layers of the architecture, which might increase the degree of adaptation of the model. Formally,

the input embeddings of the i-th layer are adapted as

$$\boldsymbol{h}^i = \left[\underbrace{h_0^i, h_1^i, ..., h_T^i}_{\text{Regular input embeddings}} \quad , \quad \underbrace{e_0^i, e_1^i, ..., e_\tau^i}_{\text{Learnable embeddings}} \right], \tag{2}$$

where $e_0^i, ..., e_\tau^i$ represents the trainable embeddings of a given layer. In this case, the set of trainable parameters θ^* is defined as $\theta^* = \bigcup_{i=1}^{L} \{e_0^i, ..., e_\tau^i\}$, where L represents the number of layers of the MLLM.

LoRA. We now turn to a different approach to adaptation, where instead of adding learnable tokens or embeddings, either at the input layer or at every layer, we aim at fine-tuning all the weights θ of the architecture. To constrain the computational complexity of the adaptation, and keep a safe regularization against overfitting, we fine-tune a low-rank adaptation of weight matrices [27] instead of directly performing a full fine-tuning.

Without loss of generality, in the following, we describe our approach for the case of a fully-connected layer, which are a key ingredient of many Transformer-based models as they build up the attention operator. Given a pre-trained layer f, with weight $W_0 \in \theta, W_0 \in \mathbb{R}^{d \times k}$ and bias $b \in \theta$, which applies a transformation $f(x) = xW_0^\intercal + b$ to its input tensor $x \in \mathbb{R}^k$, we re-parametrize its transformation during the training phase by adding a low-rank trainable component $\tilde{W}$, initialized from zero. We then fine-tune only the low-rank decomposition, leaving the rest of the layer frozen. Formally,

$$f(x) = x \underbrace{W_0^\intercal}_{\text{❄}} + x \overbrace{\tilde{W}^\intercal}^{*} + \underbrace{b}_{\text{❄}} \text{. with } \tilde{W} = BA, \tag{3}$$

where A and B provide a bottleneck that creates a low-rank decomposition which is trainable (denoted with $*$, above), with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$ and r is the rank of the decomposition. During fine-tuning, W_0 and b are kept frozen (❄) and we backpropagate gradient only on A and B. These are respectively initialized with a Gaussian and zero initialization, so that, at the beginning of fine-tuning, $\tilde{W} = BA$ is a zero matrix and f behaves exactly as in the pre-trained state.

We apply the low-rank re-parametrization to all fully connected layers of the MLLMs. The set of trainable parameters θ^* is, therefore, defined as $\theta^* = \bigcup_i \{A_i, B_i\}$, where i runs on all fully-connected layers of the network.

DoRA. As an alternative to the low-rank adaptation mentioned above, we also investigate a weight-decomposed low-rank adaptation (DoRA) [44], which is known in the literature for outperforming LoRA also when fine-tuning MLLMs such as LLaVa [43]. Specifically, DoRA initially decomposes the pre-trained weight W_0 into its magnitude and directional components, then fine-tunes both of them. Because the directional component is larger in terms of parameter numbers, it fine-tunes it using a LoRA decomposition.

Formally, given a pre-trained weight $W_0 \in \mathbb{R}^{d\times k}$ from a layer f, the matrix weight is fine-tuned as

$$f(x) = x \cdot \overbrace{m}^{*} \frac{\overbrace{W_0}^{\text{❄}} + \overbrace{BA}^{*}}{\| \underbrace{W_0}_{\text{❄}} + \underbrace{BA}_{*} \|}, \tag{4}$$

where m indicates a trainable magnitude component and, again, the $*$ notation indicates the only trainable weights. Clearly, DoRA adds an additional term to the set of trainable weights for the overall architecture, leading to $\theta^* = \bigcup_i \{m_i, A_i, B_i\}$, where i runs on all fully-connected layers of the network.

4 Experimental Evaluation

4.1 Datasets and Evaluation Metrics

In our experiments, we employ a set of five commonly used datasets for the image captioning task, namely COCO [38], nocaps [2], CC3M [56], VizWiz [24], and TextCaps [57]. All our training experiments are conducted on COCO, while evaluations are reported on all considered datasets. For nocaps, CC3M, VizWiz, and TextCaps, we report the results on the validation set of each dataset.

COCO [38]. It is the standard dataset for the task and contains more than 120,000 images, each of them annotated with five different captions. In our experiments, we follow the splits provided by Karpathy *et al.* [30], where 5,000 images are reserved for the validation set and 5,000 for the test set.

nocaps [2]. This dataset consists of 15,100 elements from the Open Images [32] validation and test sets, annotated with human-annotated captions. Images are divided into validation and test splits, respectively with 4,500 and 10,600 images.

CC3M [56]. It is a large-scale image captioning dataset composed of roughly 3.3 million images sourced from the web. The validation set is composed of approximately 14,000 elements. Each image is paired with an alt-text description, that usually focuses on the main concept of the image.

VizWiz [24]. This dataset aims to test the ability of image captioning models to assist blind people. It features 39,000 images taken by visually impaired users, each paired with five ground-truth captions. Images are grouped into training, validation, and test sets with 23,431, 7,750, and 8,000 elements each.

TextCaps [57]. It includes over 145,000 captions for more than 28,000 images, where each caption requires understanding and interpreting the textual content present in the image. The images are split into training, test, and validation sets, with respectively 21,953, 3,166, and 3,289 elements.

Evaluation. For what concerns the evaluation metrics, we employ standard captioning scores, namely BLEU-4 [50], METEOR [9], ROUGE [37], CIDEr [62], and SPICE [4]. Additionally, we report the results in terms of CLIP-Score [26] that does not rely on ground-truth captions and tends to favor a high degree of descriptiveness at the expense of grammatical fluency and correctness [18, 48].

4.2 Implementation and Training Details

In our experiments, we focus on the smallest versions of the LLaVA models, selecting LLaVA-v1.5-7B [41] and LLaVA-v1.6-7B [42], both equipped with the Vicuna-7B [17] language model and CLIP ViT-L/14@336 [51] as visual encoder. To fine-tune the considered MLLMs, we utilize the standard token-level cross-entropy loss for all training experiments. Unless otherwise specified, the CLIP-based visual encoder, the MLP adapter, and the LLM layers are kept frozen.

Both considered MLLMs are trained using the personalization strategies described in Sec. 3.2. All results are compared against the original MLLM (*i.e.* the MLLM without any fine-tuning stages), tested in zero-shot using a fixed prompt[3] to generate the output image description. For prompt learning, we implement a slight variation of the usual framework. Specifically, we initialize the learnable prompt with the same sentence used for the original MLLM, concatenated with the standard system prompt of the model, which is then optimized during training. This results in 48 learnable tokens. In prefix tuning, we employ a fixed prompt identical to the one used for the zero-shot evaluation and concatenate a single learnable token at the beginning of the input for each decoder layer of the LLM, trimming the extra output token at each layer. For LoRA and DoRA, we use identical parameters (*i.e.* the rank r is set to 128 and the scaling parameter α is equal to 256) and keep the same input prompt used in the other settings.

To ensure fair comparison among different fine-tuning methods, we maintain a consistent training setup across all tests. In particular, all training experiments are performed on a single node with four 64GB NVIDIA A100 GPUs. During each training phase, the model undergoes four epochs on the COCO dataset, with the model exhibiting the lowest validation loss being selected at the end of each run. We use a batch size of 32 for all experiments, employing gradient accumulation steps as needed for memory constraints. The standard SGD optimizer is utilized, along with a cosine learning rate scheduler, with a maximum learning rate equal to 2×10^{-2} and a minimal value of 1×10^{-5}.

4.3 Experimental Results

As previously mentioned, we consider two existing MLLMs (*i.e.* LLaVA-v1.5 and LLaVA-v1.6) and evaluate their performance across different captioning datasets. In addition to the results using the original model, we fine-tune each MLLM using four different personalization strategies, namely prompt learning, prefix tuning, LoRA, and DoRA (cf. Sec. 3.2). Moreover, we report the results of each MLLM fine-tuned by directly optimizing all parameters of both the vision-to-language adapter and all layers of the LLM. To have a direct comparison with a standard captioning model not based on MLLMs, we also consider the performance of a standard Transformer-based encoder-decoder model trained from scratch on the COCO dataset. Specifically, we follow the architecture of the CLIP-Captioner proposed in [10] and train it using the same visual encoder used

[3] In our experiments, we employ "`Briefly caption the image`" as input prompt.

Table 1: In-domain results on the COCO dataset. Bold font indicates the best results for the same MLLM, while underline indicates the overall best scores.

Model	PEFT	B-4	M	R	C	S	CLIP-S
CLIP-Captioner [10]	-	.	29.9	58.2	126.2	22.6	0.752
LLaVA-v1.5-7B [41]	-	18.7	22.4	46.6	53.9	24.8	**0.806**
	Prompt Learning	31.9	22.4	53.5	96.3	23.1	0.774
	Prefix Tuning	27.3	22.3	52.0	85.8	23.4	0.782
	LoRA	36.1	23.2	56.0	105.7	24.7	0.777
	DoRA	36.4	23.3	56.3	106.1	24.5	0.778
	Full Fine-tuning	**38.2**	**23.5**	**57.3**	**111.4**	**25.1**	0.771
LLaVA-v1.6-7B [42]	-	6.8	16.2	31.2	16.4	12.5	0.755
	Prompt Learning	33.0	22.8	54.5	100.0	24.1	0.774
	Prefix Tuning	25.5	19.4	48.4	74.1	19.7	**0.784**
	LoRA	36.9	23.3	56.6	108.6	24.8	0.771
	DoRA	36.1	23.1	56.1	106.4	24.7	0.777
	Full Fine-tuning	**38.5**	**23.4**	**57.5**	**112.3**	**25.2**	0.774

in the considered LLaVA models (*i.e.* CLIP ViT-L/14@336). In the following, we first report the results on the standard COCO dataset, thus following an in-domain evaluation. Then, we analyze the generalization capabilities to out-of-domain datasets showing the results on nocaps, CC3M, VizWiz, and TextCaps.

In-Domain Evaluation. Table 1 presents the in-domain results on the COCO dataset for the models under consideration. As indicated, the highest scores for each MLLM are highlighted in bold, while the overall best scores across all models and methods are underlined.

Firstly, examining the results of LLaVA-v1.5, it can be noticed that the highest scores are achieved by the full fine-tuning strategy with a CIDEr score of 111.4 points. Similar results are achieved by the LLaVA-v1.6 model where full fine-tuning again yields the highest scores in almost all metrics with a CIDEr score equal to 112.3 points. Among the other fine-tuning strategies, LoRA and DoRA are the ones that achieve the best results on both LLaVA versions. Overall, these results are not surprising: training a larger number of parameters using image-caption pairs from a specific dataset and evaluating the results on other pairs from the same dataset naturally leads to the best performance. This is further confirmed when taking into account the results achieved by the CLIP-Captioner, which are generally higher than all the others reported in the table. Since this model is trained from scratch on the COCO dataset, directly assessing the performance on the test set of the same dataset leads to the best scores on standard captioning metrics.

Conversely, the best results in terms of CLIP-S are achieved by the original LLaVA-v1.5 model tested in a zero-shot manner on COCO. This underscores the descriptive capabilities of MLLMs which can generate highly detailed and usually long captions describing a given image. The descriptive style of common MLLMs, however, is far from the one present in standard captioning benchmarks like COCO which contains concise and timely descriptions, as demonstrated by the CIDEr scores of both zero-shot LLaVA-v1.5 and LLaVA-v1.6 models (*i.e.* 53.9

Table 2: Out-of-domain results on nocaps and CC3M datasets. Bold font indicates the best results for the same MLLM, while underline indicates the overall best scores.

		nocaps						**CC3M**					
Model	**PEFT**	B-4	M	R	C	S	CLIP-S	B-4	M	R	C	S	CLIP-S
CLIP-Captioner [10]	-	7.3	15.9	32.4	77.1	19.7	0.693	1.9	9.0	16.7	29.1	9.5	0.651
LLaVA-v1.5-7B [41]	-	6.5	**21.3**	31.2	61.8	23.6	**0.793**	1.2	**11.0**	14.1	15.1	10.3	**0.743**
	Prompt Learning	8.4	20.1	33.2	85.9	24.2	0.763	**2.0**	10.4	**14.9**	**22.9**	10.8	0.699
	Prefix Tuning	**9.0**	20.7	33.3	84.1	24.6	0.772	1.8	10.5	14.4	21.3	**10.9**	0.720
	LoRA	8.5	20.0	33.2	86.2	24.8	0.751	1.7	10.1	14.3	21.6	10.6	0.714
	DoRA	8.5	20.2	**33.4**	**87.3**	**24.9**	0.758	1.8	10.2	14.5	22.0	10.8	0.717
	Full Fine-tuning	8.7	20.1	33.2	86.8	24.6	0.752	1.7	10.0	14.3	20.8	10.4	0.720
LLaVA-v1.6-7B [42]	-	2.3	14.8	19.6	18.1	10.8	0.749	1.2	9.8	11.7	9.6	7.5	0.731
	Prompt Learning	8.8	**20.4**	**34.1**	**91.0**	**25.5**	0.758	**1.8**	10.1	14.4	20.7	**11.0**	0.729
	Prefix Tuning	6.4	17.5	29.4	65.6	22.3	0.752	1.4	9.7	12.1	15.1	8.6	**0.738**
	LoRA	8.8	**20.4**	33.7	89.8	25.3	0.762	**1.8**	**10.2**	14.4	21.7	10.2	0.718
	DoRA	**8.9**	**20.4**	33.8	**91.0**	25.4	**0.763**	**1.8**	**10.2**	**14.5**	**21.8**	10.6	0.723
	Full Fine-tuning	8.5	20.1	33.2	86.6	24.6	0.753	1.6	9.9	14.2	20.7	10.1	0.724

Table 3: Out-of-domain results on VizWiz and TextCaps datasets. Bold font indicates the best results for the same MLLM, while underline indicates the overall best scores.

		VizWiz						**TextCaps**					
Model	**PEFT**	B-4	M	R	C	S	CLIP-S	B-4	M	R	C	S	CLIP-S
CLIP-Captioner [10]	-	17.3	19.5	40.8	35.7	9.4	0.650	14.9	17.2	35.9	34.3	11.6	0.651
LLaVA-v1.5-7B [41]	-	15.4	16.1	40.3	41.1	**15.0**	**0.758**	15.0	**18.2**	38.5	43.5	**19.3**	**0.802**
	Prompt Learning	**20.2**	15.4	**42.5**	**49.9**	14.0	0.742	**20.5**	16.2	40.4	51.5	17.0	0.758
	Prefix Tuning	19.0	**15.5**	41.9	47.2	14.2	0.744	19.6	17.0	**40.5**	**51.7**	17.8	0.773
	LoRA	19.2	14.8	42.0	43.9	13.4	0.728	17.1	15.0	38.0	40.2	15.2	0.730
	DoRA	20.1	15.1	42.4	47.4	13.7	0.733	18.5	15.5	39.0	43.5	16.0	0.738
	Full Fine-tuning	19.5	14.7	41.8	43.7	14.7	0.725	17.1	15.0	38.1	39.3	15.2	0.728
LLaVA-v1.6-7B [42]	-	6.9	13.0	28.6	16.9	13.7	0.725	5.9	14.1	26.4	20.9	12.0	**0.765**
	Prompt Learning	**20.9**	**15.4**	**43.0**	**50.4**	**14.4**	**0.740**	18.6	**15.5**	39.3	44.4	**16.5**	0.739
	Prefix Tuning	14.5	13.5	36.7	35.7	11.9	0.724	12.9	13.8	33.8	31.8	13.8	0.737
	LoRA	20.7	15.2	42.6	48.0	13.9	0.736	**20.7**	15.2	**42.5**	**47.8**	13.8	0.737
	DoRA	20.6	15.2	42.4	47.6	13.8	0.737	18.1	**15.5**	39.0	43.7	16.0	0.746
	Full Fine-tuning	19.8	14.8	41.9	44.1	13.3	0.721	16.8	14.9	38.3	38.9	15.1	0.726

and 16.4, respectively) which are significantly lower than those obtained by all fine-tuned versions.

Generalization to Out-of-Domain Settings. Tables 2 and 3 present the results on out-of-domain settings, including nocaps and CC3M datasets (Table 2) and VizWiz and Textcaps benchmarks (Table 3). Also in this case, for both LLaVA-v1.5 and LLaVA-v1.6, we compare different fine-tuning strategies with the MLLM tested in zero-shot on the considered datasets and also include the results from the CLIP-Captioner approach.

As it can be seen, the overall trend is significantly different from the one observed for in-domain evaluation with the standard captioning model trained from scratch on COCO achieving lower results than almost all fine-tuning strategies. The only exception is the CC3M dataset that, however, contains less curated captions than the other datasets, thus leading to less interpretable patterns. Fine-tuning the entire LLM does not lead to the best results in this case, underscoring the need to find viable fine-tuning alternatives to preserve good gen-

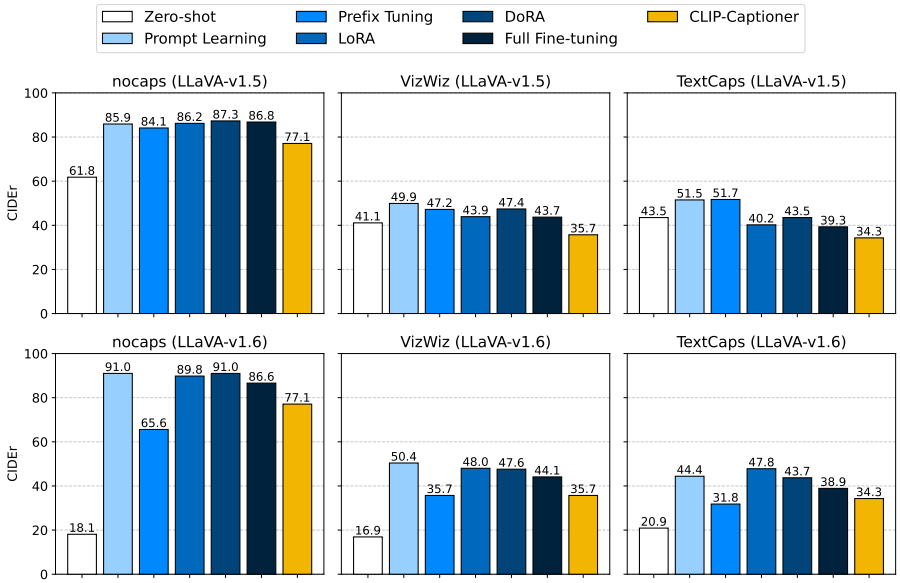


Fig. 2: Comparison between CIDEr scores achieved by the different versions of LLaVA-v1.5 (first row) and by those of LLaVA-v1.6 (second row) on out-of-domain datasets including nocaps, VizWiz, and TextCaps.

eralization capabilities in out-of-domain settings. Among the PEFT techniques under consideration, prompt learning is the one achieving the best results on average on all datasets and both LLaVA versions. Similar performances are obtained by LoRA and DoRA fine-tuning strategies, which however fail to preserve high results, especially on the VizWiz dataset.

Also in these settings, captions generated by zero-shot MLLMs are confirmed to be far from ground-truth image descriptions contained in each considered dataset, as demonstrated by the low scores in terms of standard captioning metrics achieved by these models. These results highlight the need of proper fine-tuning strategies to adapt MLLMs for the task of image captioning and the necessity of novel evaluation protocols that take into account the different descriptive styles of the textual descriptions generated by these models.

A different visualization of the results is shown in Fig. 2 where we compare the CIDEr scores of the considered models on nocaps, VizWiz, and TextCaps. Notably, the scores achieved by the zero-shot MLLMs are always below all other fine-tuned versions. This is particularly evident with LLaVA-v1.6 which tends to generate longer captions and is more prone to hallucinations. Overall, prompt learning better generalizes across different out-of-domain datasets, always achieving the best or second-best results in almost all settings and considering both LLaVA-v1.5 and LLaVA-v1.6.

Qualitative Results. Finally, we report some qualitative results in Fig. 3 and 4. Specifically, in Fig. 3 we compare captions generated by the zero-shot MLLM

GT: A girl in a pink shirt standing near a blue metal sculpture.
Zero-shot: A woman and children are standing in front of a blue fence, with the woman holding a child. They are near a pole and a sign.
Prompt learning: A woman and children are posing for a picture.

GT: The TV screen has a man who is reporting live.
Zero-shot: A man is standing in front of a TV, which is displaying a basketball game. The TV screen shows the score and the teams playing.
Prompt learning: A man on a television screen talking about the Rockets and the Grizzlies.

GT: A bottle of wine is labeled with the name MAGNE.
Zero-shot: A bottle of wine is placed in a wooden box, with the label showing that it is a 2013 vintage.
Prompt learning: A bottle of wine with a label that says Magne.

GT: Bike riders passing Burger King in city street.
Zero-shot: A group of people riding bicycles down a street, with a man on a bike in front of a Subway sandwich shop.
Prompt learning: A group of people riding bikes on a city street.

GT: A poster for the book named in god I trust.
Zero-shot: Colorful poster featuring Psalm 46:10, "Be still, and know that I am God".
Prompt learning: A poster that says "In God I Trust".

GT: A sign that is green and says "Welcome to Burnaby".
Zero-shot: Welcome to Burnaby: A city where snowflakes are a symbol of the season, and graffiti is a common sight.
Prompt learning: A sign that says "Welcome to Burnaby" with graffiti on it.

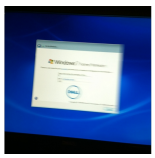
Fig. 3: Sample image descriptions generated by the zero-shot MLLM in comparison with those generated by the MLLM fine-tuned with prompt learning. For reference, we also report a ground-truth caption (GT) associated to each image.

(both LLaVA-v1.5 and LLaVA-v1.6) with those generated by the MLLM fine-tuned with prompt learning, which demonstrates to be one of the best fine-tuning solutions for the image captioning task. For completeness, we also include a sample ground-truth caption. Notably, while captions generated by the zero-shot MLLM are generally longer and more detailed, they often contain hallucinations or fail to well describe the visual content of the input image. For example, in the second row-left sample, the zero-shot MLLM correctly identifies the bottle of wine but also reports the vintage year which however is not shown in the image. Similarly, in the third row-right sample, the zero-shot MLLM correctly reads the text written in the image but provides details on the city whose name appears in the written text. These additional details, however, do not help to better describe the visual content appearing in the scene. In both cases, instead, the fine-tuned version of the MLLM can generate a concise caption, while still describing the key concepts depicted in the images.

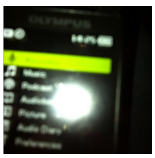
In Fig. 4, we report additional qualitative results, in this case comparing the MLLM fine-tuned with prompt learning with the predictions generated by the CLIP-Captioner approach and those generated by the MLLM after a full fine-tuning stage. As it can be seen, fine-tuning the MLLM with a PEFT-based solution leads to captions that are generally more detailed, while still preserving the concise and timely style of standard captioning benchmarks. On the contrary, training from scratch a captioning model on COCO or directly optimizing all MLLM parameters causes a loss of generality, especially when the model should

CLIP-Captioner: A man with mustaches.
Full fine-tuning: A man and a woman holding up pink and brown objects.
Prompt learning: A woman and a man with fake mustaches on their mouths.

CLIP-Captioner: A woman in uniform standing in front of a computer screen.
Full fine-tuning: A woman is walking down a runway with a dress on.
Prompt learning: A woman walking down a runway in a dress with polka dots.



CLIP-Captioner: A screenshot of a computer screen.
Full fine-tuning: A computer screen with a Windows logo on it.
Prompt learning: A computer screen with a blue background and a Windows error message.



CLIP-Captioner: A screenshot of my computer.
Full fine-tuning: A digital music player with a yellow menu.
Prompt learning: A phone screen with a music app and a podcast app.

CLIP-Captioner: A watch on a table.
Full fine-tuning: A watch sitting on a table with a broken glass.
Prompt learning: A Rolex watch is laying on a table.

CLIP-Captioner: A black and white page.
Full fine-tuning: A page of a book with a picture of a machine.
Prompt learning: A page from a magazine that is about Burman gearboxes.

Fig. 4: Qualitative results on sample images from nocaps (first row), ViZWiz (second row), and TextCaps (third row). We compare captions predicted by the CLIP-Captioner model [10], the MLLM after full fine-tuning, and the MLLM after prompt learning.

describe objects or concepts that are not present in the training dataset. These results confirm from a qualitative point of view the effectiveness of using appropriate fine-tuning strategies to adapt an existing MLLM to the image captioning task and show that utilizing a full fine-tuning of the model is not the preferable choice in this setting.

4.4 Computational Analysis

Finally, in Table 4 we present a computational and energy consumption analysis for the fine-tuning strategies under consideration on LLaVA-v1.5. For each PEFT strategy, we report the number of trainable parameters along with the energy consumed during training, measured in Kilowatt-hours (kWh). Energy consumption is detailed both for the entire training process on the COCO training split (*i.e.* four epochs in our experiments) and for the epochs up to the best checkpoint, selected based on validation loss. As it can be seen, prompt learning and prefix tuning are the least computationally demanding strategies. Furthermore, while training with LoRA or DoRA consumes a similar amount of energy as full fine-tuning when considering all epochs, DoRA generally converges in fewer iterations, leading to lower overall energy consumption.

5 Conclusion

This paper has explored the intersection of image captioning and the rapidly evolving landscape of Multimodal LLMs, assessing their potential as effective

Table 4: Computational analysis in terms of trainable parameters and energy consumed during training. Energy consumption is reported for the entire training process as well as for the epochs up to the best checkpoint (with the latter shown in parentheses). All experiments have been conducted on four 64GB NVIDIA A100 GPUs.

Model	PEFT	Trainable Params	Energy Consumption (kWh)
LLaVA-v1.5-7B [41]	Prompt Learning	19.7k	12.4 (3.1)
	Prefix Tuning	13.1k	12.1 (3.0)
	LoRA	319.8M	59.5 (28.5)
	DoRA	321.2M	58.2 (14.5)
	Full Fine-tuning	7B	64.3 (32.1)

replacements for specialized image captioning networks. Through comprehensive experiments and analyses across multiple image description benchmarks, we have demonstrated the limitations of common MLLMs when applied to this task without specific training. In fact, captions generated by standard MLLMs are often prone to hallucinations and struggle to adhere to the concise, grammatically correct, and object-focused description style characteristic of standard image captioning datasets. While standard fine-tuning schemes can improve performance to some extent, they often come at the cost of reduced generalization. To bridge this gap, we have analyzed the effectiveness of various PEFT techniques for adapting MLLMs to the image captioning task, showing that these solutions generally yield better results in terms of both coherence with ground-truth captions and generalization to out-of-domain settings. Our findings suggest the necessity for further research to design effective strategies for adapting existing MLLMs in this domain, mainly focusing on improving their ability to generate accurate, concise, and hallucination-free image captions while maintaining generalization across different domains.

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” and by the PRIN 2022-PNRR M4C2-I1.1 project “MUCES - a MUltimedia platform for Content Enrichment and Search in audiovisual archives” (CUP E53D23016290001), both funded by EU - Next-Generation EU.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: Novel object captioning at scale. In: ICCV (2019)

3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a Visual Language Model for Few-Shot Learning. In: NeurIPS (2022)
4. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation. In: ECCV (2016)
5. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
6. Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023)
7. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv preprint arXiv:2308.01390 (2023)
8. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966 (2023)
9. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshops (2005)
10. Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis. In: CVPR Workshops (2022)
11. Barraco, M., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In: ICCV (2023)
12. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
13. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The Revolution of Multimodal Large Language Models: A Survey. In: ACL Findings (2024)
14. Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In: CVPR Workshops (2024)
15. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced Projector for Multimodal LLM. In: CVPR (2024)
16. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: MiniGPT-v2: Large Language Model As a Unified Interface for Vision-Language Multi-task Learning. arXiv preprint arXiv:2310.09478 (2023)
17. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (2023)
18. Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., Bansal, M.: Fine-grained Image Captioning with CLIP Reward. In: NAACL (2022)
19. Cornia, M., Baraldi, L., Cucchiara, R.: SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In: ICRA (2020)
20. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining Transformer-based Image Captioning Models: An Empirical Analysis. AI Communications **35**(2), 111–129 (2022)

21. Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N.C., Franzon, F., Baroni, M.: Cross-Domain Image Captioning with Discriminative Finetuning. In: CVPR (2023)
22. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs. In: NeurIPS (2023)
23. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv preprint arXiv:2304.15010 (2023)
24. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning Images Taken by People Who Are Blind. In: ECCV (2020)
25. Hambardzumyan, K., Khachatrian, H., May, J.: Warp: Word-level Adversarial Reprogramming. arXiv preprint arXiv:2101.00121 (2021)
26. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)
27. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2021)
28. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: ICCV (2019)
29. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021)
30. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
31. Kornblith, S., Li, L., Wang, Z., Nguyen, T.: Guiding Image Captioning Models Toward More Specific Captions. In: ICCV (2023)
32. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Ferrari, V.: The Open Images Dataset V4. IJCV **128**(7), 1956–1981 (2020)
33. Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning. In: EMNLP (2021)
34. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv preprint arXiv:2301.12597 (2023)
35. Li, X.L., Liang, P.: Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv preprint arXiv:2101.00190 (2021)
36. Li, Y., Pan, Y., Yao, T., Mei, T.: Comprehending and Ordering Semantics for Image Captioning. In: CVPR (2022)
37. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: ACL Workshops (2004)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in context. In: ECCV (2014)
39. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. arXiv preprint arXiv:2311.07575 (2023)
40. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning Large Multi-Modal Model with Robust Instruction Tuning. In: ICLR (2024)
41. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: CVPR (2024)
42. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)

43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: NeurIPS (2023)
44. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: DoRA: Weight-Decomposed Low-Rank Adaptation. In: ICML (2024)
45. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: GPT understands, too. AI Open (2023)
46. Mokady, R., Hertz, A., Bermano, A.H.: ClipCap: CLIP Prefix for Image Captioning. arXiv preprint arXiv:2111.09734 (2021)
47. Moratelli, N., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Are Learnable Prompts the Right Way of Prompting? Adapting Vision-and-Language Models with Memory Optimization. IEEE Intelligent Systems (2024)
48. Moratelli, N., Caffagni, D., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization. In: BMVC (2024)
49. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
50. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: ACL (2002)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
52. Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation. In: CVPR (2023)
53. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-Critical Sequence Training for Image Captioning. In: CVPR (2017)
54. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In: CVPR (2023)
55. Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues. In: ECCV (2024)
56. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: ACL (2018)
57. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: TextCaps: A Dataset for Image Captioning with Reading Comprehension. In: ECCV (2020)
58. Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
59. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., et al.: Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv preprint arXiv:2406.16860 (2024)
60. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In: CVPR (2024)
61. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971 (2023)
62. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-Based Image Description Evaluation. In: CVPR (2015)
63. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)

64. Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al.: VIGC: Visual Instruction Generation and Correction. In: AAAI (2024)
65. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: CogVLM: Visual Expert for Pretrained Language Models. arXiv preprint arXiv:2311.03079 (2023)
66. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-Encoding Scene Graphs for Image Captioning. In: CVPR (2019)
67. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2t: Image parsing to text description. Proceedings of the IEEE **98**(8) (2010)
68. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv preprint arXiv:2304.14178 (2023)
69. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination Correction for Multimodal Large Language Models. arXiv preprint arXiv:2310.16045 (2023)
70. Zhao, B., Wu, B., Huang, T.: SVIT: Scaling up Visual Instruction Tuning. arXiv preprint arXiv:2307.04087 (2023)