LMCAP: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting

Rita Ramos[†] **Bruno Martins**[†] **Desmond Elliott**[⋆]

[†]INESC-ID, Instituto Superior Técnico, University of Lisbon
[⋆]Department of Computer Science, University of Copenhagen
`ritaparadaramos@tecnico.ulisboa.pt`

Abstract

Multilingual image captioning has recently been tackled by training with large-scale machine translated data, which is an expensive, noisy, and time-consuming process. Without requiring any multilingual caption data, we propose LMCAP, an *image-blind* few-shot multilingual captioning model that works by prompting a language model with retrieved captions. Specifically, instead of following the standard encoder-decoder paradigm, given an image, LMCAP first retrieves the captions of similar images using a multilingual CLIP encoder. These captions are then combined into a prompt for an XGLM decoder, in order to generate captions in the desired language. In other words, the generation model does not directly process the image, instead processing retrieved captions. Experiments on the XM3600 dataset of geographically diverse images show that our model is competitive with fully-supervised multilingual captioning models, without requiring any supervised training on any captioning data.

1 Introduction

The task of image captioning has witnessed impressive performance gains with the trend of large-scale encoder-decoder models and vision-and-language pre-training (Li et al., 2022; Wang et al., 2021; Hu et al., 2022; Wang et al., 2022). Despite all of this progress, existing models are mostly available on English or are specialised for other high-resource languages. This limits the access to the technology for a broader range of languages that exist in the world. Moreover, the current mainstream trend results in design decisions and methods that may only work well for English-centric datasets or the couple of languages for which captioning data is available (Ruder, 2020). There is a need to develop multilingual image captioning models that can serve speakers of different languages.

Still, scaling captioning models to a wide variety of languages involves different challenges. One major limitation is the lack of multilingual image-caption pairs of clean labelled data for training the models. One possible solution is to automatically translate the existing English datasets (Thapliyal et al., 2022). While effective, this approach can result in models that learn translation artefacts, and perpetuates an English-centric perspective instead of encouraging the use of geographically diverse concepts that are not overly specific to the western culture (Liu et al., 2021). Moreover, with or without automatic translations, training captioning models with multilingual data can be expensive, given the amount of data and number of parameters needed to mitigate the *curse of multilinguality* (Conneau et al., 2019; Goyal et al., 2021).

This paper presents LMCAP, an *image-blind* multilingual image captioning model that does not require any training specific for image captioning. We propose an efficient method that reuses a pre-trained multilingual language model and adapts it to the vision-and-language captioning setting. Our work is motivated by the recent "Socratic Models" framework (Zeng et al., 2022), in which different models can be combined through text prompting (e.g., image captioning can be achieved by prompting a language model with a set of visual concepts extracted from the predictions of a vision model). Different from the original Socratic Models, our approach is inspired by retrieval-augmented generation (Lewis et al., 2020; Izacard et al., 2022). Specifically, a multilingual language model generates captions given a prompt consisting of the captions retrieved from similar images, and a demonstration of how to produce a caption in the desired language. We note here that this is an *image-blind* approach, i.e. the language model producing the caption does not actually process the image.

Our main contributions are as follows: (1) We propose a few-shot multilingual image captioning approach named LMCAP, that re-uses pre-trained models without requiring any training specific for

image captioning; (2) To the best of our knowledge, LMCAP is the first captioning model with retrieval-augmented generation in a multilingual setting, and in a few-shot setting of captioning; (3) We report on experiments with the XM3600 benchmark (Thapliyal et al., 2022) of human-authored captions and geographic diverse images, demonstrating that LMCAP exhibits strong few-shot performance on a wide variety of languages; (4) We further show that LMCAP performs substantially better than the original Socratic Models. Moreover, instead of only achieving competitive performance against other zero-shot models, LMCAP can also compete with a large-scale supervised state-of-art captioning model.

2 Background and Related Work

Image Captioning: The task of automatically generating textual descriptions for input images has been largely explored in English, while multilingual image captioning has only been addressed in a couple of studies (Gu et al., 2018; Thapliyal et al., 2022; Chen et al., 2022). Like in most recent work on image captioning (Li et al., 2022; Wang et al., 2021, 2022), studies addressing multilingual setups have also focused on scaling the size of encoder-decoder models and the amount of training data, resorting to machine translated versions of multimodal data to accommodate multiple languages (Thapliyal et al., 2022). Differently from training a large-scale encoder-decoder model, we follow a few-shot setting with an *image-blind* approach based on prompting.

Few-Shot and Zero-Shot Approaches: Performing few-shot learning by prompting a language model with examples and demonstrations of a task (Brown et al., 2020; Radford et al., 2019; Schick and Schütze, 2020) is an efficient and effective alternative to update model parameters. Similarly to other NLP tasks, recent work in the vision-and-language domain has used prompt-based learning by building on top of pre-trained language and vision models, although usually also involving extra multimodal training (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Jin et al., 2021). In our work, we follow a similar few-shot prompting approach to the recent Socratic Models (Zeng et al., 2022) that do not involve any multimodal training, as described next. In image captioning, there have also been zero-shot methods that similarly to our approach do not involve any training, by relying

on prompts or adaptations over the decoding algorithms, such as ZeroCap (Tewel et al., 2021) and ConZic (Zeng et al., 2023). However, these models work for English and not for the multilingual captioning setting.

Socratic Models: Zeng et al. (2022) proposed the Socratic Models (SMs) framework, where different multimodal pre-trained models communicate via zero-shot or few-shot prompting. For the task of image captioning, SMs generate captions by prompting a language model (i.e., GPT-3 (Brown et al., 2020)) with information about the input image obtained with another pre-trained model (i.e., CLIP (Radford et al., 2021)). The visual information is in this way represented into a language-based prompt, containing the number of people presented in the image, the places, objects, and what is the type of image. We explore a similar approach in the multilingual setting by reusing multilingual models, and through a retrieval-based prompt.

Retrieval-augmentation: The knowledge from language models can be adapted and expanded by combing non-parametric knowledge from datastores (i.e., external memories) (Khandelwal et al., 2019; Lewis et al., 2020; Izacard et al., 2022; Ram et al., 2023). The success of conditioning generation with retrieved information, in several different NLP tasks, has inspired some recent studies in image captioning (Ramos et al., 2023a; Fei, 2021; Sarto et al., 2022; Ramos et al., 2023b). The study that is most closely related to our captioning model is SmallCap (Ramos et al., 2023b), an encoder-decoder model that is prompted with retrieved captions as well. However, in image captioning, retrieval-augmentation has mostly being explored with supervised learning and not few-shot learning. Moreover, retrieval-augmentation remains unexplored in the multilingual scenario.

3 Model

Language Model Prompt-based Captioning (LMCAP) is a few-shot multilingual captioning model augmented with retrieval. It involves prompting a Language Model (LM) with captions retrieved from a datastore by a Vision-and-Language Model (VLM). Captions are generated in an *image-blind* manner, without actually processing the visual contents of the input image, instead using a prompt containing the retrieved captions. The method works as follows: first,

given an input image, the VLM is used to find relevant captions in the datastore. Second, the retrieved captions are converted to a language prompt, which is encoded by the multilingual LM to generate captions in a desired language, conditioning the generation on the prompt. Finally, the set of generated captions can be scored by the VLM against the input image, to select the best caption. The main aspects of our approach are shown in Figure 1 and fully detailed next.

Image-Text Retrieval: The input image and a datastore of captions are encoded by a multilingual CLIP (Carlsson et al., 2022), i.e. a VLM that can be used to calculate image-text similarity. In this way, given the encoded data, M-CLIP is used to retrieve the K most similar captions from the datastore. The datastore contains captions associated to diverse images, which can be in English or another language. The retrieved captions will serve to guide a language model as an example of what the predicted caption should resemble, through the use of a prompt and as described next.

Retrieval-augmented Prompting: The retrieved captions, which represent the visual information about the image, are formatted into a prompt for the language model. The prompt starts with fixed N-shot examples and ends with the retrieval information about the input image, to guide the language model. Each shot is a demonstration of how to generate a caption in a desired language for an image, given a set of retrieved captions. After these N-examples, the prompt terminates with the retrieved information about the actual input image. An example of the format of the prompt can be seen in Figure 1 and in more detail in Appendix D. We note that the retrieved captions, either from the fixed N-shot examples or those corresponding to the input image, can be presented in any language or in multiple languages.

Prompting Multilingual Text Generation: The aforementioned prompt is used as input for an XGLM (Lin et al., 2021) pre-trained multilingual autoregressive LM, to generate captions in a given language. XGLM is applied in a few-shot setting, which means that LMCAP does not require any training (i.e., the captions are generated by providing the prompt at inference time to XGLM). Captions are generated in the desired language by including an example in the N demonstrations in the prompt, as shown in Figure 1.

Multilingual Reranking: After the LM generates a set of captions, the multilingual VLM performs a final image–text similarity step to find the caption that best describes the input image. This is based on the same M-CLIP model used for the initial image–text retrieval.

4 Evaluation

In this section, we describe the evaluation of LM-CAP. We describe the experimental setup and results, and we also present ablation studies and further discussions about our approach.

4.1 Experimental Setup

Model: LMCAP uses two pre-trained multilingual models, namely the autoregressive XGLM language model `facebook/xglm-2.9B`, and the multilingual M-CLIP vison-and-language model `xlm-roberta-large-ViT-H-14`, respectively available on HuggingFace (Wolf et al., 2020) and OpenCLIP[1]. Our approach does not require any training, generating captions at inference time using a single NVIDIA V100S 32GB GPU.

To generate a caption in a desired language, XGLM is prompted with retrieved captions extracted by the M-CLIP model. For caption retrieval, the input image and a set of captions from a datastore are both encoded by M-CLIP to perform direct image-text search. The datastore contains English captions from the COCO training set and is indexed offline with the nearest-neighbour search library named FAISS (Johnson et al., 2017), using the index `IndexFlatIP` that does not involve training. A set of K=4 retrieved captions are used in the prompt for the input image, along with a fixed set of N=3-shot examples, as described in Appendix D. Conditioned on the prompt, XGLM generates captions using beam-search decoding with a beam of 3. A set of c=3 candidate captions are re-ranked using M-CLIP, to select the final generated caption in the desired language. The code for LMCAP is made freely available[2].

Datasets: We mainly evaluate our approach on XM3600, i.e. a multilingual image captioning dataset (Thapliyal et al., 2022) featuring geographically-diverse images, collected from Open Images with basis on the regions of 36 languages. For each language, 100 images were selected and annotated with human generated cap-
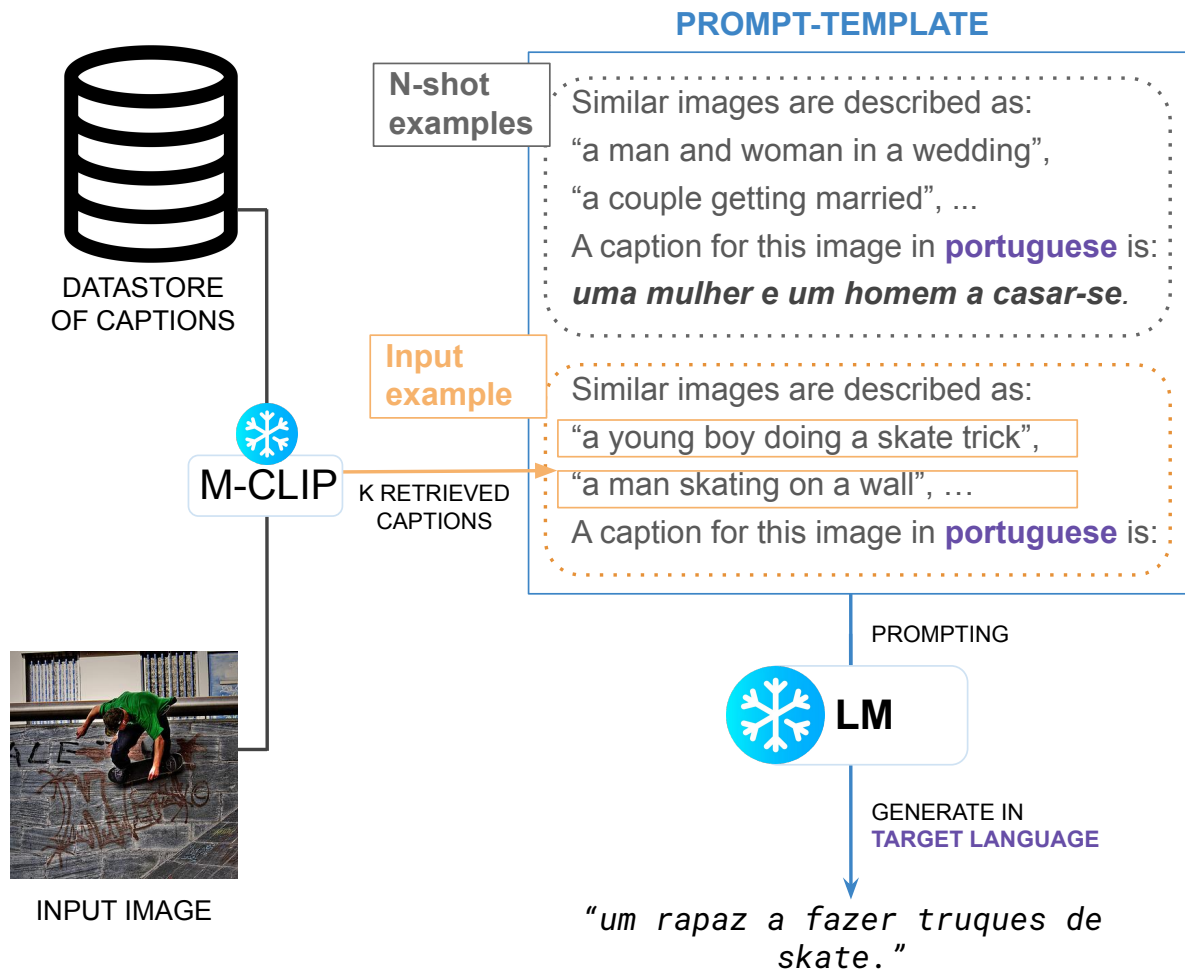
Figure 1: Illustration of the key aspects of LMCAP, a few-shot multilingual image captioning approach that re-uses pre-trained unimodal models without requiring any training. In our *image-blind* approach, a multilingual language model (XGLM) is prompted with information retrieved with a multilingual CLIP model. The prompt contains a set of N-shot examples and K retrieved captions, to guide caption generation in a desired language.

tions, resulting in a total of 3600 images and 261375 captions across the 36 languages. XM3600 does not contain training or validation splits.

For validation and hyperparameter tuning, we relied on the COCO (Chen et al., 2015) validation split (COCO-DEV) from the standard Karpathy splits (Karpathy and Fei-Fei, 2015). For "reference captions", we machine translate the English captions into Spanish, Hindi, and Chinese, using the M2M-100 model (Fan et al., 2021), similarly in spirit to Thapliyal et al. (2022) who used the Google Translate API[3]. We make this development set available to the community at https://github.com/RitaRamo/lmcap. As previously mentioned, we also use the captions from the COCO training set to build the datastore. The datastore simply contains the original English cap-

tions from COCO without incurring in an expensive and noisy machine translation process, unlike in the study from Thapliyal et al. (2022).

Model Assessment and Comparison: We compare LMCAP with the four multilingual models proposed by Thapliyal et al. (2022). These models combine different mT5 (Xue et al., 2020) and ViT (Zhai et al., 2022) versions and are trained in a fully-supervised fashion on COCO-35L and CC3M-35L, i.e., Google's machine translation API versions of the original COCO and CC3M datasets (Chen et al., 2015; Sharma et al., 2018). Specifically, BB+CC combines mT5-base and ViT-B/16 pretrained on CC3M-35L and finetuned on COCO-35L; BB is trained on COCO-35L; Bg switches to the ViT-g/14 model; and Lg uses mT5-large and and ViT-g/14, also trained with COCO-35L. For reference, Thapliyal et al. (2022) spent 5000 TPU

---

[3]https://cloud.google.com/translate

1638

hours to train their models, while our method can be used out-of-the-box for inference, i.e., 45 minutes for the X3600 benchmark per language.

Following Thapliyal et al. (2022), results are reported with the CIDEr (Vedantam et al., 2015) metric for English, Spanish, Hindi, and Chinese, with other languages covered in Section 4.4. CIDEr is a standard captioning metric that computes how well the generated caption matches the consensus of the reference captions, based on Term Frequency–Inverse Document Frequency (TF-IDF). In Appendix A, we included more generation metrics for holistic evaluation. To compute the metrics, we used the COCO evaluation package [4], and the SacreBLEU tokenization (Post, 2018).

4.2 Results

XM3600: Following Thapliyal et al. (2022), we report results on XM3600 for English, Spanish, Hindi, and Chinese, in Table 1. We can see that LMCAP outperforms all supervised approaches on Chinese, and achieves competitive performance on the other languages, despite being *image-blind* and not being trained on any image captioning data. For English, Spanish, and Hindi, we note that LMCAP is only outperformed by the large-scale supervised variant BB+CC, pre-trained on CCM3 and fine-tuned on COCO, jointly on English and the other 35 languages for the two datasets, i.e., with 123M captions. For the other variants that are only trained on COCO-35L, our model has a substantially larger performance on the CIDER metric across all four languages. We also show that our model can further benefit from increasing the datastore (LMCAP$_+$), as described in more detail over Section 4.3.

COCO: For completeness, we also report results on the machine translated COCO-DEV set in Table 2. In the top half of the table we show the performance of the 4 SOTA models on COCO-DEV via Google's machine translation API. Since this dataset was not provided by the authors, we perform as well automatic machine-translation but using the M2M-100 model (Fan et al., 2021), which gives an approximation for model comparison on COCO. As expected, LMCAP is outperformed on COCO since all the 4 variants were trained on it across 36 languages, with a large number of trainable parameters. Our model still reaches impressive performance, considering it was not trained on

COCO for any of those languages, neither was it trained on any multimodal data. This is especially the case for English, where our model reaches a similar CIDEr score, although it only reaches about half the performance for the other languages. In Appendix B, we also compare LMCAP with prompt-based captioning methods that were specially designed for English.

Model	en	es	hi	zh
Multilingual Captioning Supervised Learning				
BB+CC	**0.584**	**0.425**	**0.197**	0.202
BB	0.297	0.194	0.098	0.087
Bg	0.337	0.232	0.112	0.110
Lg	0.343	0.220	0.111	0.099
Few-shot Learning				
LMCAP	0.452	<u>0.329</u>	<u>0.132</u>	<u>0.221</u>
LMCAP$_+$	<u>0.526</u>	0.326	0.078	**0.251**

Table 1: Results on the geographically-diverse XM3600 benchmark. We compare our few-shot LMCAP model against large-scale supervised multilingual and multimodal SOTA models proposed by Thapliyal et al. (2022). Best results in bold and second-best underlined.

Model	$|\theta|$	en	es	hi	zh
COCO-DEV-GOOGLE					
BB+CC	766	0.980	0.962	0.759	0.748
BB	1230	0.856	0.844	0.671	0.659
Bg	1691	0.851	0.835	0.718	0.695
Lg	2241	0.875	0.859	0.624	0.656
COCO-DEV-M2M100					
LMCAP	N/A	0.767	0.453	0.334	0.584

Table 2: CIDEr performance on the COCO dataset. The top of the table presents SOTA results on the COCO validation split, translated via the GOOGLE API. The bottom rows of the table shows our model performance on COCO-DEV, translated via the M2M-100 model. $|\theta|$ corresponds to the number of trainable parameters in the model (in millions).

4.3 Ablation Studies

To better understand the design choices of LMCAP, we report a series of ablation tests on COCO-DEV, to avoid direct tuning on the XM3600 benchmark.

Prompt: Given that LMCAP works by prompting a language model with K retrieved captions

and N-shot examples, we study the effect of our prompt when varying K and N. Table 3 shows the importance of not depending on a single retrieved caption across the 4 languages. This is similar to previous findings in retrieval-augmented captioning studies focusing on English (Sarto et al., 2022; Ramos et al., 2023b), which showed that a large K makes the model more robust to mismatched captions. We further see that English and Spanish benefit from encoding a larger set of retrieved captions, while Hindi and Chinese work better with a smaller K. We select $K = 4$ since it has close-to-optimal performance for each of the languages. We then explore varying the number of N-shot examples, and found $N = 3$ to be the optimal value on all the four the languages. We thus use $K = 4$ and $N = 3$ in the prompt of LMCAP.

Setup	en	es	hi	zh
Varying K-Captions				
K=1, N=1	0.622	0.380	0.240	0.522
K=2, N=1	0.654	0.400	**0.269**	0.562
K=3, N=1	0.695	0.414	0.211	**0.565**
K=4, N=1	0.711	0.415	0.229	0.554
K=5, N=1	**0.734**	**0.424**	0.205	0.529
Varying N-Shot				
K=4, N=1	0.711	0.415	0.229	0.554
K=4, N=2	0.735	0.440	0.247	0.583
K=4, N=3	**0.767**	**0.454**	**0.334**	**0.584**
K=4, N=4	0.757	0.424	0.318	0.580

Table 3: The effect of using different numbers of K retrieved captions and N few-shot examples. Results reported on COCO-DEV with best results in bold.

Datastore: We also studied different contents for the datastore beyond the English captions from the COCO training set, shown in Table 4. Given that our model reaches much better performance on English, we hypothesise that our model can better generate captions in a desired language when having the retrieved captions in that same language. This could be validated using translations from COCO in the other languages, but since those are not available, we instead used a machine translated version of the Conceptual Captions dataset (CCM3) from Qiu et al. (2022). We used the English, Spanish, and Chinese versions of the CCM3 training set, respectively for each of the corresponding languages

(CCM3-L). We found that performance deteriorates on the COCO-DEV dataset, which might be explained by the difference between the COCO and CCM3-L datasets. Even combining the two datasets (COCO + CCM3-L) is worse than using only the COCO dataset.

In an attempt to cover more diverse concepts, we augmented COCO with three large web datasets (Conceptual Captions (Sharma et al., 2018), Conceptual 12M (Changpinyo et al., 2021), and SBU captions (Ordonez et al., 2011)), using their noise-free versions (Li et al., 2022). We refer to this dataset as CCS, and it contains synthetic model-generated texts for the web images. Using CCS leads to an improvement compared to just using COCO, except for Hindi. In Table 1, we also report results on XM3600 with this best datastore configuration, for which the performance again decreases for Hindi, but has a substantial improvement on English and Chinese. The benefits of including a more diverse collection of captions are further shown in Apprendix E with some qualitative examples (e.g., LMCAP was now able to generate the french concept *macarons* in English). Notice that the retrieved captions from CCS are still on English. Thus, although there is lack of multilingual image-caption pairs with clean labelled data, it would be interesting to pursue further work on incorporating retrieved information from other languages, in order to improve performance to levels similar to those for English.

Datastores	en	es	hi	zh
COCO	0.711	0.415	0.229	0.554
CC3M-L	0.387	0.309	-	0.337
COCO + CC3M-L	0.601	0.359	-	0.481
COCO + CCS	**0.713**	**0.431**	0.212	**0.563**

Table 4: Datastore ablations on COCO-DEV, where captions are retrieved from different sources of data. CC3M-L corresponds to machine translated version of Conceptual Captions proposed in Qiu et al. (2022) (Hindi not available), while CCS refers to the Conceptual Captions, Conceptual 12M, and SBU datasets (Li et al., 2022).

Model Size: In Table 5, we show the importance of using a language model that has a sufficiently large number of parameters. Both XGLM-562M and XGLM-1.7B are unable to generate captions beyond English. On the other hand, the 7.5B variant can lead to a stronger performance, but large-

scale LMs require more GPU memory, which limits the size of the prompt that can be encoded with modest hardware[5]. LMCAP uses the more efficient XGLM-2.9B version. These results are in line with previous findings, which suggest that stronger few-shot performance is achieved when the prompt is encoded by large LMs (Brown et al., 2020).

Params	Config.	RAM	en	es	hi	zh
564M	K=4, N=3	6G	0.411	0.094	0.030	0.146
1.7B	K=4, N=3	12G	0.637	0.143	0.066	0.272
2.9B	K=4, N=3	16G	0.767	0.454	0.334	0.584
7.5B	K=4, N=3	22G	**0.787**	**0.489**	**0.365**	**0.644**

Table 5: CIDEr performance on COCO-DEV, across the different variants of XGLM, to show the scaling behaviour of the LM used in LMCAP. RAM corresponds to the GPU memory consumption.

4.4 Additional Discussion

We now discuss the performance of LMCAP across the 36 languages, taking into consideration the data that was used for pre-training the LM. We also compare our approach with SMs and a simple baseline of retrieval plus translation. To support quantitative evaluation, we show some qualitative examples.

Multilingual Pre-training: In Table 6, we report the results of LMCAP on XM3600 for all the 36 languages considered in the dataset, ordered by the percentage of pre-training data used in XGLM for each language. LMCAP shows strong few shot performance on the diverse set of languages in which XGLM was pre-trained on. Similarly to BB+CC and Lg models, which are limited to the 36 languages they were trained on, our model is also dependent on the LM pre-training data, although there is potential to replace XGLM by another large LM, in order to generalize to other languages.

Comparision with Socratic Models: Since LMCAP is inspired in Socratic Models (SMs), we compare them against our approach. For this, XGLM receives the Socratic prompt that includes the image type, the number of people, places and object categories[6], instead of our retrieved captions. Results are reported in Table 7. Compared to either zero-shot or few-shot SMs, we can see that our model largely outperforms SMs, with a noteworthy

---

	BB+CC	LG	LMCAP
en	0.584	0.343	0.452
ru	0.194	0.089	0.134
zh	0.202	0.099	0.221
de	0.224	0.130	0.153
es	0.425	0.220	0.329
fr	0.410	0.217	0.260
ja	0.254	0.141	0.161
it	0.321	0.168	0.226
pt	0.380	0.202	0.283
el	0.199	0.101	0.136
ko	0.288	0.152	0.157
fi	0.177	0.089	0.112
id	0.307	0.167	0.151
tr	0.232	0.122	0.103
ar	0.227	0.106	0.107
vi	0.336	0.182	0.265
th	0.418	0.226	0.166
hi	0.197	0.111	0.132
bn	0.200	0.133	0.022
sw	0.319	0.151	0.085
te	0.196	0.099	0.042
Languages not in XGLM pretraining data			
cs	0.313	0.139	0.005
da	0.329	0.192	0.020
fa	0.311	0.155	0.002
he	0.230	0.098	0.001
hr	0.224	0.085	0.001
hu	0.175	0.096	0.006
mi	0.405	0.243	0.015
nl	0.441	0.232	0.082
no	0.385	0.230	0.025
pl	0.236	0.108	0.003
ro	0.188	0.100	0.007
sv	0.370	0.225	0.077
uk	0.189	0.081	0.006
AVG*	0.290	0.154	0.176

Table 6: Results for the 36 languages on the XM3600 benchmark, ordered by the pre-training language ratio of XGLM. BB+CC and LG are full-supervised state-of-the-art approaches from Thapliyal et al. (2022). AVG* corresponds to the average performance across the languages on which XGLM was pre-trained.

CIDER improvement of more than 39.1% on English, 20.0% on Spanish, 11.5% on Hindi, and of 21.4% Chinese. This confirms the effectiveness of our retrieval-augmented LM prompting approach.

Figure 2: Examples of captions generated by LMCAP for English, Spanish, Hindi, and Chinese, on XM3600 images and based on retrieved captions regarding each *blind-input* image.

Model	en	es	hi	zh
Socratic	0.067	0.045	0.001	0.031
Socratic N=1	0.454	0.280	0.176	0.340
Socratic N=2	0.344	0.215	0.141	0.268
Socratic N=3	0.376	0.254	0.219	0.370
LMCAP	**0.767**	**0.454**	**0.334**	**0.584**

Table 7: Comparison to Socratic Models (SMs) on the COCO-DEV dataset. LMCAP clearly outperforms SMs, as highlighted by bold.

Baseline of Retrieval with Translation: We also compared our approach against a baseline that retrieves the nearest caption on English and translates it into other languages in Table 8, using the M2M-100 model. This is to quantify the impact of prompting the language model compared to performing direct translation on retrieved captions. On COCO-DEV, we see that LMCAP only outperforms these results on English. Notice, however, that the references on COCO-DEV for the other languages rely on the M2M-100 distributions, as the baseline, promoting to an inequitable CIDEr. When evaluating on human-labeled data, as is the case with the XM3600 dataset, we see the benefits of prompting with retrieval information.

Notice also both LMCAP and the retrieval baseline outperform the BB model (the later also com-

petitive to the other 3 SOTA variants), despite training with large-scale multimodal machine translated data for hours. This shows the clear benefits of using retrieval-augmentation in multilingual image captioning, not just for result quality but to avoid high computation costs as well.

Model	en	es	hi	zh
COCO-DEV				
LMCAP	**0.767**	0.454	0.334	0.584
Baseline M2M-100	0.590	0.563	0.548	0.714
XM3600				
LMCAP	**0.452**	**0.329**	**0.132**	**0.221**
Baseline M2M-100	<u>0.333</u>	<u>0.205</u>	<u>0.120</u>	<u>0.170</u>
BB: COCO-35L	0.297	0.194	0.098	0.087

Table 8: Comparison to direct translation on retrieved captions (Baseline), on COCO-DEV and XM3600.

Qualitative Results: Figure 2 shows examples of captions generated in different languages by LMCAP, together with the retrieved captions that are provided in the prompt regarding each *blind-input* image. Qualitative examples tend to show diversity in the generation across the languages, with the retrieved information being itself diverse. For instance, in the first example, for English and Spanish, LMCAP focuses on describing that a man is

in front of microphones (i.e., based on the first two retrieved captions). In turn, for Hindi and Chinese, the man is in front of a laptop (i.e., from the first example), and the captions can also mention that he is ready to give a speech in Chinese (i.e., given the last two retrieved captions). In the second image, we can see that LMCAP can simply copy a retrieved caption to generate in English, while for the other languages the model may come up with terms not directly present in the retrieved captions (e.g., "snow slope" in Spanish). The last image is a negative example, where incorrect retrieved captions led the model into errors in English and Chinese, showing that there are also limitations in our *image-blind* approach. For more examples, see Appendix C.

5 Conclusions

This paper proposes LMCAP, an *image-blind* few-shot multilingual image captioning model. LM-CAP is based on prompting a language model with N-shot examples and retrieved captions extracted by a vision-and-language model, to condition caption generation in a desired language with a multilingual language model. On XM3600, i.e. a human-labelled massively multilingual multimodal benchmark, LMCAP performs competitively against the state-of-the-art without involving expensive training with large-scale translated multimodal data, or with any captioning data. Experimental results further demonstrate that LMCAP largely outperforms Socratic Models (Zeng et al., 2022), showing that retrieval augmentation plays a crucial role in our prompting approach. As future work, we plan to further assess the use of multilingual data in the datastore, as well as the impact of directly promoting diversity (Ye et al., 2022; Levy et al., 2022) in the captions used in the prompt.

Limitations

Image captioning and multilingual image captioning studies tend to focus on the COCO dataset, which was shown to contain gender imbalance. Previous research has also showed that models trained on COCO tend to amplify this bias (Hendricks et al., 2018; Zhao et al., 2017). While our model is not trained on COCO or in any captioning data, it relies on a pre-trained language model, which is known to suffer from different sources of bias and fairness issues (Bommasani et al., 2021; Sheng et al., 2021; Schramowski et al., 2022).

Our model also involves retrieval-augmentation with captions extracted by a vision-and-language model, also pre-trained in an unsupervised manner. Like in the case of other retrieval-augmented generative models (Lewis et al., 2020), LMCAP has inherently a bias towards the retrieved information. Notwithstanding, by conditioning on information from a datastore with clean and curated text, LMCAP has potential to ameliorate some of the generation issues of the language model (e.g., elude hateful or violent language). To have insights on the biases presented in LMCAP, we recommend analysing the retrieved captions used by the model, since they provided cues to the predictions, as shown in Figure 2. We argue that it can be much harder to have a direct interpretation for captioning models that are not retrieval-augmented.

Another limitation of our model relates to it following a full *image-blind* approach, which heavily depends on information from similar captions instead of the visual content from the actual input image. To address this limitation, future work could additionally include concepts extracted from the image in the prompt, as proposed in Socratic Models, combined with the retrieved information.

Ethics Statement

The datasets supporting the evaluation of LMCAP are publicly available for academic purposes. We also plan to release our code, and the additional resources that were built to support the experiments.

We emphasise that LMCAP challenges the efficiency of most current captioning approaches, in terms of resource usage and development/deployment effort, while at the same time promoting more equitability and inclusion, exemplified here by attempting to balance language representation at low computational costs.

We further note that while our model attempts to advance research beyond English-centric captioning, by considering captioning for a wide variety of languages, it is important to address and pay

more attention to low-resource languages as well (i.e., languages beyond those covered in our tests). Evaluating LMCAP with additional datasets, covering an even larger set of languages and concepts, would be desirable.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. volume 33, pages 1877–1901.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Zhengcong Fei. 2021. Memory-augmented image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1317–1324.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision*, pages 503–519.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2021. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

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

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2212.06800*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. volume 24.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Chen Qiu, Dan Oneata, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. *arXiv preprint arXiv:2210.13134*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023b. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. *arXiv preprint arXiv:2207.13162*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

A Standard Evaluation Metrics

In the paper, comparison between models is performed using the CIDEr metric by following Thapliyal et al. (2022). For holistic captioning evaluation, we provide here the performance of LMCAP on XM3600 across additional standard automatic metrics. Specifically, Table 9 reports performance with BLEU-1 (B-1) and BLEU-4 (B-4) (Papineni et al., 2002), ROGUE-L (Lin, 2004) and METEOR (Denkowski and Lavie, 2014).

	B-1	B-4	ROGUE-L	METEOR
en	0.387	0.067	0.299	0.129
es	0.364	0.052	0.256	0.126
hi	0.258	0.015	0.182	0.220
zh	0.318	0.053	0.231	0.105

Table 9: LMCAP performance on the XM3600 benchmark across different evaluation metrics.

B Additional Results on COCO

Table 10 provides additional results on COCO, comparing LMCAP against against other prompt-based captioning models that do not involve training, including two previously proposed zero-shot captioning methods that are English-specific (i.e., ZeroCap (Tewel et al., 2021) and ConZic (Zeng et al., 2023)). Results show that LMCAP outperforms both. We also notice that unlike these models, LMCAP works for the multilingual setting, advancing research beyond English-centric captioning.

Model	B-4	METEOR	CIDEr
ZeroCap	0.026	0.115	0.146
ConZIC	0.013	0.115	0.128
LMCAP	0.199	0.220	0.759

Table 10: Results on the COCO test set. We compare our few-shot LMCAP model against English-specific captioning models that do not involve supervising training as well.

C More Qualitative Examples

We provide several additional examples of captions generated from XM3600 images in Figure 3.

Figure 3: More examples of captions generated by LMCAP for XM3600 images, with retrieval from COCO.

D Prompt-Template

We follow the Socratic template, where instead of including different categories (objects, places, number of people, etc), we replace them by the retrieved captions. By following the same template, in place of a completely different one, we can assess the impact of including retrieval compared to the original Socratic framework. Our template is:

```
I am an intelligent image captioning
bot. Similar images have the following
captions:  <caption 1> <caption 2>
<caption 3> <caption 4>. A creative short
caption I can generate to describe this
image in <language> is:
```

Between the retrieved captions we use the special end of sentence token (i.e., </s>) of XGLM. Notice also that our prompt starts with 3 fixed shot examples from images in the training dataset (i.e., the same prompt is repeated multiple times to encode the n-shot examples). We share the N-shot examples and the set of K retrieved captions used in our prompt, together with the code at https://github.com/RitaRamo/lmcap. The following text is a concrete example of the prompt provided for the first image of XM3600.

```
I am an intelligent image captioning
```

```
bot. Similar images have the following
captions: a horse grazing in a grassy
field next to a barn</s> a brown horse
grazing in its pen and a red barn and
water</s> a pretty brown horse eating
some grass in a bare field</s> a horse
is eating grass next to a barn in the
middle of a pasture</s> A creative short
caption I can generate to describe this
image in spanish is: Un caballo marrón
es grasa cerca de una casa roja</s>
I am an intelligent image captioning
bot. Similar images have the following
captions: a teal toilet is the center of
this bathroom photo</s> a small bathroom
with brightly painted blue walls</s> the
bathroom has a splash of color with
the blue tiles</s> the sink is above a
turquoise tile sink</s> A creative short
caption I can generate to describe this
image in spanish is: Un baño muy limpio
y bien decorado</s> I am an intelligent
image captioning bot. Similar images have
the following captions: a woman and child
focus on a pink device in public</s> a
woman holding a small child while standing
near a crowd</s> a very cute lady posing
with a small kid</s> a young child with
a cell phone and an adult</s> A creative
```

1647

```
short caption I can generate to describe
this image in spanish is: Una mujer se
acercó a mirar en su teléfono mientras
está listo para tomar una foto</s> I am an
intelligent image captioning bot. Similar
images have the following captions:  a
brown chicken is walking around outside
with another hen</s> a couple of roosters
standing in a field</s> a hen pecks the
ground while another looks off in the
distance</s> a couple of roosters are in
a field</s> A creative short caption I
can generate to describe this image in
spanish is:.
```

E   Augmented Datastore Examples

In this appendix, we provide qualitative examples
on XM3600 when the datastore is augmented with
CCS, i.e., with large and diverse data. In Figure
4, we can see generation improving for English,
where LMCAP correctly mentions the french con-
cept of *macarons*, available in the retrieved cap-
tions. In line with the quantitative results provided
in Section 4.3, we can also see a possible explana-
tion for why generation degraded for Hindi, that
has a lower pre-training language ratio than En-
glish: LMCAP seems to have copied the last 3-shot
example provided in prompt, described above in
Section D), maybe due to presence of more noise in
the CCS data. Another example can be seen in Fig-
ure 5, where LMCAP is more specific in generating
the flower type *orchid*.

Figure 4 (top, COCO):

- some different colored pastry treats on a white tray
- a large assortment of sweet, rich, colorful deserts
- a colorful pastry is displayed on a table
- a close up view of some very tasty looking pastries

COCO

en: *"a large assortment of sweet, rich, colorful deserts"*

es: *"un* montón de pasteles de colores*"* (a lot of colour cakes)

hi: *"几种不同颜色的甜点放在一个白色托盘"* (few different colour desserts are placed on a white plate)

zh: *"一个年轻的男子站在他的电脑前,他准备开始演讲"* (a young man standing in front of his computer and he is ready to start a speech)

Figure 4 (bottom, COCO + CCS):

- an afternoon afternoon is incomplete for some macarons
- some colorful macarons on a tray
- several macarons on a tray
- a tray with many different types of colorful macarons

COCO + CCS

en: *"a tray of colorful macarons"*

es: *"un pastel de cumpleaños"* (a birthday cake)

hi: *"कभी-कभी आप एक तस्वीर लेने के लिए तैयार हो जाते हैं जब आप एक तस्वीर लेते हैं"* (Sometimes you get ready to take a picture when you take a picture)

zh: *"一个下午的下午,没有一个完整的午餐"* (one afternoon without a full lunch)

Figure 4: An example of captions generation by LMCAP conditioned on captions retrieved from COCO (top) compared to augmenting the datastore with CCS (bottom).

Figure 5 (top, COCO):

- a close-up of an exotic flower on a large stem
- a view of a flower from very close, it appears to be fully bloomed
- a pink, white and red orchid in a vase in front of a window
- a pink and white orchid is in a small black vase'

COCO

en: *"a close-up of an exotic flower on a large stem"*

es: *"una flor exótica en un tallo grande"* (an exotic flower in a big shell)

hi: *"एक फूल का एक बड़ा हिस्सा"* (a large portion of a flower)

zh: *"一种美丽的花朵在巨大的花中"* (a beautiful flower in a huge flower)

Figure 5 (bottom, COCO + CCS):

- a large white flowered orchid with two pink tipped spots
- an orchid with red and white markings
- the pink and white orchid is on display at show
- an orchid that won the individual's trophy

COCO + CCS

en: *"a large white flowered orchid"*

es: *"una flor de orchid"* (a flower of orchid)

hi: *"फूलों की एक बड़ी लता"* (a large flower)

zh: *"一个白色的花朵,有两个红色的斑点"* (a white flower with two red spots)

Figure 5: An example of LMCAP generation based on retrieval from COCO (top) or COCO augmented with CCS (bottom).

1649

ACL 2023 Responsible NLP Checklist

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C   ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.