PAELLA🥘: Parameter-Efficient Lightweight Language-Agnostic Captioning Model

Rita Ramos[†] **Emanuele Bugliarello**[♮] **Bruno Martins**[†] **Desmond Elliott**[⋆]
[†]INESC-ID, Instituto Superior Técnico, University of Lisbon
[♮]Google Research
[⋆]Department of Computer Science, University of Copenhagen
ritaparadaramos@tecnico.ulisboa.pt

Abstract

We introduce PAELLA, a **Pa**rameter-**E**fficient **L**ightweight **L**anguage-**A**gnostic image captioning model designed to be both parameter and data-efficient using retrieval augmentation. The model is trained by learning a small mapping network with 34M parameters between a pre-trained visual model and a multilingual language model that is conditioned on two types of input: (i) the image itself, and (ii) a set of retrieved captions in the target language. The retrieved examples play a key role in guiding the model to generate captions across languages. Through retrieval, the model can be lightweight in terms of the number of trainable parameters, which only exist in its mapping network, and also in the amount of multilingual training data that is required. Experiments on the XM3600 dataset, featuring 36 languages, show that PAELLA can outperform or compete against some models with 3–77× more learned parameters and 35–863× more data, particularly in low-resource languages. We also find that PAELLA can be trained on only monolingual data and still show strong zero-shot abilities in other languages.[1]

1 Introduction

We tackle the problem of multilingual image captioning, aiming to provide textual descriptions of visual contents that can serve speakers of different languages, in contrast to most captioning models that only generate English captions. While significant progress has been made in recent years, training image captioning models has become more expensive due to the trend of scaling both data and model size (Hu et al., 2022; Wang et al., 2022). This trend is even more prominent in multilingual approaches (Chen et al., 2023b; Thapliyal et al., 2022), given the need for training data covering each target language, and the need of even larger models to mitigate the *curse of multilinguality* (Conneau et al., 2020; Goyal et al., 2021).

Some recent research has focused on minimizing the cost of multilingual training, such as PALI-3 (Chen et al., 2023a) with 5B trainable parameters, and mBLIP (Geigle et al., 2023) with only 124M trainable parameters. Both these approaches use pre-trained multimodal language models or pre-trained visual encoders that are kept frozen, reducing the number of trainable parameters. Nevertheless, both of these models still rely on training with millions or billions of examples, including in the context of image captioning alone.

This paper describes a **Pa**rameter-**E**fficient **L**ightweight **L**anguage-**A**gnostic captioning model (PAELLA). The model is designed to be efficient, not only in terms of the number of trainable parameters, but also lightweight in the amount of multilingual training data required. PAELLA has only 34 million trained parameters, and the model can be trained using just 566K examples, i.e., the size of the English COCO dataset.

PAELLA is based on frozen pre-trained models that are augmented with retrieved examples. The only learned parameters are in a compact mapping network of cross-attention layers between a frozen CLIP image encoder and a frozen XGLM multilingual language model. The model is trained to generate captions in the desired language using a prompt in that language. Furthermore, the retrieved examples assist the model in generating meaningful captions, by providing examples of what the predicted caption should resemble. The use of retrieved examples positively contributes to reducing both the number of trainable parameters, and the required amount of multilingual data.

We conduct experiments on XM3600 (Thapliyal et al., 2022), an established multilingual captioning benchmark that covers geographically diverse images with human-annotated captions in 36 languages. Experiments show that PAELLA can out-

[1] Code and model available at https://github.com/RitaRamo/paella.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3549–3564
June 16-21, 2024 ©2024 Association for Computational Linguistics

perform or compete with models that are more demanding in terms of trained parameters or training data. The performance of our model in low-resource languages is particularly noteworthy, in contrast to concurrent models like mBLIP, that often excel in English and related languages but struggle to generalize effectively to underrepresented languages.

Results also show that PAELLA demonstrates zero-shot multilingual capabilities when trained only with monolingual data such as the English COCO dataset. PAELLA achieves language transfer through retrieval, solemnly by retrieving captions in the target language during inference. Ablation studies further demonstrate the benefit of our retrieval-augmented approach.

2 Related Work

2.1 Image Captioning

In the last years, image captioning has witnessed impressive performance improvements through end-to-end Vision-and-Language Pretraining (VLP), considering the use of large-scale models and large image-text datasets in English (Wang et al., 2021; Hu et al., 2022; Li et al., 2022).

In an effort to alleviate the increasing computation costs, recent studies have adopted off-the-shelf pre-trained encoder and decoder models that remain frozen during training (Mokady et al., 2021; Luo et al., 2022; Ramos et al., 2023b; Mañas et al., 2023). For instance, several studies have used CLIP (Radford et al., 2021) as the visual encoder, and GPT-2 (Radford et al., 2019) as the language decoder, keeping one or both of the models frozen during training, and instead learning a mapping network to align the two modalities. Having the models frozen speeds up training and reduces GPU memory usage (Mokady et al., 2021). Besides reducing computational costs, this is also a means to seamlessly integrate powerful unimodal models (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023), including large-scale pre-trained (Brown et al.; Zhang et al., 2022; Touvron et al., 2023) and instruction tuned language models (Wei et al., 2021; Chung et al., 2022; Taori et al., 2023), which would otherwise be impractical with end-to-end training, and could result in the loss of generalization from catastrophic forgetting (McCloskey and Cohen, 1989).

In the realm of multilingual image captioning, instead of expensive end-to-end training from scratch (Thapliyal et al., 2022; Yang et al., 2020), recent models have also opted for frozen pre-trained visual encoders and/or language decoders. Examples include mBLIP (Geigle et al., 2023) or PALI-3 (Chen et al., 2023a). In contrast to these studies, we use a frozen pre-trained encoder and a frozen language model, that are augmented with retrieved examples to further reduce the number for trainable parameters, as well as the need for extensive multilingual training data.

2.2 Retrieval Augmention

Retrieval-augmented language generation conditions the generation process by enhancing the input with information retrieved from an external datastore (Lewis et al., 2020). Retrieval augmented models have gained increased popularly (Khandelwal et al., 2020; Izacard et al., 2022; Shi et al., 2023; Yu et al., 2023), including in image captioning (Zhao et al., 2020; Xu et al., 2019; Ramos et al., 2021; Sarto et al., 2022; Ramos et al., 2023b; Yang et al., 2023).

The work that more closely resembles ours is SmallCap (Ramos et al., 2023b), a lightweight English captioning model that uses pre-trained encoder and decoder models, and that also uses prompting with retrieved captions. In this paper, we explore how retrieval augmentation can help to reduce not just the number of trainable parameters but also the amount of training data. Another key difference between the approaches is that PAELLA is based on a pre-trained multilingual language model instead of a monolingual English model. We explore how the prompt and retrieved captions should be designed to enable generation across different languages, instead of only English.

We note that retrieval augmentation remains largely unexplored in the multilingual image captioning scenario. Until now, only the multilingual LMCap (Ramos et al., 2023a) model has used retrieval augmentation, but solely in a training-free manner based on prompting a multilingual language model in an image-blind approach. In our work, we instead show the potential of retrieval augmentation in contributing to the training of a multilingual image captioning model.

3 Proposed Approach

The **P**arameter-**E**fficient **L**ightweight **L**anguage-**A**gnostic (PAELLA) captioning model uses retrieval augmentation to generate captions in multi-

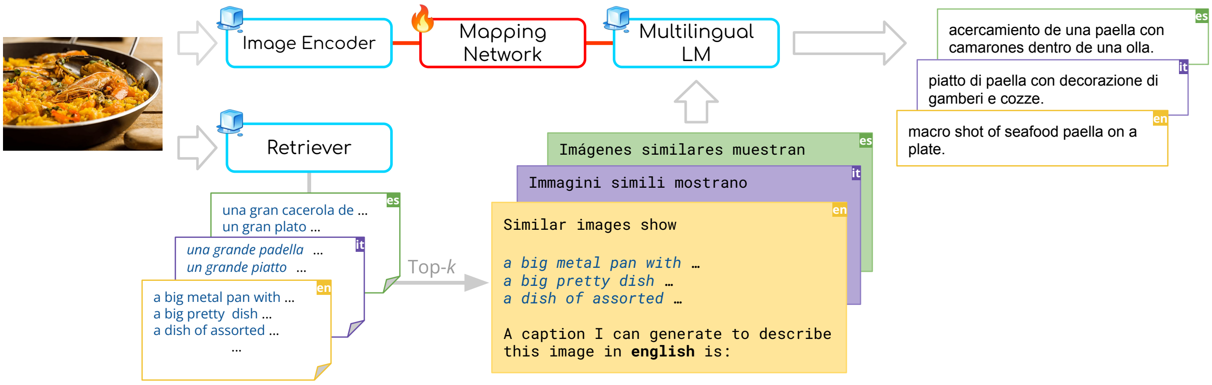


Figure 1: PAELLA uses a frozen pre-trained image encoder and a frozen multilingual decoder, connected with a trainable mapping network. The decoder generates a multilingual caption conditioned on the encoded image, together with retrieved captions given as input within a prompt in the desired language.

ple languages. An overview of the model architecture can be seen in Figure 1.

We follow a similar design to the monolingual SMALLCAP model (Ramos et al., 2023b), by building on top of powerful pre-trained unimodal models. We also use CLIP (Radford et al., 2021) as the visual encoder, but instead of GPT-2 or OPT as the decoder, we use a multilingual auto-regressive language model, i.e. XGLM (Lin et al., 2021). Both the encoder and the decoder are kept frozen during training, except for a newly added mapping network of cross-attention layers, that allows the decoder to attend to the visual inputs. PAELLA generates captions conditioned on the image and on a set of k retrieved captions[2] from similar images. The retrieved captions are used to prompt the model to generate in the desired target language. The prompt follows a fixed-template which first includes examples of the k retrieved captions and ends with an instruction for the multilingual decoder to generate a caption in a desired language. The English prompt is:

```
Similar images show [retrieved caption₁] ... [retrieved captionₖ]. A caption I can generate to describe this image in [language] is: ...
```

The prompt and captions can be tailored to different languages, by having both these parts in the desired language (see some examples of the prompts for other languages in Appendix A).

The parameters in the mapping network θ_M are trained by minimizing the sum of the negative log-likelihood of predicting the ground truth image caption for each token in the sequence $y_1 \dots y_M$, conditioned on the image $\mathbf{V}$ and the retrieval-augmented prompt $\mathbf{L}$:

$$L_{\theta_M} = -\sum_{i=1}^{M} \log P_\theta(y_i | y_{<i}, \mathbf{V}, \mathbf{L}). \quad (1)$$

We quantitatively show in Section 5 that our retrieval-augmented approach has these properties:

Parameter-efficiency: Only the cross-attention layers between a frozen encoder and a frozen decoder need to be trained. To compensate for the small number of trainable parameters, the model is guided with examples of retrieved captions.

Data-efficiency: Through retrieval, the model does not need a huge amount of multilingual data for training, since it benefits from retrieved examples that demonstrate how to generate in the target language. We thus alleviate the data hunger of existing multilingual models, that are often trained with the same image associated to captions in multiple languages, having to repeatedly translate entire English captioning datasets for each language (e.g., COCO to COCO-35L (Thapliyal et al., 2022)).

Zero-shot Multilinguality: Our model demonstrates multilingual capabilities even when trained only on monolingual image captioning data. It can be trained on the specific in-domain distribution from the available data in a high-resource language, and still generate in different languages. This by relying exclusively, at inference time, on retrieval augmentation in the target language from an available multilingual captioning dataset.

[2] See Section 4 for details on the retrieval system.

4 Experimental Setup

4.1 Implementation and Training Details

We release our code and model at `https://github.com/RitaRamo/paella`. PAELLA is implemented using the HuggingFace Transformers library (Wolf et al., 2020). The backbone of the model is based on the pre-trained CLIP model `openai/clip-vit-base-patch32`, and the pre-trained XGLM `facebook/xglm-2.9B`.

The input image $\mathbf{V}$ is encoded by the CLIP encoder, and the language-based prompt $\mathbf{L}$, which includes the k retrieved captions, is processed by XGLM to generate a caption in the target language.

🧊 **Encoder:** CLIP is a powerful multimodal model that was pre-trained to encode images and text into a shared embedding space, using contrastive learning (Radford et al., 2021). We use CLIP-ViT-B/32 to encode the input image, producing a sequence of N=50 visual features V=$\{v_1, ..., v_N\}$, each with an embedding size of 768 dimensions. This encoder has 86M million parameters, which are kept frozen during training.

🧊 **Decoder:** XGLM is a multilingual autoregressive language model that can generate in a diverse set of 30 languages[3] (Lin et al., 2021). In PAELLA, we use the variant with 2.9B parameters, which are frozen during training.

🧊 **Retrieval:** CLIP is also used for image-text retrieval. Specifically, it is used to encode both the candidate captions into a datastore, and each given input image. For each given image, the k nearest captions are retrieved from the caption datastore. The datastore is indexed efficiently through the FAISS library (Johnson et al., 2017), specifically with the `IndexFlatIP` index that does not require any training, allowing for offline retrieval. The images are also encoded with CLIP, using the visual backbone, to retrieve the captions that are most similar based on cosine similarity. We select the top $k = 4$ retrieved captions, in-line with previous findings which indicate that this is the optimal number of captions in both monolingual and multilingual setups (Ramos et al., 2023a,b).

🔥 **Mapping Network:** The only part of PAELLA that is trained is the mapping network between the frozen encoder and decoder. The mapping network consists of randomly initialized cross-attention layers (Vaswani et al., 2017) added to each of the 48 layers of XLGM, so the decoder can attend to the encoder outputs. In order to have a smaller number of trainable parameters, we use low rank cross-attention layers by reducing the original dimensionality d of the projection matrices from 128 to 8, as in Ramos et al. (2023b). Accordingly, this amounts to only 34M trainable parameters (see Appendix G). These parameters are trained by predicting the tokens in the target caption, as shown in Equation 1.

Training Requirements: PAELLA is trained for 3 epochs with an initial learning rate of 1e-4, using the AdamW optimizer (Kingma and Ba, 2014) and a batch size of 16 with 4 gradient accumulation steps, on a single NVIDIA RTX A6000 GPU. In an effort to promote accessibility, our model can be trained in a day on a single GPU, unlike other multilingual image captioning models. With the CLIP-ViT-B/32 encoder and the XGLM-2.9B decoder, PAELLA takes 23h for training the 34M trainable parameters, occupying 46G RAM. If using instead XGLM-1.7B, it takes 14h and 29G RAM. For XGLM-564M, it only takes 7h and 19G RAM[4]. Moreover, we exclusively use publicly available datasets, as described next.

4.2 Data

We now describe the data used in our experiments, covering the benchmark we evaluate our model on and its training data, as well as the dataset used for the retrieval datastore.

Evaluation Data: We assess the performance of our model on the well-established XM3600 dataset (Thapliyal et al., 2022), that covers geographically-diverse images from 36 languages (L_{36}), including the core set of languages defined by Thapliyal et al. (2022): en, es, hi and zh (L_{CORE}), and a set of low-resource languages (L_5): *bn*, *quz*, *mi*, *sw*, *te*. Each language is represented by 100 images from Open Images, chosen based on the area the language is spoken. In total, XM3600 has 3600 images with 261375 human-annotated captions. Each image has at least 2 captions/language.

Most human-annotated captioning datasets are predominantly on English. Following Thapliyal et al. (2022), we extend the evaluation to include the COCO-35L dataset (Thapliyal et al., 2022),

[3] *en*, *ru*, *zh*, *de*, *es*, *fr*, *ja*, *it*, *pt*, *el*, *ko*, *fi*, *id*, *tr*, *ar*, *vi*, *th*, *bg*, *ca*, *hi*, *et*, *bn*, *ta*, *ur*, *sw*, *te*, *eu*, *my*, *ht*, *qu*.

[4] See the performance with these models in Appendix D.

which is automatically translated from the original English COCO dataset (Chen et al., 2015). COCO-35L has 5000 images for validation, and 113k images for training, each with 5 reference captions per language. The translations were obtained with the Google Translate API[5], covering all the 36 languages in XM3600, with the exception of Cusco Quechua (*quz*), not supported by the API.

Training Data: Given the scarcity of multilingual human-annotated captions, multilingual models typically resort to training on machine translated data. The standard approach (Thapliyal et al., 2022) involves training on the aforementioned COCO-35L dataset, which contains 566K training captions translated into 35 languages, resulting in a dataset with 20.3M captions. Existing multilingual models (Thapliyal et al., 2022; Geigle et al., 2023; Chen et al., 2023b) also benefit from large-scale pre-training, using datasets such as the machine translated CC3M-35L (Thapliyal et al., 2022), built from the CC3M dataset (Sharma et al., 2018), which contains 3M image-caption pairs for training, amounting to 105M translations.

In contrast, we only train on a subset of COCO-35L, which is downsampled to match the size of the original English COCO dataset (i.e., 565K examples instead of 20.3M examples). The subset is created by sampling captions from the COCO-35L dataset according to a uniform distribution across languages, using the same language for the 5 captions associated to each image. The exploration of other sampling strategies is left for future work.

Retrieval Data: The datastore of our model contains the training captions of the COCO dataset using the Karpathy splits (Karpathy and Fei-Fei, 2015). The English captions are indexed with their corresponding IDs. In this way, we apply image–text search based on CLIP-ViT-bigG-14[6] by retrieving, for each image, the $k = 4$ caption IDs from the nearest-neighbor images[7]. Given the retrieved caption IDs, we can readily integrate either the corresponding English captions from COCO, or use the associated translations from any of the other 35 languages, by cross-referencing the IDs with COCO-35L depending on the target language.

We emphasize that our retrieval system is monolingual. The datastore only contains the English COCO captions, without demanding the scale of the entire COCO-35L dataset. We only use COCO-35L for cross-referencing the retrieved IDs to obtain the captions in the language that we desire.

4.3 Evaluation Metrics

Following previous work, we mostly evaluate multilingual captioning performance with CIDEr (Vedantam et al., 2015). CIDEr calculates the agreement between the generated caption and the consensus of the reference captions, computed through a similarity function that uses Term Frequency times Inverse Document Frequency (TF-IDF) weights. In contrast to previous multilingual captioning studies that solely report the CIDEr metric as per Thapliyal et al. (2022), our work extends the evaluation scope to a diverse set of captioning metrics, specifically BLEU-1, BLEU-4, ROGUE, and METEOR (see Appendix C). We used the COCO evaluation package[8] with SacreBLEU tokenization (Post, 2018) to compute the metrics. During evaluation, captions are generated by our model using beam search decoding with a beam size of 3.

4.4 Model Variants

We evaluate PAELLA alongside two additional variants, each trained on a more limited set of languages in order to assess the cross-lingual transfer abilities of our approach. Model selection is based on maximizing the average CIDEr across the $\mathrm{L}_{\mathrm{CORE}}$ languages in the COCO-35 validation dataset. Here we detail the model variants we compare.

PAELLA: This is our main model, trained to generate for the 35 languages in COCO-35L. In this case, we sampled uniformly from COCO-35L to ensure the scale of the COCO English dataset.

PAELLA$_{\mathbf{core}}$: This model is trained to generate for $\mathrm{L}_{\mathrm{CORE}}$, i.e. the core set of 4 languages proposed in the XM3600 dataset (en, es, hi and zh). We also sample uniformly from COCO-35L to maintain a scale consistent with the COCO English dataset, but within this restricted language set $\mathrm{L}_{\mathrm{CORE}}$.

PAELLA$_{\mathbf{mono}}$: This model is trained to generate only on English. In this case, we use the original COCO English dataset.

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

[6] See Appendix B for a discussion on the design choice of using this specific encoder for the retrieval component.

[7] We do not retrieve captions of the input image itself.

[8] https://github.com/tylin/coco-caption

5 Results

We first compare PAELLA against state-of-the-art models. We then discuss the performance of our other two variants trained on a smaller set of languages, i.e., PAELLA$_{\text{core}}$ and PAELLA$_{\text{mono}}$.

5.1 Parameter- and Data-efficient Training

Table 1 shows that PAELLA performs competitively against state-of-the-art multilingual models, despite training with a fraction of their trainable parameters and with considerably less data. With just 34M trainable parameters and only 566K training instances, PAELLA achieves a CIDEr score of 26.2 on average across all the 36 languages, and a CIDEr of 28.2 across the languages on which the XGLM backbone was pre-trained. Also, our model is able to yield 20.7 CIDEr points across the set of low-resource languages L$_5$ (*bn*, *quz*, *mi*, *sw*, *te*)[9].

PAELLA surpasses Lg (Thapliyal et al., 2022), i.e. a fully-supervised model trained with 2.6 billion parameters in the entire COCO-35L dataset (86x more trainable parameters, and 35x more training examples), largely outperforming across the set of core languages and on average. PAELLA is also competitive against BB+CC, another model from Thapliyal et al. (2022) that is pre-trained on 135M examples in the combination of CC3M-35L and COCO-35L. Although PAELLA does not outperform BB+CC on average, it reaches better performance in 3/4 of the core languages, noteworthy considering their model was trained with 238x more data than our model.

PAELLA also competes with multilingual models that were trained on diverse multimodal data from different vision-and-language tasks, such as mBLIP (Geigle et al., 2023). Akin to our model, mBLIP leverages a pre-trained multilingual language model with an effort on computational and data efficiency. Our model surpasses these efforts by having significantly fewer parameters and operating on considerably less data (e.g., in the context of captioning data, mBLIP trains on machine translations of COCO alongside a diverse set of 2.3 million examples from the synthetic Web CapFilt dataset (Li et al., 2022)). PAELLA outperforms mBLIP BLOOMZ-7B by 2.8 CIDEr points on average, and has less 2.1 points than mBLIP mT0-XL. The mBLIP mT0-XL model demonstrates strong performance on English, yielding 80.2 CIDEr, yet we see a large gap in low-resource languages, with 13.4 CIDEr points while our model achieves 20.7 points. In Section 6.1, we discuss more extensively the performance across languages.

Similarly to other multilingual captioning models, PAELLA performs significantly worse than the large-scale 17B parameter PaLI model (Chen et al., 2023b) that is trained on 12 billion examples using the private WebLI dataset. The same holds for the recent PALI-3 (Chen et al., 2023a), which makes efforts towards a more efficient model, but still trains billions of parameters on billions of multilingual data. This is still notably costly and impractical for many applications. From a research perspective, our model can be trained in a single day in consumer hardware with a public dataset.

Lastly, we see a 15.2 CIDEr points improvement compared to LMCap (Ramos et al., 2023a), which is a few-shot retrieval-augmented approach that has no training. With minimal multilingual training, our model further closes the gap towards large-scale multilingual captioning models.

Overall, the results on XM3600 demonstrate the efficacy of our approach for efficient multilingual captioning, contributing to the reduction of both trainable parameters and data requirements. For a more comprehensive evaluation, we also report results on COCO-35L in Table 2, where we observe again that our model can outperform the fully-supervised models of Thapliyal et al. (2022). See qualitative examples in Appendix H.

5.2 Zero-shot Cross-lingual Transfer

In Table 1, we observe that PAELLA$_{\text{core}}$ (trained on *en*,*es*,*hi*,*zh*) and PAELLA$_{\text{mono}}$ (trained only on *en*) have strong zero-shot performance in other languages, showing that our approach does not require captioning data for each of the languages during training. The generation can be conditioned on a different language beyond the training set, by providing the prompt and retrieved captions in the desired output language, solely at inference time.

We further observe that PAELLA is outperformed by PAELLA$_{\text{mono}}$ on English, and by PAELLA$_{\text{core}}$ on English and Spanish. This can be partially explained by the fact that PAELLA was pre-trained on a uniform sample of all 35 languages in COCO-35L, while these variants were pre-trained on a uniform sample of only those languages, i.e. with more English captions. Both the Core and Mono variants, on the other hand, are less able to generate captions for languages out-

[9] See Appendix I for the performance on all languages.

Model	Data	Train θ	Total θ	en	es	hi	zh	L$_5$	L$_{36}$
Training-free									
LMCap	-	0	2.9B	45.2	32.9	13.2	22.1	0.0	11.0
Large-scale Training									
PALI	12B	17B	17B	98.1	-	31.3	36.5	-	53.6
PALI-3	12B	5B	5B	94.5	-	-	-	-	46.1
mBLIP mT0-XL	489M	124M	4.9B	80.2	62.6	16.1	14.7	7.9	28.3
mBLIP BLOOMZ-7B	489M	124M	8.3B	76.4	60.0	24.9	14.7	6.7	23.4
BB+CC	135M	0.8B	0.8B	58.4	42.5	19.7	20.2	22.4	28.5
Lg	19.8M	2.6B	2.6B	34.3	22.0	11.1	9.9	12.5	15.0
Data & Parameter-efficient Training									
PAELLA	**566K**$_{35L}$	**34M**	3B	57.3	44.9	20.8	25.9	20.7	26.2 (28.2*)
PAELLA$_{\text{core}}$	**566K**$_{en,es,hi,zh}$	**34M**	3B	58.2	45.0	20.4	25.4	11.8	16.8 (24.9*)
PAELLA$_{\text{mono}}$	**566K**$_{en}$	**34M**	3B	58.2	42.2	17.1	23.5	12.1	15.5 (23.9*)

Table 1: CIDEr performance on XM3600, a multilingual benchmark with geographically-diverse images across 36 languages. We compare our model, PAELLA, and its two variants, PAELLA$_{\text{core}}$ (trained on *en*,*es*,*hi*,*zh*) and PAELLA$_{\text{mono}}$ (trained only on *en*) against other state-of-the-art multilingual models. L$_5$ represents the average performance across the set of low-resource languages (*bn*, *quz*, *mi*, *sw*, *te*), and L$_{36}$ over all the 36 languages. (*) corresponds to the average across the languages on which the XGLM decoder was pre-trained. We highlight in bold that our model has the lowest number of trainable parameters and requires the least amount of training data.

Model	en	es	hi	zh
BB+CC	98.0	96.2	75.9	74.8
Lg	87.5	85.9	62.4	65.6
PAELLA	113.6	113.9	86.2	123.3
PAELLA$_{\text{core}}$	118.5	120.3	94.7	130.7
PAELLA$_{\text{mono}}$	120.8	91.48	45.9	59.1

Table 2: CIDEr scores on COCO-35L validation data. The fully-supervised models from Thapliyal et al. (2022) are shown on top, with our model variants at the bottom.

side those in the XGLM pre-training data, resulting in an average decrease of 9.4 and 10.7 points of CIDEr across all 36 languages, compared to PAELLA, respectively. Despite this limitation, we emphasize the performance of PAELLA$_{\text{mono}}$, that achieved a 15.5 CIDEr score on average, especially considering its training was exclusively on English. PAELLA$_{\text{mono}}$ even outperforms Lg across the set of 4 core languages and on average, even though this model had end-to-end large-scale training across the various languages with the complete COCO-35L dataset.

Our approach's capability for zero-shot cross-lingual transfer holds particular importance with the predominance of English-centric captioning datasets. We note we did not use multilingual in-domain data in the retrieval datastore. The retrieved captions from COCO-35L have a different distribution than the XM3600 benchmark, that contains geographically diverse images and concepts. We also stress that the entire prompt (including the retrieved captions) needs to be in the target language for this zero-shot cross-lingual ability to emerge. Otherwise the PAELLA$_{\text{mono}}$ model defaults to English, as a result of having been exclusively exposed to this language and thus having a strong tendency to generate in English.

6 Discussion

We discuss PAELLA's performance across languages in relation to the different writing systems. We then conduct ablations studies, first discussing the monolingual data required to train PAELLA$_{\text{mono}}$, followed by the importance of the retrieved information. These ablation studies were performed on the validation split of COCO-35L because XM3600 only contains evaluation data.

6.1 Writing Systems

In Figure 2, we observe the performance of PAELLA across the diverse writing systems of the 36 languages, alongside the mBLIP mT0-XL model for comparison. mBLIP has a notable performance on English and languages that share the

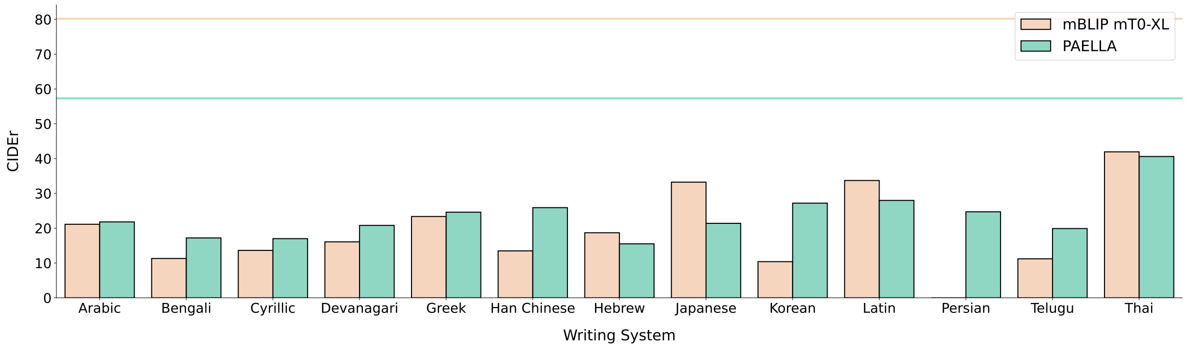


Figure 2: Performance by writing system. Horizontal lines denote corresponding English performance.

Latin script writing system. This specialization results in poor performance for some writing systems, for instance Persian and Korean. In contrast, our model demonstrates a more balanced performance across the various writing systems beyond the high-resource Latin script, achieving a better performance on the Arabic, Bengali, Cyrilic, Deveganari, Greek, simplified Chienese, Korean, Persian, and Tegulu writing systems.

6.2 Monolingual Supervision

We previously saw that our multilingual captioning model could also be trained on monolingual data (see Section 5.2). We now discuss whether PAELLA$_{\text{mono}}$ works when trained with languages other than English. As seen in Table 3, PAELLA$_{\text{mono}}$ exhibits zero-shot multilingual capabilities with the other 3 core languages as well. Surprisingly, training on Spanish yields better generalization to the other core languages compared to training on English. When trained on Chinese, on the other hand, the model loses its ability to generate captions in Hindi. Additionally, we investigated the model's behavior when trained with a language falling outside the pre-training of the XGLM decoder, such as Danish. Here, the model is able to generate captions in Danish, yet we see the interesting behaviour that this breaks the generalization to other languages.

6.3 Retrieval as PAELLA's Key Ingredient

We now study the importance of augmenting with retrieved examples, the key component of our approach. We start by ablating the retrieval component, by training without including the retrieved captions in the prompt.[10] As seen in Figure 3, the performance drops 24 CIDEr on average across the 4 core languages without retrieval (noRAG), compared to PAELLA. We also ablate the visual encoder by training on empty input images,[11] and we see again a loss of performance (i.e., 13.4 CIDEr over the 4 languages), confirming that PAELLA does indeed attend to the image and not merely rephrases the retrieved captions. Moreover, we observe that the NoRAG model performs worse than the image-blind approach with retrieved captions, reinforcing the benefit of training multilin-

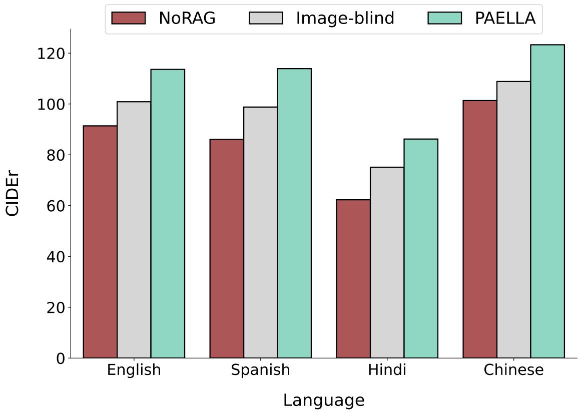


Figure 3: Ablation results on the COCO-35L validation data, reported with CIDEr metric. We ablate the retrieval (NoRAG) and the visual encoder (image-blind).

Model	en	es	hi	zh	da
PAELLA$_{\textbf{en}}$	**120.8**	91.5	45.9	59.1	2.7
PAELLA$_{\textbf{es}}$	93.3	**125.3**	52.6	95.3	2.9
PAELLA$_{\textbf{hi}}$	70.4	68.1	**99.3**	80.9	0.1
PAELLA$_{\textbf{zh}}$	65.0	49.9	1.4	**130.6**	0.4
PAELLA$_{\textbf{da}}$	5.1	1.2	2.8	4.1	**107.5**

Table 3: CIDEr results for the mono variants on the COCO-35L validation data. We denote in subscript and in bold the language each variant was trained on.

[10] The prompt only includes the last part: `A caption I can generate to describe this image in [language] is.`

[11] Setting the visual features from the encoder to zero.

gual image captioning with retrieval-augmentation. In Appendix F, we additionally discuss results for PAELLA$_{\text{mono}}$, where retrieval is shown to be crucial to generate captions in languages that substantially diverge from the English supervision. We also discuss the importance of having the retrieved captions in the target language, in Appendix H.

7 Conclusions and Future Work

We proposed PAELLA, an efficient multilingual captioning model with retrieval-augmentation. Contrary to previous studies, PAELLA is lightweight to train, both in the number of parameters and multilingual data demands. Results demonstrate competitiveness across languages, including low-resource languages. PAELLA also exhibits strong zero-shot multilingual capabilities. In the future, we plan to further investigate cross-lingual transfer with monolingual supervision.

8 Acknowledgements

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by Fundação para a Ciência e Tecnologia (FCT), through the project with reference UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) and the Ph.D. scholarship with reference 2020.06106.BD.

Limitations

While our model aims to contribute to research beyond English-centric captioning, it has limitations in that the results are conditioned on retrieved captions from machine translated data from COCO, which is English-centric and lacks coverage of geographically diverse concepts (Liu et al., 2021). Previous research has also shown that COCO has significant gender imbalance, and using this data can further amplify the bias (Zhao et al., 2017; Hendricks et al., 2018). For instance, models can become more prone to generate *woman* in kitchen settings than *man*. For a better understanding of the biases PAELLA exhibits, we suggest an analysis of the retrieved captions used by the model, as illustrated in the figures within Appendix H.

Another limitation relates to our models' coverage of languages and concepts. Expanding the range of covered languages would be desirable to accommodate more diverse speakers. Additionally, our model was evaluated on a limited number of datasets, similarly to other concurrent models, due to the scarcity of multilingual resources for assessing image captioning results.

PAELLA was only designed for the task of image captioning. In future work, we would like to investigate approaches to extend PAELLA to a range of multilingual multimodal tasks, such as those covered in IGLUE (Bugliarello et al., 2022).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the International Conference on Machine Learning*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023a. PaLI-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. PaLI: A jointly-scaled multilingual language-image model.

In *Proceedings of the International Conference on Learning Representations*.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mBLIP: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the Workshop on Representation Learning for NLP*.

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 (ECCV)*.

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*.

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*.

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*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. *arXiv preprint arXiv:2010.00710*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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 *Proceedings of the Workshop Text Summarization Branches Out*.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. 2022. I-tuning: Tuning language models with image for caption generation. *arXiv preprint arXiv:2202.06574*.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

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*.

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.

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 *Proceedings of the International Conference on Machine Learning*.

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

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023a. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. Findings of the Association for Computational Linguistics.

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*.

Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, and Bruno Martins. 2021. Retrieval augmentation for deep neural networks. In *Proceedings of the International Joint Conference on Neural Networks*.

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

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*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

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*.

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*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

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*.

Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wangrong Cheng, and Jinwen Tian. 2019. A unified generation-retrieval framework for image captioning. *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, and Vicente Ordonez. 2020. Using visual feature space as a pivot across languages. In *Findings of the Association for Computational Linguistics*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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*.

Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, and Jiaxuan Zhang. 2020. Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, 10(18).

A Prompt

To generate captions across different languages, we customize our prompt and the retrieved captions to be in the selected language. In Figure 4, we give examples in Spanish, Hindi, and Chinese, respectively. The prompts for the other languages are included in our code.

B Retrieval

Ramos et al. (2023b) has shown in the SmallCap retrieval-augmented captioning model that CLIP-ViT-B/32 is suitable as an encoder for text generation, but when used as a retrieval encoder it performs poorly. We thus pick the state-of-the-art version of CLIP, CLIP-ViT-bigG-14, for retrieval. We refrain from using that larger version in the model's encoder too, since that would significantly slow down training time.

C Standard Evaluation Metrics

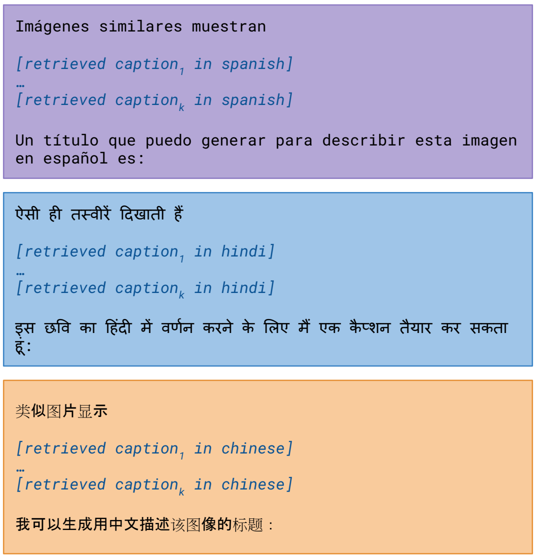
For a more comprehensive evaluation, we report the performance of our model with additional automatic metrics, including BLEU-1 (B-1), BLEU-4 (B-4) (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Denkowski and Lavie, 2014). We report these metrics both for the XM3600 dataset and the COCO-35L validation split, as seen in Table 4 and Table 5, respectively.

	B-1	B-4	ROUGE-L	METEOR
en	45.1	10.3	34.6	14.5
es	43.2	7.8	30.1	15.1
hi	29.3	2.7	21.1	21.9
zh	32.1	6.9	24.6	10.9

Table 4: PAELLA performance on the XM3600 dataset, across different evaluation metrics.

	B-1	B-4	ROUGE-L	METEOR
en	76.2	33.6	55.9	26.7
es	76.3	35.9	54.5	27.5
hi	74.9	26.5	51.0	33.7
zh	77.2	40.0	56.4	28.8

Table 5: PAELLA performance on the COCO-35L validation split, across different evaluation metrics.



Imágenes similares muestran

[retrieved caption$_1$ in spanish]
…
[retrieved caption$_k$ in spanish]

Un título que puedo generar para describir esta imagen en español es:

ऐसी ही तस्वीरें दिखाती हैं

[retrieved caption$_1$ in hindi]
…
[retrieved caption$_k$ in hindi]

इस छवि का हिंदी में वर्णन करने के लिए मैं एक कैप्शन तैयार कर सकता हूं:

类似图片显示

[retrieved caption$_1$ in chinese]
…
[retrieved caption$_k$ in chinese]

我可以生成用中文描述该图像的标题：

Figure 4: Examples of prompts in Spanish, Hindi and Chinese, respectively, shown from the top.

D Scalability

In Table 6, we see how PAELLA performs with different XGLM versions in the decoder. The larger-scale XGLM-2.9B has stronger performance, which aligns with previous findings regarding the scaling behaviour of LMs. Notwithstanding, the XGLM-1.7B and XGLM-564M versions are viable alternatives, considering that they can be trained in even less time and occupy less GPU memory. We also report performance on the validation split of COCO-35L in Table 7.

XGLM	Time	RAM	en	es	hi	zh
2.9B	23h	46G	57.3	44.9	20.8	25.9
1.7B	14h	29G	55.8	41.0	20.1	24.6
564M	7h	19G	51.7	40.0	18.0	23.8

Table 6: CIDEr results on the XM3600 dataset. We report performance for different XGLMs used in the decoder component of PAELLA.

XGLM	Time	RAM	en	es	hi	zh
2.9B	23h	46G	113.6	113.9	86.2	123.3
1.7B	14h	29G	108.7	107.7	82.2	116.6
564M	7h	19G	103.2	103.1	76.6	111.2

Table 7: CIDEr results on the validation set of COCO-35L, across the different decoders used in PAELLA.

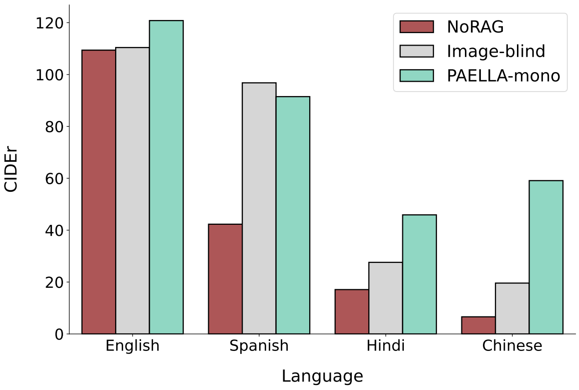


Figure 5: Ablation results on the COCO-35L dataset, reported with the CIDEr metric for the mono variant. We ablate the retrieval (NoRAG) and the visual encoder (image-blind), and compare with PAELLA$_{\text{mono}}$.

E Monolingual Retrieval

We study the behavior of our model when the retrieved captions are provided in English instead of the target languague, as seen in Table 8. We can see that our model benefits from having the retrieved examples in the same language as the target output language. In this manner, the captions can guide the process of generating content in the target language, by providing a reference for what the predicted caption should resemble.

RAG	en	es	hi	zh
Multi	113.6	113.9	86.2	123.3
En	114.1	103.8	76.8	121.3

Table 8: Performance of using either retrieved captions in the target language (multi) or in English, measured through CIDEr on the COCO validation set.

F Retrieval Impact on PAELLA$_{\text{mono}}$

Similarly to the findings for PAELLA in Section 6.3, we observe in Fig 5 that retrieval augmentation plays a key role in PAELLA$_{\text{mono}}$ as well. Indeed, retrieval is especially important for the monolingual variant. This happens because the model relies even more on the retrived examples to generate captions in languages that significantly differ from the English training data, as evidenced by the substantial drop in performance with NoRAG for Hindi and Chinese. We also see that the image-blind variant makes PAELLA$_{\text{mono}}$'s performance decline, demonstrating that our model uses not just the information from the retrieved captions, but also the image itself. The image-blind variant has to generate captions solely with retrieved information, which proves challenging for Hindi and Chinese. It can be difficult to figure out how to combine and summarize the information from the four retrieved captions into a cohesive single output, particularly for these languages with very distinct characteristics from the English supervision. Conversely, the model effortlessly uses the retrieved information for Spanish at inference, achieving better performance through straightforward rephrasing. Moreover, the image-blind approach outperforms the NoRAG model across all four languages, further emphasizing the importance of conditioning generation with retrieved examples.

G Cross-attention

Our model has 34M trainable parameters corresponding to the cross-attention layers. Here, we provide insight into the cross-attention setup, featuring an encoder hidden size of 768, and a decoder hidden size of 2048, with 16 attention heads and a stack of 48 layers. We reduce the size of the cross-attention projection matrices, denoted as d, from the standard 128 (2048/16) to 8, in order to achieve parameter efficient training. Consequently, the total parameter count is calculated as follows:

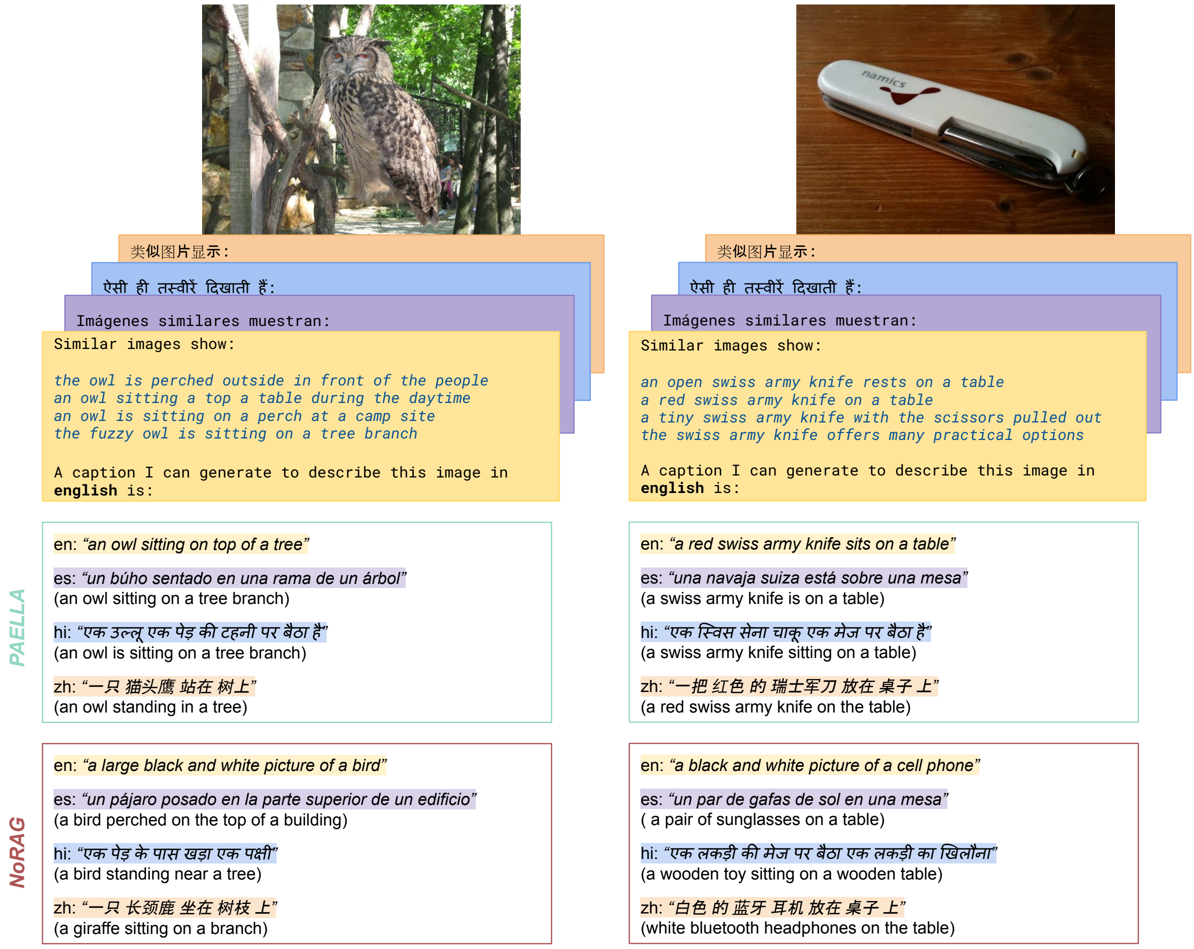
- Key Weight Matrix size: $[768, 8]$ (i.e., $enc\_d \times d$)
- Value Weight Matrix size: $[768, 8]$ (i.e., $enc\_d \times d$)
- Query Weight Matrix size: $[2048, 8]$ (i.e., $dec\_d \times d$)
- Total parameters for one layer attention with 16 heads: $16 \times (2 \times 768 \times 8 + 2048 \times 8)$
- Dense weight for projection after concatenation of heads: $[16 \times 8, 2048]$ $(h \times d \times dec\_d)$

Total number of layers is 48.

Total number of parameters: $48 \times (16 \times (2 \times 768 \times 8 + 2048 \times 8) + 16 \times 8 \times 2048) \approx 34M$

H Qualitative Results

In Fig 6, we provide examples of captions generated by PAELLA, conditioned on both the image and its retrieved captions, and captions generated by the variant without retrieval (NoRAG). In the



```
类似图片显示:
ऐसी ही तस्वीरें दिखाती हैं:
Imágenes similares muestran:
Similar images show:

the owl is perched outside in front of the people
an owl sitting a top a table during the daytime
an owl is sitting on a perch at a camp site
the fuzzy owl is sitting on a tree branch

A caption I can generate to describe this image in
english is:
```

PAELLA

en: *“an owl sitting on top of a tree”*

es: *“un búho sentado en una rama de un árbol”*
(an owl sitting on a tree branch)

hi: *“एक उल्लू एक पेड़ की टहनी पर बैठा है”*
(an owl is sitting on a tree branch)

zh: *“一只 猫头鹰 站在 树上”*
(an owl standing in a tree)

NoRAG

en: *“a large black and white picture of a bird”*

es: *“un pájaro posado en la parte superior de un edificio”*
(a bird perched on the top of a building)

hi: *“एक पेड़ के पास खड़ा एक पक्षी”*
(a bird standing near a tree)

zh: *“一只 长颈鹿 坐在 树枝 上”*
(a giraffe sitting on a branch)

```
类似图片显示:
ऐसी ही तस्वीरें दिखाती हैं:
Imágenes similares muestran:
Similar images show:

an open swiss army knife rests on a table
a red swiss army knife on a table
a tiny swiss army knife with the scissors pulled out
the swiss army knife offers many practical options

A caption I can generate to describe this image in
english is:
```

en: *“a red swiss army knife sits on a table”*

es: *“una navaja suiza está sobre una mesa”*
(a swiss army knife is on a table)

hi: *“एक स्विस सेना चाकू एक मेज पर बैठा है”*
(a swiss army knife sitting on a table)

zh: *“一把 红色 的 瑞士军刀 放在 桌子 上”*
(a red swiss army knife on the table)

en: *“a black and white picture of a cell phone”*

es: *“un par de gafas de sol en una mesa”*
(a pair of sunglasses on a table)

hi: *“एक लकड़ी की मेज पर बैठा एक लकड़ी का खिलौना”*
(a wooden toy sitting on a wooden table)

zh: *“白色 的 蓝牙 耳机 放在 桌子 上”*
(white bluetooth headphones on the table)

Figure 6: Qualitative examples for the captions generated by PAELLA, compared with the results generated with an ablated model that does not use retrieval augmentation.

first image, our model correctly captures the concept of owl across the different core languages, as present in the retrieved captions. PAELLA also demonstrates some robustness to potential misinformation that can occur in the retrieved captions (e.g., the second retrieved caption mentions an owl in a table). In contrast, the NoRAG variant generates incorrectly the captions for the 4 languages, struggling with identifying the bird, even misclassifying it as a giraffe for Chinese. On the second image, we present a negative example where the retrieved captions can mislead our model. PAELLA generates captions mentioning a red Swiss Army knife, likely influenced by the color present in the retrieved captions (and partially in the knife itself, although it is mainly white). Nonetheless, our model successfully generates the concept of a Swiss knife, while the NoRAG variant encounters difficulty by generating unrelated objects (e.g., either a cell phone, sunglasses, a toy or headphones for English, Spanish, Hindi, and Chinese, respectively).

I Performance Across the 36 Languages

In Table 9, we report XM3600 performance across all the 36 languages. We show results for our model and its variants, together with state-of-art multilingual models that have the performance for each language in the respective publications too.

Lang.	mBLIP mT0-XL	BB+CC	Lg	Mono	Core	PAELLA
en	80.2	58.4	34.3	58.2	58.2	57.3
ru	27.3	19.4	8.9	21.4	20.9	20.7
zh	13.5	20.2	9.9	23.5	25.4	25.9
de	32.5	22.4	13.0	21.7	22.1	21.5
es	62.6	42.5	22.0	42.2	45.0	44.9
fr	57.6	41.0	21.7	36.1	38.9	40.6
ja	33.2	25.4	14.1	13.0	18.6	21.4
it	45.2	32.1	16.8	29.3	32.5	33.2
pt	53.1	38.0	20.2	38.7	40.0	41.0
el	23.4	19.9	10.1	23.3	21.7	24.6
ko	10.4	28.8	15.2	21.7	21.2	27.2
fi	16.8	17.7	8.9	15.6	16.9	18.1
id	38.5	30.7	16.7	34.0	34.3	31.6
tr	22.6	23.2	12.2	19.0	19.3	21.5
ar	21.1	22.7	10.6	17.3	19.0	21.8
vi	39.2	33.6	18.2	39.3	38.7	38.0
th	41.9	41.8	22.6	20.8	22.1	40.4
hi	16.1	19.7	11.1	17.1	20.4	20.8
bn	11.3	20.0	13.3	18.8	16.5	21.7
sw	11.8	31.9	15.1	23.0	22.8	28.5
te	11.2	19.6	9.9	17.2	15.3	19.9
quz	1.1	0.0	0.0	0.2	0.7	0.8
Languages not in XGLM pre-training data						
cs	31.8	31.3	13.9	0.5	0.2	21.6
da	44.2	32.9	19.2	1.0	1.0	27.3
fa	0.0	31.1	15.5	1.5	1.5	24.7
fil	17.7	35.3	18.5	1.7	2.2	26.6
he	18.7	23.0	9.8	0.0	0.0	15.5
hr	5.2	22.4	8.5	0.3	0.2	16.0
hu	21.5	17.5	9.6	0.4	0.1	11.5
mi	4.1	40.5	24.3	1.1	3.6	33.4
nl	55.7	44.1	23.2	1.9	2.5	36.5
no	46.2	38.5	23.0	1.0	1.8	31.0
pl	31.2	23.6	10.8	0.4	0.2	17.9
ro	21.7	18.8	10.0	0.8	1.2	15.3
sv	48.4	37.0	22.5	1.0	2.0	31.6
uk	0.0	18.9	8.1	2.8	2.5	13.3
AVG	28.3	28.5	15.0	15.5	16.8	26.2
AVG*	30.5	27.7	14.7	23.9	24.9	28.2

Table 9: CIDEr results on the XM3600 benchmark across the 36 languages, ordered by the pre-training language ratio of the XGLM decoder. AVG indicates the average performance across the 36 languages, whereas AVG* indicates performance across the languages on which XGLM was pre-trained.