CONCAP: Seeing Beyond English with Retrieval-Augmented Captioning

George Ibrahim[1] Rita Ramos[2] Yova Kementchedjhieva[1]
[1]Department of Natural Language Processing, MBZUAI
[2]INESC-ID, Instituto Superior Técnico, University of Lisbon
{george.ibrahim, yova.kementchedjhieva}@mbzuai.ac.ae

Abstract

Multilingual vision-language models have advanced image captioning but still lag behind English models due to limited multilingual training data and expensive model scaling. Retrieval-augmented generation (RAG) offers a promising alternative by conditioning caption generation on retrieved example captions, reducing the need for large-scale multilingual training. However, RAG models often rely on English-translated captions, which can cause linguistic and cultural bias. We introduce CONCAP*, a multilingual image captioning model that combines retrieved captions with image-specific concepts to better contextualize the image and improve cross-lingual grounding. Experiments on XM3600 show that* CONCAP *achieves strong performance with much less training data. These results highlight the value of concept-based retrieval in multilingual captioning and open avenues for cultural adaptation.*

1. Introduction

Vision-language models (VLMs) have achieved strong performance in image captioning, largely driven by model scaling and extensive training data [9, 10, 21]. While effective, most VLMs remain English-centric and struggle to generate high-quality captions in other languages. Growing efforts focus on building multilingual datasets and training multilingual VLMs [3, 23]. Yet, even with costly large-scale training and high parameterization, recent models still exhibit a large gap between English and other languages [23].

To reduce the need for expensive multilingual training, recent work has explored retrieval-augmented generation (RAG) as a promising approach [14, 18]. In this setup, models are conditioned on both the image and a set of retrieved captions for similar images, enabling multilingual captioning with less training data. However, a key limitation of RAG models is that retrieved captions often come from translated datasets [20] and Western-centric image repositories [1], which may miss nuances of the target language and culture. Additionally, retrieved captions are from images *similar* to the input but still differ in various aspects, inherently introducing noise into the generation process.

We introduce CONCAP, a multilingual image captioning model that integrates retrieved concepts (CON) and captions (CAP). CONCAP builds on the mBLIP architecture [3], augmenting generation with retrieved information from captions similar to the input image, and concepts uniquely relevant to it. This enriches the captioning process with informative and precise context. Fine-tuning CONCAP on just 0.6M multilingual image-caption pairs yields considerably better results on the XM3600 multilingual captioning benchmark [20] as compared to the original mBLIP model trained on 4M samples ($\Delta = 5.9$), and a 7B state-of-the-art multilingual VLM trained on 6M samples [23] ($\Delta = 2.4$). Through ablation studies of the caption and concept retrieval augmentations, we show that each component makes a valuable contribution, and the two are complementary.

The concept retrieval, in its base form referred to above, is done using a lexicon built from the training data, which again, is translated. The natural next step is the integration of culturally-representative lexicons. We present early results in this direction showing that this is not a trivial task. Concepts retrieved from enriched lexicons prove noisier and hamper the accurate generation of image captions. Oracle experiments with cultural concepts show that retrieval improves captioning, but this fades as training progresses. The model increasingly depends on its learned vocabulary, limiting integration of new cultural concepts.

2. Related Work

Culturally-aware vision-language modeling has gained prominence in recent literature, largely driven by new benchmarks for the task [11, 19, 23]. One concurrent modeling effort, Pangea, provides extensive synthetic training resources through PangeaIns, an instruction-tuning dataset featuring diverse tasks, including a culture-oriented subset of visual question-answering and image captioning [23].

Retrieval-augmented generation (RAG) methods provide additional context to the model by incorporating information retrieved from an external datastore [6]. Closely related

to our work, Ramos et al. [16] proposes SmallCap, showing that augmenting an image captioning model with retrieved captions not only improves captioning performance in English but also reduces the number of trainable parameters and facilitates adaptation to out-of-domain settings through enrichment of the retrieval datastore. Meanwhile, Li et al. [8] introduces the EVCap model, which incorporates retrieved concepts (e.g., object names) instead of full captions to avoid redundancy and mitigate misleading information in the retrieved text. Zeng et al. [24] also explores augmenting image captioning with key concepts related to the image, but in a zero-shot setup.

While RAG has gained traction in image captioning, its use in multilingual image captioning has been limited: caption-based retrieval was shown to be data- and parameter-efficient [18] and effective even in a training-free setup [15]. Concept-based augmentation, on the other hand, is yet to be explored in a multilingual context. We go a step further and explore the combination of retrieved captions and concepts to enhance multilingual caption generation.

3. Proposed Approach

ConCap is an efficient multilingual retrieval-augmented image captioning model that combines retrieved captions and retrieved concepts to generate more accurate and contextually grounded captions across languages. The proposed method is applicable to any vision-language model (VLM), as it only modifies an input-specific prompt passed to the language decoder of the model.

3.1. Retrieval

ConCap builds on standard retrieval strategies [17, 18], which use vision-language representation models [13, 25] to align images and text in a shared representation space.

Captions Given a corpus of image captions, the text encoder of a CLIP-style model is used to pre-compute and cache the caption embeddings into an index. This index is then queried with input images represented with the image encoder of the same model, and cosine similarity to find the top-n captions most similar to the image.

Concepts Concept retrieval is performed analogously, but instead of full sentences, it concerns individual lexical items. We contextualize those for the purposes of retrieval, using a short language-specific template (e.g., "a photo of a {concept}"). The top-m most similar concepts to an input image are retrieved.

The intuition behind combining retrieved captions and concepts is simple. The retrieved captions serve as an example of what a caption should look like and mention some relevant concepts, but also some irrelevant ones. Retrieved concepts, on the other hand, provide no syntactic or stylistic cues, but being highly relevant to the contents of the image, they can counteract noise from the retrieved captions and fill in possible gaps in the coverage of the image contents.

3.2. Prompt Format

The final prompt for the language decoder of a VLM combines both retrieved captions and concepts to provide informative and semantically relevant context for caption generation. The prompt is organized into three segments:

`Similar images show:` `caption_1, caption_2, ..., caption_n.`
`This image might contain:` `concept_1, concept_2, ..., concept_m.`
`Caption in {lang}:`

The first segment provides n similar captions, offering sentence-level context. The second lists m highly-relevant concepts. Together, they help the model generate accurate, grounded captions in the target language. In our experiments, the prompt is always in English and mentions the full name of the target language (e.g., "Caption in Chinese").

The prompt is used as an additional conditioning context in the generation of image captions. The proposed method is supervised and relies on image captioning training data.

4. Experiments

4.1. Experimental Setup

Base Model ConCap adopts the mBLIP architecture and its initialization strategy. The vision encoder, Q-Former, and projection layer are initialized from BLIP-2 [7], and the language decoder from mT0-XL [12]. Following Geigle et al. [3], we freeze the vision encoder and language decoder, and insert LoRA layers [4] into the decoder. As a result, only the Q-Former, projection layer, and LoRA layers are updated during training. This setup maintains training efficiency with ∼111 million trainable parameters.

Training and Evaluation Data For training, we use the COCO-35L dataset [20], which comprises 19.8M image-caption pairs in 35 languages. Following Ramos et al. [18], we subsample the training set to 566K image-caption pairs with equal representation across languages. Training is done with teacher forcing and cross-entropy loss.

We evaluate on XM3600 [20], a human-annotated benchmark with 3,600 images captioned in 36 languages. Images are sourced from countries where the languages are spoken, with each image averaging two captions per language. To enable fast yet reliable multilingual evaluation, recent work [23] introduced XM100, a subset of 100 representative images from XM3600, including all languages.

Caption Retrieval Captions are always retrieved from the COCO-35L training set. We build separate indexes for each language and retrieve the top-4 captions per image, following prior work [17]. The implementation of the retrieval

Models	Train θ	Total θ	Dataset	en	es	hi	zh	L_5	L_{36}
mBLIP	111M	4.84B	2.71M	**80.2**	62.6	16.1	13.5	7.9	28.3
PAELLA	34M	3B	566K	57.3	44.9	20.8	25.9	20.7	26.9
BB + CC	0.8B	0.8B	135M	58.4	42.5	19.7	20.2	**22.4**	29.3
Pangea	7B	7B	6M	75.9	**64.6**	16.2	**29.0**	12.5	31.8
CONCAP	111M	4.84B	566K	72.4	58.6	**24.4**	21.7	18.2	**34.2**

Table 1. CIDEr scores on the XM3600 evaluation set. Total θ: model size, Train θ: number of trainable parameters, Dataset: size of training datasets.

Models	en	es	hi	zh	L_5	L_{36}
CONCAP	**72.4**	**58.6**	**24.4**	**21.7**	**18.2**	**34.2**
NoRAG	66.0	48.9	20.4	17.4	17.2	26.9
ConRAG	70.9	55.0	19.5	20.9	17.1	30.3
CapRAG	66.2	53.3	23.9	20.2	16.9	31.4
ConRAG_R	71.4	51.2	19.2	17.9	16.5	28.6
CapRAG_M	38.3	30.4	16.4	13.9	13.1	20.4

Table 2. CIDEr scores on XM3600 across ablations.

is distinct between training and evaluation to match dataset characteristics. COCO-35L primarily consists of images from English-speaking contexts originally captioned in English with translations to other languages. We therefore follow Ramos et al. [18] and adopt an English-as-a-pivot retrieval strategy, using CLIP [13] to retrieve captions from the English subset and mapping them to target languages via shared IDs (see §4.4 for target-language retrieval experiments). XM3600 includes geographically diverse images, so at inference, we use the multilingual retriever mSIGLIP [25], which should better represent diverse image inputs.

Concept Retrieval We extract unique concepts for each language from COCO-35L captions and apply language-specific templates with prefixes or suffixes (e.g., `"a photo of a " + <concept>` in English) to improve their alignment with the captioning task.

We maintain a separate wordlist for each language, and use mSigLIP [25] for retrieval. After hyperparameter tuning over $m = 4, 10, 20$, we select $m = 10$. For fast nearest neighbor search over dense embeddings, we use FAISS [2].

Evaluation We evaluate performance using the CIDEr score [22], with a special focus on the four XM3600 languages that Thapliyal et al. [20] defines as a representative core: English, Spanish, Hindi, and Chinese ($\mathbf{L}_4$), as well as five low-resource languages of special interest: Bengali, Maori, Quechua, Swahili, and Telugu ($\mathbf{L}_5$). For inference, we use beam search of size 5, a length penalty of 1, and a maximum generated caption length of 25 tokens.

4.2. Main Results

Averaged Performance ($\mathbf{L}_{36}$) Table 1 shows that on average, CONCAP outperforms all baselines by a large margin, exceeding the next best (Pangea) by 2.4 points. This is notable given the large gap in the trainable parameters (111M vs. 7B) and the amount of vision-language training data (566K vs. 6M). Comparing CONCAP to mBLIP, where both models share the same architecture and number of trainable parameters, we observe a gap of nearly 4 CIDEr points, despite an almost five-fold reduction in training data with CONCAP: combined concept and caption retrieval augmentation enables highly data-efficient training.

4.3. The Contribution of Concepts and Captions

Having established CONCAP's strong overall performance, we now analyze the contribution of its components by comparing models trained without retrieval (NoRAG), with retrieved captions (CapRAG), and with retrieved concepts (ConRAG). Results are shown in Table 2.

We find that CONCAP 's impressive performance is indeed attributed to the retrieval augmentation, as its non-augmented counterpart, NoRag, performs on par with the weaker baselines from Table 1. Looking at the two forms of retrieval augmentation in isolation, we find that each improves performance over NoRAG on average, with CapRAG being slighly more effective than ConRAG. While the gain from caption retrieval corroborates prior work [18], it is interesting to see that concepts alone can also provide a highly effective signal in this multilingual setting. The most insightful finding here is that CONCAP considerably outperforms both ConRAG and CapRAG, indicating that the gains from these two forms of retrieval augmentation are *additive*: captions help guide generation with fluent language patterns, while concepts ensure broader content coverage and more accurate grounding in the input image.

4.4. The Role of English as a Pivot

Previous work shows that retrieving captions via English works well on COCO-35L, but this relies on the availability of parallel multilingual captions, which may not be available in real-world scenarios. To address this, we experiment with target-language caption retrieval using mSigLIP in the CapRAG setting (Table 2, CapRAG_M). We find that this approach considerably underperforms English-as-a-pivot retrieval (CapRAG) across all languages and even lags behind NoRAG, indicating that in this case, retrieval degrades performance. This highlights the critical role of retriever quality in reteival-augmented multilingual captioning.

In this sense, retrieval augmentation with concepts seems more robust (ConRAG), as it consistently improves performance despite possible limitations in the mSigLIP retriever.

5. Concept Enrichment

One advantage of concept retrieval is that lexicons are easier to obtain than high-quality image captions. En-

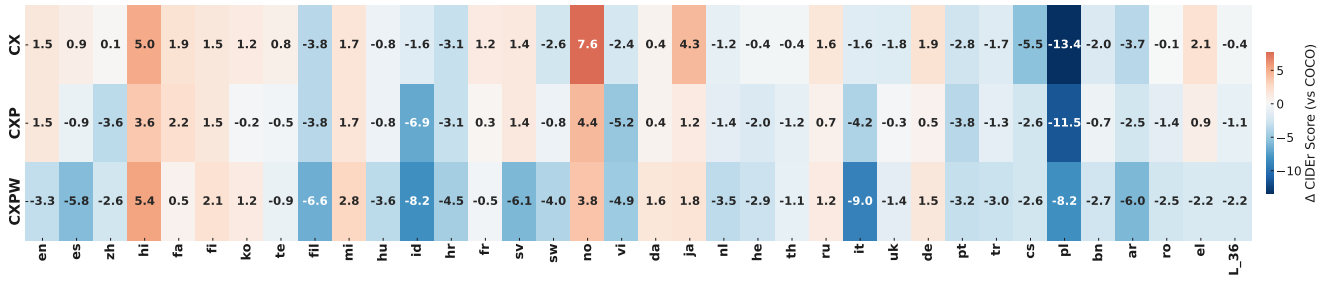

Figure 1. Change is CIDEr score on the XM100 test set when enriching the concept-retrieval lexicon. Positive values indicate an improvement over the baseline lexicon. Lexicon enrichment appears to have a mostly negative impact on performance.

riched concept retrieval could improve generalization and out-of-domain coverage, as well as the cultural awareness of VLMs. We test this by enriching the original COCO-35L lexicon with lexical items from the cultural subset of PangeaIns [23], aiming to enrich ConCap's geographic coverage and cultural awareness. The results in Table 2 (ConRAG$_R$) show a 1.7 CIDEr drop compared to ConRAG.

To better understand this finding, we test ConRAG on XM100 with varying concept list configurations, focusing on sensitivity to index size and makeup. We compare: (1) **CX**, which adds filtered XM3600 lexicons (excluding the XM100 captions); (2) **CXP**, which includes PangeaIns cultural terms; and (3) **CXPW**, which adds Wikipedia and Common Crawl entries for broader but less focused coverage. Per-language results are presented in Figure 1, as a change in CIDEr score from the base lexicon (COCO-35L) to the different augmented lexicon configurations.

This addition of the filtered XM3600 lexicon (CX) shows a mixed effect, with roughly half of the languages seeing improvements and the other half experiencing a performance drops. Expanding this setup with a culturally-relevant lexicon (CXP) leads to a further decline in performance for a larger portion of the languages. Finally, incorporating a broad web-based lexicon (CXPW), results in the majority of languages showing a degradation in performance. While expanding the lexicon pool could, in theory, improve the coverage of retrieved concepts with geographically diverse and culturally relevant terms, in practice it seems to add noise which distracts the generation process and yields lower-quality captions.

Lastly, we conduct an oracle experiment on 200 images from the JEEM dataset [5], providing the ConRAG model with only highly relevant cultural concepts. The images were manually selected by a native Arabic speaker based solely on visual content, independent of the captions, to ensure cultural relevance. This setup improves performance, achieving a CIDEr score of **17.9**, compared to **13.9** when using a general-purpose concept wordlist (only COCO-35L). A qualitative example from JEEM is shown in Figure 2.

However, this gain occurs only at earlier checkpoints. At later checkpoints, the model struggles to incorporate the provided concepts, suggesting that supervised fine-tuning shifts the decoder's output distribution, making unseen lexical items harder to predict, even when prompted.

Figure 2. ConRAG's strength in cultural contexts.

6. Conclusion

We introduced ConCap, a multilingual image captioning model that enhances caption generation by integrating retrieved captions with image-specific concepts. This approach improves caption quality while reducing the need for extensive multilingual training. Experiments on XM3600 show that ConCap outperforms strong baselines with far fewer training resources. Ablation studies confirm the additive benefits of caption and concept retrieval. While we showed that retrieved concepts can support cultural settings, their effectiveness depends on concept relevance, integration strategy, and the model's training stage, with earlier checkpoints more receptive to novel concepts. Future work should explore strategies for concept enrichment that promote semantic diversity and reduce redundancy.

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[2] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024. 3

[3] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1, 2

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[5] Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. Jeem: Vision-language understanding in four arabic dialects, 2025. 4

[6] 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. 1

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[8] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2024. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[11] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding, 2024. 1

[12] et al. Niklas Muennighoff. Crosslingual generalization through multitask finetuning, 2023. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3

[14] Rita Ramos, Bruno Martins, and Desmond Elliott. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1635–1651, Toronto, Canada, 2023. Association for Computational Linguistics. 1

[15] Rita Ramos, Bruno Martins, and Desmond Elliott. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. Findings of the Association for Computational Linguistics, 2023. 2

[16] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. SmallCap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[17] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation, 2023. 2

[18] Rita Ramos, Emanuele Bugliarello, Bruno Martins, and Desmond Elliott. PAELLA: Parameter-efficient lightweight language-agnostic captioning model. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. 1, 2, 3

[19] Florian Schneider and Sunayana Sitaram. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA, 2024. Association for Computational Linguistics. 1

[20] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. 1, 2, 3

[21] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1

[22] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 3

[23] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages, 2025. 1, 2, 4

[24] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110, 2024. 2

[25] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 2, 3