

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 **Parameter-Efficient Lightweight Language-Agnostic** 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 $3\text{--}77\times$ more learned parameters and $35\text{--}863\times$ 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

¹Code and model available at <https://github.com/RitaRamo/paella>.

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 **Parameter-Efficient Lightweight Language-Agnostic** 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-

PAELLA: Tham số-Hiệu quả Nhẹ Ngôn ngữ-Không phụ thuộc Mô hình chú thích

Rita Ramos[†] Emanuele Bugliarello Bruno Martins[†] Desmond Elliott

[†] INESC-ID, Instituto Superior Técnico, Đại học Lisbon

Nghiên cứu của Google

Khoa Khoa học máy tính, Đại học Copenhagen
ritaparadaramos@tecnico.ulisboa.pt

Tóm tắt

Chúng tôi giới thiệu PAELLA, một sản phẩm hiệu quả về mặt tham số Mô hình chú thích hình ảnh nhẹ không phụ thuộc vào ngôn ngữ được thiết kế để có cả tham số và sử dụng dữ liệu hiệu quả bằng cách tăng cường truy xuất. Mô hình được đào tạo bằng cách học một mạng lưới ảnh xạ nhỏ với 34M tham số giữa một mô hình trực quan được đào tạo trước và mô hình ngôn ngữ đa ngôn ngữ được điều kiện hóa trên hai loại của đầu vào: (i) bản thân hình ảnh và (ii) một tập hợp các chú thích được lấy lại bằng ngôn ngữ đích. các ví dụ được lấy lại đóng vai trò quan trọng trong việc hướng dẫn mô hình tạo chú thích trên nhiều ngôn ngữ. Thông qua việc lấy lại, mô hình có thể nhẹ về số lượng có thể đào tạo các tham số, chỉ tồn tại trong ảnh xạ của nó mạng lưới, và cũng trong số lượng đa ngôn ngữ dữ liệu đào tạo cần thiết. Các thí nghiệm về bộ dữ liệu XM3600, có 36 ngôn ngữ, chứng minh rằng PAELLA có thể vượt trội hơn hoặc cạnh tranh với một số mô hình có hiệu suất cao hơn từ 3-77 lần các thông số đã học và dữ liệu nhiều hơn 35-863 lần, đặc biệt là trong các ngôn ngữ có ít tài nguyên. Chúng tôi cũng thấy rằng PAELLA chỉ có thể được đào tạo trên dữ liệu đơn ngữ và vẫn cho thấy khả năng bắn trúng đích mạnh mẽ bằng các ngôn ngữ khác.1

1 Giới thiệu

Chúng tôi giải quyết vấn đề chú thích hình ảnh đa ngôn ngữ, nhằm mục đích cung cấp mô tả bằng văn bản về nội dung trực quan có thể phục vụ cho người nói ở nhiều trình độ khác nhau ngôn ngữ, trái ngược với hầu hết các mô hình chú thích chỉ tạo ra phụ đề tiếng Anh. Mặc dù đã có những tiến bộ đáng kể trong những năm gần đây, đào tạo mô hình chú thích hình ảnh đã trở nên nhiều hơn nữa do xu hướng mở rộng cả dữ liệu và kích thước mô hình (Hu et al., 2022; Wang et al., 2022). Xu hướng này thậm chí còn nổi bật hơn trong các phương pháp tiếp cận đa ngôn ngữ (Chen et al., 2023b; Thapliyal et al., 2022), cho thấy nhu cầu về dữ liệu đào tạo bao gồm từng ngôn ngữ đích và nhu cầu thậm chí

1Mã và mô hình có sẵn tại [https://github.com/](https://github.com/RitaRamo/cd_m_thap_cam_Paella)

RitaRamo/cd m thap cam Paella.

các mô hình lớn hơn để giảm thiểu tác hại của đa ngôn ngữ (Conneau và cộng sự, 2020; Goyal và cộng sự, 2021).

Một số nghiên cứu gần đây tập trung vào việc giảm thiểu

chi phí đào tạo đa ngôn ngữ, chẳng hạn như PALI-3 (Chen và cộng sự, 2023a) với 5B tham số có thể đào tạo được, và mBLIP (Geigle và cộng sự, 2023) chỉ với 124M

các thông số có thể đào tạo được. Cả hai cách tiếp cận này đều sử dụng các mô hình ngôn ngữ đa phương thức được đào tạo trước hoặc các bộ mã hóa hình ảnh được đào tạo trước để giữ nguyên, làm giảm số lượng các tham số có thể đào tạo được. Tuy nhiên, cả hai mô hình này vẫn dựa vào đào tạo với hàng triệu hoặc hàng tỷ ví dụ, bao gồm cả trong chỉ dựa vào ngữ cảnh chú thích hình ảnh.

Bài báo này mô tả một Tham số-Hiệu quả

Mô hình chú thích nhẹ không phụ thuộc vào ngôn ngữ (PAELLA). Mô hình được thiết kế để có hiệu quả, không chỉ về số lượng tham số có thể đào tạo mà còn nhẹ về số lượng

dữ liệu đào tạo đa ngôn ngữ được yêu cầu. PAELLA có chỉ có 34 triệu tham số được đào tạo và mô hình có thể được đào tạo chỉ bằng 566K ví dụ, tức là kích thước của tập dữ liệu COCO tiếng Anh.

PAELLA dựa trên các mô hình được đào tạo trước đông lạnh được bổ sung bằng các ví dụ đã thu thập được.

Chỉ các tham số đã học mới nằm trong một ảnh xạ nhỏ gọn mạng lưới các lớp chú ý chéo giữa một đông lạnh

Bộ mã hóa hình ảnh CLIP và mô hình ngôn ngữ đa ngôn ngữ XGLM đông lạnh. Mô hình được đào tạo để tạo phụ đề bằng ngôn ngữ mong muốn bằng cách sử dụng nhắc nhở bằng ngôn ngữ đó. Hơn nữa, đã lấy lại các ví dụ hỗ trợ mô hình trong việc tạo ra các chú thích có ý nghĩa, bằng cách cung cấp các ví dụ về những gì chú thích dự đoán nên giống như. Việc sử dụng các ví dụ được lấy lại góp phần tích cực vào việc giảm cả số lượng các tham số có thể đào tạo được và lượng dữ liệu đa ngôn ngữ cần thiết.

Chúng tôi tiến hành thí nghiệm trên XM3600 (Thapliyal et al., 2022), một chuẩn mực phụ đề đa ngôn ngữ đã được thiết lập bao gồm nhiều vùng địa lý khác nhau hình ảnh có chú thích do con người chỉ thích bằng 36 ngôn ngữ. Các thí nghiệm cho thấy PAELLA có thể vượt trội hơn

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, solely 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 Pre-training (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 (Tsimploukelli 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 Augmentation

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 popularity (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 **Parameter-Efficient Lightweight Language-Agnostic** (PAELLA) captioning model uses retrieval augmentation to generate captions in multi-

thực hiện hoặc cạnh tranh với các mô hình có nhiều hơn đôi hỏi khắt khe về các tham số được đào tạo hoặc dữ liệu đào tạo. Hiệu suất của mô hình của chúng tôi trong các ngôn ngữ có ít tài nguyên đặc biệt đáng chú ý, trong trái ngược với các mô hình đồng thời như mBLIP, thường xuất sắc trong tiếng Anh và các ngôn ngữ liên quan như lại gặp khó khăn trong việc khái quát hóa hiệu quả đối với các ngôn ngữ chưa được đại diện đầy đủ ngôn ngữ.

Kết quả cũng cho thấy PAELLA chứng minh khả năng đa ngôn ngữ không cần bản khi được đào tạo chỉ với dữ liệu đơn ngữ như tiếng Anh Bộ dữ liệu COCO. PAELLA đạt được sự chuyển giao ngôn ngữ thông qua việc truy xuất, một cách long trọng bằng cách truy xuất các chú thích trong ngôn ngữ đích trong quá trình suy luận. Các nghiên cứu cắt bỏ tiếp tục chứng minh lợi ích của chúng tôi phương pháp tăng cường khả năng truy xuất.

2 Công trình liên quan

2.1 Chú thích hình ảnh

Trong những năm gần đây, việc chú thích hình ảnh đã chứng kiến những cải tiến hiệu suất ấn tượng thông qua đào tạo trước về Tầm nhìn và Ngôn ngữ (VLP) toàn diện, xem xét việc sử dụng quy mô lớn mô hình và tập dữ liệu hình ảnh-văn bản lớn bằng tiếng Anh (Wang và cộng sự, 2021; Hu và cộng sự, 2022; Li và cộng sự, 2022).

Trong nỗ lực giảm bớt chi phí tính toán ngày càng tăng, các nghiên cứu gần đây đã áp dụng các phương pháp có sẵn các mô hình mã hóa và giải mã được đào tạo trước vẫn được giữ nguyên trong quá trình đào tạo (Mokady et al., 2021; Luo và cộng sự, 2022; Ramos và cộng sự, 2023b; Mañas và cộng sự, 2023). Ví dụ, một số nghiên cứu đã sử dụng CLIP (Radford et al., 2021) là bộ mã hóa hình ảnh và GPT-2 (Radford và cộng sự, 2019) là bộ giải mã ngôn ngữ, giữ nguyên một hoặc cả hai mô hình trong quá trình đào tạo, và thay vào đó là học cách lập bản đồ mạng lưới để sắp xếp hai phương thức. Có các mô hình đóng băng tăng tốc độ đào tạo và giảm GPU sử dụng bộ nhớ (Mokady et al., 2021). Bên cạnh việc giảm chi phí tính toán, đây cũng là một phương tiện để tích hợp liền mạch các mô hình đơn thức mạnh mẽ (Tsimploukelli và cộng sự, 2021; Alayrac và cộng sự, 2022; Li et al., 2023; Dai et al., 2023), bao gồm cả quy mô lớn được đào tạo trước (Brown và cộng sự; Zhang và cộng sự, 2022; Touvron và cộng sự, 2023) và ngôn ngữ được điều chỉnh hướng dẫn mô hình (Wei et al., 2021; Chung et al., 2022; Taori et al., 2023), nếu không thì sẽ không thực tế với đào tạo toàn diện và có thể dẫn đến mất khả năng khái quát hóa do quên lãng thảm khốc (McCloskey và Cohen, 1989).

Trong lĩnh vực chú thích hình ảnh đa ngôn ngữ, thay vì đào tạo toàn diện tốn kém từ đầu

(Thapliyal và cộng sự, 2022; Yang và cộng sự, 2020), gần đây các mô hình cũng đã lựa chọn bộ mã hóa hình ảnh được đào tạo trước đồng lạnh và/hoặc bộ giải mã ngôn ngữ. Ví dụ bao gồm mBLIP (Geigle et al., 2023) hoặc PALI-3 (Chen và cộng sự, 2023a). Ngược lại với những nghiên cứu này, chúng tôi sử dụng một bộ mã hóa được đào tạo trước đồng lạnh và một bộ mã hóa đồng lạnh mô hình ngôn ngữ, được tăng cường với các dữ liệu đã thu thập được ví dụ để giảm thêm số lượng các tham số có thể đào tạo, cũng như nhu cầu mở rộng dữ liệu đào tạo đa ngôn ngữ.

2.2 Tăng cường truy xuất

Điều kiện tạo ngôn ngữ được tăng cường truy xuất điều kiện hóa quá trình tạo ra bằng cách tăng cường đầu vào với thông tin được lấy từ kho dữ liệu bên ngoài (Lewis và cộng sự, 2020). Việc truy xuất được tăng cường các mô hình đã trở nên phổ biến hơn (Khandelwal et al., 2020; Izacard et al., 2022; Shi et al., 2023; Yu et al., 2023), bao gồm cả chú thích hình ảnh (Zhao et al., 2020; Xu et al., 2019; Ramos et al., 2021; Sarto và cộng sự, 2022; Ramos và cộng sự, 2023b; Dư ng và cộng sự, 2023).

Công việc gần giống với công việc của chúng tôi hơn là SmallCap (Ramos et al., 2023b), một công ty nhẹ Mô hình phụ đề tiếng Anh sử dụng được đào tạo trước mô hình mã hóa và giải mã, và cũng sử dụng nhắc nhở với các chú thích đã lấy được. Trong bài báo này, chúng tôi khám phá cách tăng cường truy xuất có thể giúp ích để giảm không chỉ số lượng tham số có thể đào tạo mà còn cả lượng dữ liệu đào tạo. Một điểm khác biệt quan trọng khác giữa các phương pháp tiếp cận là PAELLA dựa trên một hệ thống đa ngôn ngữ được đào tạo trước mô hình ngôn ngữ thay vì tiếng Anh đơn ngữ mô hình. Chúng tôi khám phá cách thức nhắc nhở và lấy lại phụ đề nên được thiết kế để cho phép tạo ra bằng nhiều ngôn ngữ khác nhau, thay vì chỉ có tiếng Anh.

Chúng tôi lưu ý rằng việc tăng cường truy xuất vẫn còn phần lớn chưa được khám phá trong kịch bản chú thích hình ảnh đa ngôn ngữ. Cho đến nay, chỉ có Mô hình LMCap (Ramos et al., 2023a) đã sử dụng tăng cường truy xuất, nhưng chỉ trong một môi trường không cần đào tạo cách dựa trên việc thúc đẩy một mô hình ngôn ngữ đa ngôn ngữ theo cách tiếp cận không cần hình ảnh. Trong công việc, thay vào đó chúng tôi cho thấy tiềm năng của việc truy xuất tăng cường đóng góp vào việc đào tạo một mô hình chú thích hình ảnh đa ngôn ngữ.

3 Phương pháp tiếp cận được đề xuất

Ngôn ngữ nhẹ hiệu quả tham số-Mô hình chú thích Agnostic (PAELLA) sử dụng tăng cường truy xuất để tạo chú thích trong nhiều

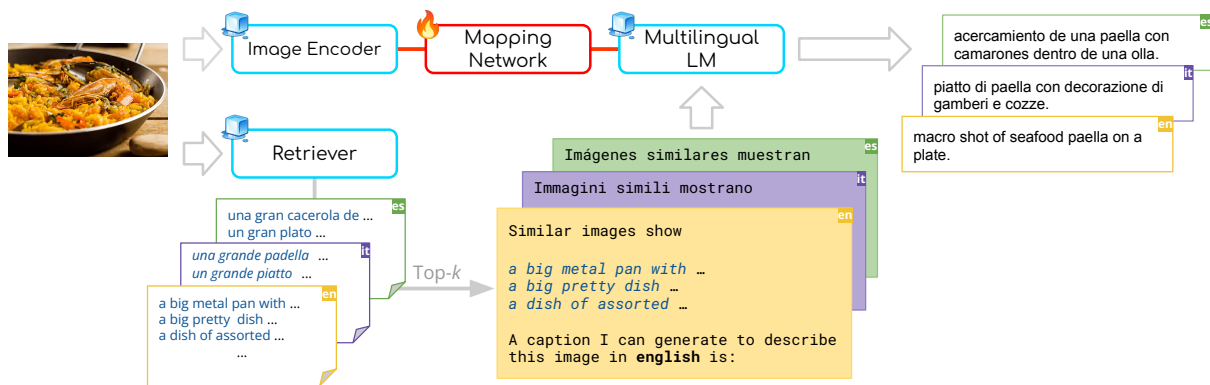


Figure 1: PAELLA uses a frozen pre-trained image encoder and a frozen multilingual decoder, connected with a trainable mapping network. The decoder generates a multilingual caption conditioned on the encoded image, together with retrieved captions given as input within a prompt in the desired language.

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 captions² 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 caption_k]. 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

²See Section 4 for details on the retrieval system.

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} :

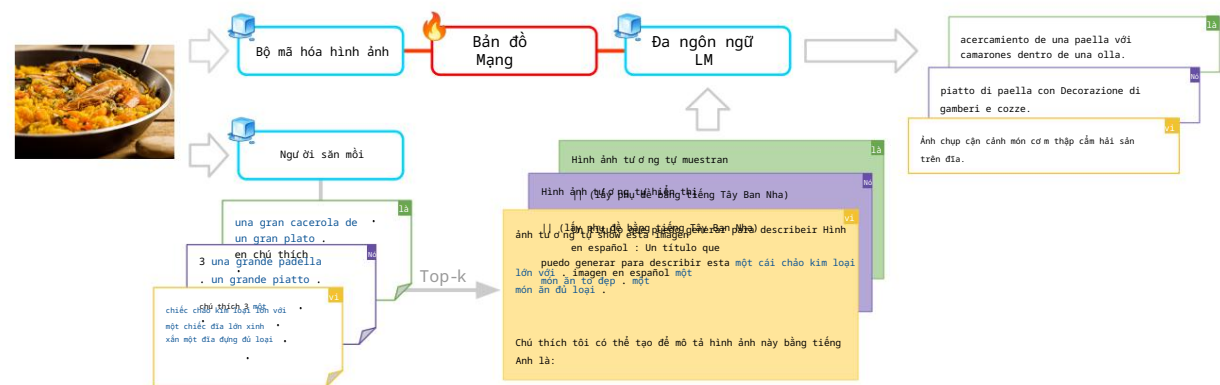
$$L_{\theta_M} = - \sum_{i=1}^M \log P_{\theta}(y_i | y_{<i}, \mathbf{V}, \mathbf{L}). \quad (1)$$

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.



Hình 1: PAELLA sử dụng bộ mã hóa hình ảnh được đào tạo trước đồng lạnh và bộ giải mã đa ngôn ngữ đồng lạnh, được kết nối với mạng lưới ánh xạ có thể đào tạo. Bộ giải mã tạo ra chú thích đa ngôn ngữ có điều kiện trên hình ảnh được mã hóa, cùng với các chú thích đã truy xuất được đưa vào làm đầu vào trong lời nhắc bằng ngôn ngữ mong muốn.

ngôn ngữ khác nhau. Tổng quan về kiến trúc mô hình có thể được thấy trong Hình 1.

Chúng tôi tuân theo một thiết kế tư ng tự như mô hình SMALLCAP đơn ngữ (Ramos và cộng sự, 2023b), bằng cách xây dựng trên các mô hình đơn thức mạnh mẽ được đào tạo trước. Chúng tôi cũng sử dụng CLIP (Radford và cộng sự, 2021) làm bộ mã hóa trực quan, nhưng thay vì GPT-2 hoặc OPT làm bộ giải mã, chúng tôi sử dụng mô hình ngôn ngữ tự hồi quy đa ngôn ngữ, tức là XGLM (Lin và cộng sự, 2021). Cả bộ mã hóa và bộ giải mã đều được giữ nguyên trong quá trình đào tạo, ngoại trừ mạng lưới ánh xạ mới được thêm vào của các lớp chú ý chéo, cho phép bộ giải mã chú ý đến thị giác

đầu vào. PAELLA tạo ra các chú thích có điều kiện trên hình ảnh và trên một tập hợp k chú thích được lấy từ các hình ảnh tư ng tự. Các chú thích được lấy được sử dụng để nhắc nhở mô hình tạo ra bằng ngôn ngữ đích mong muốn. Lời nhắc tuân theo một mẫu cố định, đầu tiên bao gồm các ví dụ về k chú thích được lấy và kết thúc bằng một hướng dẫn cho bộ giải mã đa ngôn ngữ để tạo ra một chú thích bằng ngôn ngữ mong muốn. Lời nhắc tiếng Anh là:

Hình ảnh tư ng tự hiển thị [chú thích đã lấy₁] ... [chú thích đã lấy_k]. Chú thích tôi có thể tạo để mô tả hình ảnh này bằng [ngôn ngữ] là: ...

Lời nhắc và phụ đề có thể được điều chỉnh cho phù hợp với các ngôn ngữ khác nhau bằng cách sử dụng cả hai phần này bằng ngôn ngữ mong muốn (xem một số ví dụ về lời nhắc cho các ngôn ngữ khác trong Phụ lục A).

Các tham số trong mạng lập bản đồ θ_M được đào tạo bằng cách giảm thiểu tổng số logarit âm của khả năng dự đoán hình ảnh thực tế

²Xem Phần 4 để biết thông tin chi tiết về hệ thống truy xuất.

chú thích cho mỗi mã thông báo trong chuỗi $y_1 \dots y_M$, dựa trên hình ảnh \mathbf{V} và lời nhắc tăng cường truy xuất \mathbf{L} :

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

Chúng tôi định lượng chứng minh trong Phần 5 rằng phương pháp tăng cường truy xuất của chúng tôi có các đặc tính sau:

Hiệu quả tham số: Chỉ cần đào tạo các lớp chú ý chéo giữa bộ mã hóa đóng băng và bộ giải mã đóng băng. Để bù đắp cho số lượng nhỏ các tham số có thể đào tạo, mô hình được hướng dẫn bằng các ví dụ về chú thích đã lấy.

Hiệu quả dữ liệu: Thông qua việc truy xuất, mô hình không cần một lượng lớn dữ liệu đa ngôn ngữ để đào tạo, vì nó được hưởng lợi từ các ví dụ đã truy xuất chứng minh cách tạo bằng ngôn ngữ đích. Do đó, chúng tôi giảm bớt tình trạng thiếu dữ liệu của các mô hình đa ngôn ngữ hiện có, thường được đào tạo bằng cùng một hình ảnh liên quan đến chú thích bằng nhiều ngôn ngữ, phải dịch nhiều lần toàn bộ tập dữ liệu chú thích tiếng Anh cho từng ngôn ngữ (ví dụ: COCO sang COCO-35L (Thapliyal và cộng sự, 2022)).


Đa ngôn ngữ Zero-shot: Mô hình của chúng tôi chứng minh khả năng đa ngôn ngữ ngay cả khi chỉ được đào tạo trên dữ liệu chú thích hình ảnh đơn ngữ. Nó có thể được đào tạo trên phân phối trong miền cụ thể từ dữ liệu có sẵn trong ngôn ngữ có nhiều tài nguyên và vẫn tạo ra ở các ngôn ngữ khác nhau. Điều này bằng cách chỉ dựa vào, tại thời điểm suy luận, vào việc tăng cường truy xuất trong ngôn ngữ đích từ một tập dữ liệu chú thích đa ngôn ngữ có sẵn.


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 V is encoded by the CLIP encoder, and the language-based prompt 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 $V=\{v_1, \dots, v_N\}$, 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 languages³ (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

³en, 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.

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 RAM⁴. 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),


⁴See the performance with these models in Appendix D.


4 Thiết lập thử nghiệm


4.1 Chi tiết triển khai và đào tạo

Chúng tôi phát hành mã và mô hình của mình tại <https://github.com/RitaRamo/paella>. PAELLA được triển khai bằng cách sử dụng thư viện HuggingFace Transformers (Wolf và cộng sự, 2020). Xương sống của mô hình dựa trên mô hình CLIP được đào tạo trước bởi openai/clip-vit-base-patch32 và XGLM được đào tạo trước bởi facebook/xglm-2.9B.


Hình ảnh đầu vào V được mã hóa bởi bộ mã hóa CLIP và lời nhắc dựa trên ngôn ngữ L , bao gồm k chú thích đã lấy được, được xử lý bởi XGLM để tạo chú thích bằng ngôn ngữ đích.

 Bộ mã hóa: CLIP là một mô hình đa phương thức mạnh mẽ được đào tạo trước để mã hóa hình ảnh và văn bản vào một không gian nhúng được chia sẻ, sử dụng phương pháp học tự động phân (Radford và cộng sự, 2021). Chúng tôi sử dụng CLIP-ViT-B/32 để mã hóa hình ảnh đầu vào, tạo ra một chuỗi gồm $N=50$ đặc điểm trực quan $V=\{v_1, \dots, v_N\}$, mỗi đặc điểm có kích thước nhúng là 768 chiều. Bộ mã hóa này có 86 triệu tham số, được giữ nguyên trong quá trình đào tạo.

 Bộ giải mã: XGLM là mô hình ngôn ngữ tự hồi quy đa ngôn ngữ có thể tạo ra trong một tập hợp đa dạng gồm 30 ngôn ngữ³ (Lin et al., 2021). Trong PAELLA, chúng tôi sử dụng biến thể với 2,9B tham số, được đóng băng trong quá trình đào tạo.

 Truy xuất: CLIP cũng được sử dụng để truy xuất hình ảnh văn bản. Cụ thể, nó được sử dụng để mã hóa cả chú thích ứng viên vào kho dữ liệu và từng hình ảnh đầu vào đã cho. Đối với mỗi hình ảnh đã cho, k chú thích gần nhất được truy xuất từ kho dữ liệu chú thích. Kho dữ liệu được lập chỉ mục hiệu quả thông qua thư viện FAISS (Johnson và cộng sự, 2017), cụ thể là với chỉ mục IndexFlatIP không yêu cầu bất kỳ đào tạo nào, cho phép truy xuất ngoại tuyến.

Các hình ảnh cũng được mã hóa bằng CLIP, sử dụng xương sống trực quan, để lấy các chú thích giống nhau nhất dựa trên độ tương đồng cosin. Chúng tôi chọn $k = 4$ chú thích được lấy ra hàng đầu, phù hợp với các phát hiện trước đó cho thấy đây là số lượng chú thích tối ưu trong cả thiết lập đơn ngữ và đa ngữ (Ramos et al., 2023a,b).

 Mạng lưới ánh xạ: Phần duy nhất của PAELLA được đào tạo là mạng lưới ánh xạ giữa bộ mã hóa và bộ giải mã bị đóng băng.

³En, ru, zh, De, es, Fr, Ja, nó, Pt, El, Ko, Fi, ID, Tr, Ar, Vi, th, bg, ca, hi, et, bn, ta, ur, sw, te, eu, my, ht, qu.

mạng lập bản đồ bao gồm các lớp chú ý chéo được khởi tạo ngẫu nhiên (Vaswani và cộng sự, 2017) được thêm vào mỗi lớp trong số 48 lớp của XLGM, do đó bộ giải mã có thể chú ý đến đầu ra của bộ mã hóa. Để có số lượng tham số có thể đào tạo ít hơn, chúng tôi sử dụng các lớp chú ý chéo có thứ hạng thấp bằng cách giảm chiều ban đầu của ma trận chiếu từ 128 xuống 8, như trong Ramos và cộng sự (2023b).

Theo đó, con số này chỉ tương đương với 34M tham số có thể đào tạo được (xem Phụ lục G). Các tham số này được đào tạo bằng cách dự đoán các mã thông báo trong chú thích mục tiêu, như thể hiện trong Phụ lục trình 1.

Yêu cầu đào tạo: PAELLA được đào tạo trong 3 kỷ nguyên với tốc độ học ban đầu là 1e-4, sử dụng trình tối ưu hóa AdamW (Kingma và Ba, 2014) và kích thước lô là 16 với 4 bước tích lũy gradient, trên một GPU NVIDIA RTX A6000 duy nhất. Trong nỗ lực thúc đẩy khả năng truy cập, mô hình của chúng tôi có thể được đào tạo trong một ngày trên một GPU duy nhất, không giống như các mô hình chú thích hình ảnh đa ngôn ngữ khác. Với bộ mã hóa CLIP -ViT-B/32 và bộ giải mã XGLM-2.9B, PAELLA mất 23 giờ để đào tạo 34M tham số có thể đào tạo, chiếm 46G RAM. Nếu sử dụng XGLM-1.7B thay thế, mất 14 giờ và 29G RAM.

Đối với XGLM-564M, chỉ mất 7 giờ và 19G RAM⁴. Hơn nữa, chúng tôi chỉ sử dụng các tập dữ liệu có sẵn công khai, như mô tả tiếp theo.

4.2 Dữ liệu

Bây giờ chúng tôi sẽ mô tả dữ liệu được sử dụng trong các thí nghiệm của mình, bao gồm chuẩn mực mà chúng tôi dùng để đánh giá mô hình và dữ liệu đào tạo của nó, cũng như tập dữ liệu được sử dụng cho kho dữ liệu truy xuất.

Dữ liệu đánh giá: Chúng tôi đánh giá hiệu suất của mô hình của chúng tôi trên tập dữ liệu XM3600 đã được thiết lập tốt (Thapliyal và cộng sự, 2022), bao gồm các hình ảnh đa dạng về mặt địa lý từ 36 ngôn ngữ (L_{36}), bao gồm tập hợp các ngôn ngữ cốt lõi do Thapliyal và cộng sự định nghĩa (2022): en, es, hi và zh (L_{CORE}), và một tập hợp các ngôn ngữ có ít tài nguyên (L_5): bn, quz, mi, sw, te. Mỗi ngôn ngữ được thể hiện bằng 100 hình ảnh từ Open Images, được chọn dựa trên khu vực mà ngôn ngữ đó được nói. Tổng cộng, XM3600 có 3600 hình ảnh với 261375 chú thích do con người chỉ định. Mỗi hình ảnh có ít nhất 2 chú thích/ngôn ngữ.

Hầu hết các tập dữ liệu chú thích do con người chỉ định chủ yếu là tiếng Anh. Theo Thapliyal et al. (2022), chúng tôi mở rộng đánh giá để bao gồm tập dữ liệu COCO-35L (Thapliyal et al., 2022),

⁴Xem hiệu suất của các mô hình này trong Phụ lục 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 API⁵, 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-14⁶ by retrieving, for each image, the $k = 4$ caption IDs from the nearest-neighbor images⁷. 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

⁵<https://cloud.google.com/translate>

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

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

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 package⁸ 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 L_{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_{core}: This model is trained to generate for L_{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 L_{CORE}.

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

⁸<https://github.com/tylin/coco-caption>

được dịch tự động từ tập dữ liệu COCO tiếng Anh gốc (Chen và cộng sự, 2015). COCO -35L có 5000 hình ảnh để xác thực và 113 nghìn hình ảnh để đào tạo, mỗi hình ảnh có 5 chú thích tham chiếu cho mỗi ngôn ngữ. Các bản dịch được lấy bằng Google Translate API⁵, bao gồm tất cả 36 ngôn ngữ trong XM3600, ngoại trừ tiếng Quechua Cusco (quz), không được API hỗ trợ.

Dữ liệu đào tạo: Do sự khan hiếm của các chú thích đa ngôn ngữ do con người chỉ thích, các mô hình đa ngôn ngữ thường sử dụng phương pháp đào tạo trên dữ liệu được dịch bằng máy. Phương pháp tiếp cận tiêu chuẩn (Thapliyal và cộng sự, 2022) bao gồm đào tạo trên tập dữ liệu COCO-35L đã đề cập ở trên, tập dữ liệu này chứa 566K chú thích đào tạo được dịch sang 35 ngôn ngữ, tạo ra tập dữ liệu có 20,3 triệu chú thích. Các mô hình đa ngôn ngữ hiện có (Thapliyal và cộng sự, 2022; Geigle và cộng sự, 2023; Chen và cộng sự, 2023b) cũng được hưởng lợi từ quá trình đào tạo trước quy mô lớn, sử dụng các tập dữ liệu như CC3M-35L do máy dịch (Thapliyal và cộng sự, 2022), được xây dựng từ tập dữ liệu CC3M (Sharma và cộng sự, 2018), tập dữ liệu này chứa 3M cặp chú thích hình ảnh để đào tạo, tương ứng với 105 triệu bản dịch.

Ngược lại, chúng tôi chỉ đào tạo trên một tập hợp con của COCO - 35L, được giảm mẫu để phù hợp với kích thước của tập dữ liệu COCO tiếng Anh gốc (tức là 565K ví dụ thay vì 20,3M ví dụ). Tập hợp con được tạo bằng cách lấy mẫu chú thích từ tập dữ liệu COCO-35L theo phân phối đồng đều trên các ngôn ngữ, sử dụng cùng một ngôn ngữ cho 5 chú thích liên quan đến mỗi hình ảnh. Việc khám phá các chiến lược lấy mẫu khác được để lại cho công việc trong tương lai.

Truy xuất dữ liệu: Kho dữ liệu của mô hình của chúng tôi giữ lại các chú thích đào tạo của tập dữ liệu COCO bằng cách sử dụng các phân tách Karpathy (Karpathy và Fei-Fei, 2015). Các chú thích tiếng Anh được lập chỉ mục với ID tương ứng của chúng. Theo cách này, chúng tôi áp dụng tìm kiếm hình ảnh-văn bản dựa trên CLIP-ViT-bigG-146 bằng cách truy xuất, đối với mỗi hình ảnh, k = 4 ID chú thích từ các hình ảnh lân cận gần nhất⁷. Với các ID chú thích đã truy xuất, chúng tôi có thể dễ dàng tích hợp các chú thích tiếng Anh tương ứng từ COCO hoặc sử dụng các bản dịch liên quan từ bất kỳ ngôn ngữ nào trong số 35 ngôn ngữ khác, bằng cách tham chiếu chéo các ID với COCO-35L tùy thuộc vào ngôn ngữ đích.

Chúng tôi nhấn mạnh rằng hệ thống truy xuất của chúng tôi là đơn ngữ. Kho dữ liệu chỉ chứa tiếng Anh

⁵<https://cloud.google.com/translate>

⁶Xem Phụ lục B để thảo luận về lựa chọn thiết kế của sử dụng bộ mã hóa cụ thể này cho thành phần truy xuất.

⁷Chúng tôi không lấy chú thích của hình ảnh đầu vào.

Chú thích COCO, mà không yêu cầu quy mô của toàn bộ tập dữ liệu COCO-35L. Chúng tôi chỉ sử dụng COCO- 35L để tham chiếu chéo các ID đã truy xuất để có được chú thích bằng ngôn ngữ mà chúng tôi mong muốn.

4.3 Các số liệu đánh giá

Tiếp theo công trình trước, chúng tôi chủ yếu đánh giá hiệu suất chú thích đa ngôn ngữ bằng CIDEr (Vedantam và cộng sự, 2015). CIDEr tính toán sự thống nhất giữa chú thích được tạo và sự đồng thuận của các chú thích tham chiếu, được tính toán thông qua hàm tương tự sử dụng trọng số Tần suất thuật ngữ nhân với Tần suất tài liệu nghịch đảo (TF-IDF). Trái ngược với các nghiên cứu chú thích đa ngôn ngữ trước đây chỉ báo cáo số liệu CIDEr theo Thapliyal và cộng sự (2022), công trình của chúng tôi mở rộng phạm vi đánh giá sang một tập hợp đa dạng các số liệu chú thích, cụ thể là BLEU-1, BLEU- 4, ROGUE và METEOR (xem Phụ lục C). Chúng tôi đã sử dụng gói đánh giá COCO8 với mã thông báo Sacre-BLEU (Post, 2018) để tính toán các số liệu. Trong quá trình đánh giá, mô hình của chúng tôi tạo ra các chú thích bằng cách sử dụng giải mã tìm kiếm chùm với kích thước chùm là 3.

4.4 Các biến thể mô hình

Chúng tôi đánh giá PAELLA cùng với hai biến thể bổ sung, mỗi biến thể được đào tạo trên một tập hợp ngôn ngữ hạn chế hơn để đánh giá khả năng chuyển đổi ngôn ngữ của phương pháp tiếp cận của chúng tôi. Việc lựa chọn mô hình dựa trên việc tối đa hóa CIDEr trung bình trên các ngôn ngữ L_{CORE} trong tập dữ liệu xác thực COCO-35. Ở đây chúng tôi trình bày chi tiết các biến thể mô hình mà chúng tôi so sánh.

PAELLA: Đây là mô hình chính của chúng tôi, được đào tạo để tạo ra 35 ngôn ngữ trong COCO-35L. Trong trường hợp này, chúng tôi lấy mẫu đồng đều từ COCO-35L để đảm bảo quy mô của tập dữ liệu tiếng Anh COCO.

PAELLA_{core}: Mô hình này được đào tạo để tạo ra L_{CORE}, tức là bộ lõi gồm 4 ngôn ngữ được đề xuất trong tập dữ liệu XM3600 (en, es, hi và zh). Chúng tôi cũng lấy mẫu đồng đều từ COCO-35L để duy trì thang đo nhất quán với tập dữ liệu tiếng Anh COCO, như trong tập ngôn ngữ hạn chế này L_{CORE}.

PAELLA_{mono}: Mô hình này được đào tạo để chỉ tạo ra tiếng Anh. Trong trường hợp này, chúng tôi sử dụng tập dữ liệu tiếng Anh COCO gốc.

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_{core} and PAELLA_{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 L₅ (*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_{core} (trained on *en, es, hi, zh*) and PAELLA_{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_{mono} on English, and by PAELLA_{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-

5 Kết quả

Đầu tiên chúng tôi so sánh PAELLA với các mô hình hiện đại. Sau đó chúng tôi thảo luận về hiệu suất của hai biến thể khác của chúng tôi được đào tạo trên một tập hợp nhỏ hơn ngôn ngữ, ví dụ, PAELLACore và PAELLAMono.

5.1 Đào tạo hiệu quả về tham số và dữ liệu

Bảng 1 cho thấy PAELLA có khả năng cạnh tranh với các mô hình đa ngôn ngữ hiện đại, mặc dù được đào tạo với một phần nhỏ khả năng đào tạo của họ các tham số và với dữ liệu ít hơn đáng kể. Với chỉ 34M tham số có thể đào tạo và chỉ 566K trường hợp đào tạo, PAELLA đạt được điểm CIDEr là 26,2 trung bình trên tất cả 36 ngôn ngữ và CIDEr của 28,2 trên các ngôn ngữ mà

Xu hướng sống XGLM đã được đào tạo trước. Ngoài ra, mô hình của chúng tôi có thể mang lại 20,7 điểm CIDEr trên toàn bộ ngôn ngữ tài nguyên thấp L5 (bn, quz, mi, sw, te)⁹.

PAELLA vượt trội hơn Lg (Thapliyal và cộng sự, 2022), tức là một mô hình được giám sát hoàn toàn được đào tạo với 2,6 tỷ tham số trong toàn bộ tập dữ liệu COCO-35L (nhiều hơn 86 lần các tham số có thể đào tạo và nhiều hơn 35 lần các ví dụ đào tạo), phần lớn vượt trội hơn trên toàn bộ tập hợp của các ngôn ngữ cốt lõi và trung bình. PAELLA là cũng cạnh tranh với BB+CC, một mô hình khác từ Thapliyal et al. (2022) được đào tạo trước trên 135M ví dụ trong sự kết hợp của CC3M-35L và COCO-35L. Mặc dù PAELLA không vượt trội hơn BB+CC trung bình, nhưng nó đạt hiệu suất tốt hơn ở 3/4 ngôn ngữ cốt lõi, đáng chú ý là mô hình của họ được đào tạo với 238x nhiều dữ liệu hơn mô hình của chúng tôi.

PAELLA cũng cạnh tranh với các mô hình đa ngôn ngữ được đào tạo trên dữ liệu đa phương ngữ thức khác nhau từ các nhiệm vụ thị giác và ngôn ngữ khác nhau, chẳng hạn như mBLIP (Geigle và cộng sự, 2023). Tư duy tự như mô hình của chúng tôi, mBLIP tận dụng một mô hình ngôn ngữ đa ngôn ngữ được đào tạo trước với nỗ lực tính toán và hiệu quả dữ liệu. Mô hình của chúng tôi vượt qua những nỗ lực này bằng cách có ít tham số hơn đáng kể và vận hành trên ít dữ liệu hơn đáng kể (ví dụ, trong bối cảnh của dữ liệu chú thích, mBLIP đào tạo trên các bản dịch máy của COCO cùng với một tập hợp đa dạng gồm 2.3 triệu ví dụ từ Web CapFilt tổng hợp tập dữ liệu (Li et al., 2022)). PAELLA vượt trội hơn mBLIP BLOOMZ-7B trung bình hơn 2,8 điểm CIDEr và ít hơn 2,1 điểm so với mBLIP mT0-XL.

Mô hình mBLIP mT0-XL chứng minh mạnh mẽ hiệu suất về tiếng Anh, đạt 80,2 CIDEr, nhưng chúng tôi thấy một khoảng cách lớn trong các ngôn ngữ có ít tài nguyên, với

⁹Xem Phụ lục I để biết hiệu suất của tất cả các ngôn ngữ.

13,4 điểm CIDEr trong khi mô hình của chúng tôi đạt 20,7

điểm. Trong Phần 6.1, chúng tôi thảo luận rộng rãi hơn n hiệu suất trên nhiều ngôn ngữ.

Tư duy tự như các mô hình chú thích đa ngôn ngữ khác, PAELLA hoạt động kém hơn đáng kể so với mô hình PaLI tham số 17B quy mô lớn (Chen et al., 2023b) được đào tạo trên 12 tỷ ví dụ bằng cách sử dụng tập dữ liệu WebLI riêng tư. Tư duy tự giữ cho PALI-3 gần đây (Chen et al., 2023a), điều này tạo ra những nỗ lực hướng tới một mô hình hiệu quả hơn, nhưng vẫn đào tạo hàng tỷ tham số trên hàng tỷ dữ liệu đa ngôn ngữ. Điều này vẫn còn tốn kém đáng kể và không thực tế cho nhiều ứng dụng. Từ một nghiên cứu

Theo quan điểm này, mô hình của chúng tôi có thể được đào tạo trong một ngày trong phần cứng tiêu dùng với bộ dữ liệu công khai.

Cuối cùng, chúng ta thấy sự cải thiện 15,2 điểm CIDEr so với LMCap (Ramos et al., 2023a), trong đó là một phương pháp tiếp cận tăng cường khả năng truy xuất bằng một vài cú đánh không có đào tạo. Với đào tạo đa ngôn ngữ tối thiểu, mô hình của chúng tôi tiếp tục thu hẹp khoảng cách với các mô hình chú thích đa ngôn ngữ quy mô lớn.

Nhìn chung, kết quả trên XM3600 chứng minh hiệu quả của cách tiếp cận của chúng tôi cho đa ngôn ngữ hiệu quả chú thích, góp phần làm giảm cả hai các thông số có thể đào tạo và yêu cầu dữ liệu. Đối với một đánh giá toàn diện hơn, chúng tôi cũng báo cáo kết quả trên COCO-35L trong Bảng 2, trong đó chúng tôi quan sát lại rằng mô hình của chúng tôi có thể vượt trội hơn các mô hình được giám sát đầy đủ của Thapliyal et al. (2022). Xem ví dụ định tính trong Phụ lục H.

5.2 Chuyển dịch xuyên ngôn ngữ bằng Zero-shot

Trong Bảng 1, chúng tôi quan sát thấy PAELLACore (được đào tạo trên en, es, hi, zh) và PAELLAMono (chỉ được đào tạo trên en) có hiệu suất zero-shot mạnh mẽ trong các ngôn ngữ khác, cho thấy cách tiếp cận của chúng tôi không yêu cầu dữ liệu chú thích cho từng ngôn ngữ trong đào tạo. Hệ thống có thể được điều kiện hóa trên một ngôn ngữ khác ngoài bộ đào tạo, bằng cung cấp các chú thích nhanh chóng và được lấy lại trong ngôn ngữ đầu ra mong muốn, chỉ tại thời điểm suy luận.

Chúng tôi quan sát thêm rằng PAELLA bị PAELLAMono đánh bại về tiếng Anh và PAELLACore bằng tiếng Anh và tiếng Tây Ban Nha. Điều này có thể một phần được giải thích bởi thực tế là PAELLA là được đào tạo trước trên một mẫu thống nhất của tất cả 35 ngôn ngữ trong COCO-35L, trong khi các biến thể này là được đào tạo trước trên một mẫu thống nhất chỉ gồm những ngôn ngữ đó, tức là có nhiều phụ đề tiếng Anh hơn. Cả hai Mặt khác, các biến thể Core và Mono là ít có khả năng tạo phụ đề cho các ngôn ngữ khác

⁹See 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									
<i>PALI</i>	12B	17B	17B	98.1	-	31.3	36.5	-	53.6
<i>PALI-3</i>	12B	5B	5B	94.5	-	-	-	-	46.1
<i>mBLIP mT0-XL</i>	489M	124M	4.9B	80.2	62.6	16.1	14.7	7.9	28.3
<i>mBLIP BLOOMZ-7B</i>	489M	124M	8.3B	76.4	60.0	24.9	14.7	6.7	23.4
<i>BB+CC</i>	135M	0.8B	0.8B	58.4	42.5	19.7	20.2	22.4	28.5
<i>Lg</i>	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 _{core}	566K _{en,es,hi,zh}	34M	3B	58.2	45.0	20.4	25.4	11.8	16.8 (24.9*)
PAELLA _{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_{core} (trained on *en,es,hi,zh*) and PAELLA_{mono} (trained only on *en*) against other state-of-the-art multilingual models. L₅ represents the average performance across the set of low-resource languages (*bn, quz, mi, sw, te*), and L₃₆ 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
<i>BB+CC</i>	98.0	96.2	75.9	74.8
<i>Lg</i>	87.5	85.9	62.4	65.6
PAELLA	113.6	113.9	86.2	123.3
PAELLA _{core}	118.5	120.3	94.7	130.7
PAELLA _{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_{mono}, that achieved a 15.5 CIDEr score on average, especially considering its training was exclusively on English. PAELLA_{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_{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_{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

Ngư ời mẫu	Dữ liệu	Đào tạo θ	Tổng θ	en	là	CHÀO	zh	tiếng s	L36
Không cần đào tạo									
LMCap	-	0	2,9 tỷ	45,2	32,9	13,2	22,1	0,0	11,0
Đào tạo quy mô lớn									
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	4,9T	80,2	62,6	16,1	14,7	7,9	28,3
mBLIP BLOOMZ-7B	489M	124M	8,3T	76,4	60,0	24,9	14,7	6,7	23,4
BB+CC	135 triệu	0,8 tỷ	0,8 tỷ	58,4	42,5	19,7	20,2	22,4	28,5
<small>Hình thức</small>	19,8 triệu	2,6 tỷ	2,6 tỷ	34,3	22,0	11,1	9,9	12,5	15,0
Đào tạo hiệu quả về dữ liệu và tham số									
<small>Cơ m thập cảm Paella</small>	566K35L	34M	3B	57,3	44,9	20,8	25,9	20,7	26,2 (28,2)
PAELLAcore	566Ken,es,hi,zh	34M	3B	58,2	45,0	20,4	25,4	11,8	16,8 (24,9)
PAELLAmono	566Ken	34M	3B	58,2	42,2	17,1	23,5	12,1	15,5 (23,9)

Bảng 1: Hiệu suất CIDEr trên XM3600, một chuẩn mực đa ngôn ngữ với hình ảnh đa dạng về mặt địa lý trên 36 ngôn ngữ. Chúng tôi so sánh mô hình của chúng tôi, PAELLA, và hai biến thể của nó, PAELLAcore (được đào tạo trên en, es, hi, zh) và PAELLAmono (chỉ được đào tạo trên en) so với các mô hình đa ngôn ngữ hiện đại khác. L5 biểu thị mức trung bình hiệu suất trên toàn bộ các ngôn ngữ có ít tài nguyên (bn, quz, mi, sw, te) và L36 trên tất cả 36 ngôn ngữ. () tương ứng với mức trung bình trên các ngôn ngữ mà bộ giải mã XGLM được đào tạo trước. Chúng tôi đánh dấu bằng chữ in đậm rằng mô hình của chúng tôi có số lượng tham số có thể đào tạo thấp nhất và yêu cầu lượng dữ liệu đào tạo ít nhất.

Ngư ời mẫu	vi	là	CHÀO	zh
BB+CC	98.0	96.2	75.9	74.8
<small>Hình thức</small>	87,5	85,9	62,4	65,6
<small>Cơ m thập cảm Paella</small>	113,6	113,9	86,2	123,3
PAELLAcore	118,5	120,3	94,7	130,7
PAELLAmono	120,8	91,48	45,9	59,1

Bảng 2: Điểm CIDEr trên dữ liệu xác thực COCO-35L. Các mô hình được giám sát đầy đủ từ [Thapliyal et al. \(2022\)](#) được hiển thị ở trên cùng, với các biến thể mô hình của chúng tôi ở bên dưới.

bên cạnh những dữ liệu đào tạo trước của XGLM, dẫn đến mức giảm trung bình là 9,4 và 10,7 điểm của CIDEr trên tất cả 36 ngôn ngữ, so với PAELLA, tương ứng. Mặc dù có hạn chế này, chúng tôi nhấn mạnh hiệu suất của PAELLAmono, đạt được điểm CIDEr trung bình là 15,5, đặc biệt là khi xem xét việc đào tạo của họ chỉ dành riêng bằng tiếng Anh. PAELLAmono thậm chí còn vượt trội hơn Lg trên toàn bộ 4 ngôn ngữ cốt lõi và trung bình, mặc dù mô hình này có quy mô lớn từ đầu đến cuối đào tạo trên nhiều ngôn ngữ khác nhau với bộ dữ liệu COCO-35L hoàn chỉnh.

Khả năng tiếp cận của chúng tôi đối với việc chuyển giao xuyên ngôn ngữ không cần thêm có tầm quan trọng đặc biệt với sự chiếm ưu thế của phụ đề tiếng Anh bộ dữ liệu. Chúng tôi lưu ý rằng chúng tôi không sử dụng ngôn ngữ đa ngôn ngữ

dữ liệu miễn trong kho dữ liệu truy xuất. Dữ liệu đã truy xuất chú thích từ COCO-35L có sự phân phối khác với chuẩn mực XM3600, chuẩn mực này chứa các hình ảnh và khái niệm đa dạng về mặt địa lý. Chúng tôi cũng nhấn mạnh rằng toàn bộ lời nhắc (bao gồm cả lời nhắc đã lấy được) chú thích) cần phải ở ngôn ngữ đích cho mục đích này khả năng xuyên ngôn ngữ không cần bản để xuất hiện. Nếu không thì mô hình PAELLAmono mặc định là tiếng Anh, là kết quả của việc chỉ tiếp xúc với ngôn ngữ này và do đó có xu hướng mạnh mẽ tạo ra bằng tiếng Anh.

6 Thảo luận

Chúng tôi thảo luận về hiệu suất của PAELLA trên các ngôn ngữ liên quan đến các hệ thống viết khác nhau. Sau đó, chúng tôi tiến hành các nghiên cứu cắt bỏ, đầu tiên thảo luận về dữ liệu đơn ngữ cần thiết để đào tạo PAELLAmono, tiếp theo là tầm quan trọng của thông tin đã được thu thập. Những nghiên cứu cắt bỏ này là được thực hiện trên sự phân tách xác nhận của COCO-35L vì XM3600 chỉ chứa dữ liệu đánh giá.

6.1 Hệ thống chữ viết

Trong Hình 2, chúng ta quan sát hiệu suất của PAELLA trên các hệ thống chữ viết đa dạng của 36 ngôn ngữ, cùng với mBLIP mT0-XL mô hình để so sánh. mBLIP có hiệu suất đáng chú ý về tiếng Anh và các ngôn ngữ chia sẻ

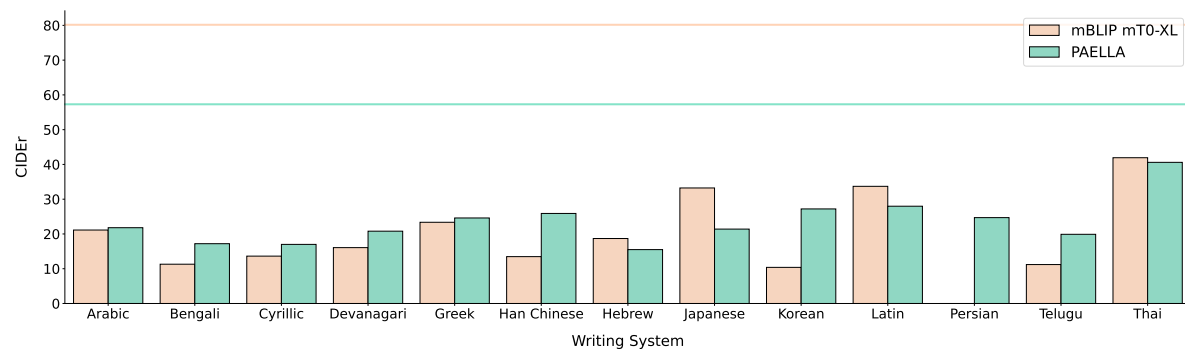


Figure 2: Performance by writing system. Horizontal lines denote corresponding English performance.

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, Cyrillic, Devanagari, Greek, simplified Chinese, Korean, Persian, and Telugu 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_{mono} works when trained with languages other than English. As seen in Table 3, PAELLA_{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

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

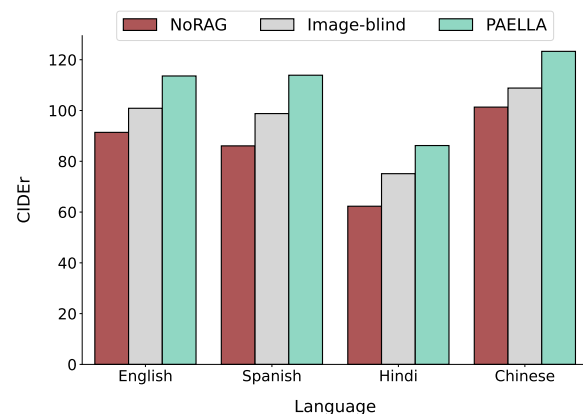


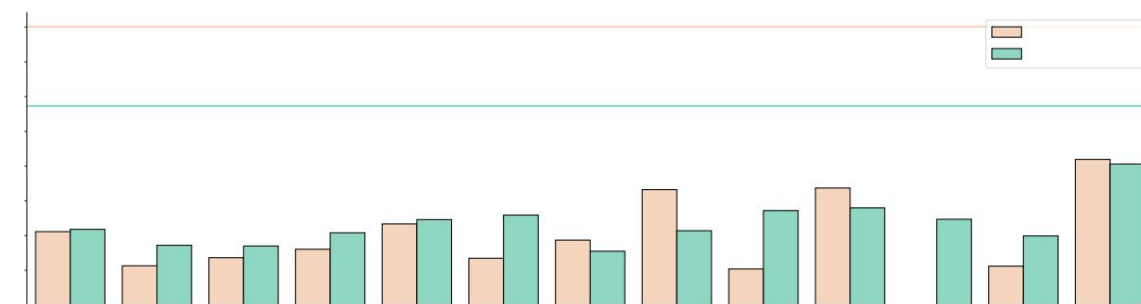
Figure 3: Ablation results on the COCO-35L validation data, reported with CIDEr metric. We ablate the retrieval (NoRAG) and the visual encoder (image-blind).

Model	en	es	hi	zh	da
PAELLA _{en}	120.8	91.5	45.9	59.1	2.7
PAELLA _{es}	93.3	125.3	52.6	95.3	2.9
PAELLA _{hi}	70.4	68.1	99.3	80.9	0.1
PAELLA _{zh}	65.0	49.9	1.4	130.6	0.4
PAELLA _{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.

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-

¹¹Setting the visual features from the encoder to zero.



Hình 2: Hiệu suất theo hệ thống chữ viết. Các đường ngang biểu thị hiệu suất tiếng Anh tương ứng.

Hệ thống chữ viết Latinh. Chuyên ngành này dẫn đến hiệu suất kém đối với một số hệ thống chữ viết, ví dụ như tiếng Ba Tư và tiếng Hàn. Ngược lại, mô hình của chúng tôi chứng minh hiệu suất cân bằng hơn trên các hệ thống chữ viết khác nhau ngoài chữ viết Latinh có nhiều tài nguyên, đạt được kết quả tốt hơn biểu diễn các hệ thống chữ viết Ả Rập, Bengal, Cyrillic, Devanagari, Hy Lạp, Trung Quốc giản thể, Hàn Quốc, Ba Tư và Tegulu.

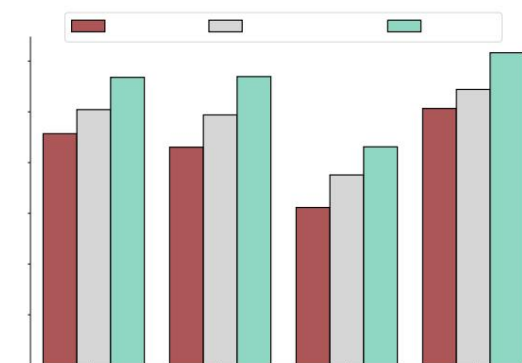
6.2 Giám sát đơn ngữ

Trước đây chúng ta đã thấy rằng mô hình chú thích đa ngôn ngữ của chúng ta cũng có thể được đào tạo trên dữ liệu đơn ngôn ngữ (xem Phần 5.2). Bây giờ chúng ta thảo luận liệu PAELLA_{mono} có hoạt động khi được đào tạo với ngôn ngữ khác ngoài tiếng Anh. Như được thấy trong Bảng 3, PAELLA_{mono} cũng thể hiện khả năng đa ngôn ngữ ngay lập tức với 3 ngôn ngữ cốt lõi khác. Đáng ngạc nhiên là việc đào tạo tiếng Tây Ban Nha mang lại khả năng khái quát tốt hơn cho các ngôn ngữ cốt lõi khác so với đào tạo về tiếng Anh. Khi được đào tạo về tiếng Trung, về mặt khác, mô hình mất khả năng tạo phụ đề bằng tiếng Hindi. Ngoài ra, chúng tôi đã điều tra hành vi của mô hình khi được đào tạo bằng một ngôn ngữ nằm ngoài quá trình đào tạo trước của bộ giải mã XGLM, chẳng hạn như tiếng Đan Mạch. Ở đây, mô hình có thể tạo chú thích bằng tiếng Đan Mạch, nhưng chúng ta thấy hành vi thú vị này phá vỡ khái quát sang các ngôn ngữ khác.

6.3 Thu hồi như là thành phần chính của PAELLA

Bây giờ chúng ta nghiên cứu tầm quan trọng của việc tăng cường với các ví dụ đã lấy lại, thành phần chính của phương pháp tiếp cận của chúng tôi. Chúng tôi bắt đầu bằng cách loại bỏ thành phần truy xuất, bằng cách đào tạo mà không bao gồm các thành phần đã lấy lại chú thích trong lời nhắc.¹⁰ Như được thấy trong Hình 3, hiệu suất giảm trung bình 24 CIDEr trên

¹⁰Lời nhắc chỉ bao gồm phần cuối cùng: Một chú thích mà tôi có thể tạo ra để mô tả hình ảnh này bằng [ngôn ngữ] là.



Hình 3: Kết quả cắt bỏ trên xác nhận COCO-35L. Chúng tôi loại bỏ dữ liệu thu thập (NoRAG) và bộ mã hóa hình ảnh (mô hình ảnh).

Ngữ ời mẫu	ví	là	cho	zh	ngày
PAELLA _{en}	120,8	91,5	45,9	59,1	PAELLA _{es} 93,3 2.7
125,3	52,6	95,3	PAELLA _{hi} 70,4	68,1	99,3 2.9
80,9	PAELLA _{zh} 65,0	49,9	1,4	130,6	0,1
PAELLA _{da}	5.1	1.2	2.8	4.1	107.5

Bảng 3: Kết quả CIDEr cho các biến thể đơn sắc trên dữ liệu xác thực COCO-35L. Chúng tôi biểu thị bằng chỉ số dư ời và in đậm là ngôn ngữ mà mỗi biến thể được đào tạo.

4 ngôn ngữ cốt lõi không cần truy xuất (noRAG), so với PAELLA. Chúng tôi cũng cắt bỏ thị giác bộ mã hóa bằng cách đào tạo trên hình ảnh đầu vào trống,¹¹ và chúng ta lại thấy sự mất hiệu suất (tức là 13,4 CIDEr qua 4 ngôn ngữ), xác nhận rằng PAELLA thực sự chú trọng đến hình ảnh và không chỉ đơn thuần diễn đạt lại các chú thích đã lấy được. Hơn nữa, chúng tôi quan sát thấy mô hình NoRAG hoạt động kém hơn là phương pháp tiếp cận mô hình ảnh với các chú thích được lấy lại, củng cố lợi ích của việc đào tạo đa dòng

¹¹Thiết lập các tính năng trực quan từ bộ mã hóa về 0.

gual image captioning with retrieval-augmentation. In Appendix F, we additionally discuss results for PAELLA_{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, Matusz 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](#).

chú thích hình ảnh theo ngôn ngữ gual với khả năng tăng cường truy xuất.

Trong Phụ lục F, chúng tôi cũng thảo luận thêm về kết quả cho PAELLAmo, nơi việc truy xuất được chứng minh là rất quan trọng để tạo ra phụ đề bằng các ngôn ngữ khác biệt đáng kể so với sự giám sát của tiếng Anh. Chúng tôi cũng thảo luận về tầm quan trọng của việc lấy lại chú thích bằng ngôn ngữ đích, trong Phụ lục H.

7 Kết luận và công việc tương lai

Chúng tôi đề xuất PAELLA, một mô hình chú thích đa ngôn ngữ hiệu quả với khả năng tăng cường truy xuất. Trái ngược với các nghiên cứu trước đây, PAELLA là nhẹ để đào tạo, cả về số lượng tham số và nhu cầu dữ liệu đa ngôn ngữ. Kết quả chứng minh tính cạnh tranh giữa các ngôn ngữ, bao gồm ngôn ngữ có ít tài nguyên. PAELLA cũng thể hiện khả năng đa ngôn ngữ mạnh mẽ không bản. Trong tương lai, chúng tôi có kế hoạch tiếp tục điều tra xuyên ngôn ngữ chuyển giao với sự giám sát đơn ngữ.

8 Lời cảm ơn

Nghiên cứu này được hỗ trợ bởi Kế hoạch phục hồi và phục hồi của người Bồ Đào Nha thông qua dự án C645008882-00000055 (tức là Trung tâm AI có trách nhiệm) và cũng bởi Fundação cho Ciência e Tecnologia (FCT), thông qua dự án có tham chiếu UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) và bằng Tiến sĩ. học bổng có tham chiếu 2020.06106.BD.

Hạn chế

Trong khi mô hình của chúng tôi nhằm mục đích đóng góp vào nghiên cứu ngoài phụ đề tiếng Anh, nó còn có những hạn chế trong đó các kết quả được điều kiện hóa trên các chú thích được lấy từ dữ liệu được dịch bằng máy từ COCO, tập trung vào tiếng Anh và thiếu phạm vi bao quát các khái niệm đa dạng về mặt địa lý (Liu và cộng sự, 2021). Các nghiên cứu trước đây cũng đã chỉ ra rằng COCO có sự mất cân bằng giới tính đáng kể và sử dụng dữ liệu này có thể khuếch đại thêm sự thiên vị (Zhao và cộng sự, 2017; Hendricks và cộng sự, 2018). Ví dụ, các mô hình có thể trở nên dễ sinh ra phụ nữ trong bếp hơn cài đặt hơn n con người. Để hiểu rõ hơn về những thành kiến mà PAELLA thể hiện, chúng tôi đề xuất một phân tích về các chú thích đã thu thập được sử dụng bởi mô hình, như được minh họa trong các hình ở Phụ lục H. Một hạn chế khác liên quan đến phạm vi bao phủ của các ngôn ngữ và khái niệm của mô hình của chúng tôi. Mở rộng phạm vi ngôn ngữ được đề cập sẽ được mong muốn thích hợp với nhiều người nói khác nhau hơn. Ngoài ra, mô hình của chúng tôi đã được đánh giá trên một số lượng hạn chế

các tập dữ liệu, tương tự như các mô hình đồng thời khác, do đến sự khan hiếm các nguồn tài nguyên đa ngôn ngữ để đánh giá kết quả chú thích hình ảnh.

PAELLA chỉ được thiết kế cho nhiệm vụ chú thích hình ảnh. Trong công việc tương lai, chúng tôi muốn nghiên cứu các phương pháp tiếp cận để mở rộng PAELLA thành phạm vi các nhiệm vụ đa phương thức đa ngôn ngữ, chẳng hạn như những thứ được bao phủ trong IGLUE (Bugliarello và cộng sự, 2022).

Tài liệu tham khảo

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 và Karen Simonyan. 2022. [Flamingo: một mô hình ngôn ngữ trực quan để học ít lần](#). Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh.

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, Đứa trẻ tái sinh, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Mùa đông, Chris Hesse, Mark Chen, Eric Sigler, Matusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever và Dario Amodei. Các mô hình ngôn ngữ là những người học ít lần. Trong những tiến bộ trong Hệ thống xử lý thông tin thần kinh.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, và Ivan Vulić. 2022. [IGLUE: Một chuẩn mực cho việc chuyển giao việc học giữa các phương thức, nhiệm vụ và ngôn ngữ](#). Trong Biên bản Hội nghị quốc tế về Học máy.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023a. [Tầm nhìn PaLI-3 mô hình ngôn ngữ: Nhỏ hơn, nhanh hơn, mạnh hơn](#). arXiv bản in trước 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 Đình, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Wei Cheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby và Radu Soricut. 2023b. [PaLI: A mô hình hình ảnh ngôn ngữ đa ngôn ngữ được chia tỷ lệ chung](#).

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*.

Trong Biên bản Hội nghị quốc tế về Học cách biểu diễn.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, và C Lawrence Zitnick. 2015. [Chú thích Microsoft COCO : Máy chủ thu thập và đánh giá dữ liệu](#). arXiv bản in trực tuyến 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 Chu, Quốc V. Lê, và Jason Wei. 2022. [Hư ớng dẫn mở rộng quy mô-điều chỉnh tinh vi mô hình ngôn ngữ](#). bản in trực tuyến arXiv arXiv:2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer và Veselin Stoyanov. 2020. [Không có giám sát học tập biểu diễn đa ngôn ngữ ở quy mô lớn](#). Trong Biên bản cuộc họp thường niên của Hiệp hội về Ngôn ngữ học tính toán.

Ôn Lư ợng Đại, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, và Steven Hoi. 2023. [InstructBLIP: Hư ớng tới mô hình ngôn ngữ thị giác mục đích chung với điều chỉnh hư ớng dẫn](#). Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh.

Michael Denkowski và Alon Lavie. 2014. [Meteor phổ quát : Đánh giá bản dịch ngôn ngữ cụ thể cho bất kỳ ngôn ngữ đích nào](#). Trong Biên bản Hội thảo về dịch máy thống kê.

Gregor Geigle, Abhay Jain, Radu Timofte, và Goran Glavaš. 2023. [mBLIP: Bootstrap-ping hiệu quả của llms thị giác đa ngôn ngữ](#). Bản in trực tuyến arXiv arXiv:2307.06930.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman và Alexis Conneau. 2021. Bộ [chuyển đổi quy mô lớn hơn n cho mô hình ngôn ngữ che dấu đa ngôn ngữ](#). Trong Biên bản Hội thảo về Đại diện Học NLP.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell và Anna Rohrbach. 2018. [Phụ nữ cũng như ván trượt tuyết: Khắc phục sự thiên vị trong các mô hình chú thích](#). Trong Biên bản Hội nghị Châu Âu về Tầm nhìn máy tính (ECCV).

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Lư u Tử Thành, Lộ Ngọc Mao và Vũ ợng Lê Quyền. 2022. [Mở rộng quy mô đào tạo trực tuyến ngôn ngữ thị giác cho hình ảnh chú thích](#). Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel và Edouard Grave. 2022. Học tập ít lần với các mô hình ngôn ngữ tăng cường truy xuất. Bản in trực tuyến arXiv arXiv:2208.03299.

Jeff Johnson, Matthijs Douze và Hervé Jégou. 2017. [Tìm kiếm sự tương đồng ở quy mô tỷ với GPU](#). arXiv bản in trực tuyến arXiv:1702.08734.

Andrej Karpathy và Li Fei-Fei. 2015. Căn chỉnh ngữ nghĩa thị giác sâu để tạo ra mô tả hình ảnh. Trong Biên bản Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer và Mike Lewis. 2020. Gần nhất dịch máy hàng xóm. bản in trực tuyến arXiv arXiv:2010.00710.

Diederik P Kingma và Jimmy Ba. 2014. Adam: A [phương pháp tối ưu hóa ngẫu nhiên](#). bản in trực tuyến arXiv 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 và Douwe Kiela. 2020. [Tạo ra khả năng tăng cường truy xuất cho các tác vụ NLP đòi hỏi nhiều kiến thức](#). Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh.

Junnan Li, Dongxu Li, Silvio Savarese và Steven Hoi. 2023. BLIP-2: Khởi động quá trình đào tạo trực tuyến ngôn ngữ-hình ảnh bằng bộ mã hóa hình ảnh đóng băng và các mô hình ngôn ngữ lớn. Bản in trực tuyến arXiv arXiv:2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong và Steven Hoi. 2022. BLIP: Khởi động quá trình đào tạo trực tuyến ngôn ngữ-hình ảnh để hiểu ngôn ngữ-thị giác thống nhất và thể hệ. bản in trực tuyến arXiv arXiv:2201.12086.

Chin-Yew Lin. 2004. ROUGE: Một gói để đánh giá tự động các bản tóm tắt. Trong Biên bản Tóm tắt văn bản hội thảo được triển khai.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, và những người khác. 2021. Học ít lần với các mô hình ngôn ngữ đa ngôn ngữ. bản in trực tuyến arXiv arXiv:2112.10668.

Lư u Phương Ngọc, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier và Desmond Elliott. 2021. [Lý luận dựa trên hình ảnh trên nhiều ngôn ngữ và nền văn hóa](#). Trong Biên bản Hội nghị năm 2021 về Phương pháp thực nghiệm trong Ngôn ngữ tự nhiên Xử lý.

Ziyang Luo, Yadong Xi, Rongsheng Zhang và Jing Ma. 2022. I-tuning: Điều chỉnh mô hình ngôn ngữ với hình ảnh để tạo chú thích. bản in trực tuyến arXiv 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 Kementchedjheva. 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, Min-joon 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, Łukasz 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*.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal và Aishwarya Agrawal. 2023. [MAPL: Điều chỉnh hiệu quả tham số của các mô hình dự đoán tạo từ dữ liệu cho lời nhắc ngôn ngữ thị giác ít lần](#). Trong Biên bản của Hội nghị của Chi hội Châu Âu thuộc Hiệp hội Ngôn ngữ học Tính toán.

Michael McCloskey và Neal J Cohen. 1989. Sự can thiệp thảm khốc trong mạng lưu trữ kết nối: vấn đề học tập tuần tự. Trong Tâm lý học về học tập và động lực, tập 24.

Ron Mokady, Amir Hertz và Amit H. Bermano. 2021. Clipcap: Tiền tố clip để chú thích hình ảnh. arXiv bản in trước arXiv:2111.09734.

Kishore Papineni, Salim Roukos, Todd Ward và Wei-Jing Zhu. 2002. BLEU: một phương pháp đánh giá tự động bản dịch máy. Trong Biên bản báo cáo Cuộc họp thường niên của Hiệp hội tính toán Ngôn ngữ học.

Matt Post. 2018. [Lời kêu gọi làm rõ trong việc báo cáo BLEU điểm số](#). Trong Biên bản Hội nghị lần thứ ba về Dịch máy: Bài nghiên cứu, trang 186–191, Brussels, Bỉ. Hiệp hội Ngôn ngữ học tính toán.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên. Trong Biên bản của Hội nghị quốc tế về máy học.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Ngôn ngữ mô hình là những người học đa nhiệm không có giám sát. OpenAI blog, 1(8):9.

Rita Ramos, Bruno Martins và Desmond Elliott. 2023a. [LMCap: Hình ảnh đa ngôn ngữ ít ảnh chú thích bằng mô hình ngôn ngữ tăng cường truy xuất nhắc nhở](#). Phát hiện của Hiệp hội Ngôn ngữ học tính toán.

Rita Ramos, Bruno Martins, Desmond Elliott và Yova Kementchedjheva. 2023b. [SmallCap: Nhẹ chú thích hình ảnh dự đoán nhắc đến với sự tăng cường truy xuất](#). Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu.

Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho và Bruno Martins. 2021. Tăng cường khả năng truy xuất cho mạng nơ-ron sâu. Trong Biên bản Hội nghị chung quốc tế về Mạng nơ-ron.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, và Rita Cucchiara. 2022. Bộ chuyển đổi tăng cường truy xuất để chú thích hình ảnh. Bản in trước arXiv arXiv:2207.13162.

Piyush Sharma, Nan Ding, Sebastian Goodman và Radu Soricut. 2018. Chú thích khái niệm: Một bản đã được làm sạch, hypernymed, tập dữ liệu văn bản thay thế hình ảnh để tự động chú thích hình ảnh. Trong Biên bản báo cáo của Cuộc họp thường niên của Hiệp hội Ngôn ngữ học tính toán.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Min-joon Seo, Rich James, Mike Lewis, Luke Zettlemoyer và Wen-tau Yih. 2023. Replug: Các mô hình ngôn ngữ hộp đen tăng cường truy xuất. arXiv bản in trước arXiv:2301.12652.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, và Tatsunori B. Hashimoto. 2023. Alpaca Stanford: Một mô hình LLaMA tuân theo hướng dẫn. https://github.com/tatsu-lab/stanford_alpaca.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, và Radu Soricut. 2022. Crossmodal-3600: Một khối lưu trữ lớn bộ dữ liệu đánh giá đa phương thức đa ngôn ngữ. arXiv bản in trước 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, và Guillaume Lample. 2023. Llama: Mở và các mô hình ngôn ngữ nền tảng hiệu quả. arXiv bản in trước arXiv:2302.13971.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Ali Eslami, Oriol Vinyals và Felix Hill. 2021. [Học tập đa phương thức với ít cú đánh với ngôn ngữ đóng băng mô hình](#). Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, tập 34, trang 200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. 2017. [Sự chú ý là tất cả bạn cần](#). Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh.

Ramakrishna Vedantam, C Lawrence Zitnick, và Devi Parikh. 2015. CIDEr: đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong Biên bản của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zichen Liu, Ce Liu, và Lijuan Wang. 2022. Git: Một hình ảnh tạo thành văn bản máy biến áp cho thị giác và ngôn ngữ. bản in trước arXiv arXiv:2205.14100.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov và Yuan Cao. 2021. SimVLM: Đơn giản mô hình ngôn ngữ trực quan dự đoán tạo từ dữ liệu với giám sát yếu. thảo luận arXiv arXiv:2108.10904.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai và Quoc V Le. 2021. Các mô hình ngôn ngữ dự đoán tinh chỉnh là những người học không cần thực hiện cú đánh nào. Bản in trước của arXiv 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, Wan-grong 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).

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Máy biến áp: Hiện đại xử lý ngôn ngữ tự nhiên. Trong Biên bản của Hội nghị về Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên: Trình diễn hệ thống.

Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wan-grong Cheng và Jinwen Tian. 2019. Một sự thống nhất khung tạo-truy xuất cho chú thích hình ảnh. Biên bản Hội nghị quốc tế ACM về Quản lý thông tin và kiến thức.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. ViLM lại: Mô hình ngôn ngữ trực quan tăng cường truy xuất cho số không và chú thích hình ảnh ít ảnh. bản in trước arXiv arXiv:2302.04858.

Ziyan Yang, Leticia Pinto-Alva, Franck Dernoncourt, và Vicente Ordonez. 2020. [Sử dụng tính năng trực quan không gian như một trục xoay giữa các ngôn ngữ](#). Trong những phát hiện của Hiệp hội Ngôn ngữ học tính toán.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang và Dong Yu. 2023. Chuỗi ghi chú : Tăng cường tính mạnh mẽ trong việc tăng cường truy xuất mô hình ngôn ngữ. bản in trước arXiv 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. Tùy chọn: Mở các mô hình ngôn ngữ chuyển đổi được đào tạo trước. bản in trước arXiv arXiv:2205.01068.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez và Kai-Wei Chang. 2017. Đàn ông cũng vậy như mua sắm: Giảm sự khuếch đại định kiến giới tính bằng cách sử dụng các ràng buộc cấp độ ngữ liệu. bản in trước arXiv arXiv:1707.09457.

Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang và Jiaxuan Zhang. 2020. [Chú thích hình ảnh tạo ra thông qua truy xuất thống nhất và dựa trên thể hệ phương pháp](#). Khoa học ứng dụng, 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), ROGUE-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	ROGUE-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	ROGUE-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, 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, in hindi] ... [retrieved caption _k in hindi] इस छवि का हिंदी में वर्णन करने के लिए मैं एक कैप्शन तैयार कर सकता हूँ:	类似图片显示 [retrieved caption, 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.

Một lời nhắc nhở

Để tạo phụ đề trên nhiều ngôn ngữ khác nhau, chúng tôi tùy chỉnh lời nhắc của chúng tôi và các chú thích đã lấy ở ngôn ngữ đã chọn. Trong Hình 4, chúng tôi đưa ra ví dụ bằng tiếng Tây Ban Nha, tiếng Hindi và tiếng Trung Quốc, từ ứng dụng. Các lời nhắc cho các ngôn ngữ khác được bao gồm trong mã của chúng tôi.

B Lấy lại

Ramos et al. (2023b) đã chỉ ra trong SmallCap mô hình chú thích tăng cường truy xuất mà CLIP-ViT-B/32 phù hợp làm bộ mã hóa để tạo văn bản, nhưng khi được sử dụng làm bộ mã hóa truy xuất thì hoạt động kém. Do đó, chúng tôi chọn công nghệ tiên tiến nhất phiên bản CLIP, CLIP-ViT-bigG-14, để truy xuất. Chúng tôi không sử dụng phiên bản lớn hơn đó trong mã hóa mô hình cũng vậy, vì điều đó sẽ đáng kể làm chậm thời gian đào tạo.

Tiêu chuẩn đánh giá C

Để đánh giá toàn diện hơn, chúng tôi báo cáo lại hiệu suất của mô hình của mình với các số liệu tự động bổ sung, bao gồm BLEU-1 (B-1), BLEU-4 (B-4) (Papineni và cộng sự, 2002), ROGUE-L (Lin, 2004) và METEOR (Denkowski và Lavie, 2014). Chúng tôi báo cáo các số liệu này cho cả bộ dữ liệu XM3600 và xác thực COCO-35L chia tách, như được thấy trong Bảng 4 và Bảng 5.

	B-1	B-4	ROGUE-L	METEOR
vì	45.1	10.3	34,6	14,5
là	43.2	7.8	30,1	15.1
Xin chào	29.3	2.7	21.1	21,9
zh	32.1	6.9	24,6	10.9

Bảng 4: Hiệu suất PAELLA trên tập dữ liệu XM3600, trên nhiều số liệu đánh giá khác nhau.

	B-1	B-4	ROGUE-L	METEOR
vì	76.2	33.6	55,9	26,7
là	76.3	35.9	54,5	27,5
Xin chào	74.9	26.5	51.0	33,7
zh	77.2	40.0	56,4	28,8

Bảng 5: Hiệu suất PAELLA trên COCO-35L phân chia xác thực, trên các số liệu đánh giá khác nhau.

Hình ảnh tương tự muestran [lấy chú thích1 bằng tiếng Tây Ban Nha] ... [lấy chú thíchk bằng tiếng Tây Ban Nha] Một tiêu đề bạn có thể tạo để mô tả hình ảnh này bằng tiếng Tây Ban Nha:	[lấy chú thích1 bằng tiếng hindi] ... [lấy chú thích bằng tiếng hindi] Bạn có thể làm điều đó.	Tải xuống hình ảnh [lấy chú thích1 bằng tiếng Trung] ... [lấy chú thíchk bằng tiếng Trung] :
---	---	--

Hình 4: Ví dụ về lời nhắc bằng tiếng Tây Ban Nha, tiếng Hindi và Tiếng Trung Quốc lần lượt được hiển thị từ trên xuống.

D Khả năng mở rộng

Trong Bảng 6, chúng ta thấy PAELLA hoạt động như thế nào với các phiên bản XGLM khác nhau trong bộ giải mã. XGLM-2.9B quy mô lớn hơn n có hiệu suất mạnh hơn n, phù hợp với những phát hiện trước đây liên quan đến hành vi mở rộng của LM. Tuy nhiên, Các phiên bản XGLM-1.7B và XGLM-564M là khả thi các giải pháp thay thế, xét đến việc chúng có thể được đào tạo trong thậm chí ít thời gian hơn n và chiếm ít bộ nhớ GPU hơn n. Chúng tôi cũng báo cáo hiệu suất trên phân chia xác thực của COCO-35L trong Bảng 7.

RAM	thời gian	XGLM	en	là	chào	zh
2,9 tỷ	23 giờ	46 giờ	57,3	44,9	20,8	25,9
1,7 tỷ	14 giờ	29 giờ	55,8	41,0	20,1	24,6
564M	7h	19G	51,7	40,0	18,0	23,8

Bảng 6: Kết quả CIDEr trên tập dữ liệu XM3600. Chúng tôi báo cáo hiệu suất cho các XGLM khác nhau được sử dụng trong thành phần giải mã của PAELLA.

RAM	thời gian	XGLM	en	là	chào	zh
2,9 tỷ	23 giờ	46G	113,6	113,9	86,2	123,3
1,7 tỷ	14 giờ	29 giờ	108,7	107,7	82,2	116,6
564M	7h	19G	103,2	103,1	76,6	111,2

Bảng 7: Kết quả CIDEr trên bộ xác thực của COCO-35L, trên các bộ giải mã khác nhau được sử dụng trong PAELLA.

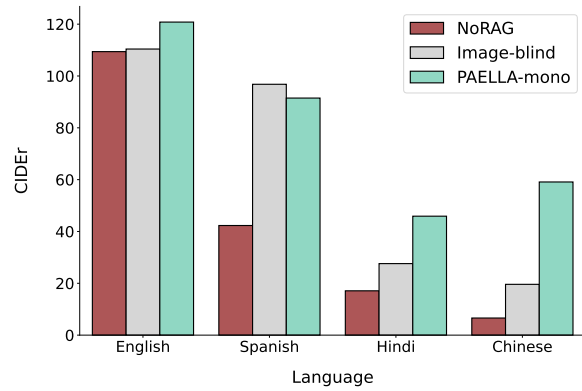


Figure 5: Ablation results on the COCO-35L dataset, reported with the CIDEr metric for the mono variant. We ablate the retrieval (NoRAG) and the visual encoder (image-blind), and compare with PAELLA_{mono}.

E Monolingual Retrieval

We study the behavior of our model when the retrieved captions are provided in English instead of the target language, 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_{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_{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_{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 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:

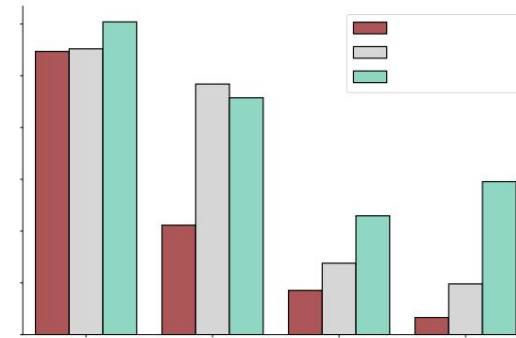
- Key Weight Matrix size: $[768, 8]$ (i.e., $enc_d \times d$)
- Value Weight Matrix size: $[768, 8]$ (i.e., $enc_d \times d$)
- Query Weight Matrix size: $[2048, 8]$ (i.e., $dec_d \times d$)
- Total parameters for one layer attention with 16 heads: $16 \times (2 \times 768 \times 8 + 2048 \times 8)$
- 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



Hình 5: Kết quả cắt bỏ trên tập dữ liệu COCO-35L, được báo cáo với số liệu CIDEr cho biến thể đơn sắc. Chúng tôi loại bỏ việc truy xuất (NoRAG) và bộ mã hóa hình ảnh (không nhìn thấy hình ảnh) và so sánh với PAELLA_{mono}.

E Truy xuất đơn ngữ

Chúng tôi nghiên cứu hành vi của mô hình khi các chú thích được lấy lại được cung cấp bằng tiếng Anh thay thế của ngôn ngữ đích, như được thấy trong Bảng 8. Chúng tôi có thể thấy rằng mô hình của chúng tôi được hưởng lợi từ việc có đã lấy các ví dụ trong cùng ngôn ngữ với mục tiêu- ngôn ngữ đầu ra. Theo cách này, các chú thích có thể hướng dẫn quá trình tạo nội dung trong ngôn ngữ đích, bằng cách cung cấp tài liệu tham khảo cho những gì chú thích dự đoán sẽ giống như sau.

RAG và	là	CHÀO	zh
Đa	113,6	113,9	86,2 123,3
En	114,1	103,8	76,8 121,3

Bảng 8: Hiệu suất sử dụng chú thích đã lấy trong ngôn ngữ đích (nhiều ngôn ngữ) hoặc bằng tiếng Anh, được đo lường thông qua CIDEr trên bộ xác thực COCO.

Tác động của việc truy xuất F lên PAELLA_{mono}

Tư duy tự như những phát hiện cho PAELLA trong Phần 6.3, chúng ta quan sát trong Hình 5 rằng việc tăng cường truy xuất cũng đóng vai trò quan trọng trong PAELLA_{mono}. Thật vậy, việc truy xuất đặc biệt quan trọng đối với biến thể đơn ngữ. Điều này xảy ra vì mô hình dựa vào thậm chí còn nhiều hơn về các ví dụ được lấy lại để tạo chú thích bằng các ngôn ngữ khác biệt đáng kể so với Dữ liệu đào tạo tiếng Anh, được chứng minh bằng sự sụt giảm đáng kể về hiệu suất với NoRAG đối với tiếng Hindi và tiếng Trung. Chúng ta cũng thấy rằng biến thể mù hình ảnh làm giảm hiệu suất của PAELLA_{mono}, chứng minh rằng mô hình của chúng tôi không chỉ sử dụng thông tin từ các chú thích đã lấy được mà còn sử dụng bằng biến thể không có sự truy xuất (NoRAG). Trong

hình ảnh bản thân. Biến thể mù hình ảnh phải tạo ra chú thích chỉ với thông tin đã lấy được, điều này thực sự là thách thức đối với tiếng Hindi và tiếng Trung. Có thể khó để tìm ra cách kết hợp và tóm tắt thông tin từ bốn thông tin đã thu thập được chú thích thành một đầu ra duy nhất gắn kết, đặc biệt đối với những ngôn ngữ này có những đặc điểm rất khác biệt so với sự giám sát của tiếng Anh. Ngược lại, mô hình sử dụng thông tin thu được một cách dễ dàng đối với tiếng Tây Ban Nha khi suy luận, đạt được hiệu suất tốt hơn thông qua việc diễn đạt lại một cách đơn giản. Hơn nữa, cách tiếp cận không dùng hình ảnh có hiệu suất cao hơn mô hình NoRAG trên cả bốn ngôn ngữ, hơn nữa nhấn mạnh tầm quan trọng của việc tạo ra điều kiện với các ví dụ được thu thập.

G Chú ý chéo

Mô hình của chúng tôi có 34M tham số có thể đào tạo tương ứng với các lớp chú ý chéo. Ở đây, chúng tôi cung cấp cái nhìn sâu sắc vào thiết lập chú ý chéo, có bộ mã hóa ẩn có kích thước 768 và bộ giải mã kích thước ẩn của 2048, với 16 đầu chú ý và một chồng 48 lớp. Chúng tôi giảm kích thước của ma trận chiếu chéo chú ý, được ký hiệu là d , từ tiêu chuẩn 128 (2048/16) đến 8, để đạt được đào tạo hiệu quả tham số. Do đó, tổng số tham số được tính như sau:

- Kích thước Ma trận trọng số $[768, 8]$ (tức là, khóa: $enc_d \times d$)
- Kích thước Ma trận Trọng số Giá trị: $[768, 8]$ (tức là, mã hóa $_d \times d$)
- Kích thước Ma trận trọng số truy vấn: $[2048, 8]$ (tức là, $dec_d \times d$)
- Tổng số tham số cho một lớp chú ý với 16 đầu: $16 \times (2 \times 768 \times 8 + 2048 \times 8)$
- Trọng lượng dày đặc cho phép chiếu sau khi nối các đầu: $[16 \times 8, 2048]$ ($h \times d \times dec_d$)

Tổng số lớp là 48.

Tổng số tham số: $48 \times (16 \times (2 \times 768 \times 8 + 2048 \times 8) + 16 \times 8 \times 2048) \approx 34M$

H Kết quả định tính

Trong Hình 6, chúng tôi cung cấp các ví dụ về chú thích được tạo bởi PAELLA, dựa trên cả hình ảnh và các chú thích đã lấy được và các chú thích đã tạo



PAELLA

NoRAG

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, Span-

ish, 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.



Hình 6: Các ví dụ định tính cho các chú thích được tạo ra bởi PAELLA, so với các kết quả được tạo ra bằng một mô hình bị loại bỏ không sử dụng khả năng tăng cường truy xuất.

hình ảnh đầu tiên, mô hình của chúng tôi nắm bắt chính xác khái niệm cú trên các ngôn ngữ cốt lõi khác nhau, như có trong các chú thích đã lấy lại. PAELLA cũng thể hiện một số tính mạnh mẽ đối với thông tin sai lệch tiềm ẩn có thể xảy ra trong các chú thích đã lấy (ví dụ: chú thích thứ hai được lấy lại đề cập đến một con cú trong một bảng). Ngược lại, biến thể NoRAG tạo ra chú thích không chính xác cho 4 ngôn ngữ, vật lộn với việc xác định loài chim, thậm chí phân loại sai nó như một con hươu cao cổ đối với người Trung Quốc. Trên hình ảnh thứ hai, chúng tôi trình bày một ví dụ tiêu cực trong đó các chú thích được lấy lại có thể gây hiểu lầm cho mô hình của chúng tôi. PAELLA tạo ra chú thích đề cập đến một con dao quân đội Thụy Sĩ màu đỏ, có khả năng bị ảnh hưởng bởi màu sắc có trong các chú thích đã lấy được (và một phần trong chính con dao, mặc dù nó chủ yếu là màu trắng). Tuy nhiên, mô hình thành công của chúng tôi hoàn toàn tạo ra khái niệm về một con dao Thụy Sĩ, trong khi Biến thể NoRAG gặp khó khăn khi tạo ra các đối tượng không liên quan (ví dụ: điện thoại di động, kính mát, đồ chơi hoặc tai nghe cho tiếng Anh, Span-

tiếng Anh, tiếng Hindi và tiếng Trung (lần lượt).

Tôi Biểu Diễn Trên 36 Ngôn Ngữ

Trong Bảng 9, chúng tôi báo cáo hiệu suất XM3600 trên tất cả 36 ngôn ngữ. Chúng tôi hiển thị kết quả cho mô hình của chúng tôi và các biến thể của nó, cùng với các mô hình đa ngôn ngữ hiện đại có hiệu suất cho từng ngôn ngữ trong các ấn phẩm tư ng ứng nữa.

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.

Lang.	mBLIP	mT0-XL	BB+CC	Lg	Mono	Core	PAELLA
vi	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
của	32,5		22,4	13.0	21.7	22.1	21,5
là	62,6		42,5	22,0	42,2	45,0	21,7
fr	57,6		41.0	36,1	38,9	14,1	13,0
Vâng	33.2		25,4	18,6	16,8	29,3	32,5
nó	45,2		32,1				33.2
pt	53,1		38.0	20.2	38.7	40.0	41.0
tel	23,4		19,9	10.1	23.3	21.7	24,6
không	10.4		28,8	15.2	21.7	21.2	27,2
có	16.8		17,7	8.9	15,6	16,9	18.1
may may	38,5		30,7	16,7	34,0	34,3	31,6
tr	22,6		23.2	12,2	19,0	19,3	10,6
ar	21.1		22,7	17,3	19,0	18,2	39,3
vi	39,2		33,6	38,7	22,6	20,8	22,1
th	41,9		41,8				40,4
	16,1		19,7	11.1	17.1	20.4	20,8
chào bạn	11,3		20.0	13,3	18,8	16,5	21,7
quốc gia của	11.8		31,9	15.1	23.0	22.8	28,5
tôi	11.2		19,6	9,9	17.2	15.3	19,9
câu đó	1.1		0.0	0.0	0,2	0,7	0,8
Ngôn ngữ không có trong dữ liệu đào tạo trước của XGLM							
cs	31,8		31.3	13.9	0,5	0,2	21,6
ngày	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
làm	17,7		35,3	18,5	1.7	2.2	26,6
Anh ta	18,7		23.0	9,8	0,0	0,0	15,5
giờ	5,2		22,4	8,5	0,3	0,2	16.0
hồ	21,5		17,5	9,6	0,4	0,1	11,5
tôi	4,1		40,5	24,3	1.1	3.6	33,4
không có	55,7		44,1	23.2	1.9	2,5	36,5
KHÔNG	46,2		38,5	23.0	1.0	1.8	31.0
xin	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
Anh quốc	0.0		18,9	8.1	2.8	2,5	13.3
TRUNG BÌNH	28.3		28,5	15.0	15.5	16.8	26,2
TRUNG BÌNH	30,5		27,7	14,7	23,9	24,9	28.2

Bảng 9: Kết quả CIDEr trên chuẩn XM3600 trên 36 ngôn ngữ, đư ợc sắp xếp theo ngôn ngữ đào tạo trư ớc tỷ lệ của bộ giải mã XGLM. AVG chỉ ra hiệu suất trung bình trên 36 ngôn ngữ, trong khi AVG biểu thị hiệu suất trên các ngôn ngữ mà XGLM đư ợc đào tạo trư ớc.