Challenges in Image-Caption Association in Portuguese: Evaluating the CLIP Model on the FM30k Dataset

Vitoria C. Benedet[1], Gustavo L. Tamiosso[1], Rafael O. Nunes[1], Dennis G. Balreira[1]
[1]Institute of Informatics, Federal University of Rio Grande do Sul
Porto Alegre, Brazil
{vcbenedet, gltamiosso, ronunes, dgbalreira}@inf.ufrgs.br

Abstract

In recent decades, multimodal models such as CLIP have achieved significant advances in associating images and texts. However, most of these advances come from models trained almost exclusively in English, which limits their effectiveness in other languages. This challenge is particularly relevant for Brazilian Portuguese, a language that still lacks dedicated multimodal resources and relies predominantly on automatic translations. This work investigates the performance of CLIP-based multimodal models in the task of associating images and descriptions written in Brazilian Portuguese. The analysis begins with a zero-shot scenario, in which different CLIP variants are directly evaluated on the FM30k dataset, composed of images and captions originally written in Portuguese. We also conducted an additional experiment with automatic translations to examine the impact of language on cross-modal retrieval tasks. Subsequently, fine-tuning is performed on the ViT-B/32 model's textual encoder, keeping the visual encoder frozen, to adapt the model to the target language. The results show that models originally trained in English perform worse in Portuguese, while linguistically adapted variants, either multilingual or Portuguese-specific, achieve superior performance. The proposed fine-tuning approach reduced this performance gap, resulting in notable improvements. In the image-to-text scenario, the model achieved an absolute increase of 27.65 percentage points in the Accuracy@1 metric, representing a 209% relative gain over the original CLIP ViT-B/32. In the text-to-image scenario, the gain was 15.47 percentage points, resulting in an even greater 385% relative improvement and contributing to a more balanced association between images and captions.

Keywords: Multimodal Models, CLIP, Brazilian Portuguese, Image Text Retrieval

1. Introduction

In recent decades, significant advances in Machine Learning and Computer Vision have enabled the development of models capable of processing and integrating multiple data modalities, such as images and text (Baltrušaitis et al., 2017). In this context, multimodal models have become essential for tasks such as information retrieval (Radford et al., 2021), automatic caption generation (Xu et al., 2016) and feature extraction (Vinícius Todescato and Luís Carbonera, 2024). This evolution reflects a fundamental premise: "Our experience of the world is multimodal — we see objects, hear sounds, feel textures, smell odors, and taste flavors. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together." (Baltrušaitis et al., 2017)

Among the proposed models, CLIP (Contrastive Language–Image Pretraining) (Radford et al., 2021), developed by OpenAI, stands out for its ability to align visual and textual representations within a shared vector space. This alignment allows for effective association between images and text, even without task-specific supervision.

However, most of these advances have been driven by models trained almost exclusively on English-language data (Li et al., 2021; Jia et al., 2021). This poses a major limitation, as the direct transfer of these solutions to other languages and cultural contexts often results in degraded performance (Rosa et al., 2021). The problem is even more pronounced for Brazilian Portuguese, which still lacks large-scale multimodal resources that reflect its linguistic and cultural richness. Capturing culturally specific visual concepts from Brazilian contexts, such as regional traditions and heritage practices, also remains a challenge for vision–language models (Cruz-Castañeda et al., 2025). Furthermore, many Portuguese datasets rely on automatic translations of English corpora, which can introduce linguistic noise and semantic imbalance.

Given this scenario, it becomes necessary to investigate **how CLIP and its derivatives perform when applied to data originally written in Brazilian Portuguese**, a low-resource and underexplored language in the multimodal domain. It is also important to assess the feasibility of adapting such models using lighter and less complex architectures, particularly in contexts such as developing countries where computational resources are often limited. This makes the development of efficient and accessible solutions even more critical.

The central goal of this research is to evaluate the performance of different CLIP variants on the task

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9409–9419
11–16 May 2026.

of associating images with descriptions in Brazilian Portuguese. To this end, we propose an experimental approach that includes: (i) a zero-shot evaluation of multiple models, including original, multilingual, and Portuguese-adapted versions; (ii) an analysis of language influence using automatically translated captions; and (iii) fine-tuning of the text encoder from the `ViT-B/32` model using the FM30k dataset (Viridiano et al., 2024), which consists exclusively of image descriptions originally written in Portuguese.

The primary contributions of this paper are:

- **A comprehensive zero-shot benchmark of prominent CLIP variants** (OpenAI originals, multilingual, and Portuguese-adapted) on the native FM30k dataset, establishing a clear baseline for performance in an authentic linguistic setting.
- **A diagnostic analysis of performance asymmetry between image-to-text and text-to-image retrieval**, empirically identifying the text encoder as the principal bottleneck for English-centric models when processing Portuguese.
- The proposal and evaluation of `CLIP-ViT/32-FM30k`, **a lightweight adaptation of CLIP** achieved by fine-tuning only the text encoder of a standard ViT-B/32 model. This "locked-image" tuning strategy is specifically designed for computational efficiency.
- **A detailed cost-benefit analysis** demonstrating that our lightweight `CLIP-ViT/32-FM30k` model achieves competitive results against much larger state-of-the-art multilingual models while being significantly more parameter-efficient, offering a practical and accessible adaptation pathway for low-resource languages.

2. Related Work

2.1. Foundational Models: CLIP

The CLIP model, developed by OpenAI, learns visual concepts from natural language supervision by training on a massive dataset of image-text pairs from the internet (Radford et al., 2021). Its architecture consists of two main components: an image encoder (e.g., a Vision Transformer or ViT) and a text encoder (e.g., a Transformer). During training, a contrastive loss function is employed to maximize the cosine similarity of the embeddings for correct image-text pairs while minimizing it for incorrect pairs within a batch. This process aligns the two modalities into a shared embedding space, enabling powerful zero-shot classification and cross-modal retrieval without task-specific training (Radford et al., 2021).

2.2. Multilingual and Adapted Models

The `mCLIP` model, developed by the LAION community, is a prominent multilingual variant that pairs the `ViT-B/32` visual encoder with the XLM-RoBERTa-base text encoder, trained on over 100 languages. Pretrained on the large LAION-5B dataset with billions of multilingual image-text pairs, mCLIP supports zero-shot retrieval across diverse languages (Schuhmann et al., 2022). Despite progress, multilingual robustness varies significantly across languages.

The `CAPIVARA` model (dos Santos et al., 2023) builds upon mCLIP to specifically adapt CLIP for Brazilian Portuguese in low-resource settings. CAPIVARA uses a pipeline where synthetic English captions are generated from images in the CC3M dataset (Sharma et al., 2018) and then automatically translated into Portuguese, creating a large-scale Portuguese image-text corpus. During fine-tuning, only the text encoder is updated, with the visual encoder frozen, and *Low-Rank Adaptation* (LoRA) is applied to reduce computational costs and speed training. CAPIVARA was evaluated on multiple Portuguese datasets, including PraCegoVer (dos Santos et al., 2022), Portuguese-translated MS COCO (Lin et al., 2015), Flickr30k (Young et al., 2014), ImageNet (Deng et al., 2009), and ELEVATER benchmarks (Li et al., 2022), focusing on zero-shot image-text retrieval.

Other monolingual adaptations exist for languages such as Italian (Bianchi et al., 2021), Chinese (Yang et al., 2023), and Korean (Ko and Gu, 2022), typically relying on fine-tuning with native corpora to address CLIP's original limitations in specific languages. Although these works extend CLIP's multilingual reach, many depend heavily on translated captions rather than native-language data, which can limit their ability to capture linguistic and cultural nuances.

Our work complements these approaches by evaluating and fine-tuning CLIP variants, including `CAPIVARA` and `mCLIP`, on FM30k, a novel dataset composed exclusively of human-written Brazilian Portuguese descriptions. Unlike corpora based on translations or synthetic captions, FM30k enables a more authentic evaluation and adaptation in a native low-resource setting, providing insights into the true capabilities of CLIP models for Brazilian Portuguese.

3. A Lightweight Approach to Multilingual Adaptation

3.1. Framed Multi30k Dataset

This study uses the *Framed Multi30k* (FM30k) dataset, an extension of the widely used Flickr30k and Multi30k datasets for image captioning and multimodal research. Flickr30k contains over 31,000 images, each with five human-written English captions (Young et al., 2014). Multi30k adds translated captions in multiple languages, while FM30k expands this by focusing on Brazilian Portuguese. The FM30k dataset consists of 158,915 original Brazilian Portuguese captions authored by native speakers, 30,104 Portuguese translations of English captions, and more than 4.5 million semantic frame annotations aligned with captions in both languages.

The original Portuguese captions were produced by 148 university students through a rigorous annotation process with quality controls (Viridiano et al., 2024). For this work, only the original Portuguese captions are used to evaluate CLIP models on native-language image-text matching.

Each image in FM30k is associated with five independent Portuguese captions, stored in a text file with the format: `image_name#caption_number caption_text`. For example, the image `1000092795.jpg` has captions such as "Two men talk in the garden near a gate"[1].

The images are part of the Flickr30k dataset, which requires request-based access via a form[2] and is subject to Flickr's terms of use. FM30k captions and annotations are publicly available on GitHub[3], maintained by the dataset's creators.

3.2. Zero-shot Evaluation

To assess the ability of CLIP-based models to associate images and texts in Brazilian Portuguese without any task-specific adaptation, we conducted a zero-shot evaluation using the FM30k dataset. This experiment served both as a baseline analysis and as motivation to explore potential improvements via fine-tuning.

3.2.1. Data Setup

The complete FM30k dataset was used without partitioning into training, validation, or test splits, since the goal was to evaluate model performance without any prior exposure to the data. Each image in the dataset is associated with exactly five independent captions in Brazilian Portuguese, which were grouped together for evaluation to preserve alignment between modalities.

3.2.2. Evaluated Models

Five CLIP-based models were evaluated to cover different architectural scales and multilingual extensions: CLIP ViT-B/32, CLIP ViT-B/16, CLIP ViT-L/14, m-CLIP (multilingual), and CAPIVARA (Portuguese-adapted). The first three models were loaded from OpenAI's CLIP implementation, while `m-CLIP` and `CAPIVARA` were loaded using OpenCLIP (Ilharco et al., 2021), with corresponding image and text preprocessing pipelines.

3.2.3. Preprocessing and Embedding Generation

Images were loaded in RGB format and preprocessed using each model's default transformation pipeline (resizing, normalization, etc.). Text captions were tokenized with the appropriate tokenizer for each model. For every image-caption pair, we computed visual and textual embeddings via the model's respective encoders. Embeddings were normalized to unit vectors, enabling cosine similarity computation via dot product.

3.2.4. Similarity Computation and Evaluation Metrics

Model efficacy is rigorously evaluated via standard cross-modal retrieval metrics, predominantly Accuracy@k (synonymous with Recall@k in this domain context). It is vital to theoretically distinguish our cross-modal retrieval objective from generative image captioning tasks. While n-gram overlap metrics (such as BLEU, ROUGE, and METEOR) and modern reference-free multi-modal metrics like CLIPScore are highly powerful for assessing the syntactic and semantic quality of *newly generated* text, they are conceptually inappropriate for this study. Our task assumes a fixed, ground-truth database where the objective is pure ranking, not synthesis. Moreover, employing a CLIP-based metric like CLIPScore to evaluate the retrieval effectiveness of fine-tuned CLIP variants would introduce problematic circular bias. Accuracy@k remains the industry standard for exact matching retrieval and is widely adopted in the literature (Radford et al., 2021; dos Santos et al., 2023), as it unambiguously quantifies the proportion of instances where the true associated item is accurately positioned within the top k ranks.

Similarity matrices were computed in both directions: image-to-text and text-to-image. For each

[1] Original: *Dois homens conversam no jardim perto do portão*.

[2] https://forms.illinois.edu/sec/229675

[3] https://github.com/FrameNetBrasil/framed-multi30k/tree/main

query, the top-k most similar elements were retrieved using `torch.topk`, and the results were used to compute standard retrieval metrics, including *Accuracy@k*.

This evaluation provides a reference point for model performance without task-specific adaptation, highlighting the baseline cross-modal alignment capabilities in Brazilian Portuguese.

3.3. Zero-shot with Different Languages

To investigate the impact of language on the CLIP model's performance, we conducted a parallel zero-shot experiment using automatic translations of the FM30k and Flickr30k captions. The underlying hypothesis is that, since CLIP was primarily trained on English text, its performance is likely to be superior when operating in that language, which may compromise its effectiveness in tasks involving captions written in Portuguese.

For this purpose, we employed the `facebook/nllb-200-distilled-600M` model, which is specialized in high-quality multilingual translation (Team et al., 2022). The translations were generated while preserving the same data structure used previously, resulting in three new vectorized versions of the captions:

- Portuguese-to-English translations of the FM30k captions;
- English-to-Portuguese translations of the Flickr30k captions;
- Original English captions from the Flickr30k dataset.

All zero-shot procedures followed the same pipeline described in Section 3.2, including the generation of text and image embeddings, similarity computation using precomputed image embeddings, and evaluation with the same metrics. However, in this multilingual setup, we used only the baseline `ViT-B/32` model for consistency and to isolate the language variable.

3.4. Fine-tuning on Brazilian Portuguese Texts

This experiment aims to evaluate the impact of adapting the text encoder of the `ViT-B/32` CLIP model to a corpus written in Brazilian Portuguese. We hypothesize that aligning the text encoder to the linguistic characteristics of the FM30k captions could improve cross-modal retrieval performance. The base model was selected for its competitive performance and relatively low computational cost (Radford et al., 2021).

3.4.1. Cross-validation Protocol

We adopted five-fold cross-validation to ensure robust generalization estimates. Each fold contains a disjoint subset of images and their associated captions. Within each round, one fold is reserved for testing, while the remaining four are split into training (80%) and validation (20%). This guarantees no image or caption leakage between partitions, enabling fair performance assessment.

3.4.2. Fine-tuning Strategy

Inspired by the *Locked-image Tuning* strategy (Zhai et al., 2022), we fine-tune only the text encoder while keeping the visual encoder frozen. This choice assumes that the pretrained visual encoder already provides strong representations, and focuses training on adapting the text encoder to Brazilian Portuguese. This approach also reduces computational cost by avoiding gradient computation in the image branch.

The training was performed using the `AdamW` optimizer with a learning rate of 10^{-5} and a weight decay of 0.01. We used a batch size of 512 and trained for up to 15 epochs, with early stopping typically halting the process between epochs 5 and 8. These hyperparameters were selected based on preliminary experiments to ensure stable convergence while preventing catastrophic forgetting of the original multimodal alignment.

Although Parameter-Efficient Fine-Tuning (PEFT) methodologies such as Low-Rank Adaptation (LoRA) have emerged as standard practices for adapting Large Language Models with billions of parameters, our proposed architecture deliberately employs a full fine-tuning mechanism constrained strictly to the text encoder. The ViT-B/32 text encoder contains roughly 63 million parameters—a size sufficiently small to perform full fine-tuning with negligible memory constraints (easily accommodated on standard GPU hardware). Recent spectral analyses indicate that while LoRA excels at proximal domain adaptations by injecting low-rank decomposition matrices, its constrained intrinsic dimensions can struggle to fully overwrite representational structures when bridging massive structural and linguistic gaps, such as the complete transition from English semantics to native Brazilian Portuguese. By executing a full parameter update on the text encoder whilst keeping the heavier visual encoder rigidly locked, we grant the textual representations the necessary tensor flexibility to organically conform to the Portuguese language's latent space without incurring the out-of-distribution forgetting vulnerabilities occasionally observed in restrictive low-rank adaptations.

The training data consist of (image, caption) pairs created by replicating each of the five captions per

image. These pairs are processed through CLIP's standard image and text preprocessing routines.

The model is trained using the original CLIP contrastive loss (Radford et al., 2021), treating cross-modal matching as a symmetric multi-class classification task. Embeddings are normalized to ensure cosine similarity and stable optimization. The loss is computed in both directions (image-to-text and text-to-image), and averaged.

3.4.3. Regularization and Early Stopping

To mitigate overfitting, we apply early stopping based on validation loss. Training halts if no significant improvement ($> 10^{-4}$) is observed after three consecutive epochs. The model with the best validation performance is saved for evaluation.

3.4.4. Final Evaluation

After each fold, the fine-tuned model (`CLIP-ViT-B/32-FM30k`) is evaluated on the corresponding test set. The evaluation procedure follows the same steps described in Section 3.2: embedding generation, similarity computation, and retrieval metric calculation.

This allows direct comparison with the zero-shot models, including those listed in Section 3.2.2. Final performance is reported as the mean and standard deviation across the five test folds, allowing us to assess the effectiveness of domain-specific text encoder adaptation for improving image-text alignment in Brazilian Portuguese.

3.5. Experimental Environment

The experiments were conducted on a high-performance computing cluster equipped with Intel(R) Core(TM) i9-14900KF processors running at 3.20 GHz, featuring 24 physical cores and 32 threads, 128 GB of DDR5 RAM, and NVIDIA GeForce RTX 4090 GPUs. This infrastructure provided the necessary computational resources to support model evaluation and fine-tuning with reliable performance and reproducibility.

The development environment was configured with Python version `3.11.2`, along with the main libraries required for the experiments: `torch`, `sklearn`, `clip`, `open_clip`, `tqdm`, `numpy`, `pandas`, `matplotlib`, and `seaborn`. This computational infrastructure provided the necessary resources for evaluating the models and performing fine-tuning, ensuring both performance and reproducibility of the experiments.

4. Results and Discussion

4.1. Zero-shot

This section presents the results obtained in the zero-shot inference setting, where CLIP models were evaluated directly on the FM30k dataset without any additional fine-tuning. The goal is to analyze the performance of different CLIP variants on the cross-modal retrieval task between images and captions in Brazilian Portuguese. Key aspects discussed include the influence of language on model performance, the role of the text encoder, the symmetry between retrieval directions, and the impact of model complexity on the quality of semantic associations.

Linguistic adaptation enhances CLIP model performance. The models `CAPIVARA` and `mCLIP`, both adapted to Portuguese, consistently outperformed other variants across all values of k and in both retrieval directions (image-to-text and text-to-image). This highlights the importance of training or adapting the text encoder to the target language, especially in tasks that rely heavily on precise semantic alignment between images and text. Notably, as shown in Figure 1, `CAPIVARA`, specifically optimized for Portuguese, tends to surpass `mCLIP`, which, although multilingual and including Portuguese in its training, lacks the same level of specialization.

Original CLIP models face limitations with Portuguese captions. The original OpenAI CLIP models (`ViT-B/32`, `ViT-B/16`, and `ViT-L/14`), trained predominantly on English data, showed inferior performance when applied to Portuguese captions. While these models excel in English-language benchmarks (Radford et al., 2021), their performance degrades significantly in multilingual contexts without specific adaptation or fine-tuning. This underscores their limited generalization capability across languages without targeted training.

Importance of the text encoder for Portuguese performance. It is important to note that `ViT-B/32`, `mCLIP`, and `CAPIVARA` share the same visual encoder. Therefore, performance differences arise exclusively from the textual encoder used by each model. While the original CLIP employs a text encoder trained mostly on English data, `mCLIP` and `CAPIVARA` use text encoders adapted to Portuguese based on different architectures—specifically XLM-RoBERTa, designed for multilingual understanding. This architectural choice, combined with targeted training, proved decisive for successful semantic retrieval with Portuguese captions.

Asymmetry in retrieval direction performance. Figure 1 reveals a significant performance gap between retrieval directions in the original OpenAI CLIP models. For instance, in the `ViT-`

Model	Visual Architecture	Textual Architecture	Parameters (Visual)	Parameters (Textual)	Total
ViT-B/32	ViT-Base (32px patch)	Transformer (OpenAI BPE)	87M	63M	150M
ViT-B/16	ViT-Base (16px patch)	Transformer (OpenAI BPE)	87M	63M	150M
ViT-L/14	ViT-Large (14px patch)	Transformer (OpenAI BPE)	303M	123M	426M
mCLIP	ViT-Base (32px patch)	XLM-RoBERTa-base	87M	278M	365M
CAPIVARA	ViT-Base (32px patch)	XLM-RoBERTa-base	87M	278M	365M

Table 1: Comparison among CLIP models

Metric	Compared Model	Accuracy	Accuracy of `CLIP-ViT-B/32-FM30k`	Absolute Difference	Relative Difference
accuracy@1	mCLIP	58.90%	40.87%	**-18.03%**	**-30.6%**
	CAPIVARA	61.85%	40.87%	**-20.98%**	**-33.9%**
accuracy@5	mCLIP	82.06%	67.47%	**-14.59%**	**-17.8%**
	CAPIVARA	83.81%	67.47%	**-16.34%**	**-19.5%**
accuracy@10	mCLIP	88.82%	77.39%	**-11.43%**	**-12.9%**
	CAPIVARA	90.14%	77.39%	**-12.75%**	**-14.1%**
accuracy@100	mCLIP	98.69%	95.45%	**-3.24%**	**-3.3%**
	CAPIVARA	98.83%	95.45%	**-3.38%**	**-3.4%**

Table 2: Absolute and Relative Comparison — Image-to-Text Retrieval for Adapted Models. Results report the Average Accuracy@k across test folds.

Metric	Compared Model	Accuracy	Accuracy of `CLIP-ViT-B/32-FM30k`	Absolute Difference	Relative Difference
accuracy@1	mCLIP	39.19%	19.49%	**-19.70%**	**-50.3%**
	CAPIVARA	42.56%	19.49%	**-23.07%**	**-54.2%**
accuracy@5	mCLIP	64.25%	42.19%	**-22.06%**	**-34.3%**
	CAPIVARA	68.51%	42.19%	**-26.32%**	**-38.4%**
accuracy@10	mCLIP	73.65%	53.83%	**-19.82%**	**-26.9%**
	CAPIVARA	77.77%	53.83%	**-23.94%**	**-30.8%**
accuracy@100	mCLIP	94.43%	86.30%	**-8.13%**	**-8.6%**
	CAPIVARA	95.96%	86.30%	**-9.66%**	**-10.1%**

Table 3: Absolute and Relative Comparison — Text-to-Image Retrieval for Adapted Models. Results report the Average Accuracy@k across test folds.

Metric	Compared Model	Accuracy	Accuracy of `CLIP-ViT-B/32-FM30k`	Absolute Difference	Relative Difference
accuracy@1	ViT-B/32	13.22%	40.87%	**+27.65%**	**+209%**
	ViT-B/16	15.53%	40.87%	**+25.34%**	**+163%**
	ViT-L/14	23.13%	40.87%	**+17.74%**	**+77%**
accuracy@5	ViT-B/32	28.62%	67.47%	**+38.85%**	**+135%**
	ViT-B/16	32.48%	67.47%	**+34.99%**	**+108%**
	ViT-L/14	44.58%	67.47%	**+22.89%**	**+51%**
accuracy@10	ViT-B/32	37.64%	77.39%	**+39.75%**	**+106%**
	ViT-B/16	42.02%	77.39%	**+35.37%**	**+84%**
	ViT-L/14	55.28%	77.39%	**+22.11%**	**+40%**
accuracy@100	ViT-B/32	72.29%	95.45%	**+23.16%**	**+32%**
	ViT-B/16	76.71%	95.45%	**+18.74%**	**+24%**
	ViT-L/14	86.46%	95.45%	**+8.99%**	**+10%**

Table 4: Absolute and Relative Comparison — Image-to-Text Retrieval with OpenAI CLIP Models. Results report the Average Accuracy@k across test folds.

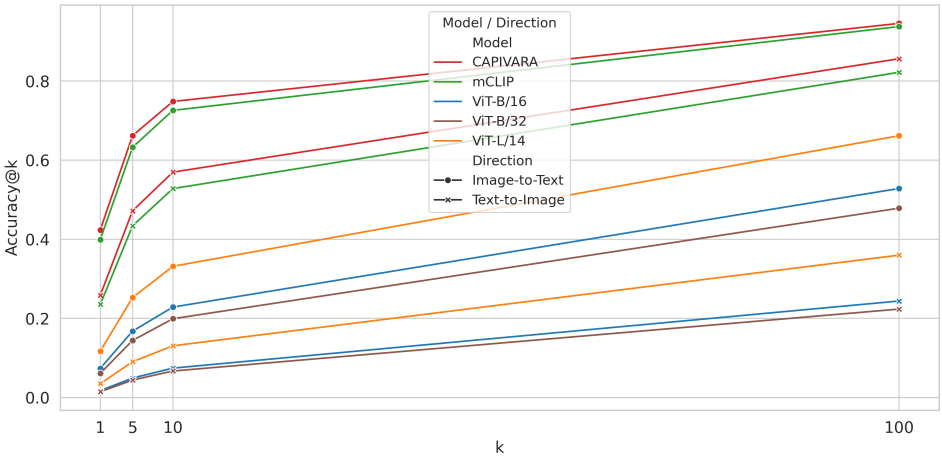

Figure 1: Accuracy@k in zero-shot scenario.

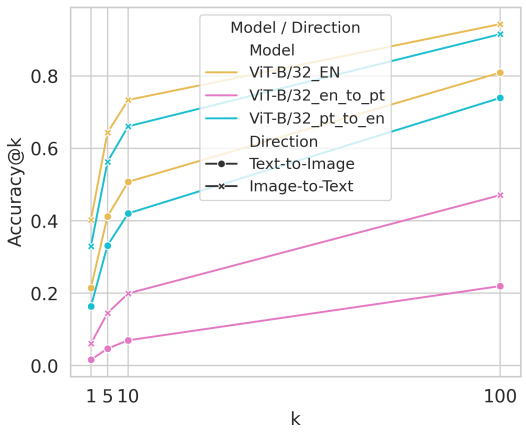

Figure 2: Accuracy@k in the zero-shot scenario with captions in different languages.

B/32 baseline, the *Accuracy@1* reaches 13.22% for image-to-text retrieval, while plummeting to a marginal 4.02% in the text-to-image task—a 3.3x disparity. This asymmetry is partly explained by the dataset structure, which provides five captions for each image. Probabilistically, it is substantially easier to retrieve at least one matching caption for a given image than it is to identify the single correct image corresponding to a specific textual query. Conversely, mCLIP and CAPIVARA show more balanced curves across both directions, reflecting more symmetric representations of images and text. This balance is attributed to the adapted text encoders. The asymmetry in original CLIP models likely stems from the textual encoder's English-only training: poor textual representation impairs image retrieval from text queries, while robust visual encoders help partially compensate when retrieving text from images.

Accuracy increases with k. As expected, Accuracy@k improves as k increases across all models. This improvement is more pronounced for models trained or adapted to Portuguese, indicating not only correct identification of image-text pairs at top ranks but also better overall ranking of relevant candidates.

ViT-L/14 outperforms ViT-B variants. Among OpenAI models, ViT-L/14 consistently surpasses the smaller ViT-B/16 and ViT-B/32 in all retrieval settings. This confirms that larger architectures tend to offer superior generalization and semantic representation capabilities. Nevertheless, ViT-L/14 remains behind mCLIP and CAPIVARA, emphasizing that linguistic adaptation has a greater impact than model size alone.

4.2. Zero-shot with Different Languages

Analysis of Accuracy@k values reveals a substantial language-dependent performance gap in the ViT-B/32 model (see Figure 2). Using original English captions (ViT-B/32_EN) yields the highest retrieval accuracy, reflecting strong alignment between the text encoder and English input. However, when the captions are translated from English to Portuguese (ViT-B-32_en_to_pt), the performance drops sharply in both directions, revealing a marked difficulty of the text encoder in adequately representing the content in Portuguese. Conversely, when the original Portuguese captions are translated into English (ViT-B-32_pt_to_en), the performance improves significantly compared to the Portuguese captions, although it still falls short of the results achieved with the captions originally written in English.

It is important to highlight that the model remains unchanged across all scenarios, with the language of the captions as the only variable. This indicates that the performance drop is directly related to the input language representation. These results suggest that the primary influencing factor is the textual representation provided to the model, where using English—more aligned with the latent space learned during training—yields a substantial advantage. The text-to-image retrieval task appears particularly sensitive to the quality of this representation, reinforcing the negative impact of using languages not native to the original text encoder. Additionally, although machine translation is a feasible alternative, it may introduce noise and semantic loss (Rosa et al., 2021), which can adversely affect the quality of the textual representation.

4.3. Fine-tuning

Following the analysis in the zero-shot setting, this section discusses the results obtained from adapting the CLIP model through fine-tuning. The impacts of fine-tuning on the performance of CLIP models pretrained in English are examined, including the model used as the fine-tuning base, as well as a comparison with models already adapted to Portuguese. The evaluation covers both absolute and relative gains in the *Accuracy@k* metric, enabling an analysis of the relationship between result quality and architectural complexity.

4.3.1. Overall Performance of CLIP Models

Figure 3 summarizes the performance of the evaluated models based on the *Accuracy@k* metric, considering both image-to-text and text-to-image retrieval tasks. The CAPIVARA model, trained specifically for Portuguese, achieved the best performance across all evaluated scenarios. The multilingual mCLIP model followed with strong performance, outperforming the original OpenAI models but still trailing behind CAPIVARA. The CLIP-ViT-B/32-FM30k model, which adapts the textual en-

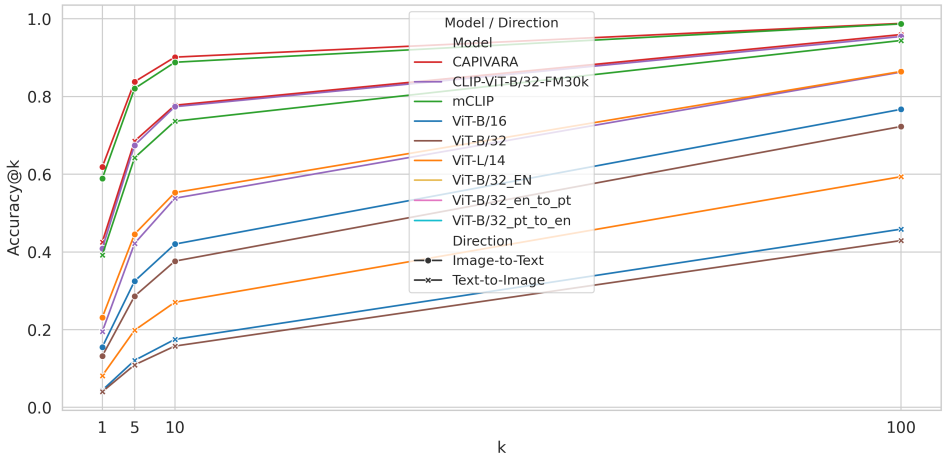

Figure 3: Average *Accuracy@k* across test folds for each model, including the proposed fine-tuning.

coder using original Portuguese captions, consistently outperformed the original OpenAI models as well, although it remained below `mCLIP` (see Tables 2 and 3).

The following subsections present differences in model complexity and analyze the cost-benefit trade-offs in terms of performance and size, highlighting the advantages of the `CLIP-ViT-B/32-FM30k` model.

4.3.2. Impact of Fine-tuning on the CLIP Model

When directly comparing the `ViT-B/32` model with the fine-tuned `CLIP-ViT-B/32-FM30k`, which uses `ViT-B/32` as the fine-tuning starting point, a substantial performance gain is observed (see Tables 4 and 5). In the image-to-text scenario, the *Accuracy@1* metric increased from approximately 13.2% to 40.87%, demonstrating a substantial absolute increase of 27.65 percentage points. This gain equates to a remarkable 209% relative improvement over the unadapted CLIP ViT-B/32, as detailed chronologically in Table 4. These improvements propagate even more aggressively within the text-to-image retrieval scenario. As outlined in Table 5, the model achieved an absolute increase of 15.47 percentage points—leaping from a marginal 4.02% base accuracy to 19.49%. This amounts to an extraordinary 385% relative improvement. The magnitude of these deltas exposes the profound representational inadequacies of the original English text encoder when processing raw Portuguese data, and underscores how robustly a native-language fine-tuning strategy realigns the bidirectional semantic boundaries. This demonstrates that the adaptation enables a more informative representation of Portuguese captions, facilitating cross-modal retrieval.

Another relevant aspect is the alignment between the two retrieval directions after fine-tuning. The `CLIP-ViT-B/32-FM30k` model exhibits growth curves with similar slopes for both image-to-text and text-to-image tasks, a behavior also observed in `CAPIVARA` and `mCLIP`. In contrast, the OpenAI models show significant disparity between directions, with drastically lower performance in text-to-image retrieval. This suggests that fine-tuning not only improved absolute metric values but also reduced semantic imbalance between textual and visual encoders—a critical factor for robust multimodal retrieval applications.

4.3.3. Model Complexity Comparison

An important aspect in the analysis of results is the comparison of the complexity of the models used, particularly regarding the number of parameters and involved architectures. Table 1 summarizes the main characteristics of each evaluated CLIP model.

As detailed in Table 1, the `CAPIVARA` and `mCLIP` models employ textual encoders based on the XLM-RoBERTa-base architecture, with approximately 278M parameters in the textual component. In contrast, our proposed `CLIP-ViT-B/32-FM30k` model, built upon the standard `ViT-B/32` backbone, utilizes the original CLIP textual architecture (OpenAI BPE Transformer) with only 63M parameters. This means our model's textual encoder is approximately **4.4 times smaller** than those used in state-of-the-art adapted models.

Despite this significant difference, the `CLIP-ViT-B/32-FM30k` model delivers competitive performance while being substantially smaller and simpler. This suggests that a lightweight adaptation focused exclusively on the textual encoder and performed with a relatively small, native corpus can be sufficient to better align Portuguese text embeddings with CLIP's multimodal space. This out-

Metric	Compared Model	Accuracy	Accuracy of CLIP-ViT-B/32-FM30k	Absolute Difference	Relative Difference
accuracy@1	ViT-B/32	4.02%	19.49%	**+15.47%**	**+385%**
	ViT-B/16	4.42%	19.49%	**+15.07%**	**+341%**
	ViT-L/14	8.16%	19.49%	**+11.33%**	**+139%**
accuracy@5	ViT-B/32	10.93%	42.19%	**+31.26%**	**+286%**
	ViT-B/16	12.09%	42.19%	**+30.10%**	**+249%**
	ViT-L/14	19.85%	42.19%	**+22.34%**	**+113%**
accuracy@10	ViT-B/32	15.78%	53.83%	**+38.05%**	**+241%**
	ViT-B/16	17.51%	53.83%	**+36.32%**	**+207%**
	ViT-L/14	27.07%	53.83%	**+26.76%**	**+99%**
accuracy@100	ViT-B/32	42.94%	86.30%	**+43.36%**	**+101%**
	ViT-B/16	45.88%	86.30%	**+40.42%**	**+88%**
	ViT-L/14	59.38%	86.30%	**+26.92%**	**+45%**

Table 5: Absolute and Relative Comparison — Text-to-Image Retrieval with OpenAI CLIP Models. Results report the Average Accuracy@k across test folds.

come has practical implications, enabling the use of lighter and more efficient models in resource-constrained environments.

4.3.4. Quantitative Comparison — CLIP Models Adapted for Portuguese

Analyzing the quantitative performance, it is observed that in image retrieval tasks the CLIP-ViT-B/32-FM30k model incurs relative losses ranging from just 3% (top-100) to approximately 34% (top-1), with more pronounced drops at stricter positions. In text retrieval, losses are more significant, varying between 8% and 54%, again with the greatest impact on top-1 and top-5 metrics. Nonetheless, the reduced computational cost of CLIP-ViT-B/32-FM30k, featuring a textual encoder with about 4.4 times fewer parameters, can justify its adoption in resource-constrained scenarios where model lightness is a priority, while still maintaining reasonable performance in multimodal tasks.

4.3.5. Quantitative Comparison — English-Pretrained CLIP Models

On the other hand, when comparing the CLIP-ViT-B/32-FM30k model with the original OpenAI CLIP models (ViT-B/32, ViT-B/16, and ViT-L/14), an inverse scenario is observed, where CLIP-ViT-B/32-FM30k consistently outperforms across all evaluated metrics. In image retrieval tasks, CLIP-ViT-B/32-FM30k surpasses both ViT-B/32 and ViT-B/16 with relative gains ranging from 24% to 209%, depending on the metric. Even compared to ViT-L/14, a significantly larger model, CLIP-ViT-B/32-FM30k maintains superior performance, with advantages between 10% and 77%. This difference becomes even more pronounced in text retrieval tasks: CLIP-ViT-B/32-FM30k achieves relative gains exceeding 100% across nearly all metrics, reaching up to a 385% increase in top-1 accuracy compared to ViT-B/32. These results clearly demonstrate that, despite its architectural simplicity, adapting the textual encoder for the Portuguese language yields highly effective outcomes.

5. Conclusion

This study investigated the performance of CLIP-based models on a native Brazilian Portuguese dataset and proposed a lightweight, cost-effective adaptation strategy. Our experiments lead to several key conclusions. First, off-the-shelf, English-centric CLIP models exhibit not only significant performance degradation but also a critical retrieval asymmetry when applied to native Portuguese, which we identified as a clear symptom of a deficient text encoder for the target language. Second, models adapted for Portuguese demonstrate vastly superior performance, confirming that linguistic adaptation is essential.

The core contribution of this work is the demonstration that a lightweight and resource-efficient fine-tuning strategy—updating only the text encoder of a standard CLIP model on a modest, native dataset—can yield a model that is not only vastly superior to its original baseline but also highly competitive with much larger and more complex multilingual architectures. Our model, CLIP-ViT-B/32-FM30k, achieves this with a text encoder that is 4.4 times smaller, presenting a viable and cost-effective path forward. Furthermore, this adaptation successfully resolved the performance asymmetry, creating a more balanced and robust bidirectional semantic space.

The primary takeaway is that lightweight, text-only fine-tuning on high-quality, native-language data is a powerful, accessible, and resource-efficient strategy for adapting large multimodal models. This approach provides a practical roadmap for researchers in other low-resource language communities to leverage the capabilities of foundational models like CLIP. Future work should focus on creating larger native datasets for Brazilian Portuguese and exploring the co-adaptation of the visual encoder to capture culturally specific concepts.

Limitations

A central limitation of this investigation is its dependence on a relatively constrained corpus scale. Although the FM30k dataset provides highly accurate, human-authored Portuguese annotations, its sheer volume is dwarfed by the billion-scale synthetic datasets deployed in the pre-training of foundational multilingual encoders. Consequently, the upper ceiling of performance for our natively adapted model may remain constrained relative to large-scale architectures. Furthermore, while the locked-image tuning protocol ensures remarkable computational efficiency, the unadapted visual encoder fundamentally relies on its zero-shot generalization capabilities. This frozen visual space may possess implicit biases towards Euro-centric or North American visual concepts, potentially hindering its capacity to flawlessly capture culturally distinct visual nuances specific to the Brazilian geographical and societal context. Future endeavors must explore efficient co-adaptation pipelines where both visual and textual encoders can be mildly adjusted via native corpora without triggering catastrophic forgetting.

Ethics Statement

The research conducted aligns thoroughly with scholarly ethics guidelines. The dataset employed (FM30k) is an academic derivation of the public Flickr30k corpus, strictly utilizing human-provided annotations compiled through ethical frameworks and voluntary participation by university contributors. The authors affirm that the fine-tuning procedures did not introduce harmful objectives, and the resultant models are intended strictly to enhance accessibility and processing equity for Brazilian Portuguese in multimodal contexts. By establishing cost-efficient adaptation frameworks, this research directly contributes to democratizing advanced AI methodologies for low-resource environments, aiming to mitigate linguistic hegemony in the vision-language domain.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also acknowledge the financial support from the Brazilian funding agencies CNPq and FAPERGS, and Petrobras. Some experiments in this work used the PCAD infrastructure (http://gppd-hpc.inf.ufrgs.br) at INF/UFRGS. Parts of this manuscript were written with the support of a generative AI tool (ChatGPT); all content was reviewed and validated by the authors.

6. Bibliographical References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language.

William Alberto Cruz-Castañeda, Marcellus Amadeus, André Felipe Zanella, and Felipe Rodrigues Perche Mahlow. 2025. From pampas to pixels: Fine-tuning diffusion models for gaúcho heritage. *Journal of the Brazilian Computer Society*, 31(1):262–270.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: a large-scale hierarchical image database. pages 248–255.

Gabriel Oliveira dos Santos, Diego Alysson Braga Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. In *Proceedings of the 3rd Workshop on Multilingual Representation Learning (MRL)*, pages 184–207, Singapore. Association for Computational Linguistics.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2022. pracegover: A large dataset for image captioning in portuguese. *Data*, 7(2).

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Byungsoo Ko and Geonmo Gu. 2022. Large-scale bilingual language-image contrastive learning.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

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

Gabriel Rosa, Lucas Bonifacio, Leandro de Souza, Rodrigo Lotufo, Rodrigo Nogueira, and John Melville. 2021. A cost-benefit analysis of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.

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 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Matheus Vinícius Todescato and Joel Luís Carbonera. 2024. Investigating performance patterns of pre-trained models for feature extraction in image classification. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1024–1031.

Marcelo Viridiano, Arthur Lorenzi, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Natália Sathler Sigiliano, Maucha Gamonal, Helen de Andrade Abreu, Lívia Vicente Dutra, Mairon Samagaio, Mariane Carvalho, Franciany Campos, Gabrielly Azalim, Bruna Mazzei, Mateus Fonseca de Oliveira, Ana Carolina Luz, Livia Padua Ruiz, Júlia Bellei, Amanda Pestana, Josiane Costa, Iasmin Rabelo, Anna Beatriz Silva, Raquel Roza, Mariana Souza Mota, Igor Oliveira, and Márcio Henrique Pelegrino de Freitas. 2024. Framed Multi30K: A frame-based multimodal-multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7438–7449, Torino, Italia. ELRA and ICCL.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2023. Chinese clip: Contrastive vision-language pretraining in chinese.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Xiaohua Zhai, Alexander Kolesnikov, Basil Mustafa, Lucas Beyer, Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Jakob Uszkoreit, Mario Lucic, and Neil Houlsby. 2022. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*.