Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5857

Brazilian Portuguese Image Captioning with Transformers: A Study on Cross-Native-Translated Dataset

Gabriel Bromonschenkel [**Instituto Federal do Espírito Santo (IFES), Serra, Brazil** | ***gabriel.mota.b.lima@gmail.com***]
Alessandro L. Koerich [**École de Technologie Supérieure (ÉTS), Montreal, Canada** | ***alessandro.koerich@etsmtl.ca***]
Thiago M. Paixão [**Instituto Federal do Espírito Santo (IFES), Serra, Brazil** | ***thiago.paixao@ifes.edu.br***]
Hilário Tomaz Alves de Oliveira [**Instituto Federal do Espírito Santo (IFES), Serra, Brazil** | ***hilario.oliveira@ifes.edu.br***]

Programa de Pós-graduação em Computação Aplicada (PPComp), Instituto Federal do Espírito Santo, Av. dos Sabiás, 330, Morada de Laranjeiras, Serra, ES, 29166-630, Brazil.

Received: 07 April 2025 • **Accepted:** 04 November 2025 • **Published:** 15 April 2026

Abstract Image captioning (IC) refers to the automatic generation of natural language descriptions for images, with applications ranging from social media content generation to assisting individuals with visual impairments. While most research has been focused on English-based models, low-resource languages such as Brazilian Portuguese face significant challenges due to the lack of specialized datasets and models. Several studies create datasets by automatically translating existing ones to mitigate resource scarcity. This work addresses this gap by proposing a cross-native-translated evaluation of Transformer-based vision and language models for Brazilian Portuguese IC. We use a version of Flickr30K comprised of captions manually created by native Brazilian Portuguese speakers and compare it to a version with captions automatically translated from English to Portuguese. The experiments include a cross-context approach, where models trained on one dataset are tested on the other to assess the translation impact. Additionally, we incorporate attention maps for model inference interpretation and use the CLIP-Score metric to evaluate the image-description alignment. Our findings show that Swin-DistilBERTimbau consistently outperforms other models, demonstrating strong generalization across datasets. ViTucano, a Brazilian Portuguese pre-trained VLM, surpasses larger multilingual models (GPT-4o, LLaMa 3.2 Vision) in traditional text-based evaluation metrics, while GPT-4 models achieve the highest CLIP-Score, highlighting improved image-text alignment. Attention analysis reveals systematic biases, including gender misclassification, object enumeration errors, and spatial inconsistencies.

Keywords: Image Captioning, Transformers, Brazilian Portuguese, Vision Encoder-Decoder, Multi-Modal Evaluation, Attention Maps, CLIP-Score, Vision-Language Models

1 Introduction

Image Captioning (IC) bridges the gap between vision and language in computational applications, enabling machines to generate coherent and meaningful descriptions of visual content. This interdisciplinary task holds potential for applications such as enhancing visual accessibility, aiding medical prescriptions and diagnoses, facilitating image-text indexing for information retrieval, and advancing human-computer interaction [Ghandi *et al*., 2023; Stefanini *et al*., 2022; Sharma and Padha, 2023].

Advances in IC models have been driven by deep learning approaches, notably Convolutional Neural Networks (CNNs) and Transformers [Stefanini *et al*., 2022; Sharma and Padha, 2023]. Despite significant progress in well-resourced (datasets and models) languages like English, those with limited resources, such as Brazilian Portuguese, remain underexplored. Captions generated in English and then translated into other languages often lack the accuracy and contextual relevance achieved by models trained directly on native datasets in the target language [dos Santos *et al*., 2022; Gondim *et al*., 2022].

The development of robust IC models for low-resource languages depends significantly on the availability and quality of both translated and native datasets. Translated datasets, such as the adapted Brazilian Portuguese version of Flickr30K, enable researchers to leverage large-scale datasets from high-resource languages for cross-lingual experimentation [Gondim *et al*., 2022; Bromonschenkel *et al*., 2024; dos Santos *et al*., 2023]. However, translations often fail to capture cultural and linguistic nuances, potentially leading to suboptimal model performance. In contrast, native datasets, such as #PraCegoVer-63K, authentically capture linguistic and contextual nuances, ensuring models better align with the specificities of the target language [dos Santos *et al*., 2022].

A preliminary study conducted by our group [Bromonschenkel *et al*., 2024] investigated Transformer-based Vision Encoder-Decoder (VED) models for IC, considering both translated and native datasets. To the best of our knowledge, it was the first comprehensive evaluation of various VED configurations for Brazilian Portuguese IC. However, the investigation was limited to a single-database evaluation, as-

sessing performance on either translated or native datasets. Additionally, the quantitative analysis did not incorporate cross-modal image-text metrics (e.g., CLIP-Score), and the qualitative assessment of generated captions lacked visual attention maps, which could improve model interpretability.

This study introduces a comprehensive analysis of the effects of dataset translation quality on image captioning performance by comparing an automatically translated version of Flickr30K with a manually annotated counterpart. The research provides valuable insights into the interplay between linguistic precision and machine learning outcomes by examining how translation quality influences model performance and caption accuracy. Additionally, we conduct extensive quantitative and qualitative evaluations to deepen our understanding of model behavior. A key aspect of this work is the application of the CLIP-Score metric, which facilitates interpretable assessments of alignment between generated captions and reference images, surpassing traditional reference-based metrics in accessibility. To further enhance interpretability, attention visualization is employed to reveal the image regions the model prioritizes during caption generation. This study builds upon our earlier investigation [Bromonschenkel *et al*., 2024], expanding its scope and depth to offer a finer perspective on these critical aspects.

This research aims to address the depicted issues in Brazilian Portuguese image captioning by presenting:

- A comprehensive evaluation of Transformer-based vision encoder-decoder models trained and evaluated on native and translated versions of the Flickr30K dataset.
- A quantitative analysis of smaller, fine-tuned VED models compared to both open-source and closed-source large pre-trained Visual Language Models (VLMs).
- A cross-source evaluation setup where models trained on machine-translated datasets are evaluated on native human-generated datasets and vice versa. Such analysis enables quantifying the impact of translation on model performance.
- The adoption of CLIP-Score within the CAPIVARA framework for cross-modal evaluation in Brazilian Portuguese, comparing its efficacy against traditional metrics such as BLEU, ROUGE, METEOR, and CIDEr.
- The integration of attention maps to visually interpret model decision-making in the Brazilian Portuguese domain, providing insights into the relationship between generated captions and corresponding image regions.

The remaining content is organized into five sections. Section 2 presents the related work containing the historical studies focused on multimodal datasets and models for Brazilian Portuguese IC. Section 3 shows the main architectural approach of this work, the vision encoder-decoder models. Section 4 explains the datasets, performance metrics, and experimental setup. Section 5 presents the results obtained from experiments and explains them more deeply with qualitative analyses. Finally, Section 6 synthesizes the results and findings of our work and suggests new avenues for further research.

2 Related Work

Early deep learning approaches relied predominantly on CNNs for visual feature extraction and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, for sequence generation [Stefanini *et al*., 2022; Sharma and Padha, 2023]. For example, Vinyals *et al*. [2015] proposed a foundational framework combining visual and language components, which paved the way for integrating attention mechanisms and Transformer-based architectures. The introduction of Transformer-based models marked a paradigm shift in IC. Self-attention mechanisms enabled these models to effectively capture long-range dependencies between words and image features. Models such as Vision Transformer (ViT) [Dosovitskiy *et al*., 2020] and Swin Transformer [Liu *et al*., 2021b] achieved state-of-the-art performance on a variety of vision and multi-modal tasks. These models use image patches as inputs, analogous to tokens in Natural Language Processing (NLP) tasks; thereby, they can process smaller parts of the global information to extract relations between visual elements.

In Brazilian Portuguese, Gondim *et al*. [2022] evaluated a CNN-RNN with attention where the CNN block is composed of an EfficientNet of the B7 family. Their work has two crucial stages: dataset translation and a two-step performance rating, in which automatic assessment and human evaluation are conducted. In the first stage, the Flickr8K dataset is translated to Portuguese using LibreTranslate. For the evaluation stage, performance is assessed using BLEU and METEOR, followed by an evaluation with 34 human annotators on 100 captions generated by the best model. Their study demonstrated that while translations provide a practical way to adapt high-resource datasets, they often fail to capture linguistic and contextual nuances, which limits the quality of generated captions. Due to the Flickr8K dataset size, the work of Gondim *et al*. [2022] is constrained to a smaller quantity of information and context compared to larger datasets, such as the Flickr record with 30,000 examples.

de Alencar *et al*. [2024] explored the Grid-and-Region-based Image Captioning Transformer model on a Portuguese-translated version of the MSCOCO dataset. This fully Transformer-based architecture demonstrated strong performance in adapting advanced IC methodologies to Brazilian Portuguese. However, like other works using translated datasets, it faced limitations in representing the cultural and linguistic specificity of the target language.

To address the scarcity of resources in Brazilian Portuguese, dos Santos *et al*. [2022] introduced #PraCegoVer, the first large-scale dataset for image captioning in such a language. The dataset, sourced from Instagram posts tagged with the hashtag #PraCegoVer, features user-provided captions intended to promote accessibility for visually impaired individuals. This dataset highlights the challenges of captioning in Brazilian Portuguese due to its cultural and linguistic diversity. The authors trained an Attention-on-Attention Network (AoANet) on the #PraCegoVer dataset, presenting it as the

Table 1. Comparison of Brazilian Portuguese Image Captioning Works.

Work	Native Dataset	Translated Dataset	Transformer-based Models	Analyze Translation Impact	Cross-Modal Evaluation	Attention Maps Investigation
dos Santos *et al*. [2022]	✓					
Gondim *et al*. [2022]		✓				
de Alencar *et al*. [2024]		✓	✓			
Our Work	✓	✓	✓	✓	✓	✓

first model for image captioning trained on a native dataset for Brazilian Portuguese.

Recent studies have shown the effectiveness of the Vision Encoder-Decoder (VED) architecture in more specialized domains, including medical image captioning and multilingual captioning. For instance, Jnaini *et al*. [2024] explored the synergy between GPT-3 summarization and VED models for generating accurate descriptions of chest X-ray images. Their results indicated that the VED architecture effectively captures visual features and aligns them with textual descriptions, leading to improved captioning performance. Similarly, Abdelaal *et al*. [2024] investigated the use of VED models for generic image captioning tasks, demonstrating that Transformer-based vision-language models outperform traditional CNN-RNN architectures by leveraging self-attention mechanisms for better feature extraction and text generation. Furthermore, Ishan *et al*. [2023] applied a VED model to Bengali image captioning, confirming its adaptability to low-resource languages. Their work highlighted that pre-trained vision encoders and language decoders enhance caption fluency and contextual accuracy, even in languages with limited labeled datasets.

Our work differs from previous research for Brazilian Portuguese IC [Gondim *et al*., 2022; de Alencar *et al*., 2024; dos Santos *et al*., 2022] in several aspects, as shown in Table 1. Unlike prior studies, which used either translated or native datasets, we employ both in Brazilian Portuguese under Flickr30K. Furthermore, we evaluate multiple VED models, rather than focusing on a single architecture. Finally, we compare the performance of large, pre-trained VLMs without fine-tuning against our fine-tuned VED models to better understand their respective strengths and trade-offs.

Additionally, we incorporate CLIP-Score, computed using the CAPIVARA framework, for reference-free evaluation, in contrast to previous work, which relied only on reference-based metrics or conducted reviews with human annotators on a limited volume of data. For qualitative analysis, we use attention maps to better understand which focal points the model tends to use. By integrating these elements, our study provides a more comprehensive and systematic evaluation of image captioning in Brazilian Portuguese than previous approaches.

3 Transformer-based Image Captioning

Typically, IC models benefit from an encoder-decoder architecture, where the encoder processes the image to an embedding space in the format of the neural network's hidden states. The decoder converts these hidden states into natural language descriptions. The vision encoder-decoder models adopted in this work leverage pre-trained checkpoints, such as the Bidirectional Encoder Representations from Transformers (BERT) [Stefanini *et al*., 2022; Ghandi *et al*., 2023; Sharma and Padha, 2023].

3.1 Visual Encoders

We employ a Transformer-based Vision Encoder-Decoder architecture, combining visual encoders for processing image content and language decoders for generating the captions.

The three visual encoders in our Vision Encoder-Decoder configurations are state-of-the-art Transformer-based architectures for vision, which are pre-trained on ImageNet-1K at 224px resolution. The first is Vision Transformer (ViT), a pioneering transformer-based architecture for computer vision tasks. Unlike convolutional networks, ViT processes images by dividing them into fixed-size patches (e.g., 16x16 pixels), which are flattened and treated as input tokens. Each patch is linearly embedded and augmented with positional encodings to retain spatial information. These tokens are then passed through multiple self-attention layers to model global dependencies between image regions. Our work explores the ViT base version with 12 layers and 87 million parameters [Dosovitskiy *et al*., 2020].

The second vision encoder assessed is the Shifted Window (Swin) Transformer, a model with four patch processing stages. In the first stage, 4px resolution patches are linearly embedded, followed by progressive merging in subsequent stages until the resolution reaches 32px. In this way, Swin Transformer builds on the strengths of ViT by introducing a hierarchical representation with shifted windows. Images are divided into non-overlapping windows, and self-attention is applied within each window, significantly reducing computational cost compared to global attention approaches, such as ViT. The shifted window mechanism allows for cross-window interaction and captures both local and global image features. This investigation uses the Swin Transformer base version with 24 layers distributed through 4 blocks separated by patch embedding layers, containing 88 million parameters [Liu *et al*., 2021b].

The third encoder investigated is the Data-Efficient Image Transformer (DeiT), a variant of ViT designed to achieve competitive performance with fewer data and computational resources. It incorporates a teacher-student knowledge distillation approach, in which a convolutional neural network acts as a teacher to improve the training efficiency of DeiT. Additionally, it uses a classification token to aggregate global information for image understanding. This study adopts the DeiT base version with 12 layers and 86 million parameters [Touvron *et al*., 2021].

3.2 Language Models

The three language decoders assessed in our VED configurations are models designed for text generation and language processing in Portuguese. The first is BERTimbau, a BERT-based encoder model pre-trained on the Brazilian Web as Corpus (BrWaC) dataset. It uses a Masked Language Modeling (MLM) objective during pre-training, where a chunk of the input tokens is masked, and the model learns to predict them based on context. Our research examines the base version with 12 layers and 110 million parameters [Souza *et al*., 2020].

The second is DistilBERTimbau, a distilled version of BERTimbau-based, designed to reduce model size and inference latency while retaining most of its performance. The distillation process involves training the smaller model to mimic the outputs of a larger BERTimbau model. This results in a lightweight model with a fraction of the computational requirements of BERTimbau. For this work, we adopt the base version with 6 layers and 66 million parameters [Silva Barbon and Akabane, 2022; Adalberto Ferreira Barbosa Junior, 2024].

GPorTuguese-2 is the third model evaluated. It is a Generative Pre-trained Transformer (GPT)-based model derived from a fine-tuned small version of GPT-2 on Portuguese Wikipedia. The model employs a decoder-only transformer architecture with causal attention, enabling autoregressive text generation. GPorTuguese-2 consists of 12 layers and 137 million parameters [Guillou, 2020].

To visually introduce the vision encoder-decoder design, Figure 1 shows one of the nine architectural combinations of our work, using the model resulting from the merger between Swin Transformer and DistilBERTimbau, the coupling with the best performance in Table 3. The visual encoders generate feature embeddings that capture the spatial and semantic information of the image. These embeddings are then projected into 768-dimensional tensors for the language decoders, aligning them with textual representations to produce captions. The alignment process is boosted by cross-attention, positional encodings, and fine-tuning. Through cross-attention, the decoders focus on specific regions of the image embeddings, ensuring that the generated captions are contextually grounded in the visual input. Both encoders and decoders use positional encodings to preserve spatial and sequential information, which is crucial for modeling the relationships between objects in images and words in captions. The models are fully fine-tuned on either native or translated datasets, ensuring they adapt to the linguistic and contextual nuances of Brazilian Portuguese.

4 Experimental Methodology

This section presents the datasets, performance metrics, and experiments. First, the datasets are introduced, considering both translated and native contexts, while highlighting nuances in data curation and dataset characteristics. Next, we briefly explain the performance evaluation metrics. In the following, we describe the experiments for both quantitative and qualitative evaluation, concluding with the implementation details.

4.1 Datasets

This study uses two datasets derived from the Flickr30K dataset, each providing a different source of image-caption pairs in Brazilian Portuguese. One dataset consists of human-generated captions, while the other contains machine-translated captions from English.

The first dataset, Flickr30K in Native Portuguese (Flickr-Native), is based on the FM30K corpus [Viridiano *et al*., 2024], which in turn is an extension of the Flickr30K dataset, embodying more examples. It contains 31,014 images, each associated with five captions written by native Portuguese speakers. The dataset maintains the original image-caption pairs from Flickr30K but includes manually created Portuguese descriptions. These captions follow the linguistic structures of the language without translation artifacts. The FM30K is designed to be a multi-modal and multilingual dataset. It expands Multi30K (a German extension of Flickr30K) with Brazilian Portuguese descriptions and frame resources.

The second dataset, Flickr30K Translated to Portuguese (Flickr-Translated), consists of the same 31,014 images, but the original captions were translated from English using Google Translate[1]. This dataset allows an evaluation of the effects of translation on image captioning models. Automatic translations can introduce errors such as unnatural phrasing and altered word order, which may affect model performance.

Table 2 summarizes statistics for both native and translated datasets. The statistical information is generated using the SpaCy toolkit[2], configured to ignore punctuation. On average, captions in Flickr-Native are longer than those in Flickr-Translated, suggesting that manually created descriptions tend to be more detailed. The toolkit identifies 20,719 distinct words in Flickr-Translated, of which 9,412 are absent in the native source. In contrast, Flickr-Native contains 18,990 distinct words, of which 7,683 are not found in the translated source.

4.2 Performance Metrics

To assess the generated captions, we incorporated CLIP-Score in addition to the reference-based (uni-modal) metrics used in our preliminary study. CLIP-Score is a model-based multi-modal metric that quantifies the similarity between image and text content within a shared embedding space. Therefore, the metrics computed in our experiments are:

[1] `https://translate.google.com/`
[2] `https://spacy.io/`

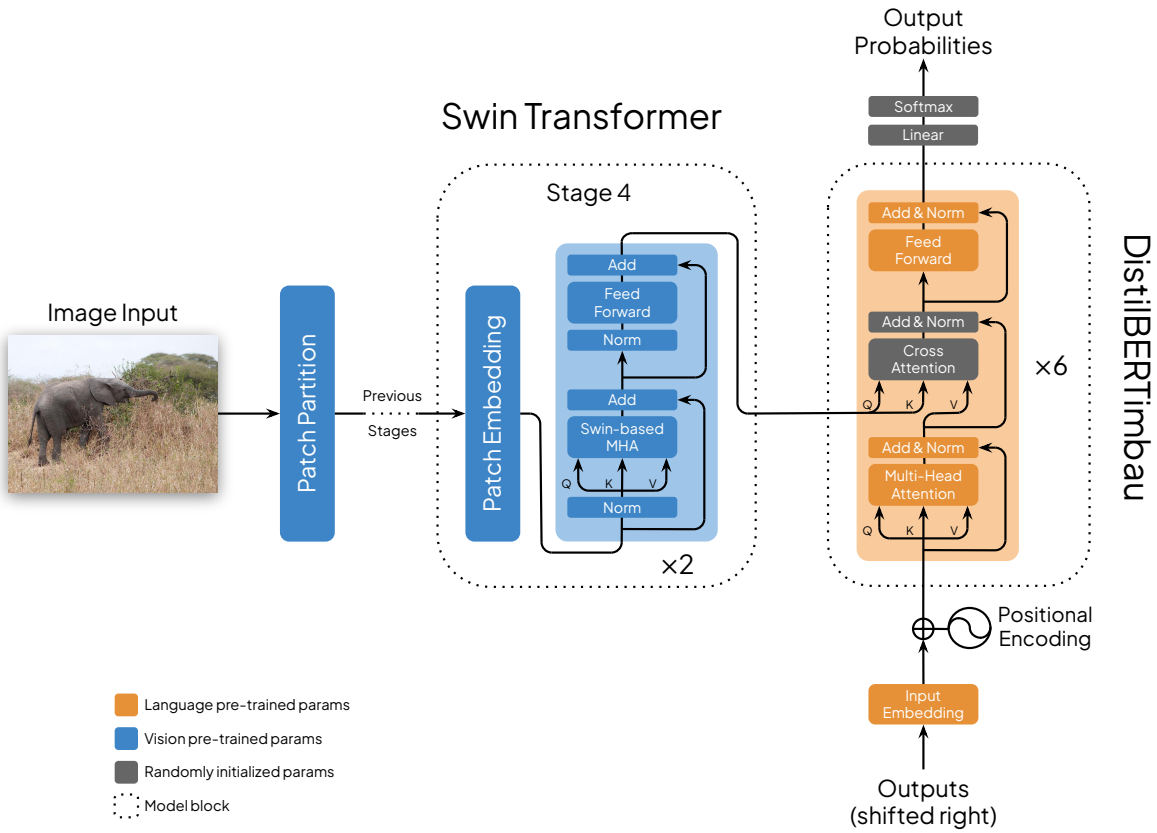

Figure 1. Illustration of the architecture on the merge of Swin Transformer and DistilBERTimbau. Swin-DistilBERTimbau is one of the nine VED combinations used in this work. MHA stands for Multi-Head Attention.

Table 2. Descriptive statistics of the datasets splits. Val. is an abbreviation for Validation.

		Avg. Caption Length (#Words)	
Split	**#Examples**	**Flickr-Translated**	**Flickr-Native**
Train	29,000	12.1 ± 5.1	13.4 ± 5.4
Val.	1,014	12.3 ± 5.3	13.5 ± 5.5
Test	1,000	12.2 ± 5.4	13.4 ± 5.4
Total	31,014	12.1 ± 5.2	13.4 ± 5.4

- **CIDEr-D**: A consensus-based metric that weights n-grams based on their Term Frequency-Inverse Document Frequency (TF-IDF) scores across reference captions. It is specifically developed for description evaluation in the image captioning task [Vedantam *et al.*, 2015].

- **BLEU-4**: A precision-based metric that evaluates n-gram overlaps up to 4-grams between generated and reference captions. It was initially designed to measure the text quality in automatic machine translation tasks [Papineni *et al.*, 2002].

- **ROUGE-L**: A metric based on the longest common subsequence between generated and reference captions. It is part of the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) group of measures initially drawn to quantify text summarization quality [Lin, 2004].

- **METEOR**: The Metric for Evaluation of Translation with Explicit Ordering (METEOR) is a recall-oriented metric that considers synonyms and word stemming for improved linguistic evaluation. It is typically used for text translation assessment [Banerjee and Lavie, 2005].

- **BERTScore**: A semantic similarity metric using contextual embeddings generated by a BERT-based model. We use the F1-score derived from BERTScore, using the BERTimbau as the embedder for Portuguese captions [Zhang *et al.*, 2019].

- **CLIP-Score**: A cross-modal metric that measures the similarity between image embeddings and text embeddings using a model based on Contrastive Language-Image Pre-training (CLIP) [Hessel *et al.*, 2021]. We leverage the CAPIVARA framework [dos Santos *et al.*, 2023] to adapt CLIP-Score's performance for Brazilian Portuguese.

Both the reference-based (uni-modal) metrics and CLIP-Score have the potential to quantify the accuracy of generated captions. While CLIP-Score captures terms and points of view not depicted in the reference captions, the reference-based metrics maintain the fidelity of the analysis anchored in the reference captions. On the other hand, as reported by CLIP-Score authors, the multi-modal evaluation is attached to the model pre-trained bias. The bias problem can insert erroneous judgment into the candidate caption assessment. For conciseness, BLEU-4, ROUGE-L, METEOR, BERTScore,

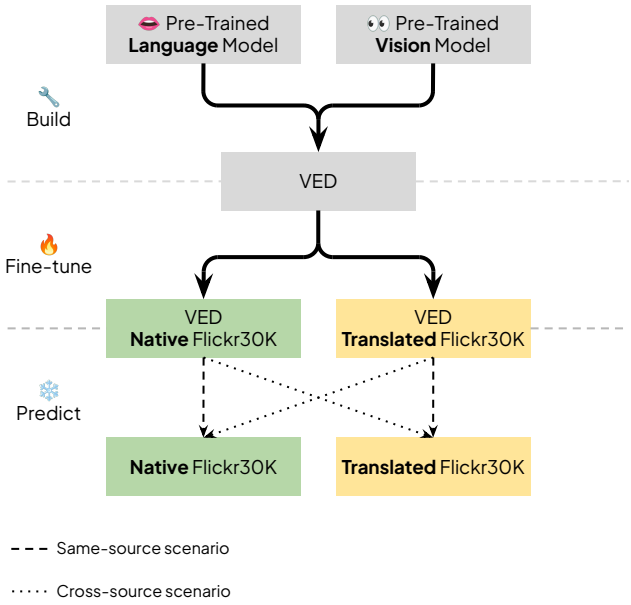

Figure 2. Simplified illustration of the experiment pipeline.

and CLIP-Score are abbreviated as B@4, RL, M, BS, and CS, respectively. The BLEU, ROUGE, METEOR, BERTScore, and CLIP-Score metrics range from 0 to 1, while CIDEr-D can exceed this range.

4.3 Experiments

VED Models Assessment. The first experiment aims to comprehensively evaluate the VED models in both same-source and cross-source setups, as depicted in Figure 2. In the same-source setup, the goal is to establish the conventional performance of the models by training and testing within the same setting (Native-to-Native and Translated-to-Translated). Cross-source evaluation assesses context shifting by considering Native-to-Translated and Translated-to-Native scenarios, allowing for an analysis of how well models generalize across human-generated and machine-translated captions. This assessment is particularly useful for quantifying the source-drift impact of training models on the translated dataset and applying them to a native dataset (Translated-to-Native), as it reduces the human effort required to produce large-scale training data in the native language.

VLMs Assessment. For a more comprehensive analysis, state-of-the-art Vision-Language Models (VLMs) were evaluated in a zero-shot setting to establish baseline performance. The fine-tuned VED models were compared with the following pre-trained VLMs (not fine-tuned on the datasets used in this study): the open-source ViTucano 1B and 2B [Corrêa *et al*., 2024], PaliGemma [Beyer *et al*., 2024], Phi-3 Vision (Phi-3 V) [Abdin *et al*., 2024], and LLaMa3.2 Vision 11B (LLaMa3.2 V) [Dubey *et al*., 2024], alongside the proprietary models GPT-4o and GPT-4o-mini [Hurst *et al*., 2024]. The ViTucano models are natively pre-trained for Brazilian Portuguese, while the others are multilingual. Caption generation followed the prompt: "Escreva uma descrição em português do Brasil para a imagem com no máximo 25 palavras." ("Write a description in Brazilian Portuguese for the image with a maximum of 25 words."), except for PaliGemma, which used the author-reported prompt "caption pt". Performance metrics were computed by comparing generated captions with reference captions from both Flickr30K Translated and Flickr30K Native, using the same 1,000-image test split. PaliGemma, Phi-3V, ViTucano 1B and 2B are used through Transformers library[3], while LLaMa3.2 V is utilized via Unsloth library[4]. PaliGemma, Phi-3 V, and LLaMa3.2 V are compressed with 4-bit quantization to fit within the available resources. The GPT-4o models are accessed via the OpenAI API[5].

Qualitative Assessment. Swin-DistilBERTimbau, the model with the highest performance across most metrics in the same-source setup, was selected for a more in-depth qualitative assessment. For this analysis, we selected examples from both Native-to-Native and Translated-to-Translated scenarios, focusing on cases with (comparatively) high performance on CLIP-Scores in view of the remaining metrics and vice versa. Additionally, we analyze the quality of the captions by examining the attention maps for the Swin-DistilBERTimbau.

4.4 Implementation Details

Training the Models. All VED models are trained for 20 epochs on an NVIDIA RTX 4090 GPU, 24GB VRAM. To improve caption generation, we use the Adam optimizer with a learning rate of 5e-5, a batch size of 16, and a beam search decoding strategy with a beam size of 5. Pre-trained checkpoints for both vision encoders (ViT, Swin, DeiT) and language decoders (BERTimbau, DistilBERTimbau, GPorTuguese-2) are leveraged for transfer learning, ensuring stable convergence and improved performance.

Attention Maps. The implementation of the attention maps is inspired on Liu *et al*. [2021a][6]. The cross-attention layer is the architectural element that associates vision and language modalities, while the other multi-head layers are self-attentive, implementing uni-modal attention. Similarly, Liu *et al*. [2021a] uses the token-patches attention abstracted in the cross-attention layer. In our study, we use the cross-attention layer of the last decoder block, which is the output block directly generating the captions. Different VED architectures have peculiarities in the internal attention states that should be regarded when computing attention maps. For instance, the Swin Transformer (the encoder in Swin-DistilBERTimbau) employs a strategy of increasing patch size through image block processing, starting with a patch size of 4 and ending with a patch size of 32 (i.e., from a grid of 56×56 to 7×7 patches). In the cross-attention layer, this architectural structure results in tensors of attention's alphas from 12 attention heads of DistilBERTimbau and 7×7 patches of Swin Transformer.

[3] `https://huggingface.co/transformers`
[4] `https://unsloth.ai/`
[5] `https://platform.openai.com/docs/overview`
[6] A similar implementation can be found in the TensorFlow tutorial at `https://www.tensorflow.org/text/tutorials/image_captioning#attention_plots`

Table 3. VED models' performance (%) for the same-source setup: Translated-to-Translated and Native-to-Native scenarios. The three highest values for each metric are highlighted in **bold**, with the cell corresponding to the highest value further highlighted in **blue**. Metrics: CIDEr-D (C), BLEU-4 (B@4), ROUGE-L (RL), METEOR (M), BERTScore (BS), and CLIP-Score (CS).

		Translated-to-Translated						Native-to-Native					
Encoder	**Decoder**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**
$DeiT_{BASE}$	$BERT_{BASE}$	49.53	19.20	36.00	39.80	69.58	49.75	47.84	17.01	34.01	39.16	69.71	50.66
	$DistilBERT_{BASE}$	50.58	19.24	35.77	39.93	69.50	49.67	50.55	18.08	34.92	40.86	70.10	49.95
	$GPT\text{-}2_{SMALL}$	50.61	19.83	36.30	40.52	69.66	49.49	48.79	21.88	**44.92**	40.00	69.03	50.51
$Swin_{BASE}$	$BERT_{BASE}$	**62.42**	**22.78**	**38.71**	**43.47**	**71.19**	**53.81**	**61.14**	20.58	37.62	**43.85**	**72.01**	**53.01**
	$DistilBERT_{BASE}$	**66.73**	**24.65**	**39.98**	**44.71**	**72.30**	**53.26**	**63.77**	**22.79**	38.06	**44.65**	**72.23**	**53.26**
	$GPT\text{-}2_{SMALL}$	**64.71**	**23.15**	**39.39**	**44.36**	**71.70**	**53.49**	**65.79**	**29.17**	**50.23**	**45.04**	**72.06**	**53.19**
ViT_{BASE}	$BERT_{BASE}$	57.32	22.12	37.50	41.72	70.63	51.93	53.66	18.15	34.87	41.38	70.40	52.84
	$DistilBERT_{BASE}$	59.32	21.19	37.74	42.70	71.15	51.84	56.97	19.49	36.36	43.08	71.38	51.86
	$GPT\text{-}2_{SMALL}$	59.02	21.39	37.68	42.64	71.03	52.44	54.17	**24.74**	**47.57**	42.59	70.53	51.47

5 Results and Discussion

This section presents and discusses the results of the conducted experiments.

5.1 VED Models Assessment

5.1.1 Same-source Setup

Table 3 shows the VED model's performance for the same-source setup on both translated and native versions of Flickr30K. In both cases, we observe that Swin-based encoders systematically outperform DeiT and ViT encoders in most metrics. Besides the configurations using Swin as the encoder, the only models among the three best are the GPT-based models for the native context.

For Flickr30K Translated, Swin-DistilBERTimbau achieved the highest performance in uni-modal metrics, while Swin-BERTimbau yielded a slightly higher CLIP-Score (53.81). On Flickr30K Native, Swin-GPorTuguese-2 dominates CIDEr, ROUGE, BLEU, and METEOR, outperforming the other encoder-decoder combinations. On the other hand, Swin-DistilBERTimbau achieved the highest values in model-based metrics, i.e., BERTScore and CLIP-Score. Model-based metrics (i.e., BERTScore, CLIP-Score) provide a better image and language context assessment of the generated captions compared to relying solely on lexical matching. Thus, these metric results indicate that, in the native context, Swin-GPorTuguese-2 achieved a better lexical match with the reference captions, while Swin-DistilBERTimbau achieved a better contextual match.

Moreover, we noticed that the models generally performed better on machine-translated data (compared to native text) for most lexical metrics. The only performance peculiarity is revealed by GPT-based models in the Native-to-Native scenario, where they achieved higher ROUGE-L and BLEU-4 scores than their equivalents in the Translated-to-Translated scenario. This gap underscores that the Flickr-Native is more challenging than Flickr-Translated in the same-dataset evaluation situation. Nonetheless, the best-performing VED configurations (Swin + DistilBERTimbau/GPorTuguese-2) showed less impact across translated and native scenarios, indicating that robust visual features and carefully pre-trained Portuguese decoders are critical to achieving accurate and fluent captions.

5.1.2 Cross-source Setup

In the cross-source setup, the models are evaluated on their ability to generalize across native and machine-translated data. Table 4 shows the results for the two scenarios under the cross-source setup: (*i*) Translated-to-Native (left side), where the model trained on the translated context is applied to the native context, and (*ii*) Native-to-Translated (right side), the counterpart.

The observed performance drop compared to the same-source setup highlights the challenges of domain shift. Nonetheless, Swin-based models achieved the highest evaluation metrics, with Swin-DistilBERTimbau assuming leadership performance across most metrics. ViT-DistilBERTimbau is a special case that achieved comparatively relevant metric estimation in ROUGE, METEOR, and BERTScore. It is noteworthy to mention that the two contexts have an information drift due to the terms and expressions that are exclusive of each dataset, as reported in Section 4.1.

Translated-to-Native vs. Native-to-Native Figure 3 depicts the mean performance drop of the models tested in Flickr-Native when the training source changes from native to translated. In this circumstance, GPorTuguese-2 models suffer greater deterioration in performance than the remaining models when the training base changes from native to translated, with declines ranging from -8.84% to -6.09%, while the other models show declines ranging from -4.66% to -2.75%.

Comparing the metrics in Tables 3 and 4, CIDEr presented the most aggressive performance drop compared to the remaining metrics. At the same time, CLIP-Score presented no performance drop for Swin-based models and ViT-GPorTuguese-2, indicating that the quality of image-text embedding alignment is comparable to the captions generated

Table 4. VED models' performance (%) for the cross-source setup: Translated-to-Native and Native-to-Translated scenarios. The three highest values for each metric are highlighted in **bold**, with the cell corresponding to the highest value further highlighted in **blue**. Metrics: CIDEr-D (C), BLEU-4 (B@4), ROUGE-L (RL), METEOR (M), BERTScore (BS), and CLIP-Score (CS).

		Translated-to-Native						Native-to-Translated					
Encoder	**Decoder**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**
$DeiT_{BASE}$	$BERT_{BASE}$	37.35	14.95	35.45	36.63	67.73	49.75	37.57	14.02	34.58	35.43	67.02	50.66
	$DistilBERT_{BASE}$	37.52	14.87	35.28	36.27	67.55	49.67	38.97	14.07	35.83	36.08	67.34	49.95
	$GPT\text{-}2_{SMALL}$	36.02	14.90	34.80	36.09	67.31	49.49	37.58	11.87	32.58	35.56	66.64	50.51
$Swin_{BASE}$	$BERT_{BASE}$	**45.37**	**17.46**	36.90	37.90	68.57	**53.81**	**46.99**	**16.99**	**38.25**	**38.37**	**69.19**	**53.01**
	$DistilBERT_{BASE}$	**48.32**	**18.17**	**38.17**	**39.34**	**69.51**	53.26	**47.54**	**17.16**	**38.41**	**39.18**	**69.17**	**53.26**
	$GPT\text{-}2_{SMALL}$	**45.41**	**17.21**	**37.75**	**39.55**	**69.04**	**53.49**	**47.25**	14.54	35.82	**39.20**	**68.80**	**53.19**
ViT_{BASE}	$BERT_{BASE}$	40.02	16.38	36.17	36.98	68.10	51.93	42.12	14.50	35.08	35.71	67.73	52.84
	$DistilBERT_{BASE}$	43.45	16.82	**37.46**	**38.98**	**69.08**	51.84	44.07	**15.15**	**37.13**	37.52	68.42	51.86
	$GPT\text{-}2_{SMALL}$	42.33	15.64	36.47	38.01	68.41	52.44	40.43	12.37	33.45	36.77	67.53	51.47

by the models trained on the Flickr-Native. METEOR and BERTScore had a regular performance drop, slightly higher than CIDEr-D. ROUGE and BLEU exhibited similar performance drops to METEOR and BERTScore, except for the GPorTuguese-2-based models, which showed a more pronounced drop than the other models.

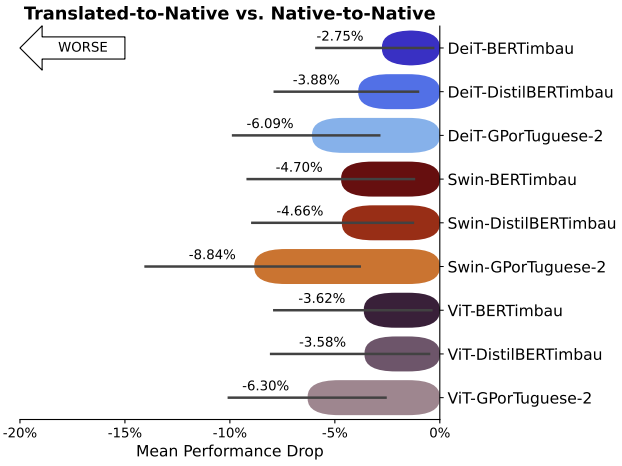

Figure 3. Illustration of the mean performance drop when switching context, using a bar chart. The chart depicts the mean percentage performance drop through the evaluation metrics. For instance, "**Translated-to-Native vs. Native-to-Native**" presents the difference between the Native evaluation of models trained on Translated and Native datasets. The numbers over the bars are their mean percentages, and the lines in the middle of the bars are the standard deviations of the mean percentages. A bar color represents each model to favor the visualization.

5.2 VLMs Assessment

Table 5 shows the results of the VLM assessment. As seen in the table, each model demonstrates unique strengths across uni-modal metrics (e.g., BLEU, CIDEr) and the multi-modal metric (CLIP-Score). For instance, ViTucano 2B demonstrates the highest CIDEr, BLEU, ROUGE, METEOR, and BERTScore in the translated dataset and the highest CIDEr and BLEU in the native dataset. The ViTucano 1B outperforms the baseline models in ROUGE, METEOR, and BERTScore in the native context. Despite the lower performance on most text-centric metrics, GPT-4 models achieved the highest CLIP-Score, indicating strong alignment between image features and caption representations. In summary, the results indicate that the natively pre-trained models generalized more effectively to the translated context than to the original language context, and the CLIP-Score consistently increased under this setting. In contrast, the complexity of the multilingual models increased, and GPT-4o was outperformed by its smaller variant across all metrics except CIDEr-D and CLIP-Score.

GPT-4o and GPT-4o-mini generated captions with an average length superior to 20 words, while the average length for PaliGemma and Phi-3 V were 8.63 and 11.39, respectively. The other models generated captions with an average length between 14.41 and 16.23. The GPT-4o and GPT-4o-mini showed the highest number of distinct words, 2,524 and 2,495, respectively, confirming their heterogeneous vocabulary. The remaining models did not exceed the 2,000-word limit for distinct words. On average, captions generated by GPT-4 models for Flickr-Translated contained more than 10 words not used in the respective reference, whereas for Flickr-Native, this number dropped to approximately 9 words. For both translated and native datasets, PaliGemma, Phi-3 V, and ViTucano models maintained a distinct word rate per caption of around 5, while LLaMa3.2 V averaged 6. We observed that captions generated by the GPT-4 family are typically descriptive and concise, with a more diverse vocabulary in some examples, contradicting the reference-based metrics.

Comparing model-based metrics such as BERTScore and CLIP-Score while considering the models' vocabulary sizes highlights the limitations of evaluation metrics in capturing the coherence, conciseness, and correctness of generated captions. The fact that GPT-4 models use words not present in the reference captions affects uni-modal metrics. BERTScore, while robust to synonyms and paraphrasing due to its use of text embeddings, may fail to recognize descriptions of situations, scenes, concepts, or entities that do not exist in the

Table 5. VLM performance (%) for zero-shot baseline VLMs on Flickr30K Translated and Flickr30K Native. The column #Params indicates the number of model parameters in billions (B). The three highest values for each metric are highlighted in **bold**, with the cell corresponding to the highest value further highlighted in **blue**. Metrics: CIDEr-D (C), BLEU-4 (B@4), ROUGE-L (RL), METEOR (M), BERTScore (BS), and CLIP-Score (CS). **Note**: The CS columns share the same values since this metric depends solely on the generated captions and image content, rather than the nature of the reference captions (translated or native).

		Pre-trained-to-Translated						Pre-trained-to-Native					
VLM	**#Params**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**	**C**	**B@4**	**RL**	**M**	**BS**	**CS**
ViTucano 1B	1.53B	**57.96**	**19.47**	**40.80**	**47.01**	**70.09**	55.96	**51.45**	**16.16**	**37.96**	**43.73**	**69.04**	55.96
ViTucano 2B	2.88B	**62.03**	**19.84**	**41.51**	**47.05**	**70.30**	56.28	**52.49**	**16.46**	**37.75**	**43.47**	**68.87**	56.28
PaliGemma	2.92B	23.47	4.88	20.76	20.33	52.37	49.67	18.69	4.65	20.59	19.78	52.15	49.67
Phi-3 V	4.15B	22.67	7.17	27.35	29.50	58.44	52.36	21.92	7.34	27.94	30.29	58.81	52.36
LLaMa3.2 V	11.70B	**34.48**	**10.20**	**31.56**	35.17	63.18	**56.95**	**34.94**	**9.81**	**30.81**	34.50	62.84	**56.95**
GPT-4o-mini	-	21.68	8.96	29.31	**41.63**	**63.83**	**61.26**	27.43	9.04	29.29	**41.96**	**64.24**	**61.26**
GPT-4o	-	25.68	7.36	25.04	38.67	61.49	**62.07**	34.02	8.93	27.36	41.18	63.35	**62.07**

reference captions. In contrast, CLIP-Score remains unaffected by vocabulary differences, as it relies on the alignment of extracted visual and language features.

VED Models vs. VLMs. This analysis compares VED models trained and tested on native data (Native-to-Native) with pre-trained VLMs evaluated on the same native data. A comparison between Tables 3 and 5 reveals that Swin-based models generally outperform VLMs in uni-modal metrics, underscoring the importance of fine-tuning in capturing both vocabulary and the expected textual structure of captions. However, it is worth noting that Swin-based models surpass only the Phi-3V and PaliGemma VLMs in terms of CLIP-Score. These results indicate that captions generated by VLMs, especially the largest ones, tend to be more aligned with image content. Nonetheless, the computational cost of such models should be considered for real-world applications. Excluding the very large GPT-4o VLMs, the most notable performance difference based on CLIP-Score is under 4 percentage points, observed when comparing the 11.70B-size VLM LLaMa3.2 V with the Swin-based models.

5.3 Qualitative Assessment

5.3.1 Comparing Reference and Generated Captions

Figures 4 to 7 illustrate four test cases under the same-source setup, presenting the reference captions, the generated captions (candidates) produced by Swin-DistilBERTimbau, and the corresponding performance metrics.

Case 1. Figure 4 shows a case from Flickr30K Translated with comparatively high performance on CLIP-Score given the uni-modal metrics. The generated caption was "Três meninos grelhando carne." ("Three boys grilling meat."). Although the elements in the generated caption are accurate, it lacks the level of detail found in the reference captions, and its length does not match that of the references. A counting error was also observed, as the model inferred three people instead of four, which corresponds to the actual scene. Finally,

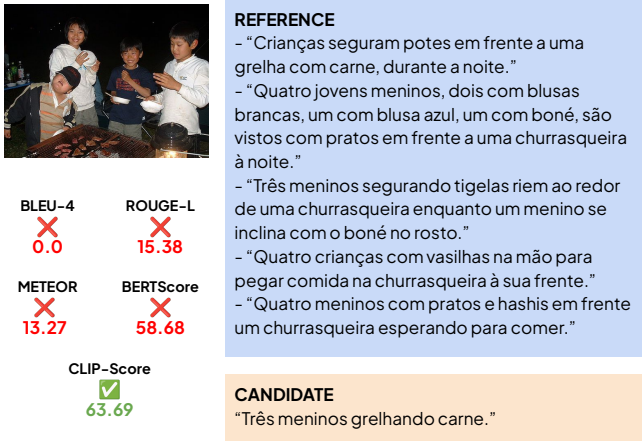

Figure 4. Case 1: Example from Flickr30K Translated with high performance on CLIP-Score compared to the uni-modal metrics.

some words in the generated caption do not appear together in a single reference caption, but are spread across distinct reference captions. For instance, the word "meninos" does not occur with "carne" or words similar to "grelhando" (e.g., "grelha"), which affects the performance of the reference-based (uni-modal) metrics.

Case 2. Figure 5 depicts a case from Flickr30K Translated with poor performance on CLIP-Score compared to the uni-modal metrics, with the generated caption "Três pessoas estão sentadas em um banco." ("Three people are sitting on a bench."). In this case, the generated caption matches one of the reference captions in terms of size and details, enabling high scores on reference-based metrics, while CLIP-Score scores are relatively lower. It is worth pointing out that the image is cropped around the middle of the people's bodies and the back support, which may impact the CLIP-Score evaluation. This cropping limits the availability of visual information, such as body positions and background elements, potentially affecting the alignment between image features and text. Even disregarding these points, the reference captions remain more informative and descriptive than the candidate caption, including descriptions of the bench (e.g., wood) and the ad with black letters on the back wall. The reference

BLEU-4 ✅ 71.11
ROUGE-L ✅ 50.00
METEOR ✅ 63.20
BERTScore ✅ 88.98
CLIP-Score ❌ 41.76

REFERENCE
- "Três pessoas sentadas no banco de madeira."
- "Três pessoas sentadas em bancos diante de uma escada rolante, com um anúncio em letras pretas gravado nas laterais dela."
- "Uma mulher de preto e uma de amarelo sentadas conversando ao lado de um homem de preto."
- "Duas mulheres conversando sentadas em um banco de madeira com um homem do lado."
- "Pessoas sentadas perto de uma escada rolante de um shopping."

CANDIDATE
"Três pessoas estão sentadas em um banco."

Figure 5. Case 2: Example from Flickr30K Translated with poor performance on CLIP-Score compared to the uni-modal metrics.

BLEU-4 ❌ 0.0
ROUGE-L ❌ 12.50
METEOR ❌ 12.33
BERTScore ❌ 56.24
CLIP-Score ✅ 65.76

REFERENCE
- "Pessoas circulam com instrumentos de percussão durante o dia."
- "Uma banda com bonés e roupas em preto e amarelo caminhando em formação enquanto carregam seus instrumentos musicais."
- "Pessoas uniformizadas de amarelo e verde andam enfileiradas segurando instrumentos em um gramado perto do mar."
- "Uma banda com várias pessoas usando roupas iguais está andando na praça com vários tambores."
- "Várias pessoas usando roupas iguais andando em fileira enquanto carregam instrumentos."

CANDIDATE
"Um grupo de homens uniformizados tocando tambor."

Figure 6. Case 3: Example from Flickr30k Native with a higher CLIP-Score value compared to the uni-modal metrics.

captions also carry some uncertain information, such as the location (e.g., shopping) and some elements in the image (e.g., escalator).

Case 3. Figure 6 shows a case from Flickr30K Native with comparatively high performance on CLIP-Score given the uni-modal metrics. The generated caption is "Um grupo de homens uniformizados tocando tambor." ("A group of uniformed men playing a drum."). We identified the same issue in Case 1 (Figure 4), where the generated caption, despite its correctness, does not reach the same level of detail and length as the reference captions. There is some uncertainty on the expression "Um grupo de homens" ("a group of men") because there may be a woman in the group. In this regard, other generic words (e.g., people) could provide a more secure alternative to gender-specific terms.

Case 4. Figure 7 shows a case from Flickr30K Native with poor performance on CLIP-Score in view of the uni-modal metrics. The generated caption was "Um homem e duas crianças brincando na grama verde." ("A man and two children playing on the green grass."), reaching a comparable level of detail and length to the reference captions. Only one reference caption includes the word "brincando" ("playing"), but the reference-based metrics achieved high values regardless of this consideration.

BLEU-4 ✅ 61.49
ROUGE-L ✅ 52.63
METEOR ✅ 63.56
BERTScore ✅ 80.60
CLIP-Score ❌ 39.64

REFERENCE
- "Pai brincando com sua filha e filho no gramado."
- "Homem e duas crianças plantando uma árvore no quintal de uma casa."
- "Duas crianças e um adulto plantando uma árvore na grama verde."
- "Um homem e duas crianças plantando uma árvore no jardim."
- "Duas crianças e um homem plantando uma árvore no chão."

CANDIDATE
"Um homem e duas crianças brincando na grama verde."

Figure 7. Case 4: Example from Flickr30K Native with a lower CLIP-Score compared to the uni-modal metrics.

5.3.2 Attention Maps

The attention scores obtained from the cross-attention layer were utilized to associate text tokens with the image's attention matrix. For this analysis, the same cases in the previous section were revisited. To produce a word-level visualization, token-level attention was aggregated into a word-level attention map by summing the token-level attention maps. For instance, the word "sentada" ("sitting") is the fusion of the tokens "sent" and "##ada". The double hashtag in the token means a connection to the previous token, forming a complete word. The attention visualization represents a sequence of word-image pairs (or simply word-image), where whitened rectangles highlight areas with high attention values, while darkened areas indicate low attention values.

Case 1. In Figure 8, the word-images for "grelhando" ("grilling") and "carne" ("meat") coherently highlight areas with plates, grill, and some meat. However, other regions unrelated to these concepts are also highlighted, like the night sky. Some noise is observed in the inference of the number of people in the scene, as evidenced by whitened pixels outside the expected regions and a strong focus on the child's arm.

Case 2. In Figure 9, the model attention highlights areas with people, including when it outputs "Três", inferring the number of people in the scene. In this example, we can also observe noise-whitening regions that do not contain word-related elements, such as in the words "pessoas" ("people") and "sentadas" ("sitting"). Despite the correct highlight for the word "banco" ("bench"), the previous pair "em um" ("on a"), stop words connected to "banco', yielded similar attention.

Case 3. In Figure 10, the word-image pair "homens" ("men") showed attention toward the people. In this case, the model exhibited a preference for a gender-specific term, a behavior that is a potential source of errors. Previous work, such as Hirota *et al*. [2023], also made similar observations, demonstrating that captioning models can perpetuate societal gender-related biases. The model correctly highlights the regions for the words "uniformizados" ("uniformed") and "tambor" ("drum").

Figure 8. Attention visualization for Case 1 (Figure 4). Generated caption: "Três meninos grelhando carne." ("Three boys grilling meat.").

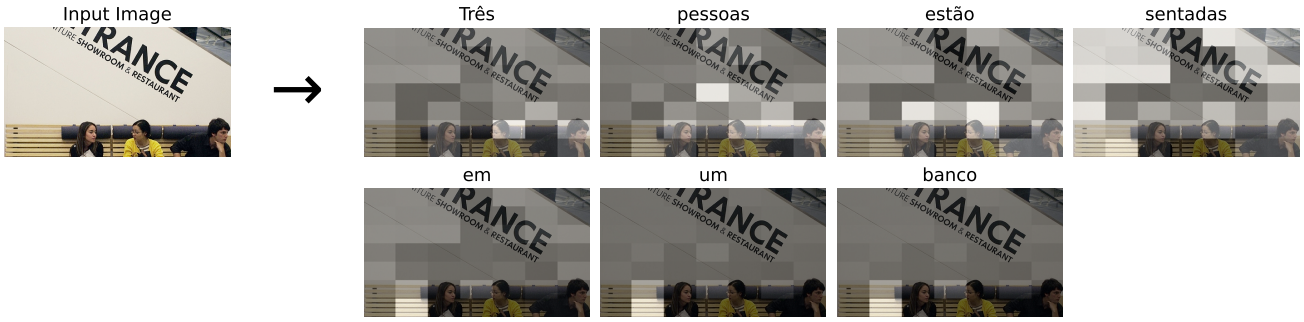

Figure 9. Attention visualization for Case 2 (Figure 5). Generated caption: "Três pessoas estão sentadas em um banco." ("Three people are sitting on a bench.").

Case 4. In Figure 11, the most coherent maps are those related to the words "Um homem" ("A man"), "duas crianças" ("two children"), and "grama verde" ("green grass"). Initially, the model highlights the man on the map with a lighter white rectangle, while the children are marked with slightly darker white rectangles. In the sequence, the attention focuses on the children, but with little more attention to the little girl besides the man. The last two words, "grama verde" ("green grass"), have maps highlighting the surroundings of the people in the image.

6 Concluding Remarks

This work evaluated the efficacy of Transformer-based VED models for Brazilian Portuguese image captioning under two main scenarios: (*i*) training and testing on the same dataset (native or translated), and (*ii*) cross-context between native and translated captions. Furthermore, we analyzed fine-tuned transformer-based VED models with up to 240 million parameters, compared with VLMs with more than 1.5 billion parameters.

Our results show that Swin encoders consistently outperform DeiT and ViT across text-centric and cross-modal metrics in Translated and Native scenarios, even when source-changing is applied. Models based on GPorTuguese-2 exhibited the most pronounced performance degradation in the Flickr-Native test set when the training source shifted from native to translated. Among the evaluation metrics, CLIP-Score was the most resilient to this change, as it operates independently of reference captions. In contrast, CIDEr exhibited the greatest performance decline.

Among the VLMs, ViTucano models were the only ones to perform comparably with VED models on evaluation metrics. The remainder of the VLMs underperformed the VED models in reference-based metrics, indicating that the generated captions of VLMs have lower lexical matching with reference captions than the captions generated by the VED models. Nonetheless, the VLMs achieved higher CLIP-Scores than the VED models, except for PaliGemma and Phi-3V, indicating greater image-text embedding congruence. Notably, native pre-trained models like ViTucano outperform larger multilingual models in reference-based evaluation metrics, including LLaMa 3.2 Vision, GPT-4o, and GPT-4o mini. Additionally, LLaMa 3.2 Vision, an open-source model with 11 billion parameters, surpasses complex closed-source models like GPT-4o in several evaluation metrics. PaliGemma and Phi-3 Vision presented the lowest performance.

The qualitative results analysis is supported by the addition of CLIP-Score and the attention maps visualization. CLIP-Score is useful for evaluation in scenarios where traditional metrics fail to capture relevant information, such as when it is spread across reference captions, making it difficult for traditional metrics to evaluate it when it co-occurs. In this situation, CLIP-Score does not replace traditional metrics for evaluating captions concisely, especially when exact lexical matches occur. Attention maps play an important role in model inference investigation, primarily to monitor whether the model links incoherent image regions to the target word. Some semantic connection mistakes were observed during the attention analysis, including incorrect attributions of gender, numerals, and adjectives due to erroneous region-word linking.

Limitations and Future Directions. This study focuses primarily on the Flickr30K dataset, which may limit generalizability to other image styles or cultural contexts. Additionally, while we compare native vs. machine-translated data, the quality of translation may vary by domain, length, or style, and further investigation into advanced translation or data augmentation strategies could prove beneficial. The implicit bias of the annotators of these datasets is an issue, as the two groups may differ in point of view, attention to detail, language vices, and mother-tongue vocabulary. Some terms in English cannot be easily translated into Portuguese while retaining their original characteristics. Finally, we did not

Figure 10. Attention visualization for Case 3 (Figure 6). Generated caption: "Um grupo de homens uniformizados tocando tambor." ("A group of uniformed men playing a drum.").

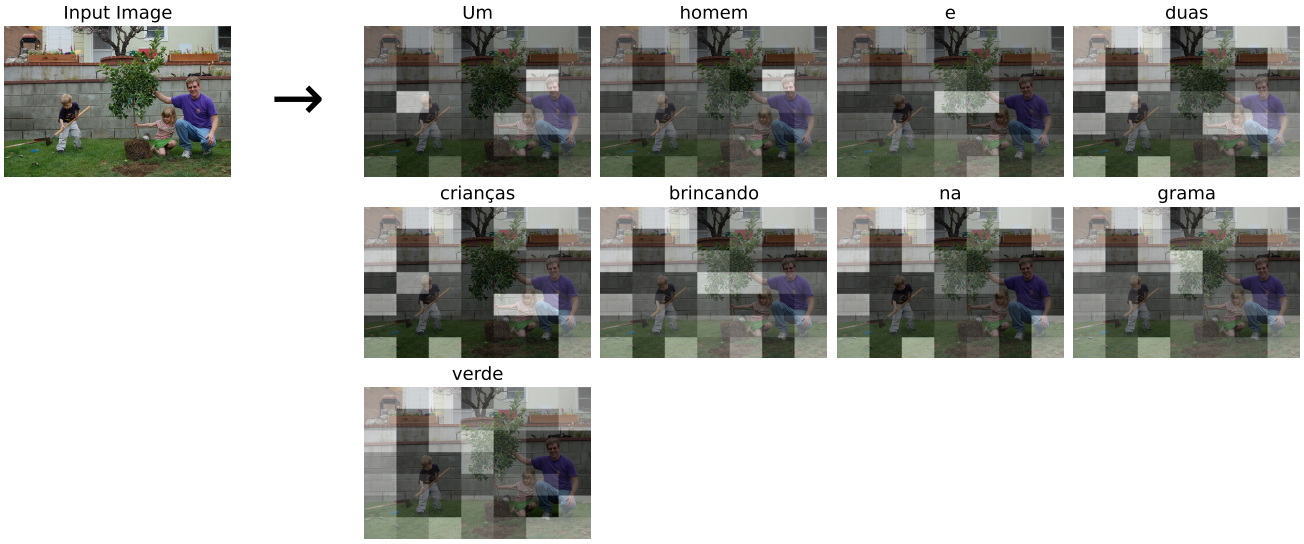

Figure 11. Attention visualization for Case 4 (Figure 7). Generated caption: "Um homem e duas crianças brincando na grama verde." ("A man and two children playing on the green grass.").

explore the impact of very large multi-modal transformers that have shown recent promise; adapting such models to Brazilian Portuguese might further bridge domain gaps. Future work will expand these experiments to other state-of-the-art VLMs and investigate more sophisticated fine-tuning or adapter approaches, particularly for low-resource languages. Furthermore, we aim to expand the investigation of VLMs for Brazilian Portuguese IC by incorporating additional native datasets, such as #PraCegoVer, and further translated sources. The application of advanced statistical analyses to the evaluation metrics [Kilickaya *et al*., 2017], the use of prompt engineering techniques to improve the performance of VLMs [Wang *et al*., 2023], and the inclusion of additional evaluation approaches, such as LLM-as-a-judge [Chan *et al*., 2023], would strengthen the present study. By addressing these limitations, we aim to bring more inclusive and effective image captioning solutions for Brazilian Portuguese speakers and beyond.

Declarations

Authors' Contributions

Gabriel Bromonschenkel contributed to the writing of the original draft, software development, data curation, and visualizations. **Alessandro L. Koerich** contributed to the writing – review & editing. **Thiago M. Paixão** was responsible for research co-supervision, and writing – review & editing. **Hilário Tomaz Alves de Oliveira** was responsible for research supervision, validation, and writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets and the models generated and analyzed during the current study are available in: `https://github.com/laicsiifes/transformer-caption-ptbr`.

References

Abdelaal, A., ELshafey, N. F., Abdalah, N. W., Shaaban, N. H., Okasha, S. A., Yasser, T., Fathi, M., Fouad, K. M., and Abdelbaky, I. (2024). Image captioning using vision encoder decoder model. In *2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, pages 101–106. IEEE. DOI: 10.1109/ICMISI61517.2024.10580628.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., *et al*. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*. DOI: 10.48550/arXiv.2404.14219.

Adalberto Ferreira Barbosa Junior (2024). distilbert-portuguese-cased (revision df1fa7a). DOI: 10.57967/hf/3041.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. Available at:`https://aclanthology.org/W05-0909/`.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., *et al*. (2024). Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*. DOI: 10.48550/arXiv.2407.07726.

Bromonschenkel, G., Oliveira, H., and Paixão, T. M. (2024). A comparative evaluation of transformer-based vision encoder-decoder models for brazilian portuguese image captioning. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. IEEE. DOI: 10.1109/SIBGRAPI62404.2024.10716325.

Chan, D., Petryk, S., Gonzalez, J., Darrell, T., and Canny, J. (2023). Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646. DOI: 10.18653/v1/2023.emnlp-main.841.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024). ViTucano: A Portuguese Vision Assitant. Available at:`https://huggingface.co/TucanoBR`.

de Alencar, R. S., Castañeda, W. A. C., and Amadeus, M. (2024). Image captioning for brazilian portuguese using grit model. *arXiv preprint arXiv:2402.05106*. DOI: 10.48550/arXiv.2402.05106.

dos Santos, G. O., Colombini, E. L., and Avila, S. (2022). #pracegover: A large dataset for image captioning in portuguese. *Data*, 7(2). DOI: 10.3390/data7020013.

dos Santos, G. O., Moreira, D. A. B., Ferreira, A. I., Silva, J., Pereira, L., Bueno, P., Sousa, T., Maia, H., Da Silva, N., Colombini, E., *et al*. (2023). Capivara: Cost-efficient approach for improving multilingual clip performance on low-resource languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 184–207. DOI: 10.18653/v1/2023.mrl-1.15.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al*. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. DOI: 10.48550/arXiv.2010.11929.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., *et al*. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. DOI: 10.48550/arXiv.2407.21783.

Ghandi, T., Pourreza, H., and Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39. DOI: 10.1145/3617592.

Gondim, J., Claro, D. B., and Souza, M. (2022). Towards image captioning for the portuguese language: Evaluation on a translated dataset. In *ICEIS (1)*, pages 384–393. DOI: 10.5220/001108000000317.

Guillou, P. (2020). Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...). Available at:`https://huggingface.co/pierreguillou/gpt2-small-portuguese`.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. DOI: 10.18653/v1/2021.emnlp-main.595.

Hirota, Y., Nakashima, Y., and Garcia, N. (2023). Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15200. DOI: 10.1109/CVPR52729.2023.01458.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., *et al*. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. DOI: 10.48550/arXiv.2410.21276.

Ishan, T. I., Al Noman, A., Rokib, R., Masum, M. I., Ahmed, S., and Shah, F. M. (2023). Bengali image captioning using vision encoder-decoder model. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. DOI: 10.1109/ICCIT60459.2023.10441125.

Jnaini, A., Shirazi, H., and Homayouni, H. (2024). Synergy of gpt-3 summarization and vision-encoder-decoder for chest x-ray captioning. In *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 476–482. IEEE. DOI: 10.1109/CCECE59415.2024.10667261.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209. DOI: 10.18653/v1/e17-1019.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Available at:`https://aclanthology.org/W04-1013/`.

Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021a). Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*. DOI: 10.48550/arXiv.2101.10804.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022. DOI: 10.1109/ICCV48922.2021.00986.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. DOI: 10.3115/1073083.1073135.

Sharma, H. and Padha, D. (2023). A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review*, pages 1–43. DOI: 10.1007/s10462-023-10488-2.

Silva Barbon, R. and Akabane, A. T. (2022). Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: a case study. *Sensors*, 22(21):8184. DOI: 10.3390/s22218184.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer. DOI: 10.1007/978-3-030-61377-8_28.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559. DOI: 10.1109/TPAMI.2022.3148210.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575. DOI: 10.1109/CVPR.2015.7299087.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164. DOI: 10.1109/CVPR.2015.7298935.

Viridiano, M., Lorenzi, A., Torrent, T. T., Matos, E. E., Pagano, A. S., Sigiliano, N. S., Gamonal, M., de Andrade Abreu, H., Dutra, L. V., Samagaio, M., *et al*. (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. DOI: 10.63317/2urtgtf4vshk.

Wang, N., Xie, J., Wu, J., Jia, M., and Li, L. (2023). Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2617–2625. DOI: 10.1609/aaai.v37i2.25360.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. DOI: 10.48550/arXiv.1904.09675.