pt-image-ir-dataset: An Image Retrieval Dataset in European Portuguese

Rodrigo Duarte[1,2(✉)], António Branco[4,5], Hugo Proença[1,6], and Ricardo Campos[1,2,3]

[1] University of Beira Interior, Covilhã, Portugal
{hugomcp,ricardo.campos,rodrigo.duarte}@ubi.pt
[2] INESC TEC, Porto, Portugal
[3] Ci2 - Smart Cities Research Center - Polytechnic Institute of Tomar, Tomar, Portugal
[4] University of Lisbon, Lisbon, Portugal
antonio.branco@di.fc.ul.pt
[5] NLX-Group, Lisbon, Portugal
[6] IT: Instituto de Telecomunicações, Covilhã, Portugal

Abstract. With the surge of multimodal models and the demand for effective image Information Retrieval (IR) systems, high-quality text-to-image datasets have become paramount. However, most existing datasets are primarily in English, limiting their applicability to multilingual settings. To address this, we introduce the `pt-image-ir-dataset`, a manually annotated resource for text-based Image IR in European Portuguese. The dataset comprises 80 diverse queries and a curated pool of 5,201 images, each annotated for relevance by multiple human judges. The proposed dataset is a step forward in supporting the development and evaluation of image IR systems for European Portuguese, addressing a clear gap in multilingual multimodal research. To this end, we have made our dataset publicly available, alongside baseline experimental results, demonstrating its suitability on the Image IR task across different retrieval paradigms, including traditional text-based lexical IR methods, semantic dense retrieval models based on language embeddings, cutting-edge vision-language models and proprietary black-box image retrieval systems. Results demonstrate that vision-language models, particularly *OpenCLIP/xlm-roberta-base-ViT-B-32*, significantly outperform other approaches (MRR = 0.610).

Keywords: text-to-image dataset · image information retrieval · vision-language models · multimodal benchmark dataset · European Portuguese

1 Introduction

The exponential growth of digital visual content has created an unprecedented demand for effective image retrieval systems capable of understanding and

R. Campos et al. (Eds.): ECIR 2026, LNCS 16486, pp. 463–478, 2026.
https://doi.org/10.1007/978-3-032-21321-1_53

matching textual queries with relevant images [4,16]. While significant advances have been made in multimodal information retrieval driven by vision-language models such as CLIP [20], ALIGN [10], and BLIP-2 [14], the vast majority of available datasets and research efforts remain focused on English-language contexts. For instance, CLIP was trained on 400 million English image–text pairs collected from the web, limiting its applicability to low-resource languages. Several other efforts [15,26,30] have been made to provide high-quality datasets to researchers, however, they also remain predominantly English-centric and typically rely on image-caption pairs, which do not necessarily reflect realistic user queries, which are often shorter and noisier. As a result, current resources fail to capture the cultural and linguistic diversity embedded in real-world multimodal search scenarios.

Portuguese, spoken by more than 250 million people worldwide, illustrates this problem particularly well. Despite its global relevance, Portuguese remains severely underrepresented in the multimodal information retrieval landscape. Existing multilingual datasets, while valuable, either omit Portuguese entirely or provide very limited coverage, typically in the form of captions rather than query-driven retrieval tasks. The challenge extends beyond mere translation, as effective systems must account for cultural nuances, domain-specific terminology, and regional linguistic variations. European Portuguese, in particular, introduces distinct lexical and syntactic characteristics that require dedicated resources for fair evaluation and robust system development. Translating an existing English dataset would fail to capture the unique visual and contextual characteristics of Portuguese institutional imagery, such as local governmental events, cultural practices, and named entities specific to Portugal. By constructing the dataset from scratch, we ensure that queries, images, and relevance judgments reflect authentic Portuguese usage and real-world search behavior, providing a resource that is both linguistically and culturally grounded.

To address these gaps, we introduce `pt-image-ir-dataset`, a dataset specifically designed for Portuguese image information retrieval research. It comprises 80 queries in European Portuguese paired with 5,201 images collected from the official Portuguese Presidency website, spanning nearly a decade of institutional content, including official ceremonies, cultural events, and governmental activities. Crucially, the dataset includes manually annotated relevance judgments from three independent annotators, ensuring high-quality ground truth.

Our contributions can be summarized as follows:

- We introduce `pt-image-ir-dataset`, the first manually annotated benchmark specifically designed for query-driven image retrieval in European Portuguese, comprising 80 queries and 5,201 images with human relevance judgments.
- We provide baseline experimental results across four retrieval paradigms: traditional lexical IR methods, semantic dense retrieval models based on language embeddings, state-of-the-art vision–language models, and proprietary black-box image retrieval systems.

– We make the dataset publicly available under a Creative Commons Attribution-NonCommercial License, ensuring long-term accessibility and facilitating reproducibility and further research.

2 Related Work

The evolution of image information retrieval (IR) has been shaped by progress in both algorithmic methods and dataset availability. While early content-based image retrieval (CBIR) systems relied on visual features such as color, texture patterns, and shape features [3,13], they struggled with the semantic gap between low-level visual features and high-level user intentions, limiting their effectiveness in realistic search scenarios. The emergence of deep learning and later transformer architectures [28] revolutionized this field by enabling richer multimodal representations. In particular, the emergence of vision-language models (VLMs) shifted the focus from purely visual similarity to semantic alignment between text and images. Prominent examples include CLIP [20], ALIGN [10] and BLIP-2 [14], which have demonstrated strong performance on diverse multimodal tasks, including zero-shot classification, caption generation, and notably, image retrieval [12,22,24]. By projecting both modalities into a common embedding space through contrastive or generative objectives, these models enable retrieval systems to rank images based on semantic similarity to user queries.

Despite these technological advances, the development of effective multimodal retrieval systems remains critically dependent on the availability of high-quality datasets that provide query-driven resources for evaluation. However, most were originally designed for image captioning, which differs fundamentally from query-driven retrieval. Moreover, existing English-language datasets focus on Western contexts, introducing linguistic and cultural biases that fail to capture region-specific visual concepts, practices, and named entities. For example, the MS COCO dataset [15], one of the most widely adopted resources in multimodal research, contains over 330,000 images with multiple English captions per image, providing rich descriptive text but lacking both the query-response pairs essential for retrieval evaluation and any coverage beyond English annotations, limiting its applicability to Portuguese and other low-resource languages. Flickr30k [30], which contains 31,000 images paired with 155,000 human-authored English captions, and its extension Flickr30k Entities [18], which links textual mentions to visual elements, face similar limitations, offering high-quality captions and grounding information but restricted to English and description-oriented tasks. Fashion200k [7], comprising 200,000 weakly labeled fashion images with concise textual descriptions of visual attributes, demonstrates the potential for fine-grained alignment between visual and textual features. However, its narrow focus on the fashion domain and the absence of query-driven structures limit its applicability to broader image retrieval research.

Efforts to broaden linguistic and cultural coverage include WIT [26], which spans 108 languages with 37 million image–text pairs (with some Portuguese

annotations), and the more recent MIRACL-VISION dataset [17], which extends document retrieval into the visual domain. While valuable for multilingual evaluation, both are constructed from Wikipedia content, leading to imbalanced language coverage (with English dominating) and lacking the explicit query–response structures needed for image retrieval. Similarly, the #PraCegoVer dataset [19] provides over 500,000 Brazilian Portuguese captions sourced from Instagram with accessibility guidelines, but it targets captioning rather than retrieval and reflects Brazilian Portuguese linguistic and cultural norms rather than European Portuguese.

In summary, while vision-language models have pushed the state of the art in multimodal IR, their effectiveness remains limited by the datasets on which they are trained and evaluated. Existing resources are largely English-centric, caption-oriented, or domain-specific, leaving a critical gap for query-driven image retrieval. Bridging this gap requires dedicated resources that combine realistic query-response images structures, relevance annotations, and culturally grounded content in European Portuguese, elements that current datasets fail to provide. To address these limitations, we introduce `pt-image-ir-dataset`, a new resource specifically designed to support Portuguese image retrieval research.

3 pt-image-ir-dataset Dataset

Building on the gaps identified in existing resources, we now present the pt-image-ir-dataset, a benchmark specifically designed to advance Portuguese image retrieval research. The dataset couples European Portuguese queries with a collection of images and corresponding human relevance judgments, providing a realistic and culturally grounded testbed for multimodal retrieval systems.

3.1 Dataset Creation

The construction of the dataset followed three main phases: content acquisition, query generation, and relevance annotation. Images were sourced from the official Portuguese Presidency website[1], which provides a rich and diverse collection of institutional content. Our decision to use the official website of the Portuguese Presidency was driven by several factors: the public availability of content, the diversity of visual contexts, the top quality of duly curated text, and the extensive temporal coverage of the published material. The website documents nearly a decade of presidential activity, capturing a broad range of events, from official ceremonies and international visits to cultural and local events, across different time periods. This longitudinal aspect adds significant value to the dataset, enabling the exploration of visual variation over time.

To extract information from the website, we developed a Python-based web scraper using BeautifulSoup for HTML parsing and Selenium for dynamic content handling. The scraper navigated through the website's structure, collected

[1] https://www.presidencia.pt/.

news articles, and extracted article titles, publication dates, and associated images. In total, 4,678 articles published between January 2016 and March 2025 were collected, yielding 42,333 images with associated textual metadata.

For query generation, we followed a hybrid approach combining automated and manual methods to ensure diversity and naturalness. Using GPT-4o mini, we generated 38 queries across four thematic categories: *events and contexts*, *facial expressions and emotions*, *interactions with the public*, and *places and environments*. These categories were selected to reflect the content diversity typically found in governmental imagery. The prompt used for query generation was:

> *"You are a journalist preparing an article about the Portuguese Presidency and need to find suitable images to illustrate it. You have a search engine available with images at your disposal. Write text-based search queries in European Portuguese that could help retrieve images related to the following category:* ***[category]****".*

To complement these model-generated queries, which tend toward verbose descriptions, we added 42 additional manually-created queries within seven categories: *public figures*, *sports and activities*, *trending topics*, *places and monuments*, *general places*, *objects*, and *others*. These queries were intentionally designed to be more concise, reflecting typical user search patterns. Example queries include a model-generated query: "Presidente de Portugal em encontros com cidadãos" (President of Portugal meeting citizens), and a manually created query: "Bombeiros" (Firefighters). This combination allows the dataset to cover a wider spectrum of query styles and complexity levels, providing a more thorough testbed for evaluating retrieval systems.

Once queries were defined, we created a pool of candidate images for annotation using a pooling [11] strategy popularized by the TREC community [8]. Pooling allows annotators to evaluate a representative subset of images per query to establish ground truth, avoiding exhaustive annotation of the entire dataset. Our strategy combined three types of retrieval paradigms: traditional text-based IR (TF-IDF and BM25 on article titles), vision-language models (several CLIP variants including *M-CLIP/LABSE-Vit-L-14, M-CLIP/XLM-Roberta-Large-Vit-B-32*, *M-CLIP/XLM-Roberta-Large-Vit-L-14*, and *OpenCLIP/xlm-roberta-base-ViT-B-32*), and proprietary black-box image retrieval systems (Google Images and Arquivo.pt, constrained to the Presidency domain). For each method, the top-10 retrieved images per query were selected to form a diverse candidate set, averaging 65 images per query. Images not included in the pooled set were assumed non-relevant, following standard pooling assumptions [29].

3.2 Annotation Process

The annotation process was carried out by three master's students who participated voluntarily, without financial compensation, following detailed guidelines to ensure consistency and reliability. Annotators were instructed to evaluate each image's relevance to its corresponding query based on two labels: *relevant* (1) if

the image directly supported, illustrated, or provided meaningful context aligned with the query's intent, and *non-relevant* (0) if it failed to do so or provided misleading information. The guidelines emphasized a systematic procedure: carefully reading and interpreting each query, assessing image content for visual elements linked to the query, and considering the context provided by article titles. In ambiguous cases, annotators were advised to default to *non-relevant*, with the requirement that subjects should be clearly visible and identifiable. To minimize subjectivity, the guidelines stressed the importance of avoiding personal bias and provided strategies for handling unknown terminology, such as using online searches. These measures were crucial for promoting inter-annotator agreement and reducing variability in judgments.

To facilitate the annotation process, we developed a dedicated web-based annotation tool with intuitive controls and keyboard shortcuts to streamline the evaluation workflow. The tool was implemented as a containerized web application, combining a Flask-based backend, a Redis database for data storage, and a responsive front-end interface. Its layout was organized into four main components: the query text at the top, a central image display, the article title below the image for contextual reference, and annotation controls at the bottom. To accelerate labeling, we integrated keyboard shortcuts (R = relevant, I = non-relevant, Z = undo) complemented by visual feedback and real-time progress tracking. Annotators submitted their judgments using these intuitive keyboard shortcuts, with each selection recorded immediately. The annotation tool also allows annotators to undo or revise previous judgments if their interpretation changes after seeing additional images. For each query, images are presented sequentially, enabling annotators to compare items for the same query. Figure 1 illustrates the clean, and intuitive interface that enabled efficient and consistent annotation.

We assessed annotation quality using Fleiss' Kappa [1], obtaining a coefficient of 0.62, which reflects substantial inter-annotator agreement. 76% of images had unanimous labels, while 11% and 13% had two annotators agreeing on relevance and non-relevance, respectively, highlighting genuine ambiguity in the task. Final relevance labels were assigned through majority voting. Both the annotation tool and the accompanying guidelines are publicly available[2].

To further examine the challenges of relevance assessment across different query types (AI-generated or manual), we analyzed inter-annotator agreement by generation method and semantic category. Figure 2 summarizes this comparative analysis, highlighting variations in annotation difficulty depending on the nature of the queries. The analysis highlights complementary strengths of AI-generated and manually-created queries, with the latter achieving higher inter-annotator agreement ($k = 0.695 \pm 0.219$) than AI-generated ones ($k = 0.494 \pm 0.259$). This difference reflects their distinct characteristics: manually-created queries, with their concise and direct nature, facilitate consistent annotation decisions, whereas AI-generated queries introduce greater semantic complexity, demanding more nuanced interpretation.

[2] https://github.com/RodrigDuarte/text2image_ir_annotation_tool.

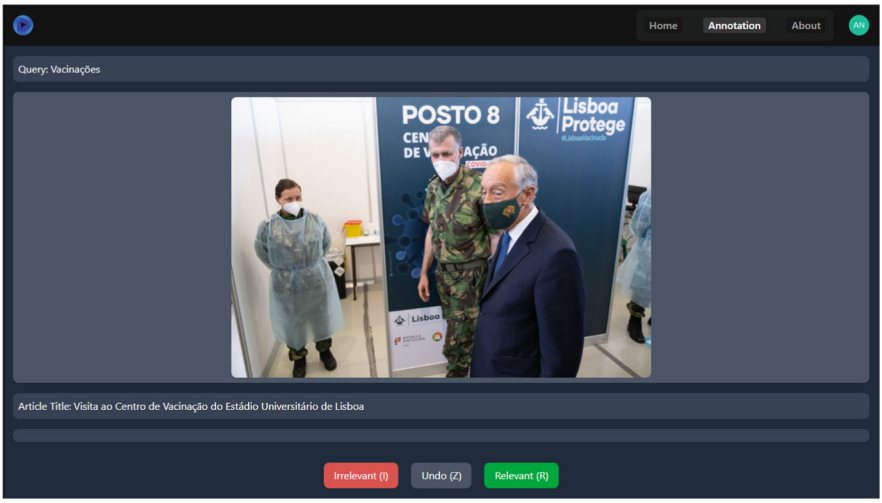

Fig. 1. Web-based annotation tool interface used for dataset labeling. The layout includes four main components: query text (top), image display (center), article title for context (below the image), and annotation controls (bottom) with keyboard shortcuts and progress tracking.

The category-based analysis shows substantial variation in annotation difficulty, with Fleiss' Kappa scores spanning a range of 0.440 across different semantic domains. Queries on *trending topics* reached the highest agreement ($k = 0.862$), reflecting their concrete and visually distinctive nature. In contrast, queries involving subjective assessments such as *interactions with the public* yielded lower agreement ($k = 0.422$), which is consistent with the context-dependent nature of human interactions. Intermediate categories, such as *places and monuments* ($k = 0.762$) and *public figures* ($k = 0.760$) also showed strong agreement, indicating that concrete, underscoring how concrete and identifiable content supports consistent judgments.

3.3 Structure and Dataset Characterization

The `pt-image-ir-dataset` comprises 80 queries and 5,201 images, providing a robust resource for Portuguese image retrieval. While the number of queries is relatively small, each query includes a substantial set of annotated images, combining AI-generated and manually-created queries to capture diverse styles and complexity levels. Images are predominantly in JPEG format with two primary resolutions: 800×533 and 1800×1200 pixels. These characteristics make the dataset particularly valuable as the first query-driven image IR resource in European Portuguese, establishing a benchmark for future research. The dataset is organized into four separate files in standard IR evaluation formats: *queries.tsv* (queries), *qrels.txt* (relevance judgments in TREC format), *images.tsv* (image metadata), and *articles.tsv* (article information with contextual metadata). The

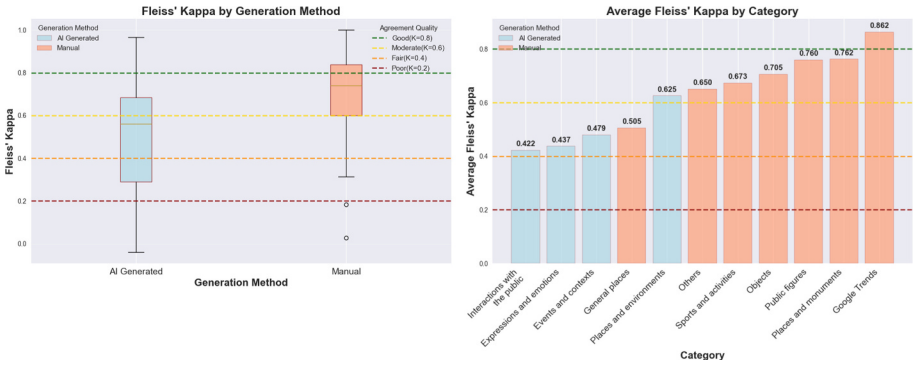

Fig. 2. Comparative analysis of Fleiss' Kappa scores across query generation methods (left) and categories (right).

dataset is made publicly available on GitHub[3] and archived on Zenodo[4] with a persistent DOI. Each query includes relevance judgments for an average of 65 images, with about 23 images labeled as relevant. Overall, 1,845 images (35%) were marked relevant and 3,356 (65%) non-relevant. The number of relevant images varies considerably across queries, reflecting natural diversity in query-driven retrieval. The 80 queries represent a deliberate balance between AI-generated and manually-created content, capturing different query characteristics and complexity levels. Figure 3 shows their distribution across categories. AI-generated queries span four categories, dominated by *events and contexts* (36.8%), followed by *places and environments* (23.7%), *expressions and emotions* (21.1%), and *interactions with the public* (18.4%). Manually created queries exhibit greater thematic diversity, covering seven distinct categories not addressed by the AI-generated set. Their distribution is more even, with *public figures* accounting for 19.0%, and *general places* and *others* each representing 16.7%. The remaining categories each contribute roughly 11.9%, reflecting the intention to capture additional semantic domains.

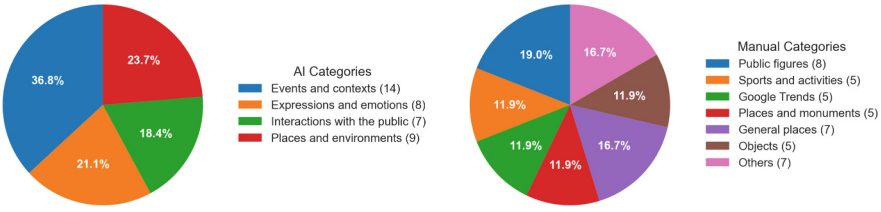

Fig. 3. Query distribution across categories: AI-generated queries (left) vs. manually-created queries (right).

[3] https://github.com/LIAAD/pt-image-ir-dataset.
[4] https://doi.org/10.5281/zenodo.15566572.

This dual approach offers important advantages for IR system evaluation. AI-generated queries are generally more verbose and descriptive (ranging from 5 to 9 words, with a median of 7 words, e.g., "Presidente de Portugal em encontros com cidadãos" - "President of Portugal in meetings with citizens"), whereas manually-created queries are shorter and more direct (ranging from 1 to 4 words, with a median of 2 words, e.g., "Bombeiros" - "Firefighters"), closely resembling real-world search behavior. By combining both types, the dataset supports evaluation across varying levels of query complexity and linguistic style.

4 Experiments and Results

To validate the usefulness and robustness of the `pt-image-ir-dataset`, we conducted a series of experiments covering a diverse set of image IR paradigms. Our goal was to demonstrate that the dataset supports multiple retrieval paradigms, establishing it as a valuable benchmark for evaluating different types of image retrieval systems. The experimental setup was organized four main categories. The first three correspond to the methods employed during the dataset's annotation phase: (1) traditional lexical IR techniques, (2) vision-language models, and (3) proprietary black-box image retrieval systems. To further broaden evaluation and assess the dataset's versatility, we introduced a fourth category: (4) semantic dense retrieval methods based on textual embeddings.

For the traditional text-based IR methods, we used TF-IDF [23] and BM25 [21], representing each image by the title of the article in which it appeared. Rankings were computed directly from term frequency and document relevance to the query. In contrast, vision-language models were employed to capture the semantic alignment between textual queries and images in a shared embedding space. We experimented with several CLIP ariants, including multilingual adaptations such as Multilingual-CLIP [5] and OpenCLIP [6,9], [20], which were trained on large-scale cross-modal datasets and are capable of handling queries in multiple languages. The results were ranked via cosine distance, Specifically, we evaluated: *M-CLIP/LABSE-Vit-L-14*; *M-CLIP/XLM-Roberta-Large-Vit-B-32*; *M-CLIP/XLM-Roberta-Large-Vit-L-14*; *OpenCLIP/xlm-roberta-base-ViT-B-32 with laion5b_s13b_b90k* along with more recent models such as BLIP-2 [14] and SigLIP [32]. For proprietary black-box retrieval, we queried two external search engines: Arquivo.pt and Google Images, with their proprietary ranking algorithms, both restricted to return images from the official Portuguese Presidency website, to constrain the search space and ensure contextual relevance. Finally, for semantic dense retrieval, we included BERTimbau Large [25] (Brazilian Portuguese) and Albertina PT-PT [2] (European Portuguese). These models were introduced only at the evaluation stage, allowing us to explore how newer architectures perform on a Portuguese language dataset. Altogether, this broad evaluation spectrum not only benchmarks diverse paradigms, but also to stress-test the dataset's capacity to generalize across retrieval settings and architectures. All systems were evaluated with standard TREC metrics using the *pytrec_eval* library [27], including Mean Average Precision (MAP), Precision@k (P@k), Recall@k (R@k), F1-Score@k (F1@k), Mean Reciprocal Rank

(MRR), and R-Precision (RP). Results are reported as macro-averages across all queries.

We now discuss the results from two complementary perspectives: first, a performance comparison across all baseline systems; and second, an analysis of how query length may impact retrieval effectiveness.

Comparative Evaluation of Baselines. Table 1 summarizes the performance of all baseline methods. While the absolute values reported in Table 1 may appear modest at first glance, they should be interpreted in light of the intrinsic difficulty of query-driven image retrieval and the specific characteristics of the `pt-image-ir-dataset`, which reflects realistic user search scenarios, featuring short, ambiguous queries and visually similar images drawn from a homogeneous institutional domain.

Table 1. Baseline performance on Portuguese image retrieval (with 95% bootstrap confidence intervals). Best scores in bold.

Method	MAP	R@10	P@10	F1@10	MRR
Traditional Text-Based IR					
TF-IDF	$0.138_{\pm 0.042}$	$0.188_{\pm 0.051}$	$0.291_{\pm 0.069}$	$0.209_{\pm 0.050}$	$0.377_{\pm 0.082}$
BM25	$0.079_{\pm 0.030}$	$0.127_{\pm 0.043}$	$0.209_{\pm 0.063}$	$0.141_{\pm 0.041}$	$0.331_{\pm 0.092}$
Portuguese Language Embedding Models					
BERTimbau Large	$0.048_{\pm 0.031}$	$0.076_{\pm 0.039}$	$0.102_{\pm 0.049}$	$0.080_{\pm 0.039}$	$0.157_{\pm 0.073}$
Albertina PT-PT 1.5b	$0.055_{\pm 0.028}$	$0.091_{\pm 0.040}$	$0.131_{\pm 0.052}$	$0.096_{\pm 0.038}$	$0.206_{\pm 0.081}$
Proprietary Black-Box Image Retrieval Systems					
Google Images	$0.076_{\pm 0.030}$	$0.125_{\pm 0.039}$	$0.190_{\pm 0.058}$	$0.128_{\pm 0.037}$	$0.336_{\pm 0.090}$
Arquivo.pt	$0.038_{\pm 0.014}$	$0.080_{\pm 0.025}$	$0.134_{\pm 0.039}$	$0.091_{\pm 0.024}$	$0.217_{\pm 0.072}$
Vision-Language Models					
M-CLIP LABSE-ViT-L-14	$0.130_{\pm 0.034}$	$0.210_{\pm 0.044}$	$0.341_{\pm 0.070}$	$0.237_{\pm 0.045}$	$0.467_{\pm 0.086}$
M-CLIP XLM-R-Large-B-32	$0.119_{\pm 0.030}$	$0.176_{\pm 0.035}$	$0.334_{\pm 0.069}$	$0.215_{\pm 0.042}$	$0.512_{\pm 0.097}$
M-CLIP XLM-R-Large-L-14	$0.158_{\pm 0.039}$	$0.245_{\pm 0.052}$	$0.376_{\pm 0.069}$	$0.266_{\pm 0.048}$	$0.491_{\pm 0.090}$
OpenCLIP xlm-roberta-base	$\mathbf{0.176}_{\pm 0.036}$	$\mathbf{0.264}_{\pm 0.049}$	$\mathbf{0.419}_{\pm 0.079}$	$\mathbf{0.288}_{\pm 0.047}$	$\mathbf{0.610}_{\pm 0.094}$
BLIP-2 ViT-G	$0.015_{\pm 0.010}$	$0.030_{\pm 0.015}$	$0.062_{\pm 0.031}$	$0.039_{\pm 0.019}$	$0.140_{\pm 0.068}$
SigLIP Base	$0.107_{\pm 0.037}$	$0.168_{\pm 0.046}$	$0.263_{\pm 0.067}$	$0.181_{\pm 0.045}$	$0.457_{\pm 0.101}$
Theoretical Upper Bound					
Oracle (Perfect Ranking)	1.000	0.544	0.926	0.613	1.000

To better contextualize the baseline results in Table 1, we also include an Oracle row, representing the best possible outcome for each metric given the available relevance judgments. The Oracle assumes a perfect ranking, where all relevant images appear at the top of the retrieved list for each query. For metrics like Precision@10, this accounts for queries with fewer than 10 relevant images.

By comparing model performance to these ceilings, it becomes clear that vision-language models, such as OpenCLIP, achieve strong early-ranking effectiveness (MRR = 0.610 vs. Oracle = 1.000, 61% of theoretical maximum) while the modest absolute scores reflect both the difficulty of the task and the limited number of relevant images per query (average of 23 relevant images per query).

To address concerns about evaluation stability given the 80-query test set, we report 95% bootstrap confidence intervals (1,000 iterations) for all methods in Table 1. The narrow confidence intervals across all metrics confirm that the observed performance differences are statistically robust and unlikely to be significantly affected by individual edge cases. For instance, OpenCLIP's MRR of $0.610_{\pm 0.094}$ demonstrates consistent strong early-ranking performance across bootstrap samples, while the substantial gap between vision-language models and text-only methods remains stable.

Overall, the results reveal substantial differences across paradigms. Vision-language models stand out clearly, with OpenCLIP achieving the best overall performance (MRR = 0.610). This score indicates good early-ranking effectiveness, with relevant images frequently appearing at the very top of the ranking, which aligns well with realistic user behavior in image search, where users often inspect only the top few results. In contrast, the comparatively low performance of traditional text-based and text-only dense retrieval methods underscores the difficulty of the task, since article titles provide only weak textual grounding. This highlights the benefits of direct visual-semantic alignment for Portuguese image retrieval.

Portuguese-specific language embedding models show only modest effectiveness. Albertina PT-PT outperforms BERTimbau Large (MRR = 0.206 vs. 0.157), suggesting that training on European Portuguese data provides limited advantages. Nevertheless, both embedding-based approaches remain far behind multimodal methods, reinforcing the importance of incorporating visual grounding for effective image retrieval. For proprietary black-box retrieval systems, performance is intermediate. Google Images achieves results comparable to traditional IR methods (MRR = 0.336) but remains well below vision–language models, while Arquivo.pt performs less competitively (MRR = 0.217), reflecting the limitations of keyword-based retrieval when applied to visually driven, query-based image search. Together, these results highlight opportunities to improve Portuguese language coverage in widely used image search engines.

Despite the clear advantages of vision–language models over other baselines, the results remain far from perfect. In particular, the modest F1@10 values and the sensitivity of performance to query formulation indicate that query-driven image retrieval in European Portuguese remains an open and challenging problem. Rather than reflecting limitations of the dataset, these results demonstrate its usefulness as a long-term benchmark, enabling future work to meaningfully compare models, analyze failure cases, and measure progress as multimodal retrieval techniques continue to evolve.

Impact of Query Length. To better understand variations across query types, we compared longer model-generated queries (38 queries) with shorter manually

created queries (42 queries). Figure 4 shows that vision-language models consistently favor concise formulations: OpenCLIP achieves 71% higher MRR (0.760 vs 0.445) and 88% higher F1@10 (0.371 vs 0.197) on short queries. This behavior reflects CLIP's original training design, optimized for image classification tasks using concise category labels rather than verbose descriptions.

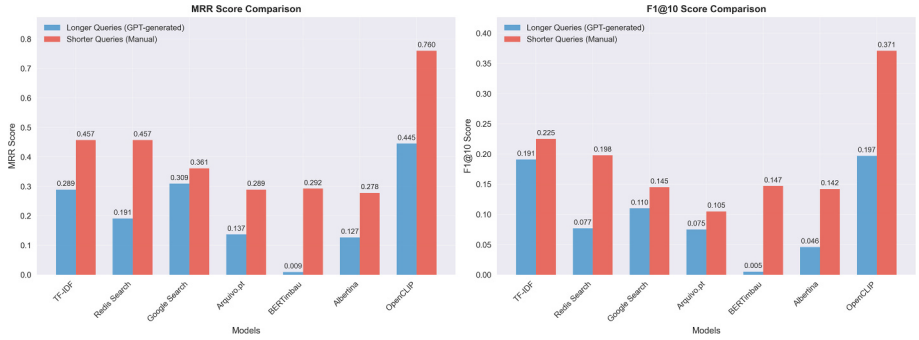

Fig. 4. Performance comparison between longer (GPT-generated) and shorter (manual) queries. Vision-language models show pronounced preference for shorter, focused queries.

In contrast, traditional text-based baselines (TF-IDF, BM25) show relatively stable performance, with differences across query lengths, with differences typically below 5% for both MRR and F1@10 metrics. This stability suggests that term-frequency approaches are less sensitive to query length variations and can effectively process both concise and verbose formulations. Overall, these results suggest that vision-language models may be less effective with naturally verbose user queries. This observation points to a potential avenue for future research: developing models that can more effectively interpret and retrieve images from longer, descriptive queries without requiring users to simplify their search behavior.

5 Conclusion and Future Work

In this paper, we introduced the `pt-image-ir-dataset`, a new resource aimed at advancing research in image IR for European Portuguese. The dataset comprises 80 diverse, thematically curated queries and a collection of 5,201 images sourced from the official website of the Portuguese Presidency. Each image was annotated for relevance by three independent human judges, following a structured set of annotation guidelines to ensure consistency and reliability. We assessed annotation quality using Fleiss' Kappa, obtaining a coefficient of 0.62, which reflects substantial inter-annotator agreement. By capturing the linguistic and cultural specificities of European Portuguese, this dataset addresses a significant

gap in text-image IR research for low-to-mid resource languages, offering a valuable tool for academic research and applied multimodal IR tasks. To evaluate the dataset's effectiveness for image retrieval tasks, we conducted a series of experiments covering four main paradigms: traditional lexical IR approaches, vision-language models, proprietary black-box image retrieval systems, and semantic dense retrieval methods based on Portuguese language embeddings. To further examine annotation quality and query characteristics, we analyzed inter-annotator agreement variations across different query generation methods and semantic categories, and investigated the impact of query length on retrieval effectiveness. Overall, the results demonstrate that vision-language models, particularly *OpenCLIP/xlm-roberta-base-ViT-B-32*, outperform other methods by a significant margin. Analyses of inter-annotator agreement and query characteristics further reveal nuanced patterns: agreement tends to be higher for manually created, concise queries, while longer, AI-generated queries are associated with lower consistency among annotators and reduced retrieval effectiveness. These findings highlight both the potential of the `pt-image-ir-dataset` as a benchmark for European Portuguese image retrieval and the importance of considering a fine-tuning strategy to the linguistic and contextual characteristics of the texts in future system development. While the current `pt-image-ir-dataset` focuses exclusively on high-quality photographs, future extensions could explore stylized or generated image variants, such as line drawings, watercolors, or black-and-white renditions, to examine how retrieval models handle different visual styles.

Limitations

While the `pt-image-ir-dataset` is a resource for Portuguese image retrieval, it is derived from an institutional source, which may limit its generalization to broader domains such as e-commerce, social media, or user-generated content. Although alternative sources (e.g., news outlets or social platforms) could provide topical diversity and more informal imagery, they pose challenges related to licensing, structural consistency, and collection. In contrast, the Presidency website offers curated and structured content, making it a starting point for Portuguese text-to-image retrieval research. From a methodological standpoint, our evaluation is restricted to single-paradigm baselines, excluding hybrid or multi-stage retrieval approaches that combine lexical and vision–language signals. While such methods are widely adopted and often achieve performance, their study is left for future work, which this dataset can support. Additionally, the dataset currently uses binary relevance annotations. Adopting a graded relevance scheme could enable evaluation and represents an extension. Finally, despite covering diverse themes, the dataset contains only 80 queries. Expanding this number in future releases would improve evaluation robustness and coverage of Portuguese image retrieval challenges.

FAIR Principles

To ensure that the dataset could be shared and used, we adhered to established ethical principles and best practices, following the FAIR (Findable, Accessible, Interoperable, and Reusable) guidelines [31]. All data was collected from publicly available sources, with attention to copyright and privacy. Specifically, images were scraped exclusively from the official Portuguese Presidency website, a public institutional resource, and each item is linked back to the original source. The dataset is distributed under a Creative Commons Attribution-NonCommercial License, which ensures proper attribution and enables research use. To further enhance accessibility and availability, the dataset is hosted on Zenodo.

Acknowledgments. Ricardo Campos is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025). The author would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC). The authors Rodrigo Duarte and António Branco would like to acknowledge ACCELERAT.AI - Multilingual Intelligent Contact Centers, funded by the covid-recovery program PRR-Plano de Recuperação e Resiliência, through IAPMEI (C625734525-00462629); PORTULAN CLARIN - Research Infrastructure for the Science and Technology of Language, funded by LISBOA2030 (FEDER-01316900); hey, Hal, curb your hallucination!, funded by FCT-Fundação para a Ciência e Tecnologia (2024.07592.IACDC). The author Hugo Proença would like to acknowledge FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier <https://doi.org/10.54499/UID/50008/2025>.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ. Psychol. Measur. **33**(3), 613–619 (1973)
2. Rodrigues, J., et al.: Advancing neural encoding of Portuguese with transformer Albertina PT-*. arXiv preprint arXiv:2305.06721 (2023)
3. Alsmadi, M.K.: Content-based image retrieval using color, shape and texture descriptors and features. Arab. J. Sci. Eng. **45**(4), 3317–3330 (2020)
4. Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-text retrieval: a survey on recent research and development. arXiv preprint arXiv:2203.14713 (2022)
5. Carlsson, F., Eisen, P., Rekathati, F., Sahlgren, M.: Cross-lingual and multilingual CLIP. In: Proceedings of the Language Resources and Evaluation Conference, pp. 6848–6854. European Language Resources Association, Marseille, France (2022)
6. Cherti, M., et al.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2818–2829 (2023)

7. Han, X., et al.: Automatic spatially-aware fashion concept discovery. In: ICCV (2017)
8. Harman, D.: Overview of the first TREC conference. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 36–47 (1993)
9. Ilharco, G., et al.: OpenCLIP. Zenodo, version 0.1 (2021). https://doi.org/10.5281/zenodo.5143773
10. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
11. Jones, K.S., Van Rijsbergen, C.J., British Library. Research and Development Department: Report on the Need for and Provision of an Ideal Information Retrieval Test Collection. University Computer Laboratory (1975)
12. Lahajal, N.K., Harini, S.: Enhancing image retrieval: a comprehensive study on photo search using the CLIP mode. arXiv arxiv:2401.13613 (2024)
13. Li, X., Yang, J., Ma, J.: Recent developments of content-based image retrieval (CBIR). Neurocomputing **452**, 675–689 (2021)
14. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
15. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
16. Luo, M., Gokhale, T., Varshney, N., Yang, Y., Baral, C.: Advances in Multimodal Information Retrieval and Generation. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-57816-8
17. Osmulsk, R., Moreira, G.S.P., Ak, R., Xu, M., Schifferer, B., Oldridge, E.: MIRACL-VISION: a large, multilingual, visual document retrieval benchmark. arXiv preprint arXiv:2505.11651 (2025)
18. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
19. dos Santos, G.O., Colombini, E.L., Avila, S.: #PraCegoVer: a large dataset for image captioning in Portuguese. Data **7**(2) (2022). Article 13
20. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
21. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: BM25 and beyond. Found. Trends® Inf. Retrieval **3**(4), 333–389 (2009)
22. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: CLIP for all things zero-shot sketch-based image retrieval, fine-grained or not. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2765–2775 (2023)
23. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
24. Schall, K., Barthel, K., Hezel, N., Jung, K.: Optimizing CLIP models for image retrieval with maintained joint-embedding alignment. arXiv arxiv:2409.01936 (2024)

25. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) BRACIS 2020. LNCS (LNAI), vol. 12319, pp. 403–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28
26. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) BRACIS 2020. LNCS (LNAI), vol. 12319, pp. 403–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28
27. Van Gysel, C., de Rijke, M.: Pytrec_eval: an extremely fast Python interface to trec_eval. In: SIGIR. ACM (2018)
28. Vaswani, A.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
29. Voorhees, E.M., Harman, D.: Overview of TREC 2001. In: TREC (2001)
30. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)
31. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
32. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986 (2023)