UNIVERSIDADE
BEIRA INTERIOR

Combining Text and Visual Modalities for Enhanced Portuguese Image Retrieval

Rodrigo Manuel Teixeira Duarte

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Ricardo Nuno Taborda Campos
Co-orientador: Prof. Doutor Hugo Pedro Martins Carriço Proença

Covilhã, outubro de 2025

Declaração de Integridade

Eu, Rodrigo Manuel Teixeira Duarte, que abaixo assino, estudante com o número de inscrição M13305 de Engenharia Informática da Faculdade de Engenharias, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 10/10/2025

Assinado por: **Rodrigo Manuel Teixeira Duarte**
Num. de Identificação: 15945963
Data: 2025.10.10 11:10:34+01'00'

Agradecimentos

This work has been carried out as part of the project "Accelerat.AI, Ref. C644865762-00000008", financed by IAPMEI and European Union — Next Generation EU Fund within the scope of call for proposals no. 02/C05-i01/2022 — submission of final Proposals for Project Development under the Mobilizing Agendas for Business Innovation of the Recovery and Resilience Plan.

I would like to express my deepest gratitude to my supervisors, Prof. Dr. Ricardo Nuno Taborda Campos and Prof. Dr. Hugo Pedro Martins Carriço Proença, for their guidance, support, and encouragement throughout this thesis. Their expertise and insights have been invaluable in shaping the direction of this research work.

I would also like to thank my family and friends for their unwavering support and encouragement throughout this journey. Their love and understanding have been a constant source of strength and motivation.

Combining Text and Visual Modalities for Enhanced Portuguese Image Retrieval

Resumo

A disponibilidade de imagens digitais na Internet tem crescido exponencialmente nos últimos anos. Isto tornou difícil para os utilizadores encontrarem imagens relevantes no contexto de tarefas de Recuperação de Informação (RI), uma vez que os motores de pesquisa frequentemente não conseguem compreender o seu conteúdo de forma precisa. Este desafio torna-se ainda maior quando se procuram imagens em línguas que não o inglês — especialmente línguas de recursos baixos a médios como o português, que frequentemente necessitam dos recursos linguísticos necessários. Para abordar estas questões, vários métodos têm sido propostos, como a utilização de modelos de linguagem multimodais que tentam compreender tanto o conteúdo da imagem como a informação textual associada. No entanto, a maioria destes modelos são afinados principalmente para a língua inglesa. Outra estratégia comum envolve modelos de tradução linguística, onde as procuras numa língua-alvo são traduzidas para inglês antes de serem processadas. Contudo, tal solução também não é perfeita, pois os detalhes da procura podem-se perder na tradução, levando a resultados subóptimos. Esta dissertação de mestrado aborda este desafio através do desenvolvimento e avaliação de abordagens multimodais para recuperação de imagens em português, com um foco específico na compreensão das limitações e oportunidades dos modelos visão-linguagem atuais. A nossa hipótese é que combinar modalidades de recuperação baseadas em texto e em imagem através de mecanismos inovadores de ajuste de pontuação, que levará a resultados mais eficazes do que abordagens que apenas utilizam uma modalidade. O objetivo principal desta investigação foi desenvolver um sistema de RI de imagens eficaz para pesquisas em português e estabelecer *baselines* de desempenho abrangentes para este domínio. Para tal, criámos um conjunto de dados de avaliação de recuperação de imagens em português com 80 pesquisas e 5.201 imagens anotadas do website da Presidência portuguesa. Desenvolvemos um algoritmo híbrido de recuperação que combina recuperação baseada em texto e em imagem através de mecanismos matemáticos de ajuste de pontuação em conjunto com o algoritmo K-Nearest Neighbors (KNN) para correspondência de similaridade. A nossa avaliação englobou métodos tradicionais de RI baseados em texto, motores de busca comerciais, modelos de linguagem específicos para português e modelos visão-linguagem estado-da-arte. Os resultados revelaram que modelos visão-linguagem multilingues, particularmente o OpenCLIP *xlm-roberta-base*, superaram substancialmente as abordagens tradicionais baseadas em texto em 62% nas pontuações MRR, alcançando 71% melhor desempenho com consultas mais curtas comparativamente a formulações descritivas mais longas. Surpreendentemente, as experiências de afinação mostraram desempenho diminuído em todas as métricas, com degradações que variaram entre 16% e 28%, sugerindo que representações multilingues prétreinadas são mais valiosas que adaptações específicas do domínio. O algoritmo híbrido proposto alcançou melhorias significativas, com um aumento de 1.8% no Mean Reciprocal Rank (MRR) em relação à melhor abordagem *baseline*.

Palavras-chave

Modelos Multimodais, Recuperação de Informação de Imagens, Recuperação de Informação Baseada em Conteúdo, Processamento de Linguagem Natural, Visão Computacional

Combining Text and Visual Modalities for Enhanced Portuguese Image Retrieval

Resumo alargado

Esta dissertação aborda o desenvolvimento de um sistema de recuperação de informação multimodal para imagens em contexto português, especificamente direcionado para superar as limitações dos sistemas de pesquisa atuais quando aplicados a línguas de recursos baixos a médios. Este estudo é motivado pela necessidade crítica de inovações tecnológicas que permitam uma pesquisa eficaz de conteúdo visual em português, uma lacuna significativa na era digital atual onde a quantidade de imagens disponíveis online cresce exponencialmente. O método proposto combina recuperação baseada em texto e em imagem através de um algoritmo híbrido inovador que utiliza mecanismos matemáticos de ajuste de pontuação. O ponto principal da metodologia envolve a aplicação de modelos visão-linguagem estado-da-arte, particularmente adaptações multilingues do CLIP, com um algoritmo de ajuste de pontuação linear projetado para equilibrar as contribuições de diferentes modalidades de recuperação, melhorando assim a precisão e a relevância dos resultados para pesquisas em português.

Para validar esta abordagem, foi desenvolvido um conjunto de dados de avaliação específico para recuperação de imagens em português, compreendendo 80 pesquisas e 5.201 imagens anotadas obtidas do website oficial da Presidência da República Portuguesa. Este *dataset* constitui uma contribuição valiosa para a comunidade de investigação, preenchendo uma lacuna na disponibilidade de recursos de avaliação para recuperação de informação visual em português europeu.

Além da introdução do algoritmo híbrido, esta dissertação aborda o estado da arte, comparando várias abordagens desde métodos tradicionais de recuperação de informação baseados em texto até modelos visão-linguagem multimodais, destacando as suas vantagens e limitações quando aplicados ao contexto específico da língua portuguesa. O sistema proposto é avaliado através de métricas padrão da área de recuperação de informação, incluindo Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), e métricas de precisão e *recall* em diferentes pontos de corte.

Uma descoberta particularmente interessante desta investigação foi que os experimentos de afinação, contrariamente às expectativas convencionais, resultaram numa diminuição do desempenho em todas as métricas avaliadas, com degradações que variaram entre 16% e 28% comparativamente ao modelo multilíngua base. Este resultado contraditório sugere que as representações multilingues pré-treinadas em larga escala são mais valiosas para a recuperação de imagens em português do que adaptações específicas do domínio, contrariando pressupostos estabelecidos sobre a necessidade de especialização de modelos.

Os resultados detalhados neste projeto de investigação demonstram que os modelos visão-linguagem multilingues, particularmente o OpenCLIP *xlm-roberta-base*, superam substancialmente todas as outras abordagens testadas, incluindo métodos tradicionais de recuperação de informação (62% melhor desempenho MRR), motores de busca comerciais e modelos de linguagem específicos para português. O algoritmo híbrido desenvolvido conseguiu alcançar melhorias significativas, com um aumento de 1.8% no Mean Reciprocal Rank comparativamente à melhor abordagem *baseline*, demonstrando a eficácia da combinação de modalidades através de mecanismos de ajuste de pontuação.

Em conclusão, esta dissertação propõe uma nova metodologia que tem o potencial de melhorar a recuperação de imagens para utilizadores de língua portuguesa. Os resultados obtidos não só validam o desempenho do algoritmo híbrido proposto mas também estabelecem *baselines* para futuras investigações na área, abrindo caminhos para o desenvolvimento de sistemas de recuperação de informação visual mais eficazes para línguas de recursos limitados.

Abstract

The availability of digital images on the Internet has grown exponentially in recent years. This has made it challenging for users to find relevant images in the context of Information Retrieval (IR) tasks, as search engines are often unable to understand their content accurately. This challenge becomes even greater when searching for images in languages other than English — especially low-to-mid resource languages like Portuguese, which often lack the necessary linguistic resources. To address these issues, several approaches have been proposed, such as using multimodal language models that attempt to understand both image content and associated textual information. However, most of these models are fine-tuned primarily for the English language. Another common strategy involves language translation models, where queries in a target language are translated into English before being processed. However, such a solution is also not perfect as the meaning of the query can be lost in translation, leading to suboptimal results. This MSc thesis tackles this challenge by developing and evaluating multimodal approaches for Portuguese image retrieval, with a specific focus on understanding the limitations and opportunities of current vision-language models. Our hypothesis is that combining text-based and image-based retrieval modalities through innovative score adjustment mechanisms will lead to more effective results than individual approaches alone. The primary objective of this research is to develop an effective image IR system for Portuguese queries and establish performance baselines for this domain. To achieve this, we created a Portuguese image retrieval evaluation dataset comprising 80 queries and 5,201 annotated images from the Portuguese Presidency website. We developed a novel hybrid retrieval algorithm that combines text-based and image-based retrieval through mathematical score adjustment mechanisms, utilizing K-Nearest Neighbors (KNN) algorithms for similarity matching. Our comprehensive evaluation encompassed traditional text-based IR methods, commercial search engines, Portuguese-specific language models, and state-of-the-art vision-language models. The results revealed that multilingual vision-language models, particularly OpenCLIP *xlm-roberta-base*, substantially outperformed traditional text-based approaches by 62% in MRR scores, achieving 71% better performance with shorter queries compared to longer descriptive formulations. Surprisingly, fine-tuning experiments showed decreased performance across all metrics, with degradations ranging from 16% to 28%, suggesting that pre-trained multilingual representations are more valuable than domain-specific adaptations. The proposed hybrid algorithm achieved meaningful improvements, with a 1.8% enhancement in Mean Reciprocal Rank over the best baseline approach.

Keywords

Multimodal Models, Image Information Retrieval, Content Based Information Retrieval, Nat-

ural Language Processing, Computer Vision

Contents

List of Figures

List of Tables

Acronym List

AI	Artificial Intelligence
ALIGN	A Large-scale ImaGe and Noisy-text embedding
BLIP	Bootstrapping Language-Image Pre-training
BPE	Byte Pair Encoding
CBIR	Content-Based Image Retrieval
CDIR	Cross-Domain Image Retrieval
CL	Contrastive Learning
CLIP	Contrastive Language-Image Pre-training
CM	Caption-Matching
CNN	Convolutional Neural Network
CV	Computer Vision
DNN	Deep Neural Network
DOM	Document Object Model
HTML	HyperText Markup Language
IAA	Inter-Annotator Agreement
IR	Information Retrieval
ITC	Image-Text Contrastive Learning
ITG	Image-Grounded Text Ceneration
ITM	Image-Text Matching
JSON	JavaScript Object Notation
KNN	K-Nearest Neighbors
LIRE	Lucene Image Retrieval
LLM	Large Language Model
MAP	Mean Average Precision
ML	Machine Learning
MRR	Mean Reciprocal Rank

NLP	Natural Language Processing
NSFW	Not Safe For Work
Q-Former	Querying Transformer
RRF	Reciprocal Rank Fusion
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference
URL	Uniform Resource Locator
ViT	Vision Transformer
VITR	ViT-Relation-Focus
WBIT	Wikipedia-Based Image Text
WebIT	WebImageText

Chapter 1

Introduction

This chapter introduces the research topic and provides the context for this thesis. Section 1.1 presents the motivation behind this work and defines its scope. Section 1.2 formally defines the research problem and highlights the limitations of current approaches. Section 1.3 details the objectives and research questions that guide this investigation. Section 1.4 outlines the methodology and tasks undertaken to achieve these objectives. Section 1.5 summarizes the key contributions of this research. Finally, Section 1.6 provides an overview of the document's structure.

1.1 Motivation and Scope

In today's digital age, visual content has become increasingly important, fundamentally changing how users interact with and consume digital media [1]. The rapid growth of images spread across various platforms, including social media, e-commerce websites, governmental archives, and educational resources, has created significant challenges in effectively retrieving relevant visual content from large collections. Traditional text-based search methods, while effective for textual information, show significant limitations when applied to visual data, making it necessary to develop specialized Information Retrieval (IR) systems that can understand and navigate the complex nature of visual content.

Recent advances in Machine Learning (ML) and Computer Vision (CV) have created new opportunities for developing sophisticated image retrieval systems that go beyond simple cataloging approaches [7]. For instance, modern image IR systems are designed to analyze visual content in detail, allowing users to retrieve images based on complex semantic attributes, contextual relationships, and specific user needs. By utilizing state-of-the-art algorithms and deep learning architectures, these systems can significantly enhance the user experience, resulting in more intuitive interfaces that better align with how humans perceive and comprehend visual information.

Moreover, effective image IR systems have important applications beyond the user's convenience. In healthcare [8, 9, 10, 11], the ability to quickly access relevant medical images can improve clinical decision-making, speed diagnostic processes, and potentially save lives. Similarly, in law enforcement [12, 13, 14], efficient visual data retrieval can enhance investigations and public safety operations. As we continue to face challenges from information overload [15, 16, 17], advanced image IR systems become crucial tools to unlock the potential of visual data and transform the way we access and use the vast amount of imagery available online.

Despite many research efforts on image retrieval, one of the most promising approaches involves the use of multimodal language models that combine the strengths of CV and Natural Language Processing (NLP) to understand both visual and textual content. However, a ma-

jor limitation remains, as most of these sophisticated models are designed and optimized for English, significantly limiting their effectiveness in other languages, including Portuguese. Although some research has tried to address this through machine translation, translating queries from Portuguese to English may introduce semantic distortions that can change the original query's meaning, leading to poor retrieval results.

These limitations are clearly visible in real-world scenarios. Figure 1.1 and Figure 1.2 show the limited performance of current systems when processing the Portuguese query "Presidente a ler um livro" (President reading a book) on both the official Portuguese Presidency website and Google Images.

Figure 1.1: Search results from the Portuguese Presidency website for the query "Presidente a ler um livro", showing limitations in image retrieval.

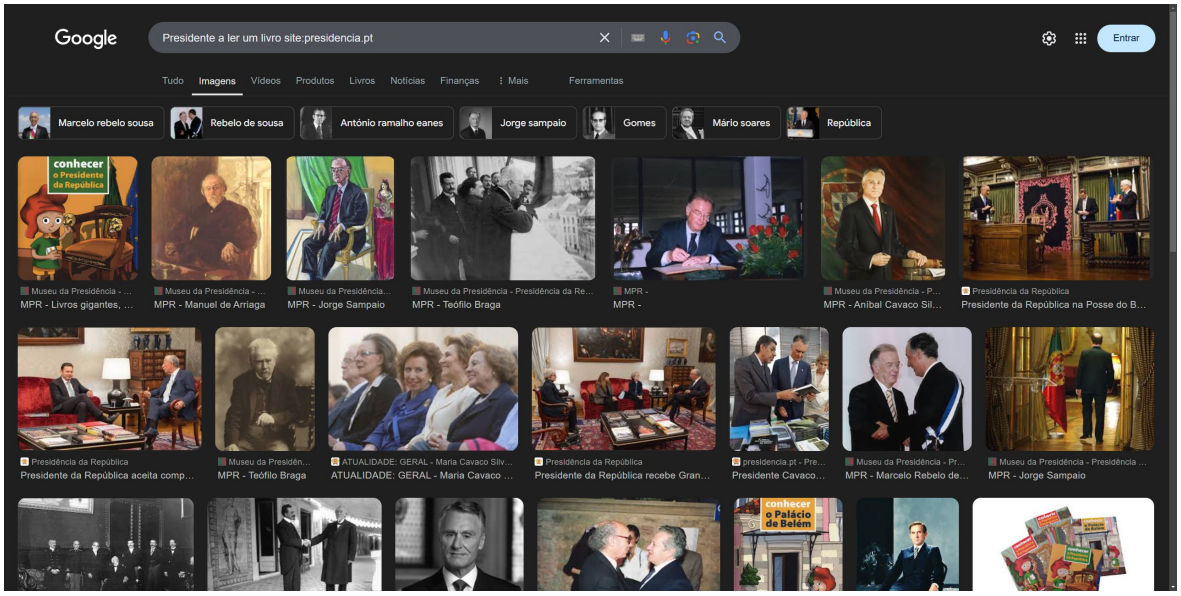

Figure 1.2: Google Images search results for the query "Presidente a ler um livro", demonstrating poor semantic alignment with the query.

The analysis of these results shows fundamental problems with the current image search methods for Portuguese queries. The official Portuguese Presidency website, which lacks a dedicated image search engine, mainly returns article-based results with images that have little to no relevance to the user's specific visual query. While these images may appear in articles somewhat related to the search terms, they fail to meet the user's clear intent to find images of the Portuguese President reading. Google Images, despite focusing exclusively on visual content, shows similar limitations in understanding context, returning images that lack a meaningful connection to the query's intended meaning. Even when limiting the search to the official Portuguese Presidency website, the system fails to identify images that accurately represent the scenario described in the Portuguese query. These failures highlight the need for developing specialized image IR systems that can process Portuguese queries with better semantic understanding and contextual accuracy.

1.2 Problem Statement

Multimodal language models have transformed the way different types of content — text, images, audio, and video can be processed together. Their integration into IR systems has significantly improved image search capabilities by allowing queries to consider both textual descriptions and visual content. However, important limitations remain when applying these models beyond English-language contexts.

The main challenge is that most current multimodal language models are designed and optimized primarily for English, severely limiting their effectiveness in other languages, particularly Portuguese. This limitation is made worse by the relative lack of research and computational resources dedicated to low-to-mid resource languages, creating a significant gap that prevents the adaptation of these advanced models to Portuguese-language image IR applications.

Current approaches try to solve this problem, typically utilizing machine translation strategies, where Portuguese queries are first translated to English before being processed by English-optimized models. However, this approach creates systematic problems because translation processes often fail to preserve nuanced meanings, cultural references, and context that are crucial for accurate image retrieval. As a result, this method frequently produces poor or irrelevant results that do not align with the user's information needs.

Furthermore, current search engines show persistent problems in retrieving images that align in a meaningful way with user queries, as demonstrated in Figures 1.1 and 1.2. These systems often rely on simple keyword matching approaches that fail to capture the semantic complexity inherent in natural language queries, resulting in outputs that may be loosely related to search terms but lack genuine relevance to user intent.

To address these limitations, this thesis proposes investigating new approaches for improving Portuguese image IR performance without relying on translation steps. Our research explores multiple strategies, including fine-tuning existing multimodal models on Portuguese-language datasets and developing hybrid retrieval algorithms that combine text-based and image-based approaches. This comprehensive approach aims to ensure better semantic un-

derstanding and context-aware search experiences while preserving the linguistic and cultural nuances inherent in Portuguese queries.

1.3 Objectives

This research aims to combine recent advances in CV, ML, and IR to design and evaluate image IR systems, with particular attention to underexplored Portuguese-language contexts.

- **OBJ1:** Investigate the application of state-of-the-art multimodal language models in Portuguese image IR systems;
- **OBJ2:** Research multilingual approaches specifically adapted for European Portuguese, evaluating existing solutions and identifying opportunities for improved performance in Portuguese-language contexts;
- **OBJ3:** Evaluate the effectiveness and limitations of current multimodal language models when applied to Portuguese image IR tasks, establishing baseline performance metrics and identifying areas for improvement;
- **OBJ4:** Analyze the impact and measure the benefits of fine-tuning multimodal language models using Portuguese-language datasets, investigating optimal training strategies and measuring performance improvements across various evaluation metrics;
- **OBJ5:** Investigate retrieval strategies that combine multimodal signals (e.g., text and images) and evaluate their suitability for European Portuguese tasks;
- **OBJ6:** Develop and deploy a demonstration web application that showcases the practical applicability and effectiveness of the proposed image IR system, using Portuguese-language scenarios as a case study.

Based on the objectives outlined above, we defined four main research questions:

- **RQ1:** What challenges arise when integrating text and image modalities for IR tasks?
- **RQ2:** What are the limitations of existing multilingual solutions for low-to-mid resource languages like Portuguese in the context of image IR?
- **RQ3:** Can fine-tuning multimodal language models for the Portuguese language improve the performance of image IR systems?
- **RQ4:** To what extent can hybrid retrieval approaches that combine text and image-based methods improve image IR performance in Portuguese-language scenarios?

1.4 Tasks Overview

To achieve the research objectives outlined in this thesis, a methodology composed of seven tasks has been developed. These tasks are organized to build on each other, ensuring a logical

progression from theoretical foundation to practical implementation and evaluation. Figure 1.3 presents the temporal organization and interdependencies of these tasks through a Gantt diagram, which illustrates the research workflow over the course of the study.

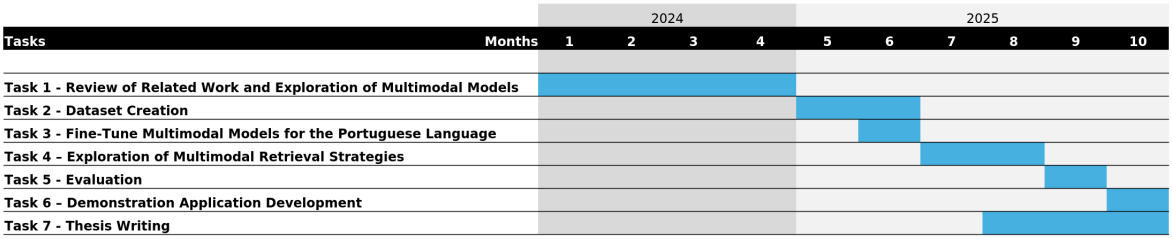

Figure 1.3: Gantt diagram showing the timeline and relationships of the research tasks.

Task 1 — Review of Related Work and Exploration of Multimodal Models

The initial phase involved a review and analysis of current research in image IR systems and multimodal language models, combined with the investigation and evaluation of state-of-the-art multimodal models for image IR applications. The objective was to provide a comprehensive theoretical framework of current methods, technological limitations, and emerging research directions. The research included a detailed analysis of model architectures, semantic representation capabilities, and performance characteristics when processing both visual content and textual information. In particular, it focused on the latest advances in multimodal architectures, cross-modal learning techniques, and their applications to image IR tasks, with particular attention to multilingual capabilities and Portuguese-language implementations. This analysis also explored how these models could be effectively adapted for Portuguese-language contexts and identified the specific modifications or training strategies needed to optimize their performance for Portuguese image IR systems.

Task 2 — Dataset Creation

This task focused on the construction of a Portuguese image retrieval evaluation dataset. Using images sourced from the official Portuguese Presidency website, we developed European Portuguese queries across multiple categories and created annotated image-query pairs with binary relevance judgments. The resulting dataset provides a comprehensive evaluation resource for assessing Portuguese image IR system performance and enables systematic comparison of different retrieval approaches.

Task 3 — Fine-Tune Multimodal Models for the Portuguese Language

This phase focused on fine-tuning selected multimodal models specifically for Portuguese-language image IR applications. The process involved developing optimal training strategies, hyperparameter optimization, and learning approaches to adapt pre-trained models to Portuguese linguistic and cultural contexts effectively. This task was essential to understand the impact of language-specific fine-tuning on retrieval performance and to establish the effectiveness of adaptation strategies for low-resource language applications.

Task 4 — Exploration of Multimodal Retrieval Strategies

This task involved the design and development of retrieval approaches that integrated text-based and image-based methods, aiming to address the limitations of relying on a single modality. The focus was on creating innovative combinations that leveraged pre-trained multilingual models to maximize effectiveness. One of the main approaches developed followed a three-phase process: dual retrieval, score adjustment, and result combination, with a critical contribution being the score adjustment mechanism that balanced the influence of both modalities through different adjustment functions. The evaluation of these methods was carried out in Portuguese-language scenarios, used as a case study to assess their applicability in real-world contexts.

Task 5 — Evaluation

The experimental phase focused on evaluating multiple approaches for Portuguese image IR using the dataset developed in Task 2. The evaluation comprises traditional text-based IR methods, commercial web search engines, Portuguese-specific language models, state-of-the-art vision-language models, a fine-tuned model, and a hybrid retrieval approach designed to combine text and image-based methods. The assessment included quantitative performance metrics across multiple evaluation measures and analysis of query characteristics, fine-tuning effectiveness, and score adjustment mechanisms.

Task 6 — Demonstration Application Development

This task involved the development and deployment of a comprehensive demonstration web application that validates the practical applicability of the hybrid retrieval approach developed in Task 3. The application serves as both a demonstration platform and a research tool, integrating the methodology and algorithms into a fully functional system that showcases the capabilities of multimodal Portuguese image retrieval. The implementation includes system architecture design, user interface development, advanced monitoring capabilities, and API integration features that enable real-world deployment scenarios.

Task 7 — Thesis Writing

The final task of this thesis involved the writing of this document. This includes the research methodology, experimental procedures, presentation of findings, the discussion of limitations, and future research directions in Portuguese image IR systems.

1.5 Contributions

In line with the research objectives described above, this thesis provides the following contributions:

Combining Text and Visual Modalities for Enhanced Portuguese Image Retrieval

- A state-of-the-art analysis of current image retrieval systems and multimodal language models, providing critical evaluation of existing methods and systematic documentation of recent advances in multimodal architectures and their applications to multilingual image IR tasks;
- A Portuguese image retrieval dataset comprising European Portuguese queries and annotated relevance judgments;
- A hybrid retrieval approach that combines text-based and image-based methods using multimodal models specifically adapted for Portuguese-language contexts;
- A demonstration web application that validates the practical applicability of the proposed hybrid retrieval approach through an intuitive user interface and real-world usage scenarios;

1.6 Document Organization

This document is organized into six chapters. The current chapter, Chapter 1 — **Introduction**, establishes the foundational context, research motivation, objectives, research questions, tasks that guide this investigation, and contributions. The structure and content of the remaining chapters are described in the following.

- Chapter 2 — **State-of-the-Art** — presents the analysis of fundamental concepts, theoretical frameworks, and current research developments in image retrieval systems and multimodal language models;
- Chapter 3 — **pt-image-ir-dataset** — details the methodology used for developing the Portuguese image retrieval dataset, including the documentation of data collection procedures, annotation protocols, quality assurance measures, and dataset characterization;
- Chapter 4 — **Proposed Hybrid Retrieval Algorithm** — describes the hybrid retrieval approach developed for image IR systems, including the methodology development process, hybrid algorithm design;
- Chapter 5 — **Experiments and Results** — presents a detailed analysis of experimental outcomes, including quantitative performance metrics, comparative evaluations, and qualitative assessment of the various baseline methods and of our own methodology;
- Chapter 6 — **Demonstration Web Application** — presents the demonstration web application that validates the practical applicability of the approach, detailing the system architecture, user interface design, deployment considerations, and integration capabilities;

- Chapter 7 — **Conclusions and Future Work** — Presents the main conclusions derived from this research and outlines possible directions for future research in Portuguese image IR systems.

Chapter 2

State-of-the-Art

This chapter examines the state-of-the-art in image IR systems, providing foundational knowledge and current approaches in the field. Section 2.1 introduces key terminologies essential for understanding the research presented in this thesis. Section 2.2 reviews existing literature in IR, CV, and multimodal networks, examining traditional retrieval methods and recent advances in vision-language models including Contrastive Language-Image Pre-training (CLIP) and Bootstrapping Language-Image Pre-training (BLIP)-2. Finally, Section 2.2.4 concludes with a summary of key insights and their relevance to Portuguese image retrieval challenges.

2.1 Terminology and Preliminary Concepts

This section defines key concepts and terminologies essential for understanding the research presented in this thesis. Each term provides the foundational knowledge necessary to understand the methodologies and approaches discussed in subsequent chapters.

- **Information retrieval** is the process of retrieving relevant information from data collections based on user queries or search terms. It involves indexing and searching through datasets to find documents, images, or other information that matches user's search criteria. Ad-hoc retrieval systems are utilized across various domains, including web search engines, digital libraries, and e-commerce platforms.

- **Deep Learning** represents a subfield of ML that focuses on training neural networks to learn patterns and representations from data. Deep neural networks enable models to perform tasks such as image classification and object recognition. This approach has transformed computer vision by enabling accurate models for processing visual data, with the ability to learn from large-scale datasets and generalize to unseen data, making them suitable for image retrieval tasks.

- **Contrastive Learning (CL)** is a ML technique that learns data representations by contrasting positive and negative examples. By training models to maximize similarity between positive examples and minimize similarity between negative examples, discriminative representations of data can be learned. This technique is used by models like CLIP to learn semantic relationships between images and text, enabling cross-modal retrieval tasks.

- **Large Language Model (LLM)** is a class of deep learning models, typically based on the Transformer architecture, trained on large-scale text corpora to capture semantic and syntactic patterns. By leveraging their scale and training data, LLMs can perform a wide range of natural language processing tasks such as text generation, translation, and sentiment analysis. LLMs can understand and generate human-like text, making

them suitable for processing textual data in image retrieval systems. The Albertina PT-PT model exemplifies an LLM used in various natural language processing tasks such as embedding generation which can be used for indirect image retrieval.

- **Zero-shot learning** is a machine learning paradigm that recognizes objects or concepts without labeled training data of the corresponding class. Instead of relying on annotated examples, zero-shot learning models generalize from seen classes to unseen classes, enabling recognition of novel objects or concepts based on semantic relationships. In image retrieval, zero-shot learning can be applied by training models on datasets of images and corresponding textual descriptions. By learning semantic relationships between visual and textual content, models can retrieve relevant images based on textual queries, even if images have not been seen during training. This approach eliminates the need for manual annotations, making it scalable and adaptable to different domains.

- **Transfer learning**, also known as knowledge distillation, transfers knowledge from a complex model (teacher) to a simpler model (student). By training the student model to mimic the teacher model predictions, knowledge, and representations learned by the teacher model are distilled into the student model, enabling it to perform complex tasks with fewer parameters and computational resources. This approach reduces the size and complexity of deep learning models while maintaining performance across various tasks.

- **Fine-tuning** is a transfer learning technique where a pre-trained model is further trained on a new dataset to adapt it to a specific task or domain. By fine-tuning a pre-trained model on a new dataset, knowledge, and representations learned on large-scale datasets can be leveraged and applied to specific problems. This approach is particularly relevant in image retrieval systems, where pre-trained models can be fine-tuned on datasets of images and corresponding textual descriptions to learn semantic relationships between visual and textual content.

- **Embeddings** are vector representations of data points in high-dimensional space, where each dimension corresponds to a feature or attribute of the data point. By embedding data points into a continuous vector space, relationships between them can be captured to perform tasks such as clustering, classification, and retrieval. In image retrieval contexts, embeddings represent images and text in a common feature space, enabling comparison and matching of images based on visual and textual content.

2.2 Related Work

This section presents a review of the existing literature related to image information retrieval systems. The review encompasses fundamental IR concepts, CV techniques, and recent advances in multimodal learning architectures. The analysis covers three main areas: text-based information retrieval foundations (Section 2.2.1), CV methodologies for visual content

processing (Section 2.2.2), and multimodal learning that bridges visual and textual content understanding (Section 2.2.3).

2.2.1 Information Retrieval

IR is a field of study focused on retrieving relevant information from data collections based on user queries or search terms. The field involves indexing and searching datasets to find documents, images, or information that match user search criteria. IR systems serve essential functions across various domains, including web search engines [18, 19], digital libraries [20, 21, 22], and e-commerce platforms [23, 24].

Building on these foundations, recent research has increasingly leveraged neural networks to improve IR system performance [25, 26, 27]. By learning powerful data representations, these methods improve the modeling of semantic relationships between queries and data, thereby advancing retrieval effectiveness.

Within this broader context, Content-Based Image Retrieval (CBIR) emerges as a specialized branch of IR dedicated to retrieving images based on their visual content. Over the last decade, significant progress has been achieved in CBIR, particularly in improving accuracy and efficiency through techniques that analyze features such as color, shape, and texture [28]. Li et al.[29] and Bhoir et al.[30] review recent advancements in this area, underscoring both the challenges and the potential of Artificial Intelligence (AI) to further enhance CBIR systems.

A notable contribution to the field is the work of Lux et al.[31], who proposed Lucene Image Retrieval (LIRE), a lightweight *Java* library for IR. It includes multiple features, such as parallel indexing, local feature aggregation, hashing, and approximate and linear search. Since it supports multiple global and local features, it enables easy comparison between new features and existing ones. Although it only works by providing an image as a search parameter, it represents a powerful tool for CBIR tasks.

In terms of image IR, Arquivo.pt[1] has developed a tool that enables searching for images in the Portuguese web archive. The authors[1] presented research and development work focused on a large-scale image search system that enables searching billions of historical images archived from the web since the 1990s. The system extracts textual metadata associated with images from web pages, analyzing the HyperText Markup Language (HTML) Document Object Model (DOM) hierarchy to identify text most closely related to images. To manage large data volumes, the system removes duplicate images across time and space, collects the oldest version of the page metadata and maintains all the image metadata captured over time, as web archive users prefer older documents. The system incorporates a Deep Neural Network (DNN) to filter potentially pornographic content, providing tags and filtering results. The image indexing workflow operates on three processing clusters for metadata extraction, Not Safe For Work (NSFW) classification, and indexing, utilizing *Apache Solr* for search functionality. *Apache SolrCloud* deploys the system, and a web interface enables users to search and view images. The overall process is presented in Figure 2.1.

[1] Arquivo.pt

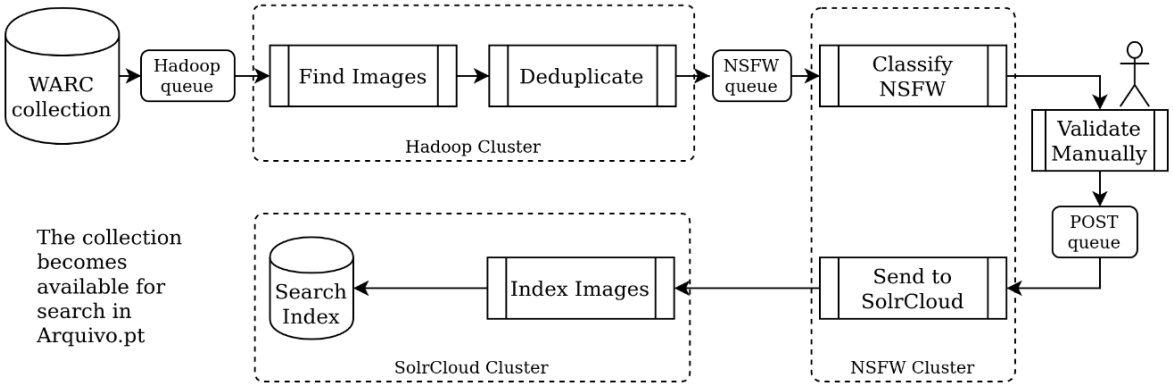

Figure 2.1: *Arquivo.pt* image indexing process. Taken from [1].

The *Arquivo.pt* image search system has proven valuable for image retrieval tasks, increasing searchable images from 22 million to 1.862 billion.

2.2.2 Computer Vision

CV is a subfield of AI that focuses on enabling computers to interpret and understand visual information from the real world. It involves developing algorithms and techniques that allow machines to analyze and process visual data, such as images and videos. With the significant advancements in deep learning and neural networks, researchers have developed powerful model architectures for processing visual data. Convolutional Neural Networks (CNNs) [32] are a class of deep learning models commonly used in CV tasks, such as image classification, object detection, and image segmentation. By leveraging the power of convolutional layers, CNNs can learn to extract features and patterns from images, making them ideal for processing visual data while maintaining relatively simple network architecture.
More recently, there has been a surge in the use of self-attention-based architectures, primarily in NLP but also increasingly in CV. The Vision Transformer (ViT) model [33] exemplifies a self-attention-based architecture that has been successfully applied to CV tasks. By treating images as sequences of patches and applying self-attention mechanisms, ViT can learn to capture long-range dependencies and relationships between image regions, enabling it to perform tasks such as image classification and object detection. This approach utilizes the transformer architecture, originally proposed by Vaswani et al.[34], which does not utilize convolution or recurrence, and focuses exclusively on attention mechanisms. Using only this mechanism, their model outperformed all previous models in text translation tasks. This later became the foundation for most modern LLMs.

2.2.3 Multimodal Learning

Multimodal learning [35] constitute a class of deep learning models that can process and analyze data from multiple modalities, such as images, text, and audio. By combining information from different sources, these networks can learn rich representations of complex data and perform a wide range of tasks, including image retrieval and image captioning.
With the surge of multimodal networks, several visual-language models have emerged and

are now fundamental components in the domain of IR. The models CLIP [2] and A Large-scale ImaGe and Noisy-text embedding (ALIGN) [36] significantly improved zero-shot capabilities in terms of generalization not only for image retrieval tasks but also for classification and visual question answering [37].

CLIP is a system that learns to associate images with their corresponding text descriptions through a CL approach. Unlike traditional CV systems that are trained to predict a fixed set of object categories, CLIP learns from raw text about images, enabling zero-shot transfer to a variety of downstream tasks. The authors of CLIP also created a new dataset called WebImageText (WebIT), which is a large-scale dataset of 400 million (image, text) pairs. It was built using publicly available images and their corresponding text descriptions from different sources on the internet. This dataset has a similar word count as the *WebText* dataset used to train *GPT-2*.

Using this extensive dataset, CLIP employs a CL approach where, instead of predicting the exact words in a caption, it learns to predict which text is paired with which image. The system begins with a pre-training task where it is trained to predict which caption corresponds to which image in a pair (image, text). Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N * N$ possible pairings actually occurred. This is achieved by maximizing the cosine similarity between the embeddings of the correct (image, text) pairs while minimizing the similarity of the embeddings of the incorrect pairs. The loss function is a symmetric cross-entropy loss over the similarity scores, and this approach was found to be more efficient than predictive objectives. Through this process, CLIP learns a multimodal embedding space by jointly training an image encoder and a text encoder. The image encoder explores using *ResNet* or *ViT* architectures, while the text encoder is a transformer-based model operating on a Byte Pair Encoding (BPE), which is a compression algorithm that replaces common words with a single token.

By using a contrastive objective and training on a large dataset, CLIP is much more efficient at zero-shot transfer compared to earlier approaches using transformer-based language models. The contrastive objective further improves the efficiency in the rate of zero-shot transfer to *ImageNet* by another 4x. Figure 2.2 illustrates the CLIP process. In the first step, named Contrastive pre-training, the image and text encoders are trained jointly to correctly predict image-text pairings. In the second step, a dataset classifier is created from label text; the figure shows some example classes such as “plane,” “car,” “dog,” and “bird.” The third and final step demonstrates zero-shot prediction, where the model is given an image and multiple text descriptions from the dataset classifier, and predicts the correct label for the image.

After the release of CLIP, Balauca et al.[38] demonstrated its usage in a new dataset with images of museum exhibits and their descriptions. The results showed that utilizing CLIP’s embeddings for image retrieval tasks outperformed traditional methods, highlighting the potential of this approach in real-world applications.

These multimodal networks also led to the development of generative models that can convert text-to-image and create images from textual descriptions. *DALL-E* and *Midjourney* are notable examples that, although not focused on image retrieval, demonstrate the benefits of multimodal networks in the field. *DALL-E* [39] is a model that can generate images from

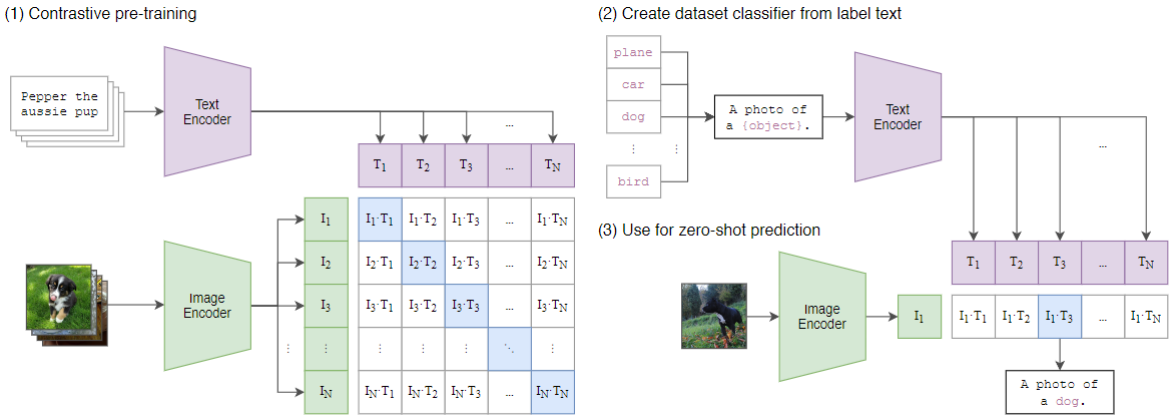

Figure 2.2: CLIP process. Taken from [2].

textual descriptions, enabling users to create novel and creative images based on their ideas. *Midjourney*[2] is a model that can generate images from textual prompts, allowing users to explore and visualize different concepts and ideas through images.

Following CLIP's success, more novel approaches were developed. BLIP-2 [3] represents a new strategy that uses off-the-shelf frozen pre-trained image encoders and frozen LLMs, the overall architecture is shown in Figure 2.3.

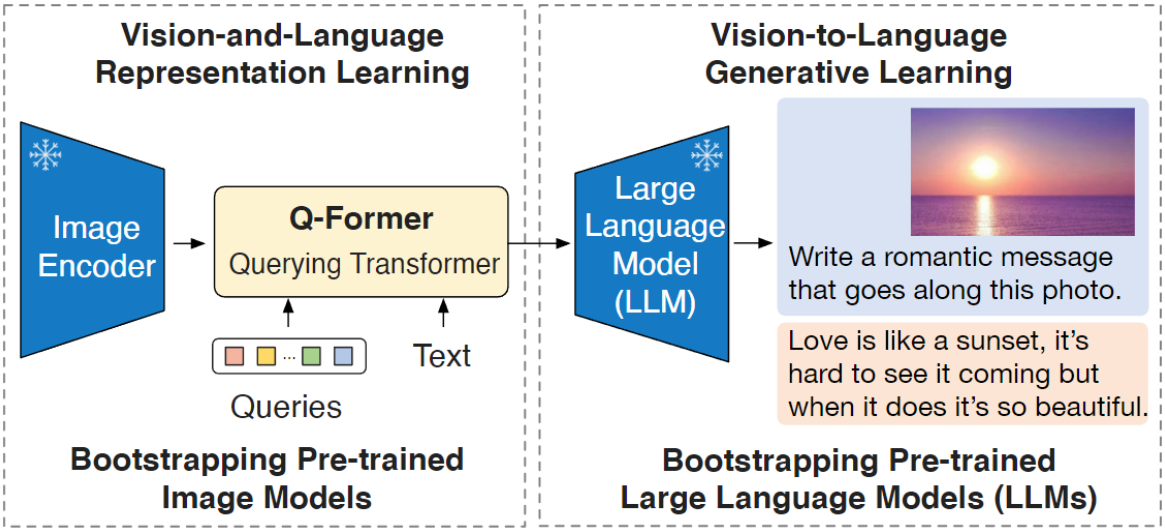

Figure 2.3: BLIP-2 architecture. Taken from [3].

In this paper, the authors proposed a Querying Transformer (Q-Former) pre-trained in two stages. In the first stage, the Q-Former is connected to a frozen image encoder and is trained to extract visual representations relevant to the text. This is achieved through the joint optimization of three pre-training objectives:

- **Image-Text Contrastive Learning (ITC)** learns to align image and text representations by maximizing their mutual information. The Q-Former learns to extract visual features that are most similar to the corresponding text. A unimodal self-attention

[2]https://www.midjourney.com/home

mask is used to prevent the queries and text from interacting with each other during this process;

- **Image-Grounded Text Ceneration (ITG)**, where the model trains the Q-Former to generate text conditioned on the input image. The queries extract visual features necessary for generating the text. A multimodal causal self-attention mask is used to control the interaction between queries and text.

- **Image-Text Matching (ITM)** aims to learn fine-grained alignment between image and text representations. The model predicts whether an image-text pair is matched or unmatched. A bidirectional self-attention mask allows queries and texts to attend to each other.

The three objectives can be summarized in Figure 2.4.

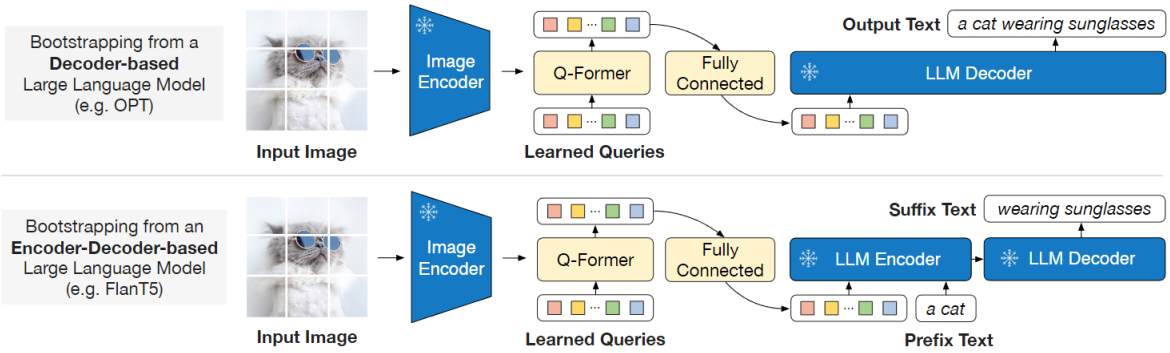

Figure 2.4: BLIP-2 first-step objectives. Taken from [3].

In the second stage the Q-Former, with the frozen image encoder, is connected to a frozen LLM. A fully-connected layer projects the output of the Q-Former into the text embedding space of the LLM. These projected embeddings are then prepended to the input text embeddings as soft visual prompts for the LLM. This stage is designed to leverage the LLM's language generation capabilities. The results obtained delivered a slight performance uplift, but the main focus was on the reduction of the number of parameters used in the model, especially in the zero-shot learning tasks.

A novel unsupervised approach to Cross-Domain Image Retrieval (CDIR) that incorporates textual context into the image retrieval was developed. Iijima et al.[4] proposed *Caption-Matching (CM)* which uses generated image captions as a domain agnostic intermediate representation to enable cross-domain searching without needing labeled data. The overall approach is illustrated in the Figure 2.5.

The *CM* method uses CLIP and BLIP-2 in its implementation. The process starts by generating captions for each image in the database, then these captions go through the text encoder from CLIP to obtain their embeddings. At the same time, the query image is encoded in the image encoder from CLIP to produce its corresponding embedding. The final step is to calculate the similarity score between the image embedding and all the text embeddings, which results in a list with similarity scores. Their results show top performance in cross-domain searches, using common datasets such as *Office-Home* and *DomainNet*, but also using AI-

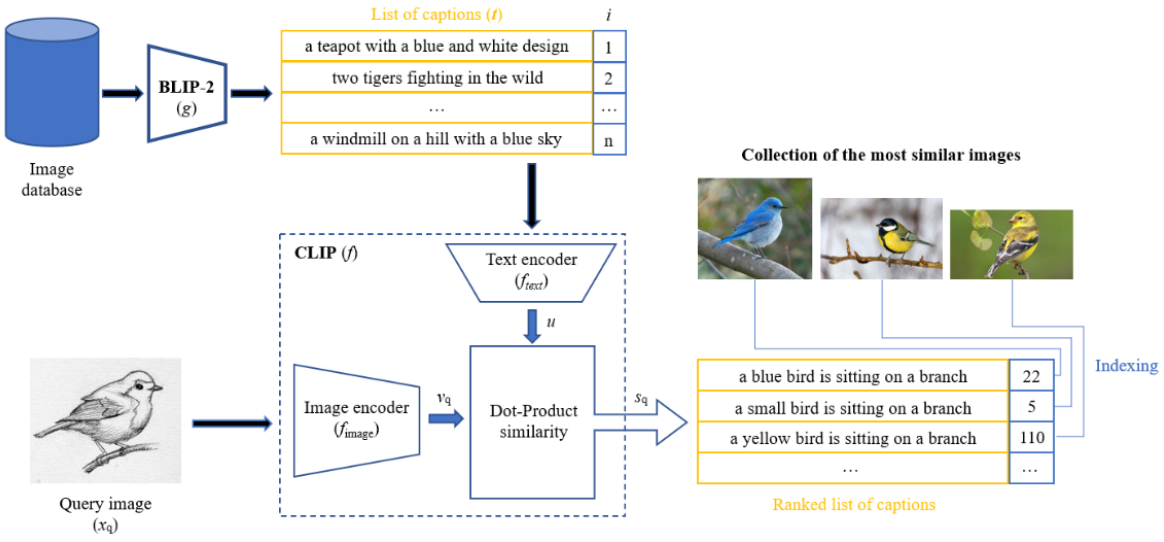

Figure 2.5: *CM* approach for cross-domain image retrieval. Taken from [4].

generated images from *Midjourney*. It demonstrates the potential in generating captions for images and using them as an intermediate representation for image retrieval tasks.

Despite its capabilities, CLIP has some limitations, one of them being the lack of support for multiple languages. To address this issue, Carlsson et al.[5] proposed a solution where they ignore the visual encoder and only train the text encoder. Using transfer learning between the original English text encoder and the target model pre-trained in a different language, the model can learn to generate embeddings for text in multiple languages. The architecture is displayed in Figure 2.6. The results are very promising and this approach was used in our work, given our focus on the Portuguese language.

Furthermore, CLIP is also not open-source, which means that researchers and developers cannot fully explore the model's architecture and capabilities. To address this limitation, *OpenCLIP* [40, 41, 42, 43] was created. This open-source implementation of the CLIP model allows researchers and developers to explore and experiment with the model. By providing access to the model's source code and pre-trained weights, *OpenCLIP* enables users to fine-tune the model on their own datasets and customize it for specific tasks and domains. This particularly benefits our research work since we can use pre-trained models that utilize multilingual datasets containing Portuguese text.

Recently, a new approach was developed using CLIP as a foundation for the implementation. Gong et al.[6] proposed ViT-Relation-Focus (VITR) which enhances ViT by extracting and reasoning about image region relations based on a local encoder. While VITR maintains the dual-encoder architecture of CLIP with separate image and text encoders, it introduces a CNN component for capturing fine-grained local image features that provide enhanced spatial detail. The architecture also includes a *fusion* module that combines outputs from both the relation reasoning process and existing pre-trained knowledge. An overview of the proposed method is shown in Figure 2.7.

Additionally, they proposed a *turbo* model for improving IR tasks. The *turbo* model has a two-stage retrieval architecture that first identifies a set of candidate images before perform-

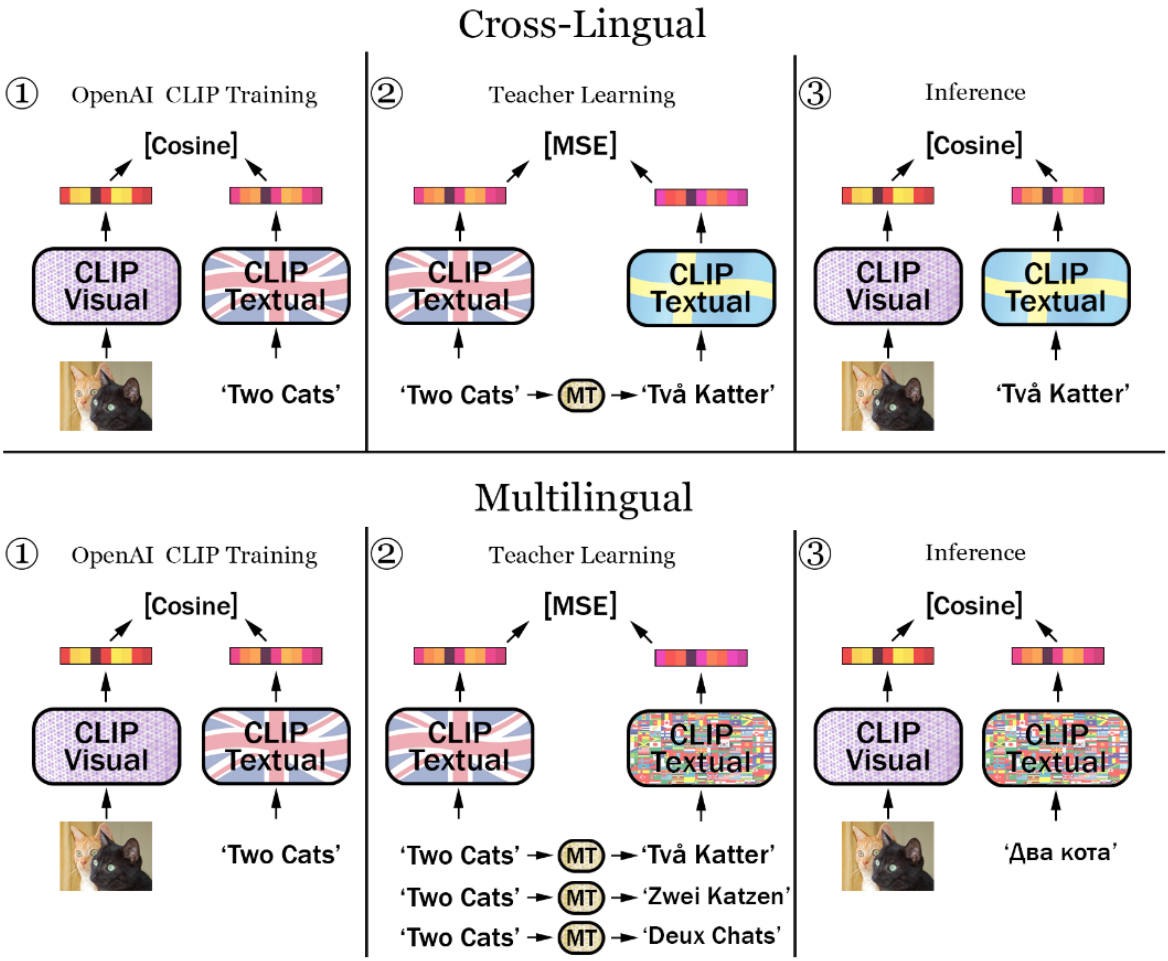

Figure 2.6: Multilingual CLIP architecture. Taken from [5].

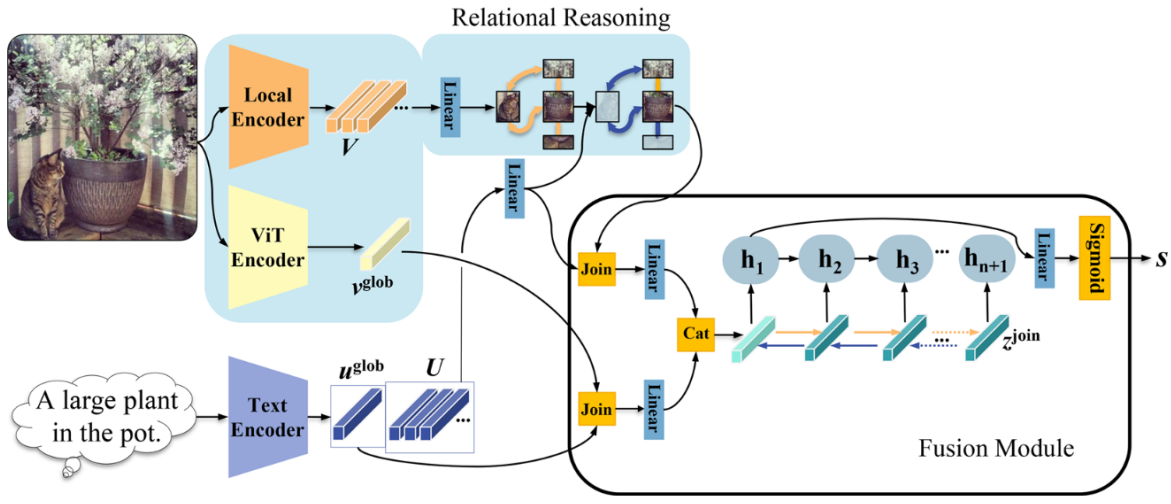

Figure 2.7: VITR architecture. Taken from [6].

ing refined ranking in the second stage. This optimization focuses specifically on enhancing retrieval speed rather than accuracy improvements. This module is shown in Figure 2.8. The suggested solution outperforms CLIP, in the *RefCOCOg* dataset, VITR achieved a 3.6% improvement in image-to-text retrieval and a 4.1% improvement in text-to-image retrieval using the Recall@1 metric. When using the *CLEVR* dataset the improvement is even more

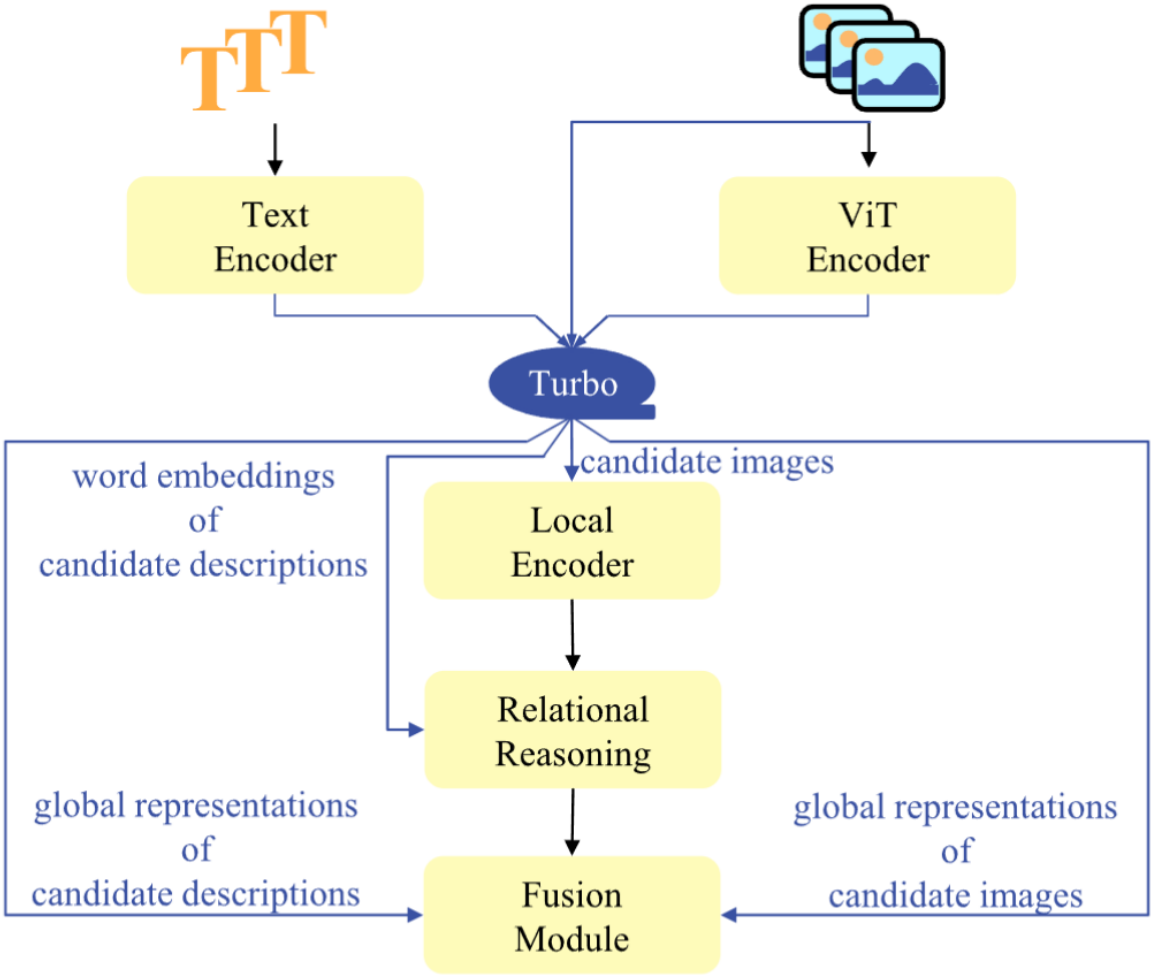

Figure 2.8: VITR turbo model. Taken from [6].

significant, for the same metric, VITR achieved a 21.5% improvement in image-to-text retrieval and a 13.9% improvement in text-to-image retrieval. Even though the results are significant, the lack of support for languages other than English, and the model's high retrieval time for a query are still aspects that need to be addressed. A text-to-image retrieval in VITR with *turbo* takes 0.1 seconds, in comparison, CLIP takes 0.04 seconds. This delay can cause problems in real-time applications, especially in web applications where users expect fast responses.

Despite the downsides of the VITR model, Gong & Cosma [7] created *Boon*, a search-engine that combines the GPT-3.5-turbo language model and the VITR model to enhance the engine's capabilities in extracting relationships in images. It provides image-to-text and text-to-image retrieval, and allows multilingual conversations about images. The use of GPT to translate the inputs and outputs to the desired language, solves the problem of multilingual support. Using the Recall@1 metric, it surpasses CLIP by 2.1% in image-to-text and 4.7% in text-to-image retrieval. However, this system is designed for conversational image exploration rather than traditional query-based retrieval, making it unsuitable for our specific use case. Additionally, the retrieval time remains high, and the reliance on GPT as a paid service creates significant cost implications for real-world applications.

Combining Text and Visual Modalities for Enhanced Portuguese Image Retrieval

An essential challenge in multimodal retrieval systems is the effective combination of results from different modalities or retrieval approaches. Reciprocal Rank Fusion (RRF), originally proposed by Cormack et al.[44], provides a parameter-free approach to rank fusion by computing combined scores based on reciprocal ranks across multiple systems. Recent advances in multimodal fusion have explored sophisticated alternatives to traditional rank fusion techniques. Jiao et al.[45] present comprehensive approaches to combining heterogeneous data sources, including early, deep, late, and hybrid fusion strategies that address the computational complexity challenges inherent in multimodal systems. Liu et al.[46] introduce efficient low-rank decomposition methods that achieve linear scalability with the number of modalities while maintaining competitive performance. Novel rank fusion variants, such as Inverse Square Rank fusion [47], demonstrate improved effectiveness through quadratic decay functions and logarithmic normalization, particularly excelling in multimodal search scenarios where different feature spaces contribute complementary information. These fusion techniques have found practical applications in e-commerce platforms, where Xu et al.[48] show how attention-based multimodal fusion can significantly improve conversion rates and user satisfaction by effectively combining textual and visual product information.

2.2.4 Summary

This chapter reviewed the evolution of image retrieval across three major areas. First, it covered traditional information retrieval and content-based image retrieval, highlighting both theoretical foundations and practical systems such Arquivo.pt. Second, it described advances in computer vision, from CNN-based feature extraction to transformer-based architectures such as Vision Transformers, which enabled richer and more flexible visual representations. Third, it examined multimodal learning, focusing on models such as CLIP, ALIGN, BLIP-2, and their extensions, as well as approaches using caption-based retrieval, generative models, and fusion strategies like Reciprocal Rank Fusion (RRF).

These developments illustrate a clear progression, from metadata-driven search, through visual feature analysis, to multimodal systems that align images and text in shared embedding spaces. Multimodal networks introduced powerful zero-shot and cross-modal capabilities, while multilingual adaptations such as Multilingual CLIP and open implementations like OpenCLIP broadened their applicability to non-English contexts. Recent research on hybrid and fusion-based retrieval approaches further demonstrated the potential of combining complementary modalities to enhance performance.

Despite significant progress, important challenges remain. Current systems still lack robust multilingual support, restricting applicability beyond English, and struggle to effectively integrate heterogeneous retrieval signals. These gaps highlight the need for multimodal and hybrid strategies specifically designed for multilingual contexts.

Chapter 3

PT-Image-IR Dataset

This chapter presents an overview of existing datasets in the image IR domain and introduces our dataset, developed as part of this research. Section 3.1 conducts a comprehensive analysis of currently available datasets in the field of image retrieval and captioning, evaluating their applicability to Portuguese-language research contexts. Section 3.2 presents the methodology and development of a novel dataset specifically designed for European Portuguese image retrieval, including detailed descriptions of the dataset creation process, annotation methodology, and dataset organization. Finally, Section 3.3 concludes the chapter with a summary of the key contributions and insights gained from the dataset development process for Portuguese image IR research.

3.1 Review of Existing Datasets

To gain a comprehensive understanding of the datasets currently available, we analyzed several resources in the fields of image retrieval and captioning. Each dataset was assessed with respect to its suitability for the requirements of our Portuguese image IR system.

The MS COCO dataset [49] represents one of the most widely adopted resources for object detection, segmentation, and captioning tasks. This large-scale dataset comprises over 330,000 images, of which more than 200,000 are annotated across 80 distinct object categories. Each image is accompanied by 5 natural language captions that describe the scene context and content. The dataset is structured into three splits: a training set containing 72% of the images, a test set with 25%, and a validation set comprising the remaining 3%. While initially considered for our research due to its widespread adoption and rich textual descriptions that could potentially support image retrieval tasks, the MS COCO dataset proved unsuitable for our specific needs. The primary limitations include the absence of query-response pairs and the exclusive use of English-language captions.

The Flickr30k dataset [50] constitutes another prominent resource in the field of image-text alignment, comprising 31,000 images paired with 155,000 captions (5 captions per image). The captions consist of English sentences authored by human annotators who describe the visual content of each image. As the dataset's name suggests, the images are sourced from the Flickr platform[1], which hosts a diverse collection of user-generated content spanning various topics and contexts. The dataset creators implemented rigorous quality control measures to ensure high relevance and accuracy of the human-generated captions. An enhanced version, Flickr30k Entities [51], extends the original dataset with additional features: coreference chains that enable tracking of identical entities across multiple images, bounding boxes for entities mentioned in captions, and entity recognition capabilities that link textual mentions to specific visual elements within images. Despite these comprehensive features and the po-

[1] https://flickr.com

tential for adaptation to retrieval tasks, both the original Flickr30k and its enhanced variant remain unsuitable for our research objectives due to the absence of query-response structures and their exclusive reliance on English-language annotations.

Addressing the need for Portuguese-language resources, the #PraCegoVer dataset [52] represents a significant contribution to accessibility-focused image description in Portuguese. Constructed from Instagram posts, this dataset comprises 533,523 image-caption pairs covering a diverse range of social media content and everyday scenarios. The dataset's primary purpose centers on assisting visually impaired users through detailed visual descriptions written in Brazilian Portuguese. These captions follow accessibility guidelines and provide comprehensive descriptions of visual elements, making them particularly rich in descriptive content. However, several limitations prevent its direct application to our research objectives. First, the dataset specifically targets the Brazilian Portuguese variant, which differs linguistically and culturally from European Portuguese in vocabulary, expressions, and contextual references. Second, the dataset structure focuses on descriptive captioning rather than query-response paradigms essential for IR evaluation.

In the domain of specialized image retrieval, the Fashion200k dataset [53] represents a notable contribution to domain-specific image-text alignment. This weakly-labeled dataset comprises 200,000 images of fashion items, each accompanied by concise textual descriptions that capture essential visual attributes such as style, color, and garment characteristics. The dataset demonstrates the potential for developing models that learn fine-grained relationships between visual features and textual descriptions within specialized domains. Fashion200k incorporates both exact matches and relative comparisons (e.g., "similar to this but in red"), making it particularly valuable for understanding compositional queries in retrieval tasks. While this dataset illustrates the importance of domain-specific approaches to image retrieval, it remains unsuitable for our research due to its focus on fashion-related content and the exclusive use of English-language annotations. Nevertheless, its methodology provides valuable insights for constructing domain-specific datasets with rich query-response structures.

Representing a significant advancement in multilingual image-text resources, the Wikipedia-Based Image Text (WBIT) dataset [54] constitutes a large-scale collection of 37 million image-text pairs spanning 108 languages. The dataset uses a structured approach where each image is associated with multiple types of descriptive text, including reference descriptions, attribution descriptions, alt-text descriptions, and contextual information. Additionally, comprehensive metadata accompanies each entry, encompassing page titles, URL references, and sectional information that provides valuable context for understanding image-text relationships. While the multilingual nature of WIT represents a substantial contribution to cross-lingual research, the dataset's coverage across 100+ languages introduces inherent imbalances in data distribution per language. As illustrated in Table 3.1, the data distribution is highly skewed: only 9 languages have more than 1 million image-text pairs, while the majority of languages (38 out of 108) contain between 14,000 and 50,000 pairs. This uneven distribution reflects the varying availability of Wikipedia content across different languages and demonstrates the challenges of creating balanced multilingual datasets. Despite containing a

considerable number of Portuguese-language image-text pairs, the WIT dataset remains unsuitable for our research objectives due to its fundamental structural limitation: the absence of query-based retrieval structures essential for IR system evaluation and training.

Table 3.1: Distribution of languages by data volume in the WIT dataset, showing how many languages fall within each range of image-text pair counts.

Number of Languages	Number of Image-Text Pairs
9	> 1M
10	500K - 1M
36	100K - 500K
15	50K - 100K
38	14K - 50K

More recently, the MIRACL-VISION dataset [55] represents a paradigm shift in multilingual retrieval benchmarking by extending the established MIRACL dataset [56], originally designed for text-only retrieval tasks, into the visual domain. Constructed from Wikipedia content, this dataset introduces the concept of visual document retrieval, which fundamentally differs from traditional image retrieval approaches. Rather than focusing on retrieving standalone images based on visual similarity or content matching, MIRACL-VISION targets visually rich documents that integrate images with substantial contextual text, such as Wikipedia articles containing embedded figures, photographs, and diagrams. This approach enables comprehensive evaluation of vision-language models in realistic cross-lingual retrieval scenarios by requiring systems to process and correlate multimodal inputs with multilingual textual content. The dataset's emphasis on document-level retrieval, combined with its multilingual scope, positions it as a valuable benchmark for assessing holistic understanding of visual-textual relationships across language boundaries. However, the Wikipedia-centric construction and document-focused paradigm limit its direct applicability to our Portuguese image retrieval objectives, which require query-specific image targeting rather than document-level retrieval capabilities.

While several datasets exist for image captioning and retrieval tasks, none fully addresses the critical need for a comprehensive, query-driven dataset specifically designed for European Portuguese. This significant gap underscores the necessity for developing a dedicated dataset that encompasses both Portuguese queries and their corresponding relevant images. Such a resource would serve as an essential foundation for our research, enabling the systematic development and rigorous evaluation of image IR systems specifically optimized for Portuguese-language contexts.

3.2 Development of the PT-Image-IR Dataset

To address the limitations of existing datasets and to support our research objectives, we propose the creation of a novel dataset specifically designed for Portuguese image IR systems. This dataset consists of a collection of images paired with queries in European Portuguese, along with their corresponding relevant images. The dataset was constructed using images

sourced from the official website of the Portuguese Presidency[2], which provides a rich and diverse set of images related to governmental activities, events, and cultural heritage.

This section is divided into three subsections. Subsection 3.2.1 describes the dataset creation process, including the selection of images and formulation of queries. Subsection 3.2.2 outlines the annotation methodology and the criteria used for selecting relevant images. Finally, Subsection 3.2.3 discusses the structure and the characterization of the dataset for better understanding and usability in image IR tasks.

3.2.1 Dataset Creation

The first step in creating the dataset involved analyzing the official website of the Portuguese presidency to identify suitable images for the dataset. The website provides a wide range of information about the presidency, including news articles, press releases, and event announcements. 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, 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. Moreover, alternative sources, such as news outlets or social media platforms, might offer broader topical diversity but often pose challenges related to licensing and structural consistency, making large-scale and reliable image acquisition more difficult. In contrast, the Presidency website provides well-curated and consistently structured data, making it a high-quality starting point for research in text-image retrieval in European Portuguese.

In order to extract all the information available on the website, a web scraper was developed. The scraper navigated through the website's structure, collect all news articles, and extract relevant information such as article titles, publication dates, and associated images. The scraper was implemented using Python with the BeautifulSoup library[3] for parsing HTML content and the Selenium library[4] for handling dynamic content and JavaScript rendering. It was configured to follow the website's pagination and retrieve all available articles, ensuring comprehensive coverage of the presidency's news content. The collected data was stored in a JavaScript Object Notation (JSON) file using a structured format, including the article title, publication date, body of text, and image Uniform Resource Locators (URLs). The images were downloaded and stored locally, using the URLs extracted from the articles. This process successfully collected 4,678 articles, each containing a title, publication date, body of text, and an associated set of images. In total, 42,333 images from articles published were collected over a time span of 9 years, from January 2016 to March 2025.

The next step in building our dataset was to generate a set of queries tailored for image IR tasks. To this end, we used GPT-4o mini via the duck.ai platform, employing a carefully

[2]https://www.presidencia.pt
[3]https://www.crummy.com/software/BeautifulSoup/
[4]https://www.selenium.dev/

designed prompt to generate queries in Portuguese across distinct categories: *events and other contexts*, *facial expressions and emotions*, *interactions with the public*, and *places and environments*. Such categories were specifically selected to capture the diversity of visual content typically present on the Portuguese presidency's website. The prompt used for query generation was as follows:

> *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:* ***insert category here***

Using this approach, 38 queries were automatically generated across the four specified categories. As model-generated queries tend to be more verbose and descriptive, we decided to complement this set with 42 additional manually-created queries, which were intentionally designed to be more concise and straightforward. For the manual query creation, we defined seven additional categories to broaden the semantic coverage of our dataset: *public figures*, *sports and activities*, *trending topics*, *places and monuments*, *general places*, *objects*, and *others*. These categories were defined to encompass domains not fully represented in the AI-generated set, including specific personalities relevant to Portuguese context, recreational activities, contemporary events of public interest, geographical locations of cultural significance, everyday environments, and tangible items that might appear in presidential contexts. The manually-created queries within these categories were formulated to reflect typical user search patterns, emphasizing brevity and directness. The following are two illustrative examples of the queries: one generated by the model - "Presidente de Portugal em encontros com cidadão" (President of Portugal meeting citizens); and one manually created - "Bombeiro" (Firefighters). The complete list of all 80 queries organized by category is available in Appendix A.

Having generated the queries, we proceeded to create a pool of candidate images for annotation. We opted to use a pooling [57] strategy popularized by the Text REtrieval Conference (TREC) community [58]. This approach enables the creation of relevance judgments for each query without the need to annotate the entire dataset by selecting a representative subset of images that annotators will evaluate to establish the ground truth. Our pooling strategy combines the results of three distinct categories of image retrieval techniques during the dataset creation phase: (1) traditional lexical IR techniques (TF-IDF and BM25), (2) vision-language models (multilingual CLIP variants), and (3) end-to-end image retrieval systems (Google Images and Arquivo.pt). For each method, we selected the top-10 retrieved images per query to construct a diverse and representative candidate set.

For the traditional text-based IR methods, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) [59] and BM25 [60] algorithms. We adopted a straightforward approach where each image is represented by its textual metadata, specifically the title of the article in which it appears. Rankings were then computed based on term frequency and document relevance to the query. In contrast, vision-language models were used to directly

capture the semantic alignment between textual queries and visual content. We used several variants of the CLIP model [2], ranking results via cosine distance in the shared embedding space. This included multilingual adaptations such as Multilingual-CLIP [61] and OpenCLIP [40, 41, 42, 43]. Specifically, we used the following models: *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* with *laion5b_s13b_b90k*. Finally, for end-to-end image retrieval, we queried two external search engines: Arquivo.pt[5] and Google Images[6]. These systems apply their proprietary ranking algorithms. To constrain the search space and ensure contextual relevance, queries were restricted to return images only from the official Portuguese Presidency website.

Subsequently, we combined the top-10 images retrieved by each method for each query into a single pool, removing any duplicates. Images that are not included in the pooled set are assumed to be non-relevant, following the standard assumption in pooling-based evaluation [62]. We acknowledge that this approach has limitations, as relevant images may exist beyond the top-ranked results from our selected methods. Nevertheless, prior work [62] has demonstrated that pooling from diverse retrieval systems helps mitigate this issue by increasing the likelihood of capturing the most relevant items across different ranking paradigms. The following section details the annotation process applied to manually assess the relevance of each image to its corresponding query.

3.2.2 Dataset Annotation

The annotation process was conducted by a team of three master's students, who systematically evaluated the relevance of each image to its corresponding query. The annotators participated voluntarily in this research effort without financial compensation. To ensure objectivity and consistency, the final ground-truth labels were determined through majority voting across all three annotators.

To achieve high-quality and consistent annotations, all annotators were provided with detailed written guidelines that clearly defined the annotation task, evaluation criteria, and decision-making processes. These guidelines were essential for ensuring inter-annotator agreement and reducing subjective interpretation variability. The complete guidelines provided to the annotators are reproduced below:

> **Annotator Guidelines**
>
> The goal of this annotation task is to determine whether an image is relevant (1) or irrelevant (0) to a given query, with relevance defined by the relationship between the query, the image, and the context provided by the article title.
>
> 1. **Understanding the query**: Read the query carefully. Identify the key terms and concepts. Consider the intent behind the query. What information is the user seeking?

[5] https://arquivo.pt/

[6] https://images.google.com/

2. **Evaluating the image**: Assess the content of the image. Does it depict something that is related to the query? Look for visual elements (people, objects, actions, places) that can be linked to the query.
3. **Context from the article title**: Read the article title provided since it may help clarify the context of the image. It helps in situations where the query might depict a specific event, person or object that is mentioned in the title.
4. **Labeling the image**: Based on your assessment, label the image as relevant (1) or irrelevant (0). Use the following criteria to guide your decision:
    - **Relevant (1)** — The image directly supports, illustrates or is related to the query. It should provide meaningful context or information that aligns with the query's intent.
    - **Irrelevant (0)** — The image does not support or relate to the query in a meaningful way. It may be unrelated, misleading, or provide no additional context to the query.
5. **Common scenarios**:
    - **Direct match**: If the image clearly depicts the subject, place, action or object mentioned in the query, it is relevant.
    - **Indirect match**: If the image depicts something that could be related to the query but not in an obvious way, consider the context provided by the article title. If the image is related to the article title and could be relevant to the query, it is relevant.
    - **Unrelated content**: If the image shows something completely different from the query and does not provide any context or information related to the query, it is irrelevant.
    - **Ambiguous cases**: If in doubt even after using the article title as context, choose the image as irrelevant. The subject of the image should be clearly shown in the image, for instance, if the image is blurry or too small to identify the subject, it is irrelevant. For people, the face should be clearly visible and identifiable, when the subject is to the side or not clearly visible, it is irrelevant.
6. **Avoiding biases**: Try to be objective and avoid personal biases. Focus on the content of the image and its relationship to the query. Do not let personal opinions or feelings about the subject matter influence your decision.
7. **Unknown terminology**: If you encounter a term or concept in the query or article title that you do not know, you can search for it online to understand its meaning. This can help you better assess the relevance of the image.

Note: If you made a mistake you can undo by clicking the "Undo" button or using the keyboard shortcut "Z". The progress is saved automatically and you can leave at any time. There is also a progress bar to track how many images you have been annotated for the current query.

To assess annotation quality and reliability, we computed the Inter-Annotator Agreement (IAA) using Fleiss' Kappa[63], a statistical measure that quantifies agreement levels among multiple annotators while accounting for chance agreement. The analysis yielded a Fleiss' Kappa score of 0.62, indicating substantial inter-annotator agreement according to standard interpretation guidelines. This level of agreement demonstrates the consistency of annotator evaluations and validates the reliability of the resulting ground-truth labels, providing confidence in the quality of our dataset annotations.

3.2.2.1 Annotation Tool

Given the large scale of the annotation task (5,201 image-query pairs) and the need for consistent, high-quality annotations across multiple annotators, we developed a dedicated web-based annotation tool to streamline the evaluation process. The annotation tool was implemented as a containerized web application comprising three main components: a Flask-based backend server, a Redis database for data storage and session management, and a responsive front-end interface. Flask was selected for the backend due to its simplicity and rapid development capabilities, while Redis was chosen for its high-performance in-memory data structures and excellent support for concurrent user sessions.

The system architecture incorporates Auth0 for user authentication and session management, providing security with support for multiple login methods. This choice ensures data integrity, enables proper attribution of annotations, and allows for individual progress tracking across sessions. The authentication system maintains detailed user session management, automatically tracking annotation progress and enabling pause-and-resume functionality. The database stores three primary entity types: images (containing metadata and article associations), articles (containing titles and image relationships), and queries (containing query text and ground truth associations). Additionally, the system maintains real-time user progress tracking, enabling annotators to seamlessly resume their work across multiple sessions.

The annotation interface was designed to minimize cognitive load and maximize annotator efficiency. The main screen features a simple, four-component layout: the query text prominently displayed at the top, a central image viewing area that maximizes visual clarity, the contextual article title below the image, and annotation controls at the bottom. The annotation controls consist of three clearly distinguished buttons (relevant, non-relevant, and undo) with distinct color coding for immediate recognition. Figure 3.1 illustrates the clean and intuitive interface design that enabled efficient annotation workflows.

To accelerate the annotation process, the interface incorporates keyboard shortcuts (R for relevant, I for non-relevant, Z for undo) that significantly reduce the time required for each annotation. Visual feedback mechanisms, including button highlighting and a real-time progress bar, provide clear confirmation of user actions and track completion status.

Annotation guidelines are integrated directly into the interface through a modal dialog that appears upon first access, covering query interpretation strategies, image evaluation criteria, and approaches for handling ambiguous cases. The guidelines remain accessible throughout the annotation process while providing a "do not show again" option for returning users.

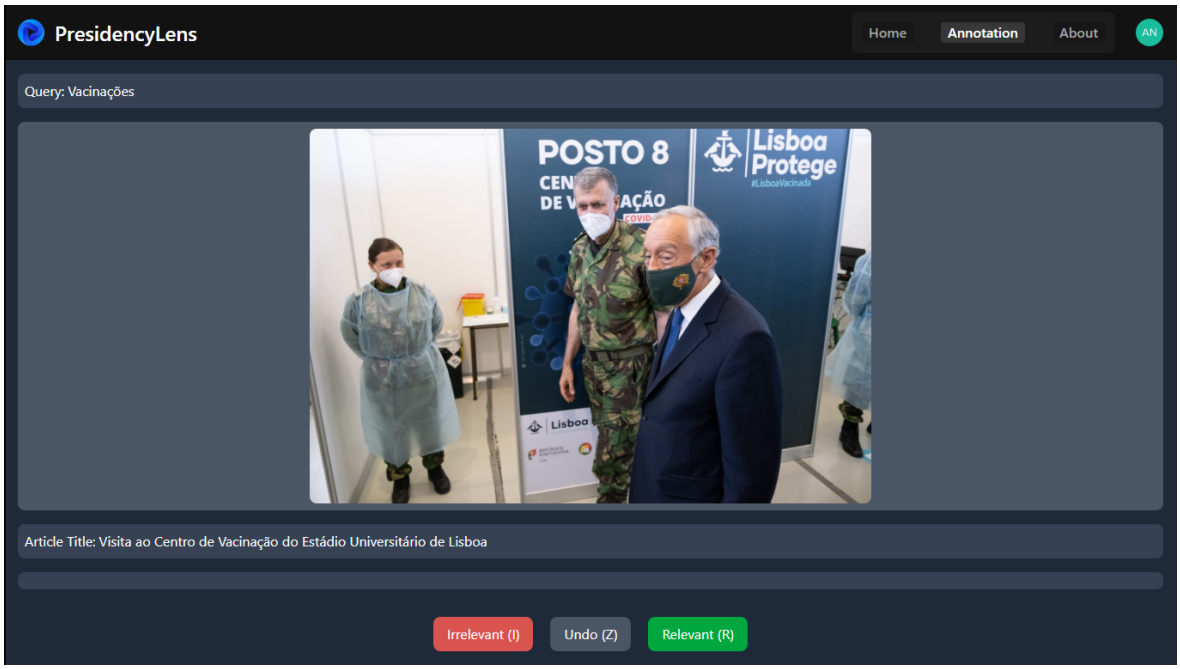

Figure 3.1: Screenshot of the annotation tool interface showing the four-component layout: query text at the top, central image viewing area, article title context, and annotation controls at the bottom with keyboard shortcuts and progress tracking.

The complete annotation tool and associated guidelines have been made publicly available on GitHub[7].

3.2.3 Structure and Characterization

The final dataset comprises 80 queries with a total of 5,201 annotated images, organized into four complementary files following standard IR evaluation formats. The dataset structure consists of:

- `queries.tsv` file containing query definitions with two fields — *id* (numerical identifier) and *query* (corresponding Portuguese query text);
- `qrels.txt` file containing relevance judgments formatted according to TREC conventions, where each line specifies the query ID, a standard placeholder value (0), the image ID, and the binary relevance score (0 or 1);
- `images.tsv` file (available on GitHub) containing image metadata with two fields — *id* (unique image identifier) and *url* (corresponding image URL from the Portuguese Presidency website);
- `articles.tsv` file (available on GitHub) containing article information with three fields — *id* (unique article identifier), *title* (article title), *content* (article content), *date* (publication date), and *images* (comma separated list with image ids from `images.tsv`) — that provides contextual metadata for the images and establishes the source context for each image.

[7] `https://github.com/RodrigDuarte/text2image_ir_annotation_tool`

The dataset exhibits balanced coverage across queries, with each query having relevance judgments for an average of 65 images. On average, approximately 23 images per query (35%) were annotated as relevant, while the remaining 42 images (65%) were labeled as non-relevant. This distribution reflects the natural sparsity of relevant documents in IR collections, where the majority of retrieved items are typically non-relevant.

The 80 queries in our dataset represent a deliberate balance between AI-generated and manually-created content, reflecting different query characteristics and complexity levels. As described in Section 3.2.1, the first 38 queries were generated using GPT-4o mini across four specific categories designed to capture the visual diversity of presidential content, while the remaining 42 queries were manually created within seven additional categories established to broaden the semantic coverage beyond the AI-generated domains. Figure 3.2 illustrates the distribution of AI-generated queries across the four categories. The largest category, *events and contexts*, comprises 36.8% of AI-generated queries, reflecting the prominence of event-based content on the Portuguese Presidency website. The remaining categories are more evenly distributed: *places and environments* accounts for 23.7%, *expressions and emotions* represents 21.1%, and *interactions with the public* comprises 18.4%.

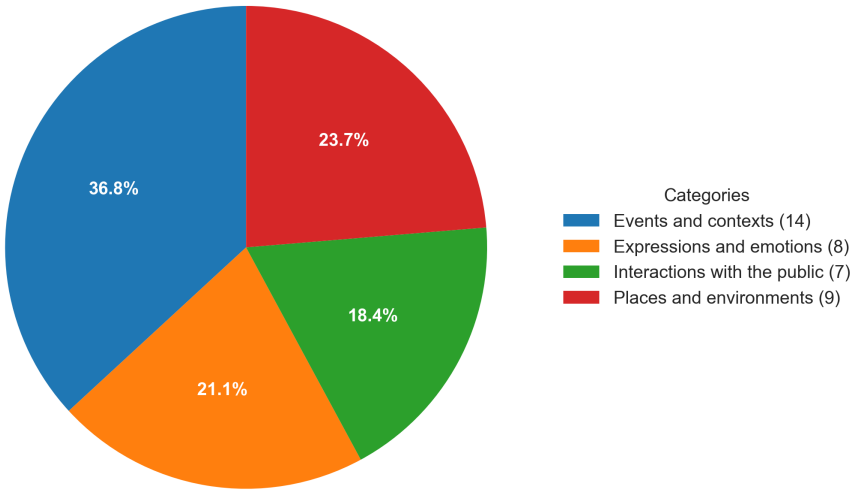

Figure 3.2: Distribution of AI-generated queries across the four predefined categories, showing the relative proportion of queries in each thematic area designed to capture different aspects of presidential visual content.

The manually-created queries exhibit greater thematic diversity, spanning the seven distinct categories established by the author as shown in Figure 3.3. This broader categorization reflects the deliberate intention to cover additional semantic domains not addressed by the AI-generated categories, encompassing specific personalities, recreational activities, contemporary topics, geographical landmarks, everyday environments, and tangible objects relevant to presidential contexts. The distribution shows *general places* and *others* as the largest categories, each representing 16.7% of manual queries. *Public figures* accounts for 19.0%, while *sports and activities*, *trending topics*, *places and monuments*, and *objects* each contribute approximately 11.9% to the manual query set. Both the AI-generated and manually-created

query sets demonstrate well-balanced distributions across their respective categories. The AI-generated queries maintain reasonable balance with an 18.4 percentage point difference between the largest and smallest categories, while the manually-created queries achieve even better balance with only a 7.1 percentage point difference, ensuring that no single semantic domain dominates either subset and providing adequate coverage for comprehensive evaluation across diverse query types.

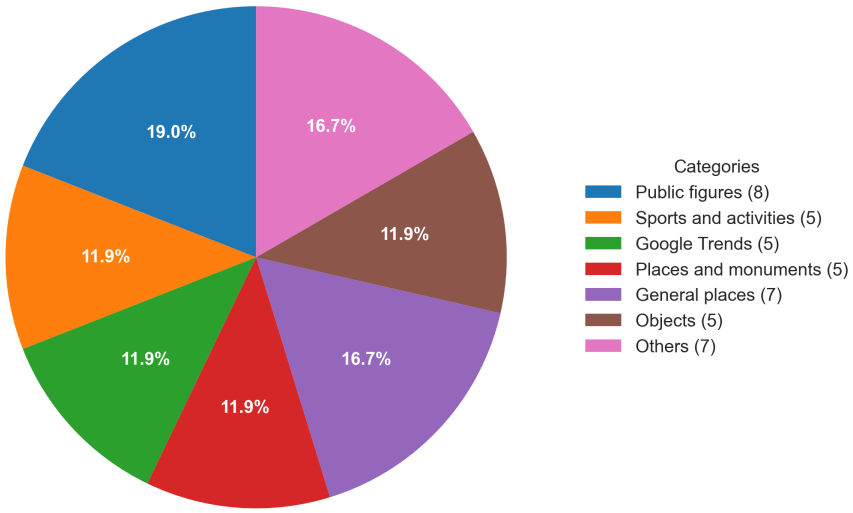

Figure 3.3: Distribution of manually-created queries across seven categories, demonstrating the broader thematic coverage achieved through author-generated queries.

This dual approach to query creation provides several advantages for IR system evaluation. The AI-generated queries tend to be more verbose and descriptive, often incorporating multiple contextual elements that reflect natural language query formulation patterns. In contrast, the manually-created queries are typically more concise and direct, resembling the succinct search patterns commonly observed in real-world information seeking behavior. Together, these two query sets enable comprehensive evaluation of IR systems across different query complexity levels and linguistic styles, providing a more robust assessment framework for Portuguese image retrieval capabilities.

To better understand the challenges inherent in relevance assessment across different query types, we conducted an analysis on annotation difficulty using the IAA data obtained during the dataset creation process. Figure 3.4 presents a comparative analysis of annotator agreement across generation methods and semantic categories. The analysis reveals complementary strengths between AI-generated and manually-created queries, with manually-created queries achieving higher agreement levels ($\kappa = 0.695 \pm 0.219$) compared to AI-generated queries ($\kappa = 0.494 \pm 0.259$). This difference reflects the distinct characteristics of each query type: manually-created queries, with their concise and direct nature, provide clear evaluation criteria that facilitate consistent annotation decisions, while AI-generated queries, though more contextually rich and descriptive, introduce additional semantic complexity that re-

quires more nuanced interpretation during annotation.

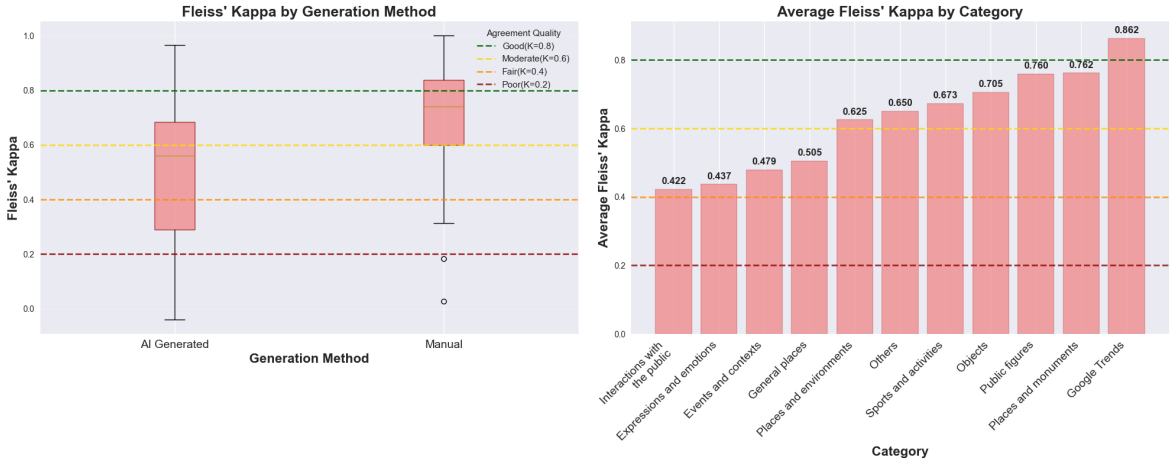

Figure 3.4: Comparative analysis of Fleiss' Kappa scores showing annotation difficulty across query generation methods (left) and categories (right).

The category-based analysis reveals variations in annotation difficulty, with Fleiss' Kappa scores exhibiting an agreement range of 0.440 across different semantic domains. Queries related to *trending topics* achieved the highest agreement levels ($\kappa = 0.862$), reflecting the concrete and well-defined nature of trending topics that provide clear visual references. In contrast, queries involving subjective assessments such as *interactions with the public* exhibited lower agreement ($\kappa = 0.422$), which is expected given the contextually dependent nature of human interaction evaluation. Categories such as *places and monuments* ($\kappa = 0.762$) and *public figures* ($\kappa = 0.760$) demonstrated strong agreement levels, indicating that concrete, identifiable content facilitates consistent annotation decisions.

These findings highlight the inherent complexity of relevance assessment in image retrieval tasks and demonstrate the value of our dual-approach query generation strategy. The observed variations in annotation difficulty across different semantic domains provide valuable insights for future dataset construction efforts and inform the development of category-specific evaluation metrics that account for the varying levels of subjectivity inherent in different types of visual content assessment. The combination of AI-generated and manually-created queries in our dataset ensures comprehensive coverage of both linguistically diverse and semantically focused query patterns, providing a robust foundation for evaluating Portuguese image retrieval systems across different complexity levels and user interaction scenarios.

The complete dataset has been made publicly available through Zenodo[8] and GitHub[9] to facilitate reproducibility and enable further research in Portuguese image IR systems. The GitHub repository includes all dataset files necessary for full replication of our experiments, while the Zenodo link provides a stable, citable version of the dataset for academic use.

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

3.3 Summary

This chapter addressed the critical gap in Portuguese image IR resources across two major areas. First, it conducted an analysis of existing datasets in image retrieval and captioning, examining prominent resources such as MS COCO, Flickr30k, #PraCegoVer, Fashion200k, WIT, and MIRACL-VISION to evaluate their applicability to Portuguese contexts. Second, it presented the development of a novel European Portuguese image retrieval dataset, detailing the methodology for automated content extraction, query generation, and annotation processes using content from the Portuguese Presidency website.

These efforts illustrate the fundamental challenges in adapting existing multilingual resources to specific language contexts and domains. The analysis revealed limitations in current datasets, such as, including absent query-response paradigms, exclusive English annotations, and cultural misalignments that prevent effective evaluation of Portuguese IR systems. The development of our specialized dataset demonstrates the necessity of domain-specific approaches while establishing methodological foundations for future Portuguese IR research.

Chapter 4

Proposed Hybrid Retrieval Algorithm

This chapter presents our hybrid retrieval approach for multimodal image IR systems, combining text-based and image-based retrieval methods through innovative score adjustment mechanisms. Section 4.1 introduces the hybrid retrieval algorithm with its three-phase process and mathematical formulations. Finally, Section 4.2 concludes the chapter with a summary of the key contributions and methodological insights of the proposed hybrid retrieval approach.

4.1 Hybrid Retrieval Algorithm

The revision of the state-of-the-art provided in Chapter 2 reveals that multimodal models have shown significant advancements in image retrieval tasks, particularly with CLIP models. While traditional text-based information retrieval systems have been effective in many scenarios, they often struggle with understanding the nuances of visual content. Vision-language models, such as CLIP, can address this challenge by leveraging both visual and textual information, enabling more accurate and context-aware image retrieval.

Various approaches have been proposed utilizing CLIP as a foundation for image retrieval tasks. Balauca et al.[38] demonstrated CLIP's effectiveness on museum exhibit datasets, showing superior performance over traditional methods. Iijima et al.[4] developed *CM*, an unsupervised approach that uses generated captions as intermediate representations for cross-domain retrieval. Additionally, Carlsson et al.[5] addressed CLIP's language limitations by proposing multilingual adaptations through teacher-student learning.

Despite these advancements, CLIP models excel at classification tasks, making them less effective for IR when queries contain complex terms or non-visual concepts. Addressing this problem using only visual content from images can lead to suboptimal results when the query is not directly related to the image's visual features. This limitation highlights the need for a more nuanced approach that combines visual and textual information effectively.

To refine the image IR process, we considered different methodological alternatives that could potentially enhance retrieval performance. Our primary focus was on leveraging the capabilities of CLIP models, particularly the multilingual adaptations already available. One natural direction was to explore fine-tuning OpenCLIP with our dataset to determine whether it could better capture the nuances of Portuguese queries in the domain of Portuguese presidency images. However, given the scale of our dataset and computational constraints, preliminary tests suggested that fine-tuning degraded performance compared to the base multilingual model. We also considered more complex strategies, such as using multiple multimodal networks combined through ensemble algorithms, but these approaches would have required substantial computational resources and introduced latency that is impractical for real-time applications. Based on these considerations, we decided to focus on a hybrid re-

trieval solution that leverages CLIP's pretrained capabilities while introducing a score adjustment mechanism to combine text-based and image-based retrieval effectively.

Our proposed algorithm implements a multimodal retrieval approach that combines text-based retrieval, direct image retrieval, and a score adjustment mechanism to balance the contributions of both retrieval modalities. Overall, it operates through three distinct phases, each designed to optimize different aspects of the retrieval process while maintaining computational efficiency and result quality. The first phase performs dual retrieval by conducting parallel text-based and image-based searches to generate complementary result sets. The second phase applies a score adjustment mechanism to harmonize the different score distributions between modalities, ensuring balanced contributions from both approaches. The third phase combines and deduplicates the adjusted results to produce a unified ranking that leverages the strengths of both retrieval methods. Figure 4.1 provides an overview of the algorithm's architecture and data flow across all three phases.

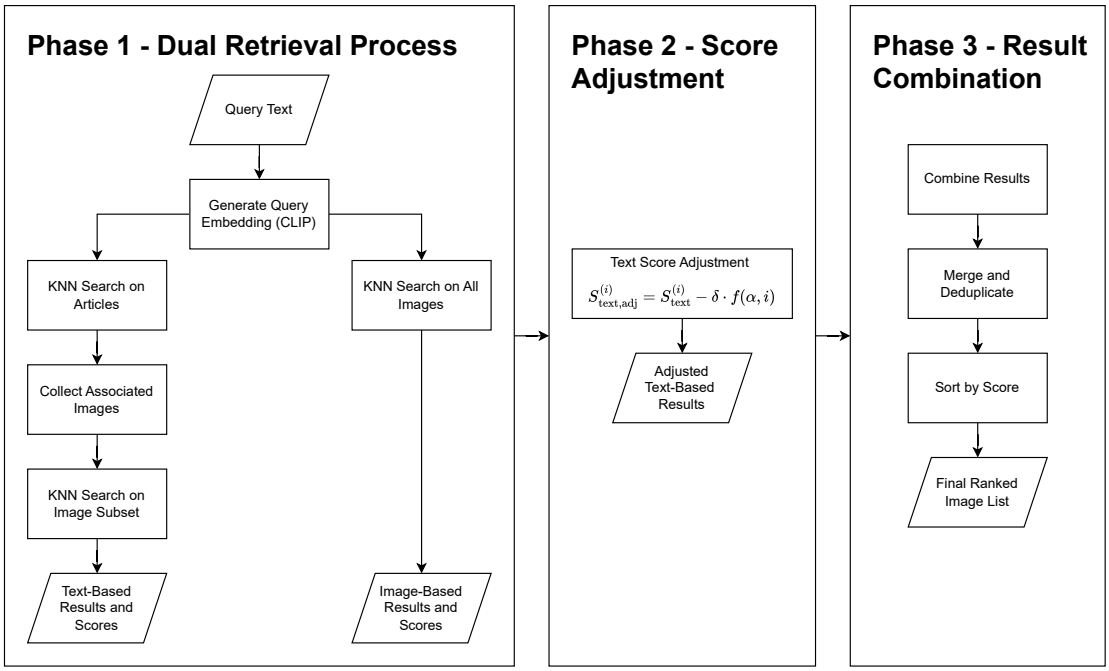

Figure 4.1: Hybrid Retrieval Algorithm Overview

4.1.1 Step 1: Dual Retrieval Process

The first step of our hybrid retrieval algorithm begins by generating a text embedding for the input query using the selected multilingual CLIP model. CLIP generates embeddings for both text and images in the same vector space, enabling direct comparison between textual queries and visual content through cosine distance. This query embedding serves as the foundation for both retrieval methods, ensuring consistency in the semantic representation used across different modalities.

Our system operates on a pre-computed database of embeddings created during a preprocessing phase. Specifically, we used the same CLIP model to generate: (1) text embeddings for all article titles from the Portuguese Presidency website, and (2) image embeddings for

all 42,333 collected images. Both types of embeddings reside in the same high-dimensional vector space, which is the key property that enables CLIP to perform cross-modal retrieval tasks.

The text-based retrieval method leverages the semantic richness of article content through a two-stage process. First, a K-Nearest Neighbors (KNN) search is performed using the query embedding against the article title embeddings. This search retrieves the top-k most semantically similar articles producing cosine distance scores that quantify how far each article title is from the query in the embedding space (lower values indicate closer alignment). The rationale behind this approach is that articles with titles semantically related to the query are likely to contain relevant images, even if those images might not be directly retrievable through visual similarity alone.

Subsequently, all images associated with these top-k retrieved articles are collected and indexed temporarily to create a focused search space. This temporary indexing is essential because performing KNN search across the entire image collection would likely exclude potentially relevant images that score poorly in the global embedding space, despite their contextual relevance to the query. By constraining the search to images from semantically relevant articles, we ensure that images with high contextual relevance but lower visual-semantic alignment can still be properly ranked and retrieved. A second KNN search is then performed using the same query embedding against the pre-computed image embeddings within this subset. This approach enables the system to generate meaningful distance scores for images that might otherwise be overshadowed by globally higher-scoring but contextually less relevant images. This yields a ranked list of images with their corresponding cosine distance scores, referred to as "text-based image results" since the images were initially filtered through textual semantic similarity (article titles) before being ranked by visual-semantic distance.

Simultaneously, the direct image retrieval method conducts a KNN search using the query embedding directly against all pre-computed image embeddings stored in the database, without any intermediate filtering. This approach yields a ranked list of images with their cosine distance scores, providing a complementary perspective that focuses purely on visual-semantic alignment between the query and image content. We refer to these as "image-based results" because the retrieval is based solely on direct visual-semantic comparison without any textual intermediation.

By explicitly producing distance scores for both modalities, Phase 1 establishes the necessary foundation for the subsequent score adjustment mechanism. These scores often differ in scale and distribution across modalities, which motivates the harmonization process in Phase 2.

4.1.2 Step 2: Score Adjustment

The second phase of our proposed algorithm addresses the critical challenge of balancing contributions from both retrieval modalities. A fundamental issue in multimodal retrieval systems is that the same retrieval mechanism applied to different search spaces often produces distance scores with different distributions and scales. Specifically, while both retrieval

pipelines use identical KNN search with cosine distance, the text-based pipeline (operating on article title embeddings followed by image filtering) and the image-based pipeline (operating directly on the complete image embedding space) generate scores in different ranges, making direct combination problematic. Although both text and image embeddings reside in the same CLIP multimodal space, the practical distribution of distance values differs significantly depending on the query type and embedding subset. Textual queries against article title embeddings typically yield lower distance scores because titles are concise, direct, and often semantically closer to the query terms. Conversely, queries against image embeddings generally produce higher scores, as the semantic relationship between text and images is more indirect and requires cross-modal interpretation. Without proper score harmonization to account for these systematic differences, one modality might systematically dominate the final ranking, potentially missing relevant results that would be highly ranked by the other modality.

This challenge reflects the core issue underlying our research question (RQ1): integrating text and image modalities for information retrieval requires aligning visual and textual representations, which exist in distinct semantic spaces and demand sophisticated mechanisms for effective cross-modal understanding.

Traditional multimodal retrieval systems commonly employ rank fusion methods, such as RRF [44], which normalize scores from different modalities to a common scale (e.g., using min-max normalization) before combining rankings. However, in our hybrid system, we adopt a different approach motivated by the asymmetric nature of our dual-retrieval architecture. Specifically, we treat image-based retrieval as a reliable baseline, since it represents direct text-to-image alignment in the CLIP embedding space. Text-based retrieval, while valuable for capturing contextual information, introduces an additional semantic layer through article titles, making it inherently less direct than the image-based method.

This methodological choice leads us to design an innovative score adjustment mechanism that modifies only the text-based scores to align with the image-based distribution, rather than normalizing both modalities equally. In doing so, we preserve the integrity of the direct visual-semantic correspondence while harmonizing the indirect textual modality. Although this departs from conventional rank fusion methodologies, it is specifically tailored to the asymmetric setup of our system, where one approach (image-based) provides direct semantic correspondence and the other (text-based) introduces beneficial contextual information through an intermediate textual representation.

To implement this mechanism, we first compute the difference between the top-ranked scores from each modality, and then apply a position-dependent adjustment to the text-based scores before combining them with image-based results.

Let $S_{\text{text}}^{(i)}$ and $S_{\text{image}}^{(i)}$ denote the distance scores for the i-th ranked result from text-based and image-based retrieval, respectively. We first compute the difference between the top-ranked scores of each modality as shown in Equation 4.1:

$$\delta = S_{\text{text}}^{(1)} - S_{\text{image}}^{(1)}. \quad (4.1)$$

This delta value serves as an estimate of the systematic offset between the two scoring distributions. By comparing the top-ranked results from each modality, we obtain a practical approximation of the overall difference between text-based and image-based similarity scores. The underlying assumption is that if the lowest-scoring result from the text-based pipeline consistently differs from the corresponding image-based result by a certain margin, this difference reflects the general scaling offset between the modalities. A positive delta indicates that text-based scores are generally higher, while a negative delta indicates the opposite tendency. This top-1 score difference is then used as the reference for harmonizing the entire score distributions across both modalities. To implement this harmonization, we propose a general score adjustment framework defined in Equation 4.2:

$$S^{(i)}_{\text{text,adj}} = S^{(i)}_{\text{text}} - \delta \cdot f(\alpha, i). \quad (4.2)$$

Here, $f(\alpha, i)$ denotes the position-dependent adjustment function, $\alpha \in [0, 1]$ controls the strength of the adjustment, and i represents the rank position (1-indexed). This formulation provides flexibility to balance the contributions of text and image modalities, allowing fine-tuning based on dataset characteristics and retrieval requirements.
We explore four distinct formulations for $f(\alpha, i)$, each with different mathematical properties and behavioral characteristics designed to address various aspects of the multimodal retrieval challenge:

1. Linear Adjustment with Zero-Indexing (Section 4.1.2.1)
2. Linear Adjustment with One-Indexing (Section 4.1.2.2)
3. Square Root Adjustment (Section 4.1.2.3)
4. Exponential Adjustment (Section 4.1.2.4)

These formulations are described in detail in the following subsections.

4.1.2.1 Linear Adjustment with Zero-Indexing

The linear adjustment with zero-indexing follows the pattern defined in Equation 4.3:

$$f(\alpha, i) = 1 - \alpha \cdot (i - 1). \quad (4.3)$$

This approach applies a simple linear decay starting from the first result (index 0). The adjustment decreases linearly with each subsequent rank, providing predictable and uniform score modification across positions. When $\alpha = 0$, no adjustment is applied, preserving the original score distributions. When $\alpha = 1$, maximum linear adjustment is applied to the first result, with the adjustment decreasing linearly for each subsequent rank position. The zero-indexing ensures that the top-ranked result receives the full delta adjustment, while lower-ranked results receive progressively smaller adjustments. This creates a smooth

transition that maintains the relative ordering within the adjusted modality while achieving cross-modal score alignment.

4.1.2.2 Linear Adjustment with One-Indexing

The linear adjustment with one-indexing is defined in Equation 4.4:

$$f(\alpha, i) = 1 - \alpha \cdot i. \quad (4.4)$$

Similar to the zero-indexed version but starting the decay from position 1. This variation applies a different adjustment pattern where even the top-ranked result receives only a partial adjustment when $\alpha < 1$. The one-indexing approach distributes the adjustment more conservatively across all positions, potentially preserving more of the original score relationships within the text-based modality. This may be beneficial when the original text-based rankings are considered highly reliable and should be minimally perturbed.

4.1.2.3 Square Root Adjustment

The square root adjustment function is expressed in Equation 4.5:

$$f(\alpha, i) = 1 - \alpha^{\sqrt{i-1}}. \quad (4.5)$$

This function provides a non-linear decay pattern that starts aggressively but tapers off more gradually than exponential approaches. The square root scaling creates a moderate adjustment curve that falls between linear and exponential behaviors. For small values of i, the adjustment is close to the full delta value, but it decreases more slowly than linear functions for larger i values. This behavior may be advantageous when both high precision for top results and reasonable recall for lower-ranked results are important, as it maintains meaningful adjustments deeper into the ranking while still emphasizing the top positions.

4.1.2.4 Exponential Adjustment

The exponential adjustment function takes the form presented in Equation 4.6:

$$f(\alpha, i) = 1 - \alpha^{e^{i-1}}. \quad (4.6)$$

This represents the most aggressive adjustment approach, where the exponential term e^{i-1} creates a rapidly diminishing effect. The function exhibits extreme focus on the top-ranked results, with adjustments becoming negligible very quickly as the rank position increases. For $i > 2$, the adjustment becomes virtually zero unless α is very close to 1. This aggressive decay pattern concentrates the score harmonization effect almost exclusively on the first few

results, making it suitable for applications where users rarely examine results beyond the top positions and maximum precision at the top of the ranking is of most importance.

4.1.2.5 Summary and Comparison of Adjustment Functions

Each adjustment function addresses different aspects of the multimodal retrieval challenge and exhibits distinct mathematical properties that make them suitable for different scenarios. Linear functions provide straightforward, interpretable adjustments that are easy to tune and understand, making them suitable for scenarios requiring predictable behavior and transparency in the adjustment process. The linear relationship between rank position and adjustment strength allows for intuitive parameter tuning and clear understanding of how the algorithm affects different ranking positions.

Square root scaling offers a balanced approach that moderately emphasizes top results while still applying meaningful adjustments to mid-ranked items. This function may be particularly useful when both precision and recall are important evaluation criteria, as it maintains a reasonable adjustment effect throughout a broader range of ranking positions compared to exponential approaches.

Exponential scaling maximizes the focus on top-ranked results, which may be valuable when users primarily examine the first few results and the application demands maximum precision at the top of the ranking. However, this aggressive focusing comes at the cost of virtually ignoring lower-ranked results, potentially reducing the algorithm's ability to improve recall metrics.

The mathematical properties of these functions also differ in terms of computational complexity and parameter sensitivity. Linear functions are computationally efficient and exhibit stable behavior across different parameter values. Square root functions introduce moderate computational overhead while maintaining reasonable parameter stability. Exponential functions, while still computationally feasible, may exhibit more sensitive behavior to parameter changes and require more careful tuning.

The choice of adjustment function depends on the specific requirements of the retrieval task, the relative importance of top-ranked versus lower-ranked results, the computational constraints of the system, and the desired balance between maintaining original modality rankings and achieving cross-modal score harmonization. Different application contexts may favor different approaches based on user behavior patterns, evaluation criteria, and system performance requirements.

To illustrate the practical differences between these adjustment functions, we present a concrete example using realistic distance scores derived from our hybrid retrieval system for the Portuguese query "Universidade da Beira Interior". The example demonstrates how each adjustment function affects the integration of text-based and image-based retrieval results in a unified ranking.

Table 4.1 shows the initial distance scores from both retrieval modalities for this query. The text-based scores represent distance values obtained from the two-stage process (article retrieval followed by image filtering), while the image-based scores represent direct visual-semantic distance between the query and images. As expected in cosine distance metrics,

lower scores. The difference between the two modalities reflects the different search spaces and semantic interpretation challenges inherent in cross-modal retrieval.

Table 4.1: Initial Distance Scores from Both Retrieval Modalities

Rank	Text-based Score	Image-based Score
1	0.785	0.730
2	0.786	0.737
3	0.792	0.739
4	0.797	0.741
5	0.799	0.746

For this example, we calculate $\delta = 0.785 - 0.730 = 0.055$ and use $\alpha = 0.2$ to demonstrate the adjustment mechanism's behavior. The relatively smaller delta value compared to our theoretical examples reflects the realistic score distributions observed in our system, where both modalities operate within the same CLIP embedding space but exhibit distributional differences.

To understand the impact of score adjustment on the final retrieval ranking, we must consider how text-based and image-based results are integrated. In our notation, T*i* represents a result originally ranked at position *i* in the text-based modality, while I*i* represents a result originally ranked at position *i* in the image-based modality. The final unified ranking combines both types of results, ordering them by their adjusted distance scores from lowest (best) to highest (worst).

Table 4.2 presents the top 10 positions in the final unified ranking for each adjustment method. Each cell contains the adjusted distance score and, in parentheses, the original source and rank of that result. This representation clearly illustrates how different adjustment functions alter the relative balance between text-based and image-based retrieval contributions in the final ranking.

Table 4.2: Combined Ranking Comparison Across Adjustment Methods ($\alpha = 0.2$, $\delta = 0.055$)

Final Rank	Original Score	Linear Zero	Linear One	Square Root	Exponential
1	0.730 (I1)	0.730 (T1)	0.730 (I1)	0.730 (I1)	0.730 (I1)
2	0.737 (I2)	0.730 (I1)	0.737 (I2)	0.737 (I2)	0.732 (T2)
3	0.739 (I3)	0.737 (I2)	0.739 (I3)	0.739 (I3)	0.737 (I2)
4	0.741 (I4)	0.739 (I3)	0.741 (T1)	0.741 (I4)	0.737 (T3)
5	0.746 (I5)	0.741 (I4)	0.741 (I4)	0.742 (T3)	0.739 (I3)
6	0.785 (T1)	0.743 (T2)	0.746 (I5)	0.743 (T2)	0.741 (T1)
7	0.786 (T2)	0.746 (I5)	0.754 (T2)	0.746 (T4)	0.741 (I4)
8	0.792 (T3)	0.759 (T3)	0.770 (T3)	0.746 (I5)	0.743 (T4)
9	0.797 (T4)	0.775 (T4)	0.786 (T4)	0.747 (T5)	0.745 (T5)
10	0.799 (T5)	0.788 (T5)	0.799 (T5)	0.785 (T1)	0.746 (I5)

The analysis of the combined ranking results reveals distinct behavioral patterns across the four adjustment methods. In the baseline configuration without adjustment, all image-based results (I1-I5) completely dominate the top 5 positions due to their systematically lower distance scores, while text-based results (T1-T5) appear exclusively in positions 6-10. This distribution demonstrates the fundamental challenge that motivated our adjustment mechanism: without proper score harmonization, one retrieval modality systematically dominates

the final ranking, potentially excluding highly relevant results from the alternative modality.

Linear zero-indexing adjustment produces the most aggressive harmonization behavior, successfully promoting T1 to the top position (rank 1) and achieving perfect score alignment with the best-performing image result. This adjustment creates the most balanced cross-modal integration, with text and image results distributed throughout the top 10 positions. The substantial repositioning of text-based results demonstrates this method's effectiveness in achieving cross-modal score harmonization while preserving meaningful quality distinctions between individual results.

Linear one-indexing adjustment implements a more conservative harmonization approach, with T1 achieving position 4, representing a significant improvement from its original position 6. This method maintains greater preservation of the original image dominance within the top positions while still providing meaningful integration of text-based results. The approach establishes a smoother transition that respects the original score distributions while achieving cross-modal balance through controlled adjustment magnitudes.

Square root adjustment generates an interesting mixed integration pattern, with T3 achieving position 5 and T2 reaching position 6. This behavior demonstrates how non-linear adjustment functions can create different promotion patterns compared to linear methods. The square root approach achieves effective cross-modal integration while maintaining sensitivity to the original ranking structure, as evidenced by T1 remaining at position 10 despite receiving adjustment treatment.

Exponential adjustment exhibits distinctly different behavior from the other methods, achieving extensive text-image integration distributed throughout the ranking structure. Notably, T2 and T3 achieve positions 2 and 4 respectively, creating the most distributed multimodal integration pattern observed. This demonstrates the exponential function's capacity to provide fine-grained score adjustments that create natural interspersion of results from both retrieval modalities.

The choice of adjustment function directly shapes how text-based and image-based scores are harmonized and, consequently, the final ranking of retrieved items. Linear zero-indexing applies the most aggressive harmonization, ensuring that both modalities contribute evenly and is suitable for scenarios where strong multimodal integration is desired. Linear one-indexing offers a more controlled integration, preserving a substantial portion of the original ranking hierarchy while still promoting cross-modal balance. Square root adjustment provides moderate integration, selectively promoting text-based results to achieve a compromise between cross-modal harmonization and ranking stability. Exponential adjustment distributes the influence of the text-based modality gradually, producing fine-grained integration that can be advantageous in exploratory search tasks or applications prioritizing the preservation of the inherent modality-specific ranking.

Overall, these conceptual distinctions highlight that the different mathematical formulations provide flexible strategies depending on the retrieval objectives: aggressive multimodal integration favors linear zero-indexing, moderate compromises can be achieved with square root or linear one-indexing, and preservation-focused approaches benefit from exponential adjustment. By carefully selecting the appropriate adjustment function, system designers

can tailor the balance between modalities to align with specific application contexts and user interaction patterns.

4.1.3 Phase 3: Result Combination

The final phase combines the adjusted text-based results with the unchanged image-based results to produce a unified ranking. This combination process involves several critical steps to ensure the integrity and quality of the final result set.

The combination algorithm begins by merging the two result lists while handling potential duplicates. Duplicate entries are identified based on image identifiers or URLs, as the same image may appear in both result sets through different retrieval methods. When duplicates are detected, the algorithm retains the result with the better (lower) distance score, ensuring that each image appears only once in the final ranking while preserving the best available score for that image.

The deduplication process is crucial for maintaining ranking quality, as it prevents artificially inflating the importance of images that happen to be retrieved through both modalities. By selecting the better score for each duplicate, the algorithm ensures that the final ranking reflects the strongest evidence for relevance from either modality.

Following deduplication, the combined results are sorted in ascending order of their distance scores to determine the final ranking presented to the user. This sorting process creates a unified ranking that seamlessly integrates contributions from both retrieval modalities, with the score adjustment mechanism ensuring that neither modality systematically dominates the results.

The combination approach addresses the fundamental challenge of score distribution differences between the two retrieval modalities while maintaining computational efficiency. The position-dependent adjustment mechanism ensures that the balancing effect is most pronounced for top-ranked results, where retrieval accuracy has the greatest impact on user satisfaction and system effectiveness. Lower-ranked results receive proportionally smaller adjustments, preserving the relative ordering within each modality while achieving the necessary cross-modal score harmonization.

4.2 Summary

This chapter presented a hybrid retrieval approach for multimodal image IR that addresses the limitations of single-modality methods through a systematic three-phase algorithm. The methodology development process explored various enhancement strategies and led to the development of a hybrid solution that combines text-based and image-based retrieval methods with a configurable score adjustment mechanism. The hybrid retrieval algorithm operates through three distinct phases: dual retrieval process, score adjustment, and result combination, systematically balancing contributions from both text-based and image-based modalities while maintaining computational efficiency and result quality. The design of the score adjustment mechanism directly addresses the core challenge underlying our research question (RQ1): integrating text and image modalities for information retrieval requires

aligning visual and textual representations, which exist in distinct semantic spaces and demand sophisticated mechanisms for effective cross-modal understanding.

Four different score adjustment functions were presented as methodological alternatives for harmonizing cross-modal distance scores, each exhibiting distinct mathematical properties and behavioral characteristics suited for different retrieval scenarios. The framework provides flexibility for adapting the algorithm to specific application requirements and user behavior patterns, ranging from linear approaches that provide interpretable and uniform adjustments, to exponential methods that focus intensively on top-ranked results. The result combination phase ensures seamless integration of multimodal evidence while maintaining ranking integrity through systematic deduplication and score-based sorting, creating a robust foundation for multimodal image retrieval that can be adapted to various domains and requirements.

Chapter 5

Experiments and Results

This chapter presents the experimental evaluation of the proposed hybrid retrieval framework. The goal is to systematically assess how different retrieval strategies perform in the context of Portuguese multimodal information retrieval, and to determine the contribution of each component of our approach. To achieve this, we first establish a solid reference using multiple baselines, then evaluate the effect of fine-tuning a vision-language model on the domain-specific dataset, and finally analyze the hybrid retrieval approach with its novel score adjustment mechanism. Section 5.1 describes the experimental configuration and evaluation methodology. Section 5.2 presents the experimental results. We begin by establishing benchmarks for baseline methods on the task at hand. We then examine whether different queries exhibit varying retrieval performance. Finally, we investigate the potential benefits of fine-tuning and evaluate our proposed hybrid retrieval approach. Finally, Section 5.3 concludes with a summary of key findings and answers to the research questions.

5.1 Experimental Setup

The experiments are designed to provide a comprehensive assessment of the retrieval task by establishing reference baselines, applying domain adaptation, and evaluating multimodal fusion. This section introduces the dataset, the evaluation metrics, and the baseline systems against which our methods are compared.

5.1.1 Dataset

The experimental evaluation relies on the Portuguese multimodal retrieval dataset introduced in Chapter 3, which comprises 80 natural language queries in European Portuguese, manually formulated to cover a range of topics, and 5,201 annotated images, retrieved from the Presidency's official website and judged for relevance with respect to the queries.
For experiments requiring model adaptation, such as fine-tuning CLIP, we utilized the complete collection of 42,333 images from the Portuguese Presidency website. The fine-tuning dataset was split into training and validation sets using an 80/20 split: 80% of the images (33,866 images) were used for training and 20% (8,467 images) were reserved for validation. This division ensured adequate training data while maintaining the ability to evaluate performance on unseen data and detect potential overfitting.
It is important to note that all performance evaluations presented in this chapter are conducted exclusively on the separate annotated test dataset described in Chapter 3 (80 queries and 5,201 images). This evaluation dataset serves as an independent downstream task for image retrieval assessment, ensuring that all reported results reflect genuine retrieval performance rather than training or validation performance. The fine-tuning dataset (42,333

images) is used solely for model adaptation and is never used for evaluation purposes, maintaining proper experimental separation between training and testing phases.

5.1.2 Evaluation Metrics

To assess the performance of different retrieval systems, we use a diverse set of evaluation metrics that capture various aspects of retrieval effectiveness. Our evaluation includes Precision at Rank k (P@k), Recall at Rank k (R@k), F1 Score at Rank k (F1@k), Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and R-Precision (Rprec). Each metric provides unique insights into system performance.
Precision at Rank k (P@k) measures the proportion of relevant images among the top k retrieved results, defined as:

$$P@k = \frac{\text{Number of relevant images in top k results}}{k}. \quad (5.1)$$

This metric is particularly important for retrieval effectiveness evaluation, as it measures the precision of the top-ranked results returned by the system.
Recall at Rank k (R@k) measures the proportion of relevant images retrieved within the top k results relative to the total number of relevant images:

$$R@k = \frac{\text{Number of relevant images in top k results}}{\text{Total number of relevant images}}. \quad (5.2)$$

Recall metrics are essential for understanding the coverage capabilities of retrieval systems, particularly important for extensive search scenarios where users need to find multiple relevant items.
F1 Score at Rank k (F1@k) provides a harmonic mean of precision and recall at rank k:

$$F1@k = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k}. \quad (5.3)$$

The F1 score is particularly valuable for assessing overall system effectiveness as it penalizes systems that optimize one metric at the expense of the other.
Mean Reciprocal Rank (MRR) evaluates the quality of ranking by considering the position of the first relevant image for each query:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}. \quad (5.4)$$

where $|Q|$ is the number of queries and rank_i is the position of the first relevant image for query i. MRR is particularly sensitive to the ranking quality of top results, making it crucial for evaluating the effectiveness of retrieval systems in identifying relevant content.
MAP computes the average precision across all queries:

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP_i. \quad (5.5)$$

where AP_i is the average precision for query i. MAP provides a comprehensive assessment of ranking quality across the entire result list.

R-Precision (Rprec) measures precision at the rank equal to the number of relevant images for each query, automatically adapting to the specific characteristics of each query. This metric provides a balanced assessment that considers both precision and the natural cutoff point determined by relevance judgments.

For statistical robustness analysis of experimental results, we also use the Coefficient of Variation (CV), defined as:

$$CV = \frac{\sigma}{\mu} \times 100\%. \quad (5.6)$$

where σ is the standard deviation and μ is the mean of a performance metric across bootstrap samples. The coefficient of variation provides a normalized measure of dispersion that enables direct comparison of variability across different methods regardless of their absolute performance levels, with lower values indicating more consistent performance.

These metrics collectively provide a thorough evaluation that captures different aspects of retrieval performance, enabling analysis of system strengths and limitations across various operational scenarios.

5.1.3 Baselines

Our evaluation encompasses a carefully selected set of baseline systems representing different paradigms and technological approaches to image retrieval. This collection enables analysis of the state-of-the-art while establishing performance benchmarks across multiple retrieval methodologies.

Traditional Lexical IR Techniques serve as fundamental baselines, representing established approaches that have formed the foundation of IR systems for decades. We evaluate TF-IDF [59] and BM25 [60] algorithms applied to article titles associated with images, representing each image by the title of the article in which it appeared. These systems provide essential baselines for understanding the effectiveness of textual metadata-based approaches.

End-to-End Image Retrieval Systems represent sophisticated, production-ready systems that incorporate decades of optimization and proprietary ranking algorithms. We evaluate Google Images and Arquivo.pt, both constrained to return results exclusively from the Portuguese Presidency website to ensure contextual relevance. These systems provide realistic baselines that reflect real-world search capabilities.

Portuguese Language Embedding Models offer baselines designed specifically for Portuguese language understanding. We evaluate BERTimbau Large [64] (Brazilian Portuguese) and Albertina PT-PT [65] (European Portuguese) models. These models were specifically chosen to account for the linguistic differences between European Portuguese and Brazilian Portuguese, with Albertina PT-PT being particularly relevant for European Portuguese contexts.

Vision-Language Models represent the current state-of-the-art for multimodal retrieval tasks. We employed several CLIP variants, including multilingual adaptations such as Multilingual-

CLIP [5] and OpenCLIP [40, 41, 2], which were trained on large-scale cross-modal datasets and are capable of handling queries in multiple languages. 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*, and *OpenCLIP/xlm-roberta-base-ViT-B-32*, along with more recent models such as BLIP-2 [66] and SigLIP [67].

5.2 Experiments

The experiments are organized progressively: we begin with a comprehensive benchmark of baseline systems, followed by the evaluation of a fine-tuned CLIP model adapted to Portuguese, and finally assess the hybrid retrieval approach.

5.2.1 Benchmarking and Query Analysis

We first benchmark all baseline systems using the metrics defined in Section 5.1.2. Table 5.1 summarizes the results across models, providing an initial reference for subsequent experiments. The experimental results reveal clear differences between categories of baselines.

Table 5.1: Performance comparison of baseline methods on Portuguese image retrieval task. Best results in each column are highlighted in bold.

Method	MAP	P@5	R@5	P@10	R@10	F1@10	MRR	RP
Traditional Text-Based IR								
TF-IDF	0.138	0.307	0.107	0.291	0.188	0.209	0.377	0.175
BM25 (Redis)	0.079	0.195	0.068	0.209	0.127	0.141	0.331	0.108
End-to-End Image Retrieval Systems								
Google Images	0.076	0.202	0.076	0.190	0.125	0.128	0.336	0.113
Arquivo.pt	0.038	0.145	0.045	0.134	0.080	0.091	0.217	0.072
Portuguese Language Embedding Models								
BERTimbau Base	0.035	0.080	0.029	0.090	0.059	0.066	0.131	0.055
BERTimbau Large	0.048	0.088	0.035	0.103	0.076	0.080	0.157	0.065
Albertina PT-PT	0.055	0.132	0.052	0.131	0.091	0.096	0.206	0.080
Vision-Language Models								
M-CLIP LABSE-ViT-L-14	0.130	0.317	0.100	0.341	0.210	0.237	0.468	0.189
M-CLIP XLM-R-Large-B-32	0.119	0.322	0.091	0.334	0.176	0.215	0.512	0.167
M-CLIP XLM-R-Large-L-14	0.158	0.365	**0.123**	0.376	0.245	0.266	0.491	0.209
OpenCLIP xlm-roberta-base	**0.176**	**0.418**	0.121	**0.419**	**0.264**	**0.288**	**0.610**	**0.218**
BLIP-2 ViT-G	0.015	0.063	0.015	0.062	0.030	0.039	0.140	0.030
SigLIP Base	0.107	0.240	0.083	0.262	0.168	0.181	0.457	0.147

The experimental results reveal clear differences between categories of baselines. Traditional lexical approaches such as TF-IDF and BM25 perform poorly in terms of semantic coverage (MRR $= 0.377$ for TF-IDF), as expected given their lack of cross-modal capability. Portuguese embedding models (BERTimbau, Albertina PT-PT) improve retrieval effectiveness by leveraging contextual semantics, but remain limited by their inability to process visual information. Their performance ranges from MRR $= 0.131$ to 0.206, with Albertina PT-PT achieving the strongest results among them, highlighting the importance of variant-specific modeling despite the fundamental limitation of monolingual text-only approaches for multimodal retrieval.

These results directly address RQ2, revealing significant limitations of existing multilingual solutions for Portuguese image IR. Commercial systems demonstrate poor performance when constrained to Portuguese contexts (Google Images: 0.336 MRR, Arquivo.pt: 0.217 MRR), while Portuguese-specific language models show severe multimodal limitations (best: Albertina PT-PT with 0.206 MRR), and even multilingual vision-language models like M-CLIP variants underperform compared to models with comprehensive multilingual pre-training. Vision-language models substantially outperform all other approaches, confirming their ability to align queries and images in a shared embedding space. The performance gap is considerable (with OpenCLIP obtaining 0.610 MRR and TF-IDF only 0.377) illustrating the significant advantages of direct visual understanding over text-based proxy approaches. Among these, OpenCLIP *xlm-roberta-base* achieves the best results across all key metrics, with an MRR of 0.610 and F1@10 of 0.288. This demonstrates the effectiveness of large-scale multilingual pre-training and inherent vision-language capabilities in Portuguese contexts.

5.2.1.1 Query Characteristics Analysis

To better understand system behavior, we conducted a query characteristics analysis. For this purpose, we analyzed system effectiveness across two distinct query types: longer queries generated using the GPT model (38 queries, average length 6-8 words) and shorter queries created manually (42 queries, average length 1-3 words).
Figure 5.1 presents the performance comparison across these query types using MRR and F1@10 metrics. We focus on these two metrics as they provide complementary perspectives on retrieval effectiveness: MRR evaluates ranking quality by measuring how quickly systems identify the first relevant result, which is crucial for understanding immediate retrieval success, while F1@10 provides a balanced assessment of precision and recall at a practical cutoff point, capturing both accuracy and coverage within the top-10 results that users typically examine.

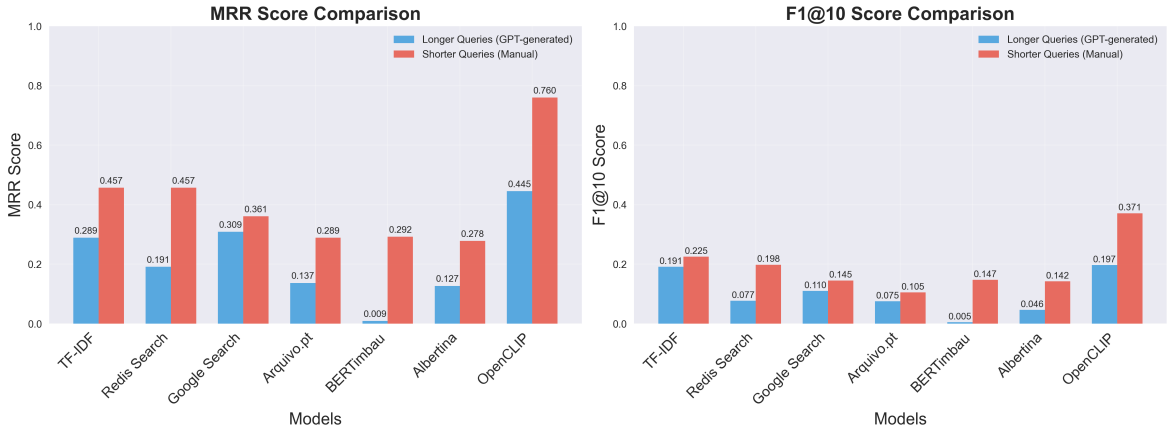

Figure 5.1: Performance comparison between short and long queries across different retrieval methods

Vision-language models achieve stronger results for shorter queries across both evaluation metrics. With OpenCLIP, MRR rises from 0.445 for longer queries to 0.760 for shorter ones, a 71% improvement in identifying the first relevant result. Similarly, F1@10 performance

increases from 0.197 to 0.371, representing an 88% improvement that demonstrates the substantial advantage of concise queries for balanced precision-recall performance.

This performance difference can be attributed to the architectural design and training methodology of vision-language models. Since these models were primarily developed for image classification tasks, optimized to associate images with short category labels rather than long descriptive sentences, they inherit an inherent preference for compact query formulations.

5.2.1.2 Statistical Robustness Analysis

To provide statistically robust performance estimates and assess the reliability of our experimental findings, we conducted a bootstrap analysis with 1,000 iterations across all evaluated methods. This analysis provides 95% confidence intervals for each performance metric, enabling assessment of result stability and statistical significance of performance differences between methods. Using the coefficient of variation as defined in Section 5.1.2 to quantify the relative variability of different methods across bootstrap samples.

The bootstrap analysis reveals several important insights about the statistical robustness and variance characteristics of our experimental findings. Starting with the weakest performing approaches, BLIP-2 ViT-G shows consistently poor performance with very tight confidence intervals, achieving MRR $= 0.139 \pm 0.034$ and F1@10 $= 0.044 \pm 0.010$. These narrow intervals indicate not only consistent poor performance but also low variance across bootstrap samples, with coefficients of variation of 24.5% for MRR and 22.7% for F1@10, confirming both the model's limitations and its predictable behavior for image retrieval tasks.

Portuguese language embedding models demonstrate moderate performance with relatively stable confidence bounds but notable variance differences. BERTimbau Base achieves MRR $= 0.203 \pm 0.038$ and F1@10 $= 0.097 \pm 0.021$, while BERTimbau Large shows improved performance at 0.221 ± 0.042 MRR and 0.116 ± 0.024 F1@10. Albertina PT-PT performs best among text-only models with MRR $= 0.257 \pm 0.044$ and F1@10 $= 0.126 \pm 0.023$. The variance analysis reveals that these models exhibit coefficients of variation ranging from 15-20% for most metrics, indicating moderate but consistent performance variability across different query subsets. Notably, Albertina PT-PT shows the highest variance among Portuguese models (17.1% coefficient of variation for MRR), suggesting that while it achieves better average performance, it is more sensitive to query characteristics than the BERTimbau variants.

Vision-language models demonstrate significantly higher performance with varying degrees of statistical variability that provide insights into model stability. The M-CLIP variants show performance ranges from 0.466 ± 0.046 to 0.509 ± 0.049 for MRR, with confidence intervals representing approximately 9-10% coefficients of variation. Interestingly, M-CLIP XLM-32 exhibits the lowest variance (9.6% coefficient of variation) despite moderate performance, while M-CLIP XLM-14 shows slightly higher variance (9.6%) but superior average performance, indicating different sensitivity patterns to query characteristics. SigLIP Base demonstrates moderate performance with MRR $= 0.457 \pm 0.041$ and F1@10 $= 0.181 \pm 0.024$, showing reasonable statistical stability with approximately 9.0% coefficient of variation for MRR. While SigLIP Base outperforms text-only Portuguese models and basic vision-language approaches like BLIP-2, it exhibits higher variance compared to the more robust M-CLIP vari-

ants, indicating moderate sensitivity to query characteristics. OpenCLIP Laion5b achieves the strongest baseline performance with MRR $= 0.609 \pm 0.048$ and F1@10 $= 0.288 \pm 0.024$, demonstrating both high effectiveness and stable performance with only 7.9% coefficient of variation for MRR, indicating exceptional consistency across bootstrap iterations.

Most significantly, the hybrid retrieval approach (Linear Zero ($\alpha = 0.1$)) demonstrates superior performance compared to the baseline OpenCLIP model while maintaining comparable variance characteristics. The hybrid method achieves MRR $= 0.617 \pm 0.051$ compared to 0.609 ± 0.048 for the baseline, representing a meaningful improvement with overlapping but shifted confidence intervals. The variance analysis reveals that the hybrid approach exhibits a slightly higher coefficient of variation (8.3% vs 7.9%), indicating marginally increased variability, which is expected given the additional complexity of multimodal fusion. For F1@10, the hybrid approach reaches 0.291 ± 0.023 versus 0.288 ± 0.024 for the baseline, showing consistent enhancement in precision-recall balance with virtually identical variance characteristics (7.9% vs 8.3% coefficient of variation).

The bootstrap analysis confirms that performance improvements from the hybrid approach are statistically meaningful rather than artifacts of specific query selections or increased variance. The confidence intervals demonstrate that the hybrid method consistently outperforms the baseline across different bootstrap samples, with the confidence bounds for MRR (0.566-0.668 for hybrid vs 0.561-0.657 for baseline) showing limited overlap in the upper performance ranges. This statistical validation supports the effectiveness of our score adjustment mechanism for image retrieval tasks without introducing substantial performance instability.

The variance analysis also reveals that vision-language models exhibit larger confidence intervals for precision-focused metrics (P@5, P@10) compared to recall metrics (R@5, R@10), indicating greater variability in top-ranked result quality. Specifically, precision metrics show coefficients of variation ranging from 8-12%, while recall metrics demonstrate more stable behavior with 6-10% coefficients of variation. This pattern suggests that while these models consistently identify relevant results, the exact ranking positions may vary depending on query characteristics, with precision at higher ranks being particularly sensitive to query formulation. The hybrid approach maintains similar variance patterns, indicating that the score adjustment mechanism preserves the statistical stability of the underlying vision-language model while achieving performance improvements.

These statistical findings provide strong evidence supporting the conclusions drawn from point estimates, confirming that the observed performance differences represent genuine improvements rather than random variations or artifacts of increased system complexity. The variance analysis demonstrates that the hybrid approach achieves meaningful performance gains without compromising system reliability or introducing excessive variability. The bootstrap analysis validates the robustness of our experimental methodology and strengthens confidence in the reported performance comparisons across all evaluated approaches. Complete bootstrapped results with detailed confidence intervals and variance statistics for all metrics are provided in Appendix C.

5.2.2 Fine-tuned Vision-language Model

The second experiment evaluates the effect of domain adaptation by fine-tuning OpenCLIP on the Portuguese dataset. The objective is to adapt a strong vision-language model to the linguistic and visual characteristics of the task.

To this regard, we fine-tuned the OpenCLIP *xlm-roberta-base* model using the 42,333 images collected from the Portuguese Presidency website with their associated article titles as labels.

Figure 5.2 illustrates the fine-tuning data preparation process with concrete examples. Each image from the Portuguese Presidency website is paired with the title of the article in which it appeared, creating image-text pairs for fine-tuning. The process involves extracting article metadata (shown in JSON format) and converting it into training pairs (shown in CSV format) where each image path is associated with its corresponding article title as the text label.

Article Data (JSON):

```
{
  "title": "Presidente da República recebe Presidente do Conselho Europeu",
  "date": "2025-01-02",
  "images": [
    "img_001234.jpg",
    "img_001235.jpg"
  ],
  "content": "O Presidente da República, Marcelo..."
}
```

Training Dataset (CSV):

```
image_path,text_label
img_001234.jpg,"Presidente da República recebe Presidente do Conselho Europeu"
img_001235.jpg,"Presidente da República recebe Presidente do Conselho Europeu"
img_001236.jpg,"Cerimónia de condecoração no Palácio de Belém"
```

Figure 5.2: Data preparation for fine-tuning OpenCLIP. Top: Article metadata in JSON format containing title, date, associated images, and content. Bottom: Training dataset in CSV format where each image is paired with its article title as the text label for contrastive learning.

The fine-tuning hyperparameters were initially based on the default configuration provided in the OpenCLIP fine-tuning examples. Starting from these default values (learning rate 5e-6, batch size 32, weight decay 0.1), several adjustments were made to improve training performance: the learning rate was increased to 1e-5, batch size doubled to 64, and weight decay reduced to 0.01. These modifications were motivated by the need to balance training stability with more effective domain adaptation, as the default parameters resulted in slow convergence for our specific dataset characteristics. Additionally, the hardware constraints of a GTX 1080 with 8GB of memory influenced the final hyperparameter selection,

particularly limiting the maximum achievable batch size and requiring careful memory management during training. Table 5.2 presents the final configuration used for the fine-tuning experiments.

Table 5.2: Hyperparameters used for fine-tuning the OpenCLIP xlm-roberta-base model.

Hyperparameter	Value
Batch Size	64
Learning Rate	1e-5
Weight Decay	0.01
Maximum Epochs	100
Validation Split	20%
Model Checkpoint Frequency	Every 20 epochs

The training process was conducted for a maximum of 100 epochs, with model checkpoints saved every 20 epochs to track performance evolution. The choice of 100 epochs was motivated by the need to balance comprehensive training exploration with computational constraints, allowing sufficient time for domain adaptation while preventing excessive training that could lead to forgetting the pre-trained representations. This approach resulted in five model snapshots at epochs 20, 40, 60, 80, and 100, enabling analysis of how training duration affects retrieval performance.

To evaluate the effectiveness of fine-tuning, we compared the performance of the fine-tuned OpenCLIP models with all baseline methods established in the previous section. Table 5.3 presents the comparison between baseline methods and fine-tuned OpenCLIP models, revealing the impact of domain adaptation on retrieval performance across all evaluation metrics.

Table 5.3: Performance comparison between baseline methods and fine-tuned OpenCLIP models

Method	MAP	P@5	R@5	P@10	R@10	F1@10	MRR	RP
Traditional Text-Based IR								
TF-IDF	0.138	0.307	0.107	0.291	0.188	0.209	0.377	0.175
BM25	0.079	0.195	0.068	0.209	0.127	0.141	0.331	0.108
End-to-End Image Retrieval Systems								
Google Images	0.076	0.202	0.076	0.190	0.125	0.128	0.336	0.113
Arquivo.pt	0.038	0.145	0.045	0.134	0.080	0.091	0.217	0.072
Portuguese Language Embedding Models								
BERTimbau Base	0.035	0.080	0.029	0.090	0.059	0.066	0.131	0.055
BERTimbau Large	0.048	0.088	0.035	0.103	0.076	0.080	0.157	0.065
Albertina PT-PT	0.055	0.132	0.052	0.131	0.091	0.096	0.206	0.080
Vision-Language Models								
M-CLIP LABSE-ViT-L-14	0.130	0.317	0.100	0.341	0.210	0.237	0.468	0.189
M-CLIP XLM-R-Large-B-32	0.119	0.322	0.091	0.334	0.176	0.215	0.512	0.167
M-CLIP XLM-R-Large-L-14	0.158	0.365	0.123	0.376	0.245	0.266	0.491	0.209
OpenCLIP xlm-roberta-base	**0.176**	**0.418**	**0.121**	**0.419**	**0.264**	**0.288**	**0.610**	**0.218**
BLIP-2 ViT-G	0.015	0.063	0.015	0.062	0.030	0.039	0.140	0.030
SigLIP Base	0.107	0.240	0.083	0.262	0.168	0.181	0.457	0.147
Fine-tuned OpenCLIP Models								
FT-20 (20 epochs)	0.131	0.282	0.116	0.291	0.205	0.207	0.513	0.173
FT-40 (40 epochs)	0.109	0.270	0.092	0.282	0.195	0.202	0.437	0.148
FT-60 (60 epochs)	0.122	0.270	0.107	0.276	0.193	0.196	0.461	0.162
FT-80 (80 epochs)	0.126	0.275	0.112	0.286	0.207	0.207	0.449	0.173
FT-100 (100 epochs)	0.129	0.273	0.106	0.298	0.213	0.215	0.462	0.176

The fine-tuning experiments reveal a significant performance degradation from the original pre-trained model across all evaluation metrics. Analyzing these results provides several critical insights about domain adaptation for Portuguese image retrieval.

All fine-tuned variants (FT-20, FT-40, FT-60, FT-80, FT-100) showed substantial performance declines across every metric when compared to the original OpenCLIP baseline. The worst performing model was FT-40 which achieved only 0.437 MRR, representing a 28% decrease from baseline, followed by FT-80 with 0.449 MRR. FT-60 and FT-100 showed slightly better but still degraded performance at 0.461 and 0.462 MRR respectively. The best performing fine-tuned model across the three metrics (F1@10, MAP, and MRR) was FT-20 (trained for 20 epochs), which achieved an MRR of 0.513, F1@10 of 0.207, and MAP of 0.131, representing decreases of 16%, 28%, and 26%, respectively, relative to the OpenCLIP baseline. This consistent deterioration across every evaluation dimension indicates systematic loss in performance rather than effects confined to specific metrics, suggesting that early training stages preserve more of the original model capabilities, while extended training leads to forgetting pre-trained representations.

The degradation varies significantly across different evaluation metrics when comparing the best fine-tuned model (FT-20) to the original baseline OpenCLIP model. Precision-focused metrics (P@5, P@10) showed moderate decreases ranging from 25% to 33% relative to the baseline, while recall-based metrics (R@5, R@10) experienced more severe drops between 35% and 43% compared to the baseline performance. F1@10 scores, which balance precision and recall, decreased by 28% to 32% from the baseline values. This pattern indicates that fine-tuning particularly damaged the model's ability to retrieve relevant results while maintaining some capability in ranking quality.

When positioned within the broader baseline comparison, the fine-tuned models still outperform traditional text-based IR methods and end-to-end retrieval systems. TF-IDF achieved 0.377 MRR and BM25 reached 0.331 MRR, while Google Images obtained 0.336 MRR and Arquivo.pt scored 0.217 MRR. However, the fine-tuned models significantly underperform compared to other vision-language models. Even the best fine-tuned variant (FT-20: 0.513 MRR) falls slightly below *M-CLIP/XLM-Roberta-Large-Vit-B-32* (0.512 MRR) and substantially behind *M-CLIP/XLM-Roberta-Large-Vit-L-14* (0.491 MRR).

The performance recovery observed in later epochs (FT-60, FT-80, FT-100) compared to FT-40 suggests that overfitting might have happened. The initial performance drop from FT-20 to FT-40 indicates rapid adaptation to the training distribution at the expense of generalization. The subsequent partial recovery suggests the model begins to find a balance between domain-specific adaptation and preservation of general multimodal capabilities, though never recovering the original baseline performance.

Figure 5.3 summarizes the key findings from the fine-tuning experiments, showing the consistent performance degradation across all training epochs when compared to the original baseline model. The visualization confirms that domain-specific fine-tuning fails to improve Portuguese image retrieval performance, with all fine-tuned variants underperforming the pre-trained model across the three primary evaluation metrics.

These findings provide an answer to RQ3: fine-tuning multimodal language models for Por-

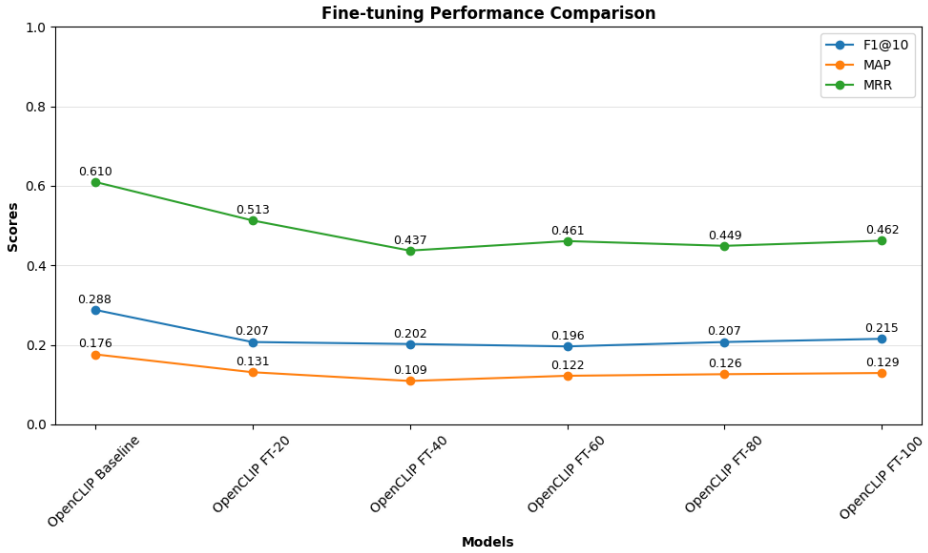

Figure 5.3: Performance evolution during fine-tuning across F1@10, MAP, and MRR metrics. The plot starts with the original pre-trained OpenCLIP baseline model, followed by fine-tuning results at epochs 20, 40, 60, 80, and 100 (FT-20, FT-40, FT-60, FT-80, FT-100).

tuguese contexts consistently degrades rather than improves performance across all evaluated metrics. This counterintuitive result suggests that the general representations learned during large-scale multilingual pre-training are more valuable for Portuguese image retrieval than domain-specific adaptations. The modest size of our fine-tuning dataset (42,333 images) appears insufficient to effectively adapt the complex CLIP architecture without causing the forgetting of rich cross-modal representations acquired during pre-training on massive multilingual datasets.

5.2.3 Hybrid Retrieval Approach

The final experiment assesses the hybrid retrieval method introduced in Chapter 4. The approach combines text-based and image-based retrieval through a score adjustment mechanism, harmonizing contributions from both modalities. The hybrid system leverages the semantic richness from article content through text-based approaches while maintaining the visual-semantic alignment capabilities of direct image retrieval.

To evaluate its effectiveness, we assess our approach through two complementary analyses: Section 5.2.3.1 evaluates different score adjustment functions to determine optimal configurations for balancing multimodal contributions, and Section 5.2.3.2 compares our hybrid approach against Reciprocal Rank Fusion (RRF) as an alternative multimodal fusion technique to demonstrate the effectiveness of our score adjustment mechanism.

5.2.3.1 Score Adjustment Function Comparison

We evaluated four different score adjustment functions to determine optimal configurations for balancing text-based and image-based retrieval contributions: Linear-Z (Linear Zero-indexed), Linear-O (Linear One-indexed), Sqrt (Square Root), and Exp (Exponential). Table

5.4 presents the performance analysis across different factor values (α), focusing on MRR and F1@10 as the primary evaluation metrics.

For space efficiency, the table presents key factor values that capture the essential behavior patterns: $\alpha = 0.0$ represents the baseline without score adjustment, $\alpha = 0.1$, 0.2, and 0.3 represent the low adjustment range where optimal performance is typically observed, $\alpha = 0.5$ represents a moderate adjustment level, and $\alpha = 0.9$ and 1.0 show the behavior at high adjustment factors. The complete table with all intermediate values ($\alpha = 0.4$, 0.6, 0.7, 0.8) is provided in Appendix B.

Table 5.4: Performance comparison of four score adjustment functions across key factor values, showing MRR and F1@10 metrics for hybrid retrieval approach

MRR Performance							
Function	**0.0**	**0.1**	**0.2**	**0.3**	**0.5**	**0.9**	**1.0**
Linear-Z	**0.622**	0.621	0.618	0.617	0.617	0.617	0.617
Linear-O	**0.622**	0.617	0.614	0.614	0.612	0.610	0.610
Sqrt	**0.619**	**0.619**	0.616	0.616	0.613	0.610	0.610
Exp	**0.622**	0.620	**0.620**	**0.620**	0.618	0.612	0.610
F1@10 Performance							
Function	**0.0**	**0.1**	**0.2**	**0.3**	**0.5**	**0.9**	**1.0**
Linear-Z	0.287	0.293	**0.294**	0.292	0.291	0.291	0.291
Linear-O	0.287	**0.293**	0.292	0.291	0.291	0.288	0.288
Sqrt	0.285	0.285	0.285	**0.289**	0.288	**0.288**	**0.288**
Exp	**0.287**	**0.287**	**0.287**	**0.287**	**0.287**	**0.288**	**0.288**

The analysis reveals distinct patterns for each adjustment function. Both square root and exponential functions exhibit lower and more variable performance for F1@10. Square root function shows modest results peaking at $\alpha = 0.3$ (F1@10=0.289), while exponential function demonstrates irregular patterns with a late peak at $\alpha = 0.9$ (F1@10=0.288), indicating limited effectiveness for precision-recall balance.

The better performing linear functions demonstrate superior and more consistent behavior. Linear one-indexed achieves good performance with optimal results at $\alpha = 0.1$ (F1@10=0.293), while the linear zero-indexed function demonstrates the best performance, achieving its peak at $\alpha = 0.2$ (F1@10=0.294) and maintaining stable performance across factor values 0.2-0.5.

For MRR, the square root function shows the weakest performance with degraded results (0.619 at $\alpha = 0.0$), while exponential and linear functions achieve identical peak performance at $\alpha = 0.0$ (MRR=0.622), indicating that minimal score adjustment provides optimal ranking quality. However, the functions differ significantly in their sensitivity to increasing factor values. Square root and exponential functions demonstrate higher sensitivity, with performance declining more rapidly as α increases, while linear zero-indexed maintains relatively stable MRR performance across all α values, and linear one-indexed shows more pronounced but still manageable degradation.

Based on the analysis of Table 5.4, we select the linear zero-indexed function with $\alpha = 0.1$ as the optimal configuration for our hybrid retrieval approach. This configuration achieves strong performance across both key metrics: MRR=0.621 (close to the peak of 0.622) and F1@10=0.293 (near the peak of 0.294). The choice of $\alpha = 0.1$ represents an optimal balance point that provides competitive ranking quality while enhancing precision-recall per-

formance through controlled integration of textual signals. This configuration maintains the strengths of the vision-language model while achieving meaningful improvements from the hybrid approach.

Table 5.5 presents the optimal factor values for each function and their corresponding peak performance levels.

Table 5.5: Optimal factor values and peak performance for each score adjustment function across all metrics

Function	MAP		P@5		P@10		F1@10		MRR		RP	
	α	Val.	α	Val.	α	Val.	α	Val.	α	Val.	α	Val.
Linear Zero	0.3	**0.181**	0.2	**0.423**	0.2	**0.424**	0.2	**0.294**	0.0	**0.622**	0.2	0.224
Linear One	0.2	**0.181**	0.1	**0.423**	0.1	**0.423**	0.1	0.293	0.0	**0.622**	0.0	0.223
Square Root	0.8	0.176	0.3	**0.423**	0.3	0.419	0.3	0.289	0.0	0.619	0.3	0.224
Exponential	0.5	0.176	0.8	0.422	1.0	0.419	0.9	0.288	0.0	**0.622**	0.0	**0.225**

The optimal factor analysis reveals distinct optimization patterns across evaluation metrics, demonstrating the complexity of balancing multimodal signals in hybrid retrieval systems. For precision-focused metrics (MAP, P@5, P@10), linear functions consistently require moderate adjustment factors. Linear zero-indexed achieves optimal MAP performance at α=0.3 (0.181), while both linear functions reach peak precision at relatively low adjustment factors: P@5 at α=0.1-0.2 (0.423) and P@10 at α=0.1-0.2 (0.423-0.424).

The analysis shows differences in optimal configurations across function types. Linear functions demonstrate consistent moderate optimization points, with linear zero-indexed requiring α=0.2 for most precision metrics and linear one-indexed preferring slightly lower values (α=0.1). Non-linear functions exhibit more variable optimization patterns: square root functions achieve best performance with moderate to high adjustment factors (α=0.3-0.8), while exponential functions require diverse optimization points ranging from α=0.5 for MAP to α=1.0 for P@10.

Most significantly, MRR optimization consistently favors minimal adjustment (α=0.0) across all functions, with three functions (linear zero-indexed, linear one-indexed, and exponential) achieving identical peak performance (MRR=0.622). Only the square root function shows degraded MRR performance (0.619), reinforcing that ranking quality benefits most from the base vision-language model capabilities without score modification.

The F1@10 metric shows that linear zero-indexed achieves the highest peak performance (0.294) at α=0.2, while other functions require different optimization points but achieve lower performance. This pattern suggests that balanced precision-recall performance benefits from controlled integration of textual signals, with linear zero-indexed providing the most effective approach.

These findings establish that linear zero-indexed functions consistently achieve superior performance across multiple metrics, making them the optimal choice for hybrid retrieval applications. The analysis supports α=0.1-0.2 for linear zero-indexed functions as providing near-peak performance across most evaluation dimensions while maintaining practical simplicity.

5.2.3.2 Comparison with Reciprocal Rank Fusion

To provide a comprehensive evaluation of our proposed score adjustment approach, we conducted a comparative analysis using RRF as an alternative method for combining text-based and image-based retrieval results. RRF represents a well-established technique in the IR community for merging multiple ranked lists, offering a parameter-free approach to rank fusion that relies on the reciprocal rank positions of documents across different retrieval systems.

The RRF method [44] computes a combined score for each document based on its rank positions across the two retrieval modalities. For a document appearing at rank r_{text} in the text-based results and rank r_{image} in the image-based results, the RRF score is calculated using Equation 5.7:

$$\text{RRF}(d) = \frac{1}{k + r_{\text{text}}} + \frac{1}{k + r_{\text{image}}}. \quad (5.7)$$

where k is a constant that prevents division by zero and provides smoothing for highly-ranked documents. We evaluated RRF with different k values (10, 30, 60, 100) to identify the optimal configuration for our Portuguese image retrieval task. Our analysis revealed that all k values produced identical results, indicating stability across parameter choices. Consequently, we present results for $k = 60$ as the representative configuration.

Table 5.6 presents the comparison between the baseline OpenCLIP model, our optimal hybrid score adjustment approach using the linear zero-indexed score adjustment function with α=0.1, and the RRF method with k=60 across all evaluation metrics, which provided stable results across different parameter values during our evaluation.

Table 5.6: Comparison between baseline OpenCLIP, hybrid approach, and RRF method

Method	MAP	P@5	R@5	P@10	R@10	F1@10	MRR	RP
Baseline OpenCLIP	0.176	0.418	0.121	0.419	0.264	0.288	0.610	0.218
Hybrid Approach	**0.179**	**0.420**	0.125	**0.423**	**0.268**	**0.293**	**0.621**	0.221
RRF Method	0.171	0.413	**0.127**	0.414	0.261	0.289	0.597	**0.222**

The analysis reveals mixed effectiveness of RRF for Portuguese image retrieval. While RRF demonstrates modest improvements in specific recall metrics (R@5: $+0.006$), it significantly degrades precision-focused metrics including MRR (-0.013), P@5 (-0.005), P@10 (-0.005), and MAP (-0.005). This performance pattern indicates that traditional rank fusion techniques provide limited benefits for multimodal retrieval scenarios where underlying modalities exhibit substantially different effectiveness levels.

The suboptimal RRF performance stems from its symmetric treatment of both retrieval modalities. The substantial performance gap between text-based and image-based methods creates an imbalanced fusion scenario where the weaker modality negatively impacts the stronger one through equal rank aggregation. RRF's reciprocal rank formulation fails to account for quality differences between direct visual-semantic matching and indirect text-based approaches, creating a trade-off where improved recall coverage comes at the cost of precision.

In contrast, our proposed hybrid approach outperforms both baseline and RRF methods across most evaluation metrics. The hybrid method achieves superior performance in MRR (0.621 vs 0.597), F1@10 (0.293 vs 0.289), P@5 (0.420 vs 0.413), P@10 (0.423 vs 0.414), and MAP (0.179 vs 0.171). The MRR improvement of $+0.011$ over the baseline contrasts favorably with RRF's decline of -0.013, demonstrating substantially better ranking quality for first relevant results.

The improved performance stems from fundamental methodological distinctions addressing multimodal retrieval challenges. Our asymmetric design recognizes image-based retrieval effectiveness while selectively incorporating textual information. By treating image-based results as the primary ranking foundation and applying position-dependent score adjustments to text-based contributions, our method preserves high-quality visual-semantic alignments while enhancing overall effectiveness. This targeted mechanism achieves consistent improvements across both coverage and precision metrics, avoiding RRF's inherent trade-offs.

The key advantage lies in operating directly on distance scores rather than rank positions alone. While RRF's reciprocal rank transformation provides effective aggregation, it potentially loses valuable similarity information encoded in original distance values. Our score adjustment mechanism preserves distributional characteristics within the CLIP embedding space, enabling optimal cross-modal harmonization that leverages both visual-semantic matching precision and contextual richness from relevant textual content.

These methodological advantages translate into concrete performance gains that validate the effectiveness of our approach. The optimal hybrid configuration achieves an MRR of 0.621, representing a 1.8% improvement over the baseline OpenCLIP performance (0.610). This improvement is particularly significant for MRR, which directly measures retrieval effectiveness by evaluating the ranking quality of the first relevant result. Unlike RRF, which trades precision for recall, our score adjustment mechanism enhances overall retrieval quality without sacrificing ranking effectiveness. These results directly address RQ4, demonstrating that the hybrid approach with score adjustment functions can effectively enhance Portuguese image retrieval performance by selectively incorporating textual information while preserving the strengths of vision-language models.

5.3 Summary and RQ Answers

This chapter presented a comprehensive evaluation of various approaches for Portuguese image retrieval, providing clear answers to our research questions and establishing the effectiveness of different methodological approaches.

The baseline evaluation revealed significant limitations of existing multilingual solutions for Portuguese image IR (RQ2), with commercial systems achieving poor performance (Google Images: 0.336 MRR, Arquivo.pt: 0.217 MRR) and Portuguese-specific language models showing severe multimodal limitations (best: Albertina PT-PT with 0.206 MRR). Our evaluation demonstrated that vision-language models, particularly OpenCLIP *xlm-roberta-base*, substantially outperform traditional approaches, achieving 62% better MRR performance than

text-based methods. The analysis revealed that these models exhibit strong preference for shorter, focused Portuguese queries over longer descriptive formulations, with 71% better performance on concise queries reflecting their training methodology optimized for classification tasks.

Fine-tuning experiments showed counterintuitive results (RQ3), with domain-specific adaptation consistently degrading performance across all metrics by 16% or more compared to the baseline multilingual model. This indicates that general multilingual representations are more valuable than domain-specific adaptations for Portuguese image retrieval tasks.

The proposed hybrid approach effectively enhanced retrieval performance (RQ4), achieving 1.8% improvement in MRR over the baseline through strategic combination of text-based and image-based retrieval modalities. The linear zero-indexed adjustment function with $\alpha = 0.1$ provided optimal balance between precision and coverage metrics, outperforming traditional rank fusion methods.

These findings establish that modern vision-language models provide substantial advantages for Portuguese image retrieval tasks, with careful hybrid approaches offering meaningful enhancements through strategic combination of multimodal signals while maintaining high ranking quality.

Chapter 6

Demonstration Web Application

This chapter presents the demonstration web application that validates the practical applicability of our hybrid retrieval approach and provides an accessible platform to evaluate the effectiveness of image IR systems on a PT built dataset. Section 6.1 provides an overview of the application and its research objectives. Section 6.2 details the system architecture and implementation design. Section 6.3 describes the user interface and interaction design. Section 6.4 presents the advanced features and configuration management. Section 6.5 discusses the search functionality and performance aspects. Section 6.6 outlines the integration and extensibility features of the application. Section 6.7 demonstrates the system's practical effectiveness through concrete search examples and comparative analyses. Finally, Section 6.8 concludes the chapter with a summary of the demonstration application's contributions.

6.1 Overview

To validate the practical applicability of our hybrid retrieval approach and provide an accessible platform for evaluating the effectiveness of image IR systems, we developed a web application that serves as both a demonstration platform and a research tool. The application integrates the methodology and algorithms described in Chapter 4 into a fully functional system that showcases the capabilities of multimodal image retrieval while providing valuable insights into system performance and user interaction patterns.

The demonstration application represents our proposed approach, incorporating the hybrid retrieval algorithm within a scalable system architecture that can handle real-world deployment scenarios. Through this application, we provide concrete evidence of the practical viability of our theoretical contributions while offering a platform for continued research and development in Portuguese image IR systems. The complete source code and implementation details are publicly available on GitHub[1], and a live demonstration of the system is accessible at `https://imageseek.inesctec.pt/`.

6.2 System Architecture and Implementation

The web application implements a modular architecture designed to support both immediate demonstration needs and future research requirements. The system is built using the *Flask* web framework, which provides the flexibility needed to integrate complex retrieval algorithms while maintaining responsive user interfaces and robust API capabilities.

The architecture consists of four primary layers that work in coordination to deliver comprehensive image retrieval functionality, as illustrated in Figure 6.1. This overview diagram

[1] `https://github.com/RodrigDuarte/image-retrieval-system`

shows the hierarchical organization of the system, with bidirectional communication enabling request-response patterns between all layers. User queries and configuration parameters flow downward from the interface layer, while search results, status updates, and system feedback flow upward. Each layer maintains clear separation of concerns while providing well-defined interfaces for inter-layer communication that support both command propagation and response delivery.

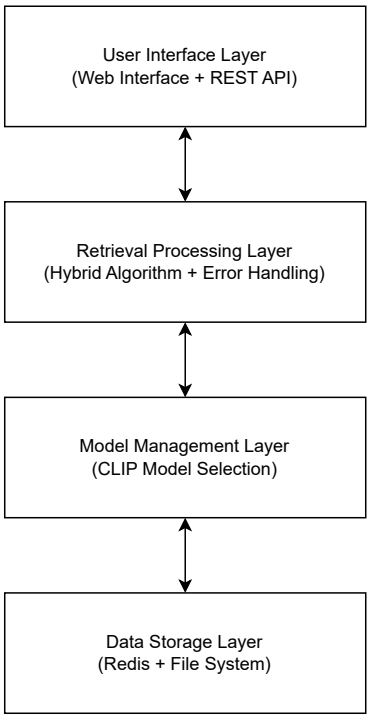

Figure 6.1: System architecture overview showing the four primary layers and their main responsibilities: User Interface Layer, Retrieval Processing Layer, Model Management Layer, and Data Storage Layer with bidirectional data flow enabling request-response patterns and real-time status updates between components.

The data storage layer utilizes *Redis* as both a traditional key-value store and a vector database, enabling efficient storage and retrieval of image embeddings, article embeddings, and associated metadata. This dual-purpose approach leverages *Redis Stack*'s vector field capabilities to support high-performance similarity searches while maintaining the simplicity and reliability of *Redis* for general data management operations.

The model management layer provides a flexible framework that supports multiple CLIP architectures and their variants, as detailed in Figure 6.2. The diagram illustrates the critical constraint that only one model can be active at any given time, ensuring consistent embedding spaces and avoiding compatibility issues between different model architectures. The system accommodates standard OpenAI CLIP implementations, multilingual OpenCLIP models trained on large-scale datasets, and specialized multilingual variants such as M-CLIP models designed specifically for cross-lingual retrieval tasks. This flexibility enables comparative evaluation of different model architectures while supporting future integration of

emerging multimodal models.

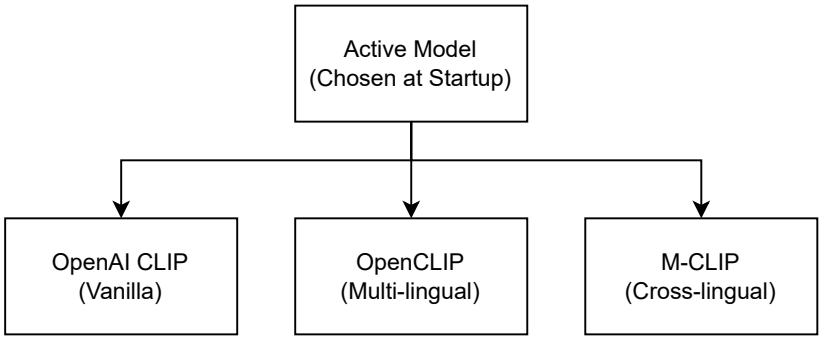

Figure 6.2: Model Management Layer detail showing the selection mechanism where one model is active at any time, chosen from supported CLIP variants: OpenAI CLIP, OpenCLIP (multilingual), and M-CLIP (cross-lingual).

The retrieval processing layer implements our hybrid algorithm with several operational enhancements, such as dynamic temporary index creation for focused retrieval, comprehensive error handling with fallback mechanisms, automatic duplicate detection and removal, input validation that checks for empty queries and malformed parameters, and systematic cleanup of temporary resources. Query preprocessing includes text normalization and encoding validation to ensure compatibility with the selected model's expected input format. The hybrid retrieval mechanism executes text-based and image-based operations efficiently through *Redis* vector search capabilities and KNN algorithms with configurable distance metrics. This layer demonstrates clear bidirectional communication by receiving queries from the user interface, requesting embeddings and model operations from the model management layer, accessing stored data from the data storage layer, and returning processed results and status updates upward through the architecture stack.

The user interface layer provides both human-accessible web interfaces and programmatic API endpoints that enable integration with external systems and automated evaluation frameworks. The interface design prioritizes usability while maintaining access to the underlying system parameters that can be used for research activities and performance analysis. This layer initiates requests by sending user queries and configuration parameters to the retrieval processing layer, while simultaneously receiving and displaying search results, progress indicators, error messages, and system status information flowing back from the lower layers.

The system incorporates automatic file system monitoring through a watchdog-based observer pattern that continuously monitors configured directories for new image additions. When new images are detected, the system automatically generates perceptual hashes and stores corresponding metadata in *Redis*, ensuring that the database remains synchronized with the underlying file system without manual intervention. This feature demonstrates the system's capability to handle dynamic content updates in real-world deployment scenarios.

As illustrated in Figure 6.3, the data storage architecture maintains a clear separation between logical data management in *Redis* and physical file storage in the file system. The diagram shows how *Redis Stack* serves as the central hub for all searchable content, including vector embeddings for similarity search, metadata for contextual information, and file

path references that link to the actual multimedia content. This architecture ensures efficient search operations while maintaining scalable storage for large image and document collections. The bidirectional nature of data storage operations enables all upper layers to both retrieve existing data and store new information, supporting read operations for search queries and write operations for caching results, storing system logs, and maintaining synchronization with the file system.

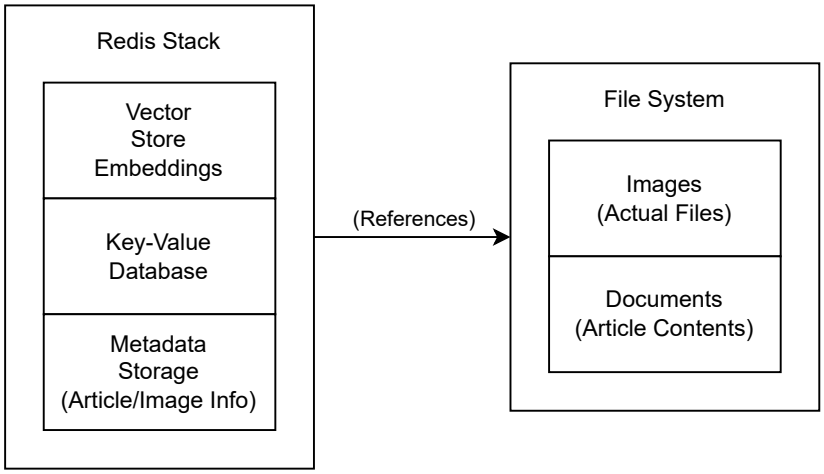

Figure 6.3: Data Storage Layer architecture showing Redis Stack managing embeddings, metadata, and file paths, while actual images and documents are stored in the file system with automatic synchronization.

Configuration management is achieved through externalized configuration files that separate application settings from model definitions, enabling environment-specific configurations without requiring code modifications. The system supports both manual and automated embedding generation scheduling, with configurable parameters for start times, interval configurations, and resource management options.

6.3 User Interface and Interaction Design

The web interface implements a simple and intuitive design that facilitates both general usage and detailed research evaluation activities. The primary search interface provides immediate access to the core functionality through a responsive design that supports Portuguese language queries and provides real-time result display mechanisms.

To facilitate user interaction and demonstrate system capabilities, the application displays a welcome section with categorized query suggestions when users first access the interface. As illustrated in Figure 6.4, these suggestions are organized into four categories: “Desportos e atividades” (Sports and activities), “Presidente em atividades” (President in activities), “Objetos” (Objects), and “Outros” (Others), providing users with diverse examples of queries in Portuguese that showcase different aspects of the image collection. Users can click on any suggestion to automatically populate the search field and execute the query, offering an intuitive way to explore the system’s functionality without requiring prior knowledge of the dataset content.

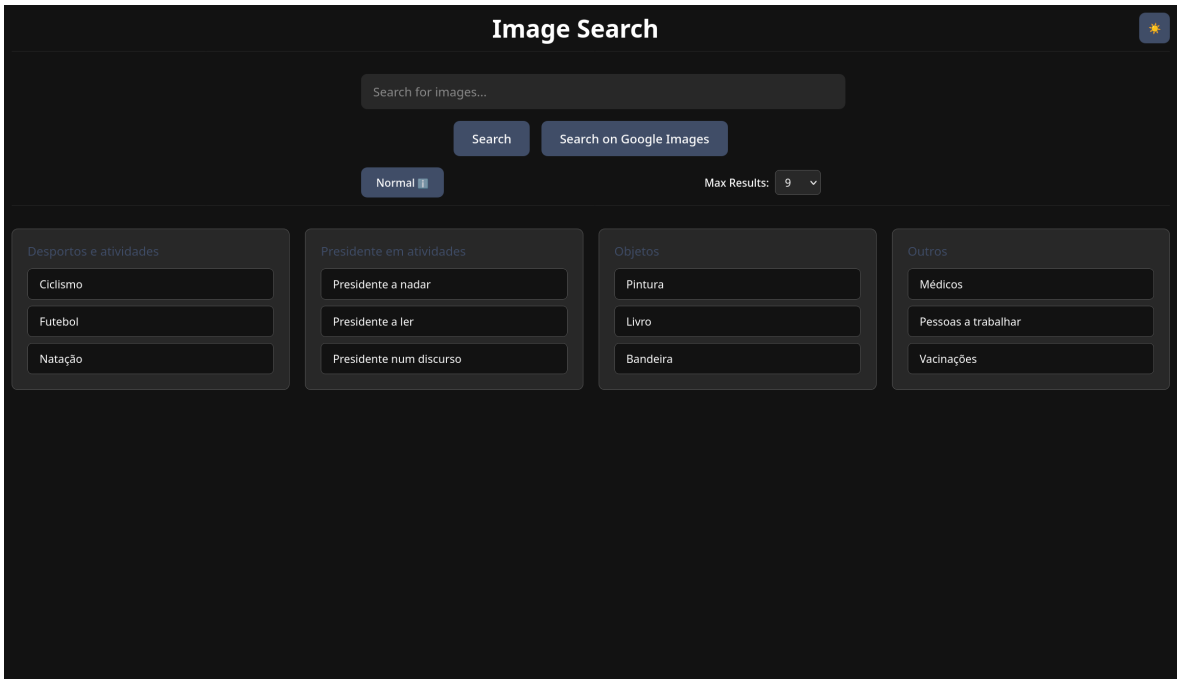

Figure 6.4: Screenshot of the demonstration web application welcome screen showing categorized query suggestions in Portuguese to help users explore the system's capabilities.

Users interact with the system through a search input field that accepts natural language queries in Portuguese, reflecting the primary focus of our research on Portuguese image retrieval. As shown in Figure 6.5, once a query is submitted, the interface provides a clean and accessible design that enables users to easily examine retrieved results. The interface provides immediate visual feedback during query processing, including loading indicators and progress information that help users understand system status during potentially complex retrieval operations.

Configuration options are integrated into the interface, enabling users to modify search parameters including result count limits and search type selection between conventional image retrieval and our hybrid approach. Additionally, the interface includes a Google Images comparison button that opens the same query in Google Images in a new browser tab, restricted to the Portuguese Presidency website domain, allowing users to directly compare retrieval results from our system against those from Google's commercial image search service using the same image collection.

Result visualization is implemented through a grid-based layout that displays image thumbnails with associated metadata derived from the retrieval process. Each result includes contextual information such as relevance scores, source article references, and direct navigation links to original content when available. The grid layout automatically adapts to different screen sizes and device orientations, ensuring accessibility across desktop computers, tablets, and mobile devices.

The interface incorporates sophisticated error handling and user feedback mechanisms that provide informative responses to various system states and potential issues. When retrieval operations encounter problems, users receive clear explanations of the issue along with suggested corrective actions. The system gracefully handles scenarios such as model loading

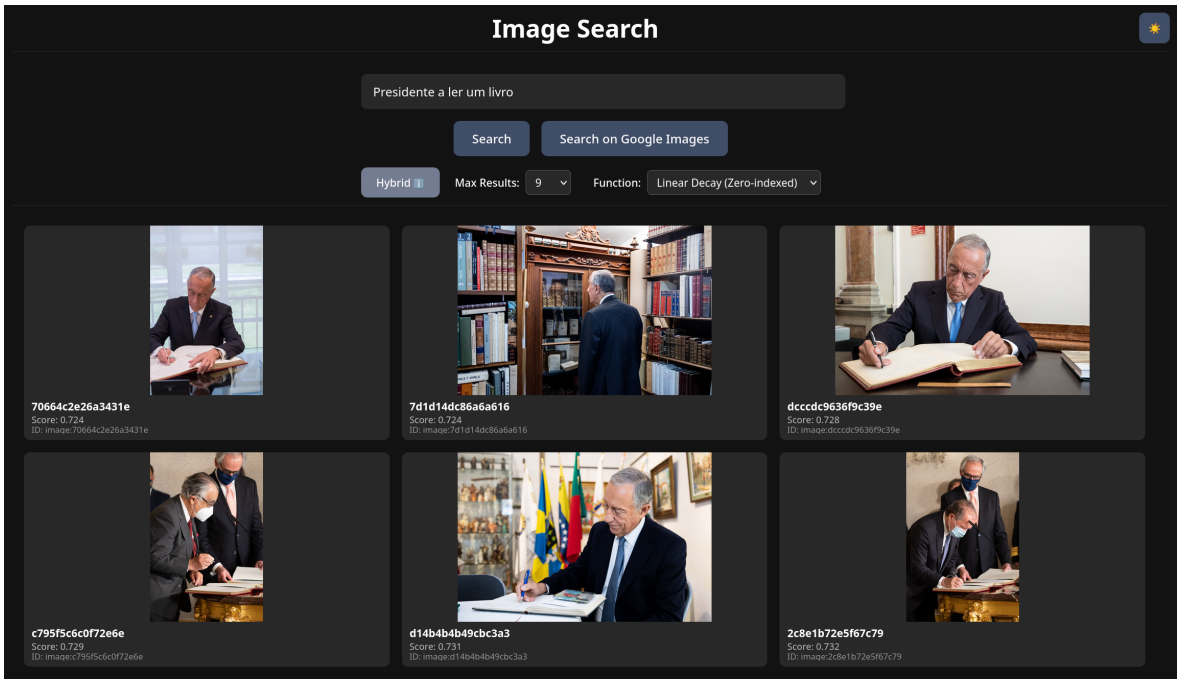

Figure 6.5: Screenshot of the demonstration web application showing the main search interface with Portuguese query input, configuration options, and image retrieval results displayed in a grid layout.

failures, network connectivity issues, and invalid query formats.

Visual themes and accessibility features enhance the user experience across different usage scenarios and user preferences. The application supports both light and dark themes, with automatic detection of user system preferences and manual override capabilities. Typography and color schemes are designed to maintain readability and visual appeal while supporting users with different visual requirements and preferences.

6.4 Advanced Features and Monitoring Capabilities

The demonstration application incorporates several advanced features that distinguish it from simple prototype implementations and demonstrate its suitability for research and operational deployment scenarios.

An integrated debug panel provides real-time access to system statistics, model status information, and performance metrics that are essential for system evaluation and research. The debug panel displays live information about image statistics including total image counts, visible versus hidden content ratios, and monitoring directory status information. Model performance metrics include detailed tracking of model loading times, memory usage patterns, query processing latency measurements, and embedding generation progress tracking. System health monitoring encompasses *Redis* connectivity status, file system accessibility verification, and continuous monitoring of critical system components to ensure reliable operation during extended usage. As illustrated in Figure 6.6, the monitoring provides immediate visibility into system performance characteristics that enable researchers to understand the relationship between system configuration, query complexity, and response performance.

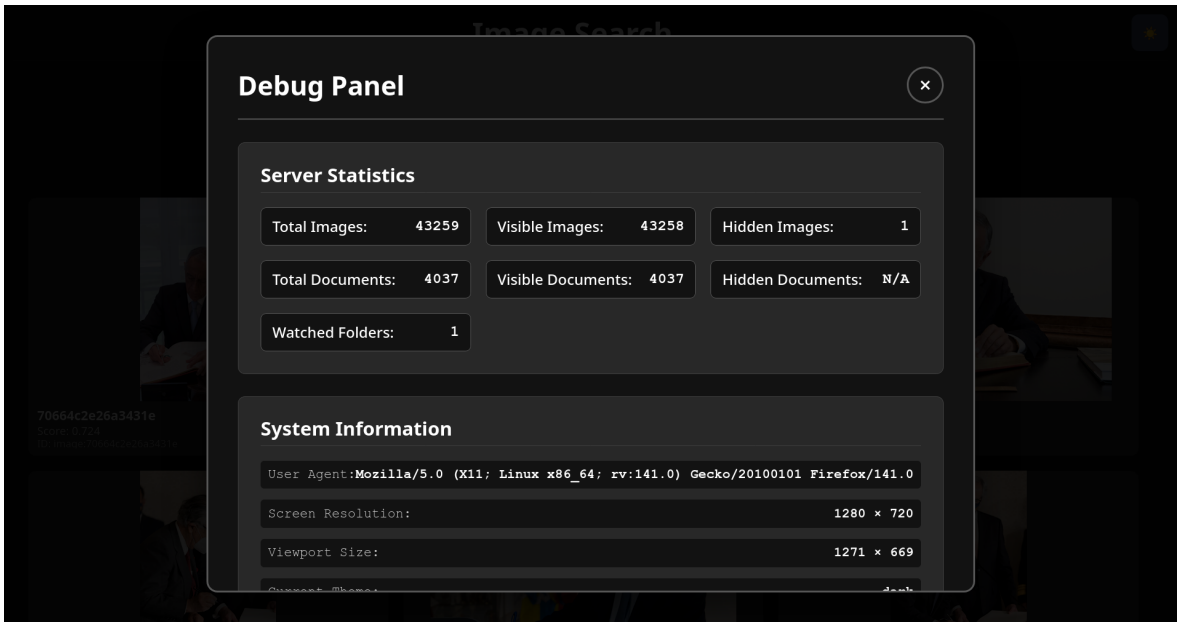

Figure 6.6: Screenshot of the integrated debug panel showing system statistics, model status information, performance metrics, and administrative controls.

The application includes a logging system that captures system-level information for debugging and performance analysis purposes. The logs record server actions such as model loading events, query processing times, database connectivity status, and error conditions, without storing any personally identifiable information such as IP addresses or user identifiers. Log data can be analyzed to understand system usage patterns, identify performance bottlenecks, and guide future system improvements. The logging system is designed to support both real-time monitoring and post-analysis of system behavior, with access restricted to system administrators for debugging and maintenance purposes.

Administrative features enable system configuration through a production mode toggle that controls access to debug information and advanced controls. When production mode is disabled, users gain access to comprehensive system statistics, embedding generation controls, and resource monitoring capabilities through the web interface. This design provides flexibility between streamlined user experiences in production environments and detailed system access for development and research activities.

The application supports multiple deployment configurations ranging from single-user development environments to multi-user research platforms. The modular architecture enables straightforward deployment across different computational environments, from development workstations to cloud-based research infrastructure. The flexible configuration system supports various hardware configurations while maintaining compatibility across different operating systems and deployment scenarios.

6.5 Search Functionality and Performance

The search functionality implements both conventional image retrieval and our proposed hybrid approach, enabling users to experience the performance differences between normal

and enhanced retrieval methods. The conventional search mode performs direct image embedding comparisons using the selected multimodal model, providing baseline functionality that demonstrates standard CLIP-based retrieval capabilities.

The hybrid search mode implements our complete algorithm including the two step retrieval process and linear score adjustment mechanism described in Section 4.1. Users can observe the effects of the hybrid approach by switching between search modes and comparing the different result sets, which demonstrates how the algorithm balances text-based and image-based retrieval contributions.

Performance optimization ensures responsive user interactions even when processing complex queries or large image collections. The system implements efficient caching strategies for frequently accessed embeddings and query results, reducing latency for repeated searches while maintaining accuracy for novel queries. Background processing capabilities enable the system to handle computationally intensive operations without blocking user interface responsiveness.

Query processing supports various Portuguese language constructs including complex descriptive phrases, specific terminology related to Portuguese political and cultural contexts, and queries that incorporate both visual and conceptual elements. The system demonstrates robust handling of Portuguese linguistic features that may not be effectively handled by English-oriented ones.

Result ranking and presentation provide clear indication of retrieval confidence and relevance scores, enabling users to understand the system's decision-making process and evaluate the quality of retrieved results. The interface displays distance scores that reflect the final ranking produced by the selected retrieval method, providing transparency into the retrieval process that supports research evaluation and system improvement activities.

The search functionality incorporates sophisticated fallback mechanisms that ensure robust operation even when individual system components encounter issues. If the hybrid retrieval process encounters problems due to missing document associations or index issues, the system automatically falls back to conventional image retrieval while providing clear indication of the operational mode to users. This design ensures consistent functionality while maintaining transparency about system capabilities and limitations.

6.6 Integration and Extensibility

The demonstration application is designed with integration capabilities that support connection with external systems and research tools. RESTful API endpoints provide programmatic access to all search functionality, enabling automated evaluation frameworks, batch processing systems, and integration with other research tools and platforms.

The API design follows modern web service standards, providing JSON-based request and response formats that facilitate integration with various programming languages and research environments. The open access design supports demonstration purposes and research collaborations, making the system readily accessible for evaluation and integration activities.

Database integration capabilities enable the system to work with existing image collections and metadata repositories without requiring extensive data migration procedures. The system can import image metadata from various formats while maintaining compatibility with existing organizational structures and naming conventions.

The demonstration application provides a straightforward RESTful API that supports programmatic access to the core search functionality. The API endpoints include basic search operations, complex hybrid search, system status monitoring, and image serving capabilities. All API responses use JSON format, making integration with various programming languages and research environments straightforward.

The system architecture supports configuration through external JSON files that separate application settings from model definitions. The model configuration system allows for easy addition of new CLIP architectures by defining their parameters in the `models.json` file without requiring code modifications. Currently supported model types include standard CLIP, multilingual CLIP, OpenCLIP, and specialized variants like BLIP-2 and SigLIP.

Administrative endpoints are available for development and research purposes, including model loading and unloading operations, embedding generation controls, and comprehensive system statistics. These features are controlled by the production mode setting, ensuring that debug capabilities remain accessible during development while being disabled in production environments.

6.7 System Demonstration and Analysis

To demonstrate the practical effectiveness of our hybrid retrieval approach, this section presents detailed analyses of search results across various query types and scenarios. These examples illustrate both the strengths and limitations of our approach through concrete cases that showcase real-world performance characteristics, including comparisons with alternative approaches when relevant.

The query “Presidente a nadar” (President swimming) represents a specific activity-based search that demonstrates the effectiveness of our hybrid approach in understanding Portuguese language constructs and retrieving contextually relevant images from the presidential archive. As illustrated in Figure 6.7, our hybrid retrieval system successfully identifies and ranks relevant images for this specific query. The first and fifth results show precisely what the user requested: images of the Portuguese President engaged in swimming activities. These results demonstrate the system’s ability to understand the semantic relationship between the Portuguese query terms and the visual content, effectively combining textual context from article metadata with visual features to identify appropriate images.

The remaining results in the top rankings, while not showing the President actively swimming, maintain thematic relevance by displaying water-related activities, aquatic environments, or ceremonial events near water bodies. This demonstrates the system’s capability to provide contextually related alternatives when exact matches are limited, offering users a broader perspective on presidential activities related to the query theme. In comparison, when the same query is submitted to Google Images with domain restriction to the Por-

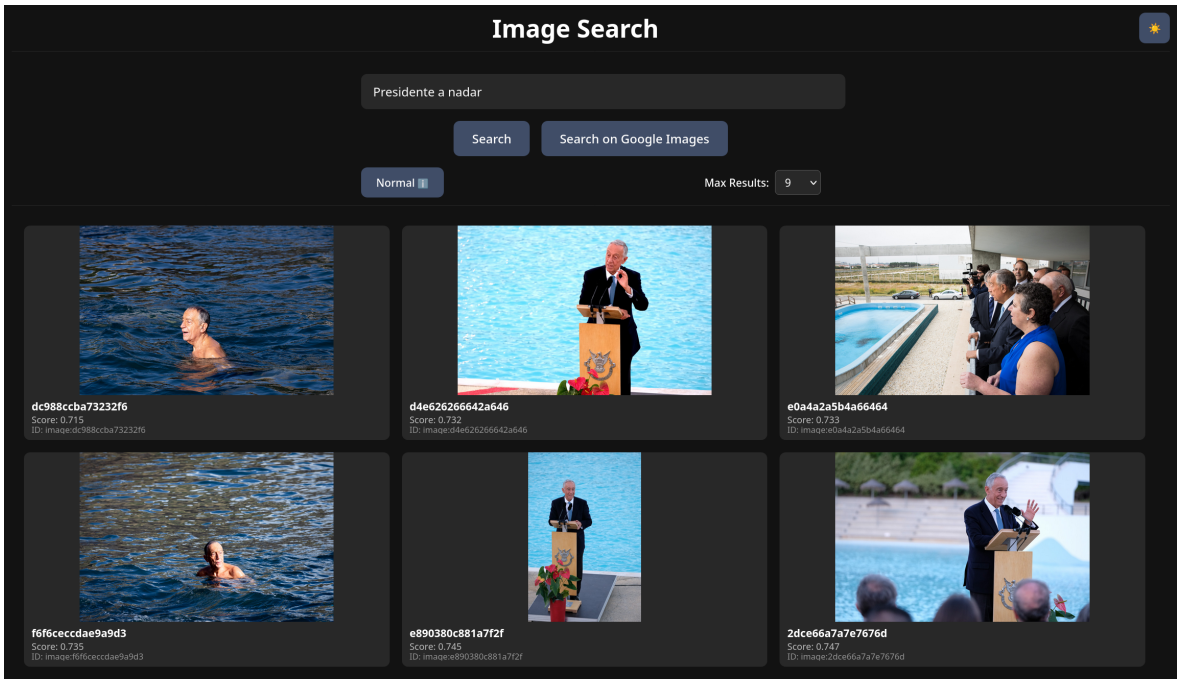

Figure 6.7: Results from our hybrid retrieval system for the query “Presidente a nadar” showing relevant swimming-related images from the Portuguese Presidency archive.

tuguese Presidency website, the results show significantly lower relevance, as illustrated in Figure 6.8. The Google search fails to return significant results that match the user’s intent, with most returned images appearing unrelated to swimming activities or showing poor relevance to the specific query terms.

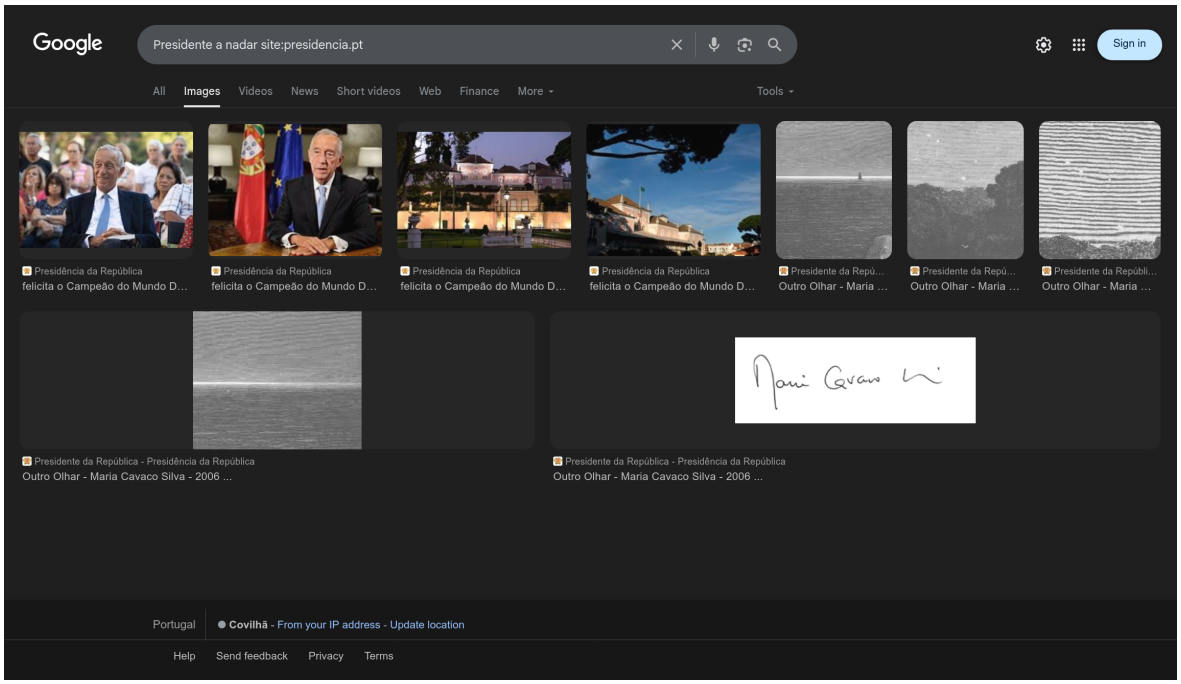

Figure 6.8: Google Images search results for the same query “Presidente a nadar” restricted to the Portuguese Presidency website domain, showing limited relevant results.

This comparison highlights a fundamental limitation of conventional image search approaches

when dealing with domain-specific content and Portuguese language queries that require understanding of both linguistic nuances and contextual relationships within specialized image collections. The superior performance of our hybrid approach in this case can be attributed to the integration of article-level semantic information that provides contextual clues about image content, combined with the score adjustment mechanism that balances text-based and image-based retrieval contributions.

A particularly illustrative comparison between conventional and hybrid retrieval modes can be observed with the query "Donald Trump". This example demonstrates the quantifiable improvement achieved through our score adjustment mechanism when searching for content that may not be optimally ranked by pure visual similarity alone. Figure 6.9 shows the results from the conventional image retrieval mode, where no relevant images appear in the visible results area. This poor ranking performance illustrates a common limitation of conventional CLIP-based image retrieval when dealing with queries that require contextual understanding beyond direct visual similarity.

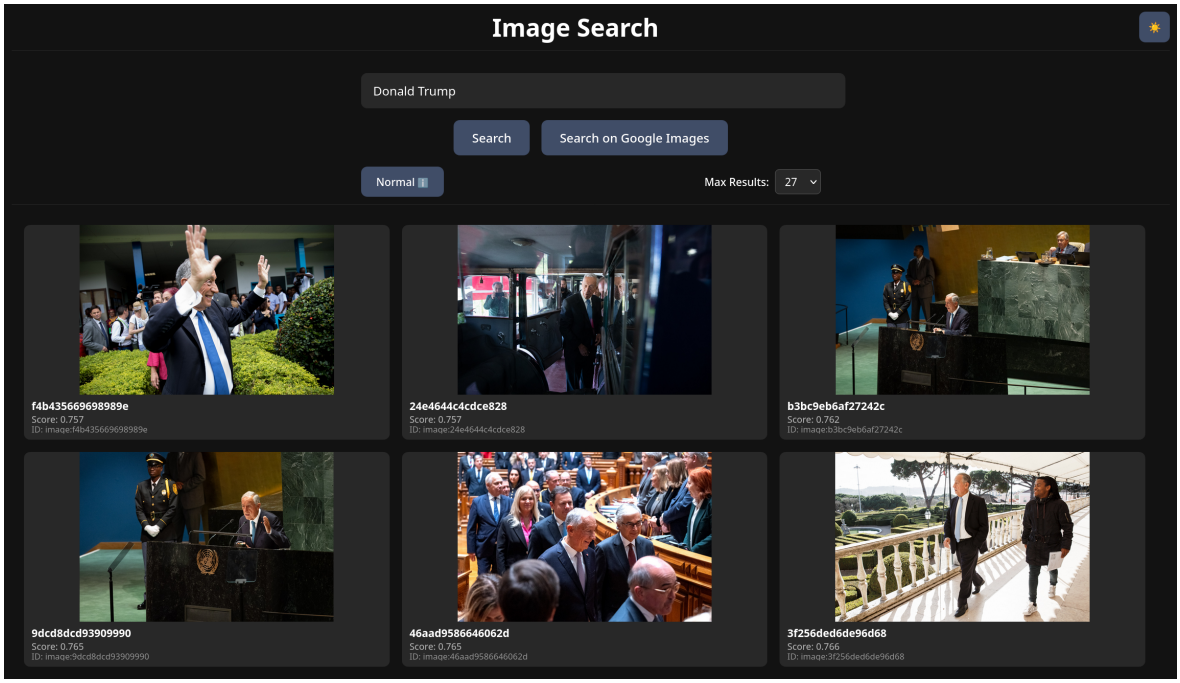

Figure 6.9: Results from conventional image retrieval mode for the query "Donald Trump" showing no relevant images in the visible results area.

In contrast, when the same query is processed using our hybrid retrieval approach, the system demonstrates significant improvement in result ranking quality, as illustrated in Figure 6.10. A relevant image that was not visible in the conventional mode results is promoted to the second position, representing a dramatic improvement in retrieval effectiveness. Additionally, another relevant result appears at the 8th position, though still outside the immediately visible area, indicating that the hybrid approach successfully reorders results to prioritize more relevant content in the top rankings.

This improvement can be attributed to the hybrid algorithm's ability to leverage textual context from article metadata and titles, which likely contain explicit references to "Donald Trump" that provide semantic clues about image content. The score adjustment mechanism

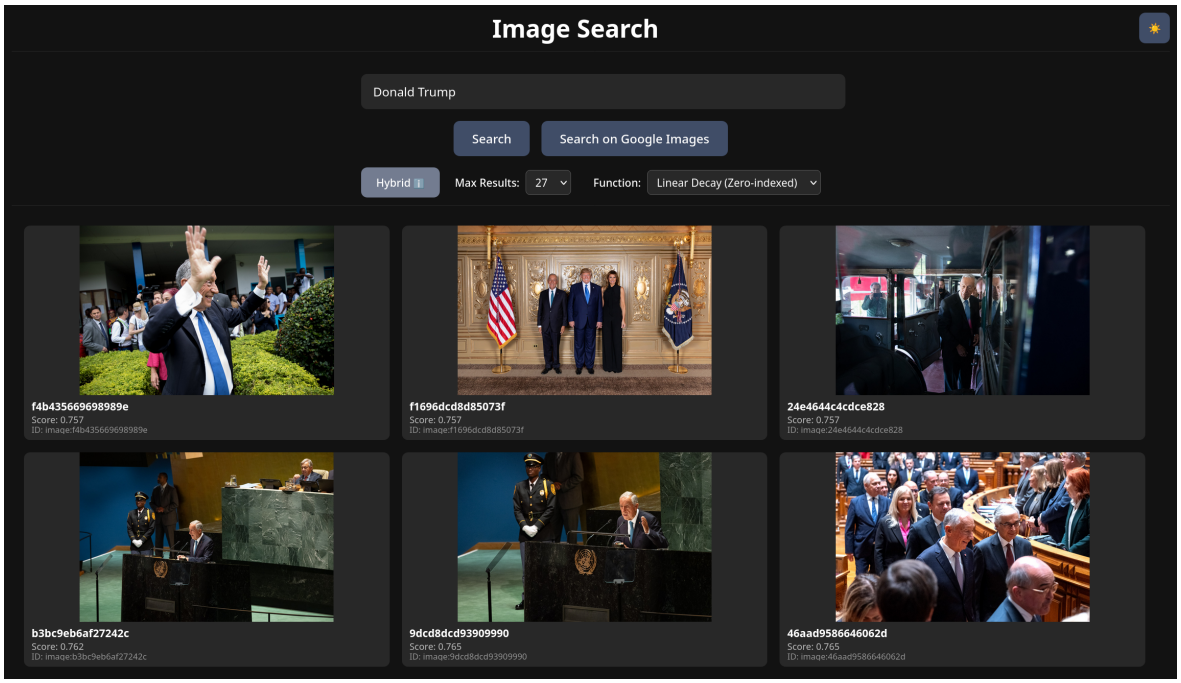

Figure 6.10: Results from hybrid retrieval mode for the same query "Donald Trump" showing improved ranking with relevant content promoted to position 2.

effectively balances these textual signals with visual distance scores, resulting in a more contextually aware ranking that better serves user search intent. This example clearly demonstrates the practical value of the hybrid approach in real-world scenarios where users expect relevant results to appear prominently in the initial result set.

However, our hybrid approach does not universally improve retrieval performance across all query types. The query "Bombeiros" (Firefighters) serves as an illustrative example of scenarios where the hybrid method provides no meaningful improvements over conventional image retrieval. As shown in Figures 6.11 and 6.12, both approaches produce identical result rankings with no discernible improvement in relevance or user experience.

The conventional image retrieval mode successfully identifies and ranks images containing firefighters and emergency response activities based purely on visual similarity matching within the CLIP embedding space. In the hybrid retrieval mode, the same images appear in identical positions, indicating that the score adjustment mechanism finds no substantial textual signals in the article metadata that would warrant reordering the results. This outcome demonstrates that when article metadata contains limited semantic information relevant to the visual query terms, the hybrid approach naturally defaults to the conventional ranking without introducing detrimental effects on result quality.

A particularly revealing limitation of CLIP-based retrieval systems becomes apparent when examining queries that involve linguistic negation or complex syntactic structures. The queries "Papa Francisco com carro" (Pope Francis with car) and "Papa Francisco sem carro" (Pope Francis without car) demonstrate how cosine distance-based retrieval fails to understand language syntax, particularly the critical distinction between "com" (with) and "sem" (without). As illustrated in Figures 6.13 and 6.14, both queries return similar results showing Pope Francis in the presence of vehicles, despite the explicit negation in the second query.

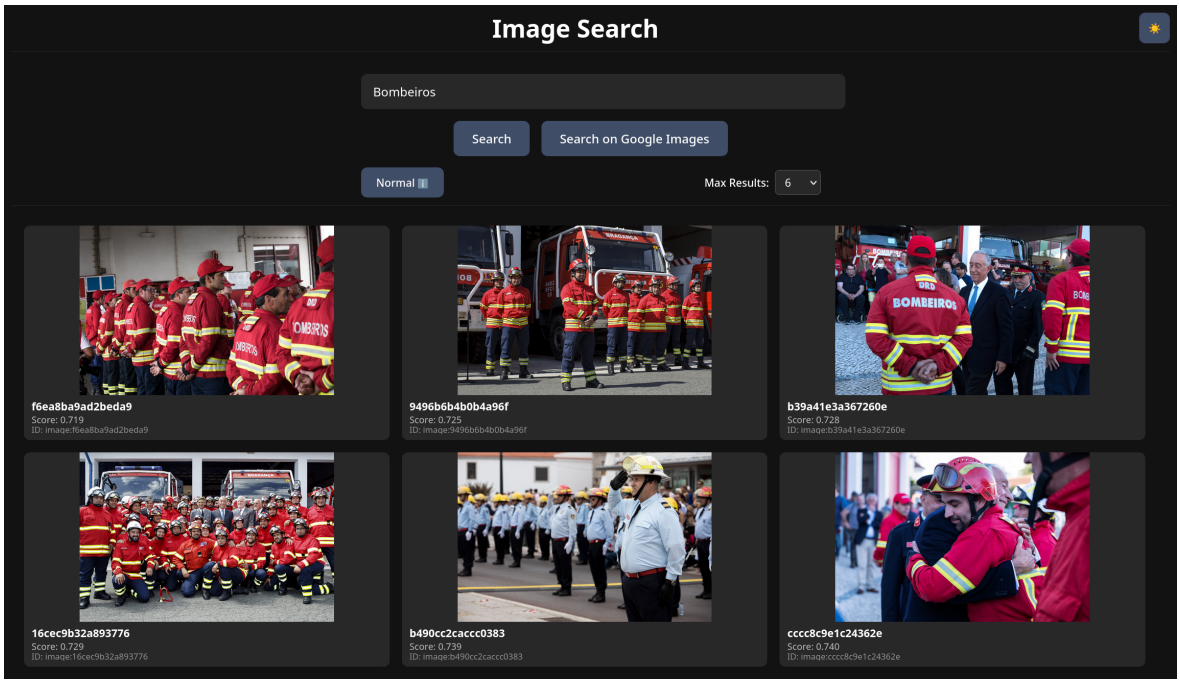

Figure 6.11: Results from conventional image retrieval mode for the query “Bombeiros” showing firefighter-related images ranked by visual similarity.

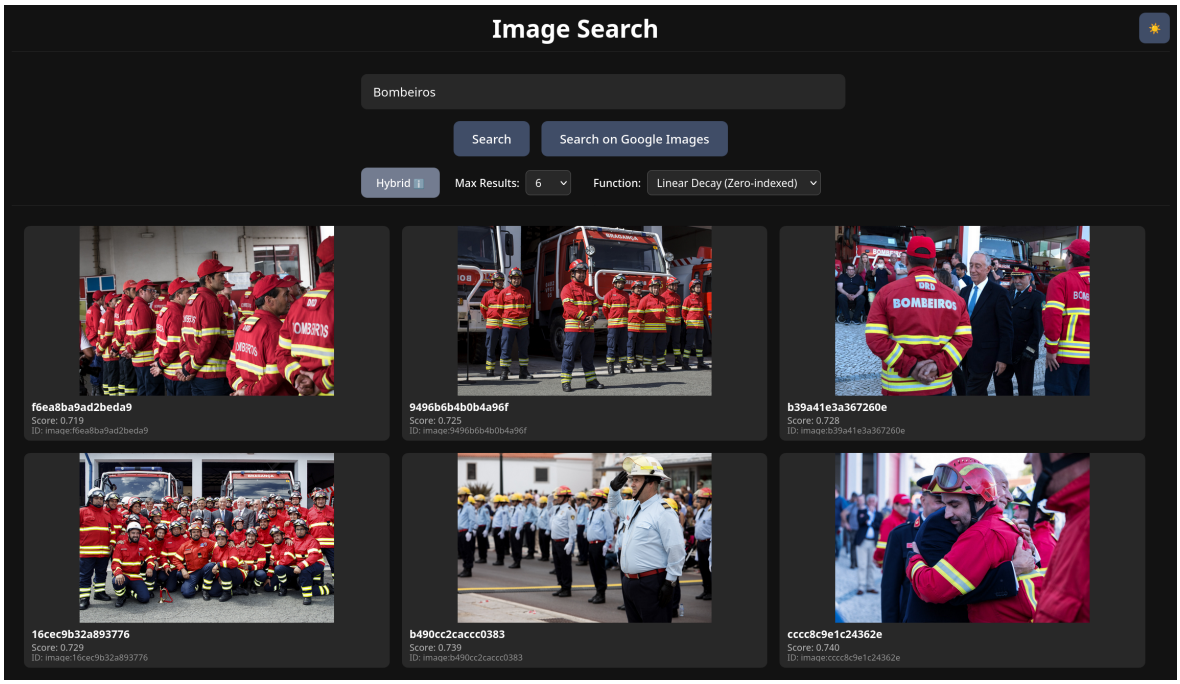

Figure 6.12: Results from hybrid retrieval mode for the same query “Bombeiros” showing virtually identical ranking with no improvement over the conventional approach.

This example exposes a fundamental limitation of current multimodal retrieval approaches that rely primarily on embedding similarity rather than compositional language understanding. The CLIP model processes the query terms independently within its embedding space, focusing on the semantic similarity between “Papa Francisco” and “carro” while failing to interpret the syntactic relationship indicated by “sem”. This results in both queries being treated as semantically equivalent searches for images containing both Pope Francis and au-

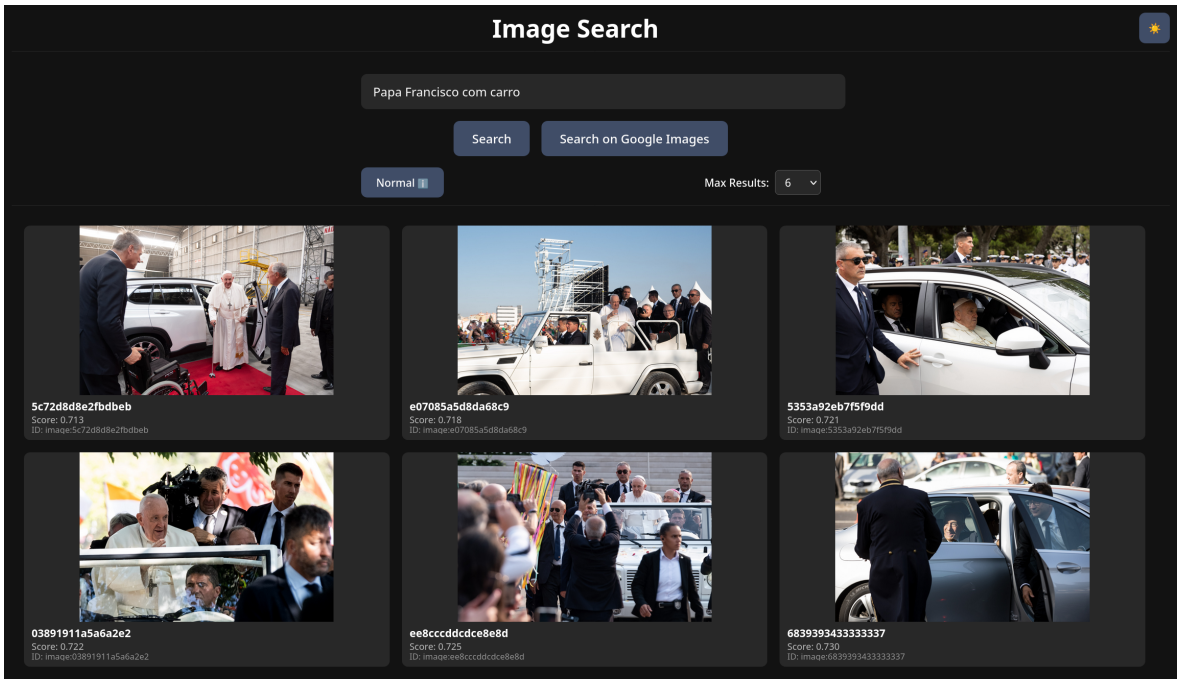

Figure 6.13: Results for the query “Papa Francisco com carro” showing images of Pope Francis with vehicles, as expected by the query semantics.

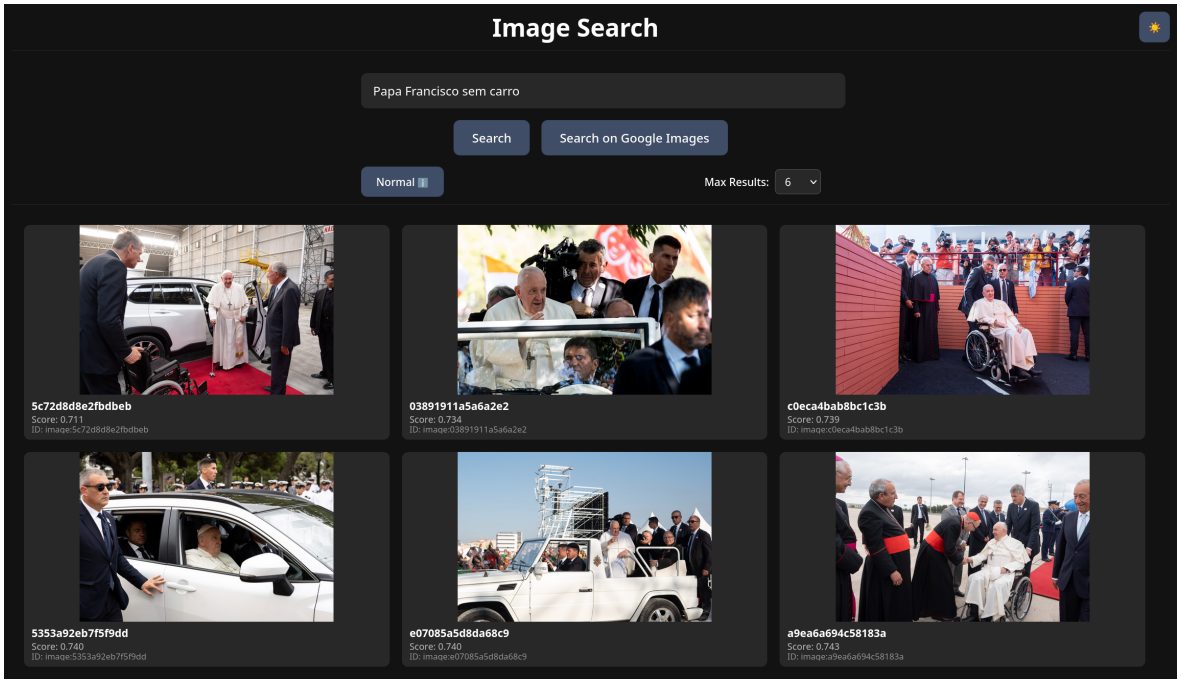

Figure 6.14: Results for the query “Papa Francisco sem carro” showing identical images with vehicles, demonstrating the system’s inability to process negation syntax.

tomotive elements, regardless of the intended exclusionary relationship expressed through negation.

6.8 Summary

This chapter presented a demonstration web application across four major components. First, it detailed the system architecture implementing a modular design with data storage, model management, retrieval processing, and user interface layers. Second, it described the user interface and interaction design featuring Portuguese query suggestions, real-time result visualization, and comparative analysis capabilities with commercial search engines. Third, it outlined advanced features including monitoring capabilities, configuration management, and automatic file system synchronization. Fourth, it demonstrated practical system effectiveness through concrete search examples comparing baseline and hybrid retrieval approaches.

These implementations illustrate the successful translation of theoretical multimodal retrieval advances into functional systems that provide tangible benefits for Portuguese language users. The application effectively validates our hybrid algorithm's real-world applicability while establishing a platform for evaluation and continued research. The modular architecture, extensive monitoring capabilities, and flexible configuration options ensure scalability and adaptability to diverse deployment scenarios.

Chapter 7

Conclusions and Future Work

This thesis addressed the fundamental challenge of image IR, presenting an investigation into multimodal retrieval systems for non-English languages. The research recognized that existing image retrieval datasets and methodologies predominantly focus on English-language contexts, creating a significant gap in understanding how state-of-the-art multimodal models perform when applied to image retrieval tasks.

The first major contribution was the creation of the Portuguese Image Retrieval Dataset, comprising 80 carefully constructed Portuguese queries and 5,201 annotated images sourced from the Portuguese Presidency website. This dataset provides a benchmark for multilingual image retrieval systems. The dataset construction methodology, incorporating both manual-generated and AI-assisted query formulation with rigorous multi-annotator evaluation, establishes a reproducible framework for creating similar resources in other languages and domains.

The second contribution was the development of a hybrid retrieval approach that strategically combines text-based and image-based retrieval methods through score adjustment mechanisms. Unlike traditional rank fusion techniques that treat all modalities symmetrically, our approach recognizes the asymmetric nature of multimodal retrieval systems. The systematic evaluation across multiple score adjustment functions demonstrated that linear zero-indexed functions with moderate adjustment factors (α=0.1-0.2) provide optimal balance, achieving 1.8% improvement in MRR and consistent enhancements across precision and recall metrics.

Our experimental evaluation revealed key insights across multiple dimensions. Multimodal vision-language models, particularly OpenCLIP *xlm-roberta-base*, vastly outperformed traditional text-based approaches by 62% in MRR scores, demonstrating the advantages of direct visual understanding. However, these models showed marked preference for shorter Portuguese queries over longer descriptive formulations, with performance differences reaching 71% between query types.

The fine-tuning experiments provided counterintuitive insights, showing that task-specific fine-tuning consistently degraded performance across all evaluation metrics by 16% to 28% compared to the baseline multilingual model. This suggests that general representations learned during large-scale multilingual pre-training are more valuable for Portuguese image retrieval than domain-specific adaptations, at least within the constraints of our dataset size.

The hybrid retrieval evaluation demonstrated superior performance over traditional RRF methods. While RRF showed modest improvements in specific recall metrics, it significantly degraded precision-focused metrics. Our asymmetric score adjustment approach achieved superior performance across most evaluation dimensions while avoiding the inherent trade-offs of symmetric fusion methods.

Finally, the development and deployment of the demonstration web application validates the practical applicability of our theoretical contributions. The application successfully integrates our hybrid retrieval algorithm within a scalable system architecture, demonstrating real-world viability. The public availability of both the live system and complete source code ensures reproducibility and enables continued development by the research community.

This work establishes image retrieval as a legitimate research area while providing methodological frameworks applicable to non-English languages. The combination of dataset creation, algorithmic innovation, and comprehensive evaluation creates a foundation for advancing multilingual multimodal information retrieval systems.

Several limitations of this research provide clear directions for future investigation. The evaluation dataset is limited to images from the Portuguese Presidency website, which may constrain the generalizability of findings to other Portuguese image collections or broader contexts. Future research should prioritize expanding the dataset to include diverse domains such as cultural heritage, e-commerce, social media, and educational content, which would enhance the scope and applicability of Portuguese image IR research. Additionally, investigating advanced query expansion and reformulation techniques could address the identified limitations with longer descriptive queries, potentially through integration of large language models specifically trained for Portuguese contexts.

The development of more sophisticated score adjustment functions represents another important direction. While our exploration of linear, square root, and exponential functions provided valuable insights, more complex adaptive mechanisms that dynamically adjust based on query characteristics or user feedback could further enhance retrieval effectiveness.

Future work should also address scalability challenges inherent in hybrid retrieval systems, particularly computational overhead associated with dual retrieval processes. Investigating approximate similarity search techniques and distributed computing approaches could enable deployment at web scale while maintaining performance benefits.

The integration of emerging multimodal architectures, such as newer CLIP variants and forthcoming vision-language models, presents opportunities to validate our hybrid approach across different foundational models. As the field evolves, our score adjustment framework provides a flexible foundation for integrating diverse model capabilities.

Finally, exploring domain adaptation strategies beyond traditional fine-tuning approaches could address the limitations identified in our experiments. Techniques such as adapter modules, prompt engineering, and few-shot learning approaches might provide more effective methods for specializing multilingual models without experiencing catastrophic forgetting. Additionally, developing automatic evaluation methods could enable more efficient assessment of system improvements beyond manual annotation approaches.

Bibliography

[1] A. Mourão and D. Gomes, "Searching images in a web archive," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2023, pp. 1–10. xv, 1, 11, 12

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. xv, 13, 14, 26, 50

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742. xv, 14, 15

[4] L. Iijima, N. Giakoumoglou, and T. Stathaki, "A multimodal approach for cross-domain image retrieval," *arXiv preprint arXiv:2403.15152*, 2024. xv, 15, 16, 35

[5] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and multilingual clip," in *Proceedings of the thirteenth language resources and evaluation conference*, 2022, pp. 6848–6854. xv, 16, 17, 35, 50

[6] Y. Gong, G. Cosma, and A. Finke, "Vitr: augmenting vision transformers with relation-focused learning for cross-modal information retrieval," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 9, pp. 1–21, 2024. xv, 16, 17, 18

[7] Y. Gong and G. Cosma, "Boon: A neural search engine for cross-modal information retrieval," *arXiv preprint arXiv:2307.14240*, 2023. 1, 18

[8] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1–23, 2004. 1

[9] L. R. Long, S. Antani, T. M. Deserno, and G. R. Thoma, "Content-based image retrieval in medicine: retrospective assessment, state of the art, and future directions," *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, vol. 4, no. 1, pp. 1–16, 2009. 1

[10] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data," *Journal of digital imaging*, vol. 26, pp. 1025–1039, 2013. 1

[11] V. V. Estrela and A. E. Herrmann, "Content-based image retrieval (cbir) in remote clinical diagnosis and healthcare," in *Encyclopedia of E-Health and Telemedicine*. IGI Global, 2016, pp. 495–520. 1

[12] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Biometric technology for human identification*, vol. 5404. Spie, 2004, pp. 561–572. 1

[13] C.-Y. Wen and C.-C. Yu, "Image retrieval of digital crime scene images," *Forensic Science Journal*, vol. 4, no. 1, pp. 45–53, 2005. 1

[14] A. K. Jain, J.-E. Lee, R. Jin, and N. Gregg, "Content-based image retrieval: An application to tattoo images," in *2009 16th IEEE international conference on image processing (ICIP)*. IEEE, 2009, pp. 2745–2748. 1

[15] H. Klinke, "Big image data within the big picture of art history," *International Journal for Digital Art History*, no. 2, Oct. 2016. [Online]. Available: https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/33527 1

[16] L. Shahrzadi, A. Mansouri, M. Alavi, and A. Shabani, "Causes, consequences, and strategies to deal with information overload: A scoping review," *International Journal of Information Management Data Insights*, vol. 4, no. 2, p. 100261, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096824000508 1

[17] H. Cho, M. T. Pham, K. N. Leonard, and A. C. Urban, "A systematic literature review on image information needs and behaviors," *Journal of Documentation*, vol. 78, no. 2, pp. 207–227, 2022. 1

[18] S.-S. Liaw and H.-M. Huang, "An investigation of user attitudes toward search engines as an information retrieval tool," *Computers in human behavior*, vol. 19, no. 6, pp. 751–765, 2003. 11

[19] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010, vol. 520. 11

[20] B. R. Schatz, "Information retrieval in digital libraries: Bringing search to the net," *Science*, vol. 275, no. 5298, pp. 327–334, 1997. 11

[21] I. Xie, *Interactive information retrieval in digital environments*. IGI global, 2008. 11

[22] Á. Tejeda-Lorente, C. Porcel, E. Peis, R. Sanz, and E. Herrera-Viedma, "A quality based recommender system to disseminate information in a university digital library," *Information Sciences*, vol. 261, pp. 52–69, 2014. 11

[23] M. R. F. Falero, L. S. Gil, and L. V. G. Tapia, "Cycle of information retrieval through color in e-commerce: Store choice," in *Advances in Information Systems and Technologies*. Springer, 2013, pp. 1033–1040. 11

[24] M. Hendriksen, M. Bleeker, S. Vakulenko, N. Van Noord, E. Kuiper, and M. De Rijke, "Extending clip for category-to-image retrieval in e-commerce," in *European Conference on Information Retrieval*. Springer, 2022, pp. 289–303. 11

[25] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 55–64. 11

[26] B. Mitra and N. Craswell, "Neural models for information retrieval," *arXiv preprint arXiv:1705.01509*, 2017. 11

[27] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, "Dense text retrieval based on pretrained language models: A survey," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–60, 2024. 11

[28] M. K. Alsmadi, "Content-based image retrieval using color, shape and texture descriptors and features," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3317–3330, 2020. 11

[29] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)," *Neurocomputing*, vol. 452, pp. 675–689, 2021. 11

[30] S. V. Bhoir and S. Patil, "A review on recent advances in content-based image retrieval used in image search engine," *Library Philosophy and Practice*, pp. 1–45, 2021. 11

[31] M. Lux, M. Riegler, P. Halvorsen, K. Pogorelov, and N. Anagnostopoulos, "Lire: open source visual information retrieval," in *Proceedings of the 7th International Conference on Multimedia Systems*, 2016, pp. 1–4. 11

[32] K. O'Shea, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015. 12

[33] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 12

[34] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017. 12

[35] C. Shang, H. Zhang, H. Wen, and Y. Yang, "Understanding multimodal deep neural networks: A concept selection view," *arXiv preprint arXiv:2404.08964*, 2024. 12

[36] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 4904–4916. [Online]. Available: https://proceedings.mlr.press/v139/jia21b.html 13

[37] H. Song, L. Dong, W.-N. Zhang, T. Liu, and F. Wei, "Clip models are few-shot learners: Empirical studies on vqa and visual entailment," *arXiv preprint arXiv:2203.07190*, 2022. 13

[38] A.-A. Balauca, D. P. Paudel, K. Toutanova, and L. Van Gool, "Taming clip for fine-grained and structured visual understanding of museum exhibits," in *European Conference on Computer Vision*. Springer, 2025, pp. 377–394. 13, 35

[39] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831. 13

[40] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021, if you use this software, please cite it as below. [Online]. Available: https://doi.org/10.5281/zenodo.5143773 16, 26, 50

[41] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818–2829. 16, 26, 50

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. 16, 26

[43] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=M3Y74vmsMcY 16, 26

[44] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 758–759. [Online]. Available: https://doi.org/10.1145/1571941.1572114 19, 38, 60

[45] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications," *Computers, Materials and Continua*, vol. 80, no. 1, pp. 1–35, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221824005216 19

[46] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne,

Australia: Association for Computational Linguistics, Jul. 2018, pp. 2247–2256. [Online]. Available: https://aclanthology.org/P18-1209/ 19

[47] A. Mourão, F. Martins, and J. Magalhães, "Inverse square rank fusion for multimodal search," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014, pp. 1–6. 19

[48] E. Xu, X. Li, Z. Zhou, J. Ji, J. Zhao, D. Miao, S. Wang, L. Liu, and S. Xu, "Advancing re-ranking with multimodal fusion and target-oriented auxiliary tasks in e-commerce search," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 5007–5014. 19

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755. 21

[50] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014. 21

[51] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649. 21

[52] G. O. dos Santos, E. L. Colombini, and S. Avila, "#pracegover: A large dataset for image captioning in portuguese," *Data*, vol. 7, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2306-5729/7/2/13 22

[53] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," in *ICCV*, 2017. 22

[54] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2443–2449. 22

[55] R. Osmulsk, G. de Souza P. Moreira, R. Ak, M. Xu, B. Schifferer, and E. Oldridge, "Miracl-vision: A large, multilingual, visual document retrieval benchmark," 2025. [Online]. Available: https://arxiv.org/abs/2505.11651 23

[56] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin, "Miracl: A multilingual retrieval dataset covering 18 diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1114–1131, 09 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00595 23

[57] K. S. Jones, C. J. Van Rijsbergen, and British Library. Research and Development Department, *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*. University Computer Laboratory, 1975. 25

[58] D. Harman, "Overview of the first trec conference," in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 1993, pp. 36–47. 25

[59] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0306457388900210 25, 49

[60] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. 25, 49

[61] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, "Cross-lingual and multilingual clip," in *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 6848–6854. [Online]. Available: https://aclanthology.org/2022.lrec-1.739 26

[62] E. M. Voorhees and D. Harman, "Overview of trec 2001." in *Trec*, 2001. 26

[63] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, no. 3, pp. 613–619, 1973. [Online]. Available: https://doi.org/10.1177/001316447303300309 28

[64] F. Souza, R. Nogueira, and R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," in *Intelligent Systems*, R. Cerri and R. C. Prati, Eds. Cham: Springer International Publishing, 2020, pp. 403–417. 49

[65] J. Rodrigues, L. Gomes, J. Silva, A. Branco, R. Santos, H. L. Cardoso, and T. Osório, "Advancing neural encoding of portuguese with transformer albertina pt-*," 2023. 49

[66] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023. 50

[67] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986. 50

Appendix A

Complete Query List

This appendix presents the complete list of 80 queries used in the pt-image-ir-dataset, organized by their respective categories. The queries are divided into two main groups: AI-generated queries (38 queries across 4 categories) and manually-created queries (42 queries across 7 categories).

A.1 AI-Generated Queries

Table A.1: AI-generated queries for Events and Contexts category.

Events and Contexts
Presidente de Portugal em conferência de imprensa
Reunião do governo português
Presidente de Portugal com líderes internacionais
Visita oficial do Presidente de Portugal
Presidente de Portugal em discurso público
Eventos culturais com o Presidente de Portugal
Presidente de Portugal a visitar escolas
Reunião do Conselho de Ministros em Portugal
Presidente de Portugal em cerimónias de tomada de posse
Ações do Presidente de Portugal na União Europeia
Encontros bilaterais do Presidente de Portugal
Eventos culturais promovidos pela presidência de Portugal
Presidente de Portugal com líderes mundiais
Visitas a instituições durante a presidência de Portugal

Table A.2: AI-generated queries for Expressions and Emotions category.

Expressions and Emotions
Expressões faciais de felicidade
Emoções de tristeza em rostos
Sorrisos de líderes políticos
Expressões de surpresa em pessoas
Rostos de raiva e frustração
Expressões de confiança em políticos
Emoções de preocupação em rostos
Expressões de empatia e compreensão

A.2 Manually-Created Queries

Table A.3: AI-generated queries for Interactions with the Public category.

Interactions with the Public
Presidência de Portugal interações com o povo
Presidente de Portugal em eventos públicos
Cidadãos a interagir com o Presidente de Portugal
Visitas do Presidente de Portugal a comunidades
Presidente de Portugal em encontros com cidadãos
Apoio do Presidente de Portugal a iniciativas locais
Reuniões do Presidente de Portugal com associações

Table A.4: AI-generated queries for Places and Environments category.

Places and Environments
Palácio de Belém
Assembleia da República
Lisboa
Cascais
Porto
Vila Nova de Gaia
Algarve
Serra da Estrela
Funchal

Table A.5: Manually-created queries for Public Figures category.

Public Figures
Papa Francisco
Donald Trump
Barack Obama
Angela Merkel
Amália Rodrigues
Cristiano Ronaldo
Marcelo Rebelo de Sousa
António Costa

Table A.6: Manually-created queries for Sports and Activities category.

Sports and Activities
Ciclismo
Futebol
Natação
Corrida
Surf

Table A.7: Manually-created queries for Trending Topics category.

Trending Topics
Jogos Olímpicos
Covid-19
Estado de Emergência
Brexit
Web Summit

Table A.8: Manually-created queries for Places and Monuments category.

Places and Monuments
Mosteiro dos Jerónimos
Universidade da Beira Interior
Universidade do Porto
Universidade de Lisboa
Fátima

Table A.9: Manually-created queries for General Places category.

General Places
Praia
Bombeiros
Hospital
Escola
Supermercado
Museu
Centro Militar

Table A.10: Manually-created queries for Objects category.

Objects
Pintura
Livro
Arma
Telemóvel
Bandeira

Table A.11: Manually-created queries for Others category.

Others
Crianças a brincar
Pessoas a trabalhar
Pessoas a estudar
Cozinheiros
Médicos
Pessoas a fazer compras
Vacinações

Appendix B

Complete Score Adjustment Performance Table

This appendix presents the complete performance comparison of the four score adjustment functions across all factor values from 0.0 to 1.0 with 0.1 increments, showing detailed MRR and F1@10 metrics for the hybrid retrieval approach.

Figure B.1: Complete performance comparison of four score adjustment functions (Linear-Z, Linear-O, Sqrt, Exp) across all factor values from 0.0 to 1.0, showing MRR and F1@10 metrics for hybrid retrieval approach

MRR Performance										
Function	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Linear-Z	**0.622**	0.621	0.618	0.617	0.617	0.617	0.617	0.617	0.617	0.617	0.617
Linear-O	**0.622**	0.617	0.614	0.614	0.613	0.612	0.612	0.612	0.612	0.610	0.610
Sqrt	**0.619**	**0.619**	0.616	0.616	0.615	0.613	0.610	0.610	0.610	0.610	0.610
Exp	**0.622**	0.620	**0.620**	**0.620**	0.619	0.618	0.618	0.615	0.614	0.612	0.610
F1@10 Performance											
Function	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Linear-Z	0.287	0.293	**0.294**	0.292	0.292	0.291	0.291	0.291	0.291	0.291	0.291
Linear-O	0.287	**0.293**	0.292	0.291	0.291	0.291	0.291	0.291	0.290	0.288	0.288
Sqrt	0.285	0.285	0.285	**0.289**	**0.289**	0.288	0.288	0.287	**0.288**	**0.288**	**0.288**
Exp	**0.287**	**0.287**	**0.287**	**0.287**	**0.287**	**0.287**	**0.287**	0.285	0.286	**0.288**	**0.288**

Appendix C

Bootstrapped Evaluation Results

This appendix presents the complete bootstrapped evaluation results with 95% confidence intervals for all evaluated methods. The bootstrap analysis was conducted with 1,000 iterations to provide statistically robust performance estimates and confidence bounds for the retrieval effectiveness metrics.

Figure C.1: Complete bootstrapped evaluation results with 95% confidence intervals (mean ± std) for all methods across evaluation metrics. Results based on 1,000 bootstrap iterations with 80 queries and top-10 retrieved results.

Method	**MAP**	**P@5**	**R@5**	**P@10**	**R@10**	**F1@10**	**MRR**	**RP**
Portuguese Language Embedding Models								
BERTimbau Base	0.054 ± 0.016	0.120 ± 0.027	0.042 ± 0.011	0.136 ± 0.028	0.086 ± 0.020	0.097 ± 0.021	0.203 ± 0.038	0.077 ± 0.018
BERTimbau Large	0.070 ± 0.018	0.137 ± 0.030	0.054 ± 0.014	0.155 ± 0.031	0.111 ± 0.025	0.116 ± 0.024	0.221 ± 0.042	0.090 ± 0.020
Albertina PT-PT	0.074 ± 0.017	0.164 ± 0.030	0.064 ± 0.013	0.170 ± 0.030	0.119 ± 0.023	0.126 ± 0.023	0.257 ± 0.044	0.104 ± 0.019
Vision-Language Models								
M-CLIP LABSE-14	0.129 ± 0.017	0.319 ± 0.035	0.100 ± 0.012	0.339 ± 0.036	0.208 ± 0.023	0.234 ± 0.024	0.466 ± 0.046	0.188 ± 0.020
M-CLIP XLM-32	0.118 ± 0.015	0.319 ± 0.038	0.090 ± 0.011	0.332 ± 0.037	0.175 ± 0.018	0.214 ± 0.022	0.509 ± 0.049	0.166 ± 0.018
M-CLIP XLM-14	0.157 ± 0.020	0.366 ± 0.039	0.123 ± 0.017	0.376 ± 0.037	0.244 ± 0.027	0.266 ± 0.026	0.490 ± 0.047	0.208 ± 0.022
OpenCLIP Laion5b	0.176 ± 0.019	0.417 ± 0.042	0.122 ± 0.013	0.417 ± 0.040	0.264 ± 0.025	0.288 ± 0.024	0.609 ± 0.048	0.218 ± 0.021
BLIP-2 ViT-G	0.017 ± 0.005	0.075 ± 0.019	0.018 ± 0.005	0.069 ± 0.016	0.034 ± 0.008	0.044 ± 0.010	0.139 ± 0.034	0.032 ± 0.008
SigLIP Base	0.107 ± 0.024	0.240 ± 0.033	0.083 ± 0.015	0.262 ± 0.035	0.168 ± 0.022	0.181 ± 0.024	0.457 ± 0.041	0.147 ± 0.020
Hybrid Retrieval								
Linear Zero ($\alpha = 0.1$)	**0.180 ± 0.017**	**0.421 ± 0.039**	**0.124 ± 0.013**	**0.422 ± 0.035**	**0.268 ± 0.025**	**0.291 ± 0.023**	**0.617 ± 0.051**	**0.224 ± 0.019**