Information Retrieval Research 1 (2025) 197-220 Submitted 10/24; Revised 03/25; Published 09/25

Evaluating Dense Model-based Approaches for Multimodal Medical Case Retrieval

Catarina Pires UP201907925@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Sérgio Nunes SSN@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Luís Filipe Teixeira LUISFT@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Editor: Ben He

Abstract

Medical case retrieval plays a crucial role in clinical decision-making by enabling healthcare professionals to find relevant cases based on patient records, diagnostic images, and textual descriptions. Given the inherently multimodal nature of medical data, effective retrieval requires models that can bridge the gap between different modalities. Traditional retrieval approaches often rely on unimodal representations, limiting their ability to capture cross-modal relationships. Recent advances in dense model-based techniques have shown promise in overcoming these limitations by encoding multimodal information into a shared latent space, facilitating retrieval based on semantic similarity. This paper investigates the potential of dense models to enhance multimodal search systems. We evaluate various dense model-based approaches to assess which model characteristics have the greatest impact on retrieval effectiveness, using the medical case-based retrieval task from ImageCLEFmed 2013 as a benchmark. Our findings indicate that different dense model approaches substantially impact retrieval effectiveness, and that applying the CombMAX fusion method to combine their output results further improves effectiveness. Extending context length, however, yielded mixed results depending on the input data. Additionally, domain-specific models—those trained on medical data—outperformed general models trained on broad, non-specialized datasets within their respective fields. Furthermore, when text is the dominant information source, text-only models surpassed multimodal models.

Keywords: Medical Search, Multimodal Retrieval, Dense Retrieval

1 Introduction

The increasing volume of digital medical records and imaging data has made medical case retrieval an important tool for clinical decision-making (Sivarajkumar et al., 2024). Physicians and researchers often need to retrieve relevant cases that share similar characteristics with a given medical query, which may include text and images. This retrieval supports comparing diagnostic outcomes, exploring treatment options, or gaining insights from historical cases. However, the multimodal nature of medical case data—often consisting of

 WWW: https://irrj.org DOI: 10.54195/irrj.19769

both textual descriptions (e.g., reports, diagnoses) and visual content (e.g., radiographs, MRIs)—poses significant challenges for retrieval systems. Traditional information retrieval methods frequently rely on text-based searches, which may fail to capture the intricate relationships between textual and visual data fully. The challenge lies in effectively integrating these modalities to improve retrieval accuracy and relevance, i.e., to retrieve results that are not only topically relevant but also visually and semantically aligned with the query.

Recent advances in deep learning, particularly in dense model-based approaches, offer new opportunities for multimodal retrieval. Multimodal models encode multiple modalities into a shared latent space, allowing for retrieval based on cross-modal semantic similarity.

In this paper, we explore whether dense multimodal models can outperform traditional retrieval methods by addressing the challenge of integrating text and images through dense representations within a multimodal framework. To evaluate the effectiveness of various dense model-based approaches, we focus on the case-based retrieval task from ImageCLEFmed 2013. The ImageCLEFmed 2013 dataset is distinguished by its multimodal collection of text and images, along with relevance judgments that are key for effectiveness evaluation. To guide our investigation, we pose the following research questions:

RQ1 Which characteristics of dense models have the greatest impact on retrieval effectiveness in multimodal search systems?

RQ2 How does the effectiveness of dense multimodal models compare to traditional search systems in medical case retrieval, and what factors influence their relative effectiveness?

To answer **RQ1**, we conduct a series of experiments, exploring different result fusion methods and dense models. These experiments analyze how factors like context length and domain specificity, particularly within the medical domain, influence retrieval effectiveness. Our findings show that dense model approaches significantly influence results, with the CombMAX fusion method yielding the best effectiveness, specialized domain-specific models surpassing general ones, context length extensions producing mixed effects, and text-based models outperforming multimodal models when text is the primary information source. To answer **RQ2**, we perform a comparative analysis against the leading submissions from the ImageCLEFmed 2013 case-based retrieval task, which predominantly employed traditional sparse approaches. The results suggest that dense retrieval holds great potential, particularly for improving semantic similarity searches across different modalities.

2 Related work

Our work explores dense retrieval models for multimodal medical ad hoc search, a field that builds upon two key areas: medical case retrieval and multimodal fusion. Medical case retrieval has evolved from traditional text-based methods, such as keyword search, query expansion, and relevance feedback, to multimodal systems that integrate various types of medical data. While systems like PubMed,[1] which use keywords, Boolean operators, and MeSH terms to refine search results, remain foundational in biomedical information

1. `https://pubmed.ncbi.nlm.nih.gov/`

retrieval (Jin et al., 2024; Lu et al., 2009), they are primarily text-centric. In contrast, real-world clinical scenarios often demand the integration of heterogeneous evidence, including both textual and visual data, to support accurate diagnosis and treatment.

Multimodal retrieval systems integrate multiple evidence types by combining information from different modalities. Early examples include MedGIFT (Group, 2009), which allows independent text or image-based searches, and img(Anaktisi) (Zagoris et al., 2009), focused on image-based retrieval across medical datasets. Later approaches, like NovaMed-Search (Mourão and Martins, 2013), fuse textual and visual data to improve case retrieval performance. The ImageCLEF medical retrieval task (Müller and Kalpathy-Cramer, 2010) introduced fusion techniques specifically for case retrieval, emphasizing the integration of textual and visual information. Comparisons of different fusion methods in medical case retrieval tasks were further explored by Garcia Seco de Herrera et al. (2015).

Multimodal fusion systems combine information from diverse sources to support decision-making by creating a context-aware representation. Fusion techniques include early fusion, late fusion, and hybrid fusion, which blend aspects of both approaches (Bayoudh et al., 2022; Zhang et al., 2021; Bruni et al., 2014).

Early fusion combines raw data or features from multiple modalities at the beginning of the data processing pipeline, allowing a more thorough cross-modal correlation analysis. However, it often requires data standardization, such as dimensionality reduction, to ensure compatibility across modalities. In contrast, late fusion integrates information from different modalities at the decision stage, where each modality is processed independently before merging results through operations like concatenation or averaging (Feng et al., 2021). This approach offers more flexibility and can better handle individual modality errors but often fails to capture complex cross-modal dependencies and interactions (Zhang et al., 2021).

Recent studies have explored deep learning-based multimodal fusion for medical tasks. Li et al. (2024) reviewed deep learning-based information fusion techniques for medical image classification, highlighting their role in enhancing medical decision-making. Similarly, Cui et al. (2023) provided a comprehensive review of deep multimodal fusion of image and non-image data for disease diagnosis and prognosis, demonstrating the growing importance of multimodal techniques in biomedical applications. These works underscore the need for optimized fusion techniques tailored to medical retrieval tasks.

In this work, we focus on late fusion strategies to aggregate individual results (e.g., derived from different modalities) into a final score. This choice is motivated by the desire to leverage the strengths of individual modality-specific models while maintaining flexibility in handling different data types and potential errors within each modality. Within late fusion, there are score-based methods (e.g., CombSUM, CombMNZ) that merge normalized scores and rank-based methods (e.g., CombMAX) that prioritize document order (Hsu and Taksa, 2005). Rank-based methods are further categorized into positional (e.g., Borda Fuse (Aslam and Montague, 2001), and Reciprocal Rank Fusion (Cormack et al., 2009)) and majoritarian (e.g., Condorcet (Montague and Aslam, 2002)), with positional approaches assigning votes by rank and majoritarian methods using pairwise comparisons between documents.

Despite extensive exploration of image–text multimodal models, their applications in the biomedical field remain under-researched, particularly in areas such as clinical case retrieval. Guo et al. (2024) provided a detailed survey on advancements in these models, emphasizing their impact on biomedical multimodal technologies. A major breakthrough

in multimodal retrieval was the introduction of CLIP (Radford et al., 2021), which learns joint representations of images and text through contrastive learning. CLIP has shown strong zero-shot and few-shot learning capabilities across diverse image-text tasks, making it a central component in many recent multimodal retrieval systems, including our work. Although more recent approaches such as ColPali (Faysse et al., 2025) have emerged, CLIP remains a foundational model that influenced the development of our system, particularly given its relevance at the time of our experiments.

Building on these insights, our study addresses the specific challenges in medical case-based retrieval tasks by evaluating various dense model-based approaches. The findings contribute to a deeper understanding of which dense multimodal approaches are most effective for biomedical applications, contributing to the broader development of image-text retrieval technologies in this domain.

3 Methodology

This section presents the methodology for achieving the research goals, including defining the retrieval task, presenting the experimental pipeline, and detailing the experimental variables considered and how they will be varied to assess their impact on the final results.

3.1 ImageCLEFmed 2013 case-based retrieval task

We selected the case-based medical retrieval task from ImageCLEFmed 2013 (Garcia Seco de Herrera et al., 2013) due to its multimodal collection of text and images and the availability of relevance judgments, which are crucial for evaluating a search system. To our knowledge, this was the only available dataset with these characteristics, making it our best choice. The task simulates a clinician's diagnostic workflow by finding articles from a vast collection of biomedical literature (PubMed Central[2]) that could aid in differential diagnosis, based on a given case description and images of a patient's case. The dataset includes 75,000 articles and 35 query topics (i.e., cases), all following a well-defined structure, and a total of 300,000 images. Each article is structured into sections such as the title, author names, abstract, full text, figures, and captions, while the query topics, also divided into sections, contain a detailed case description and several relevant images.

Our analysis of the dataset revealed significant heterogeneity in the textual component, with text sections ranging from concise titles (average of 21 tokens) to fulltexts (up to 90,605 tokens). For the visual component, it was found that images in relevant articles differ notably from query images. Article images are mostly graphs and charts, while query images are medical exam images, creating a disadvantage for visual retrieval. Compound images, which contain multiple sub-images within a single frame, also add complexity.

In ImageCLEF, the articles were evaluated for relevance based on their contribution to differential diagnosis, using a three-point scale: relevant, partly relevant, or non-relevant (Kalpathy-Cramer et al., 2015). Analysis showed that physicians prioritized textual information over visual content in their decision-making. The images in relevant articles differed significantly from the query images, suggesting that visual content played a secondary role in their relevance assessments. Consequently, it is unlikely that a system relying solely on

2. https://www.ncbi.nlm.nih.gov/pmc/

visual information could effectively retrieve relevant articles. Furthermore, Garcia Seco de Herrera et al. (2017) observed that incorporating visual data into a multimodal approach did not enhance retrieval effectiveness for the specific topics of the ImageCLEF task.

A total of 15,028 relevance judgments were made across all query topics, with only 0.57% of the collection judged per topic, as detailed in Section 5.7. The limited judgments, especially the lack of relevant documents, posed challenges, as also noted by Garcia Seco de Herrera et al. (2015). This impacted top results in the case-based task with relatively low scores, where MAP scores ranged from 0.0281 to 0.2429, depending on retrieval type (visual, mixed, or textual), as discussed in Section 5.6. Mitigation efforts are explored in Section 6.

3.2 Experimental multimodal retrieval pipeline

To systematically experiment with and evaluate dense model-based approaches, we built a functional prototype of the retrieval system. The source code is openly available on GitHub.[3] The workflow is organized into five key steps:

1. **Dataset collection and article encoding**: In this step, raw data is processed and encoded into dense embeddings.

2. **Storage and indexing of embeddings**: The encoded articles are represented in an embedding space and indexed for efficient retrieval. We used Faiss[4] (Johnson et al., 2021) version 1.8.0 with the GPU implementation, employing its HNSW index with squared Euclidean (L2) distance, which preserves ranking while improving efficiency by avoiding square root calculations.

3. **Query encoding**: Query documents are transformed into embeddings followed by a similarity search, using Faiss, against the pre-computed indexed embeddings of the articles for retrieval.

4. **Results fusion**: Results from multiple retrieval approaches are combined, matching each section of the query with corresponding sections in the article, which may contain textual or visual information.

5. **Retrieval**: A ranked list of documents based on the fused results is produced.

3.3 Experimental variables

With the functional pipeline established as the foundation for experimentation, we have the flexibility to adjust and refine various elements of the system. Through this process, we systematically assess how different changes impact the effectiveness of the retrieval task.

3.3.1 Results fusion

Since the documents consist of multiple distinct sections, each section of the article document must be compared with each section of the topic document, resulting in multiple

3. https://github.com/catarinaopires/eval-multimodal-medical-case-retrieval
4. https://faiss.ai

ranked lists that require results fusion. For instance, when comparing a topic description with article images, we need to compute the similarity between the textual description and each article image, generating multiple ranked lists. To obtain a single final ranking, we then apply one of the result fusion methods.

We experiment with CombSUM, CombMAX, and CombMNZ, given their demonstrated effectiveness within the Comb family, introduced by Shaw and Fox (1994). For each document i, the score after fusion can be computed as:

$$CombSUM(i) = \sum_{k=1}^{N(i)} S_k(i), \tag{1}$$

$$CombMAX(i) = max(S), \forall S \subset D_i, \tag{2}$$

$$CombMNZ(i) = N(i) * CombSUM(i), \tag{3}$$

where $S_k(i)$ is the score of the i-th document in the k-th result list, $N(i)$ refers to the number of times a document appears in the result lists, and D_i denotes the set of scores (S) assigned to document i across all result lists in which it appears.

3.3.2 Models

Our study explores the use of dense models in multimodal search, focusing on how various model architectures and capabilities impact retrieval effectiveness, particularly when handling both text and image data. The models employed, primarily using HuggingFace[5] implementations in Python, include CLIP's ViT-B/16 variant (Radford et al., 2021), Long-Clip B/16 (Zhang et al., 2024), a fine-tuned version of CLIP that extends the token capacity from 77 to 248 for longer text–image pairs, PubMedCLIP (Eslami et al., 2021), a fine-tuned CLIP model for medical image-caption tasks, as well as Llama 3 (Grattafiori et al., 2024) with 8 billion parameters and LLaVA-1.5 (Liu et al., 2023) with 7 billion parameters.

One of the challenges in multimodal retrieval is managing different modalities. A key decision is whether to use a multimodal model that processes both text and images together or to handle each modality separately using unimodal models. While CLIP effectively manages visual data, it struggles with long texts due to its limited token capacity. Aggravating this problem, Zhang et al.'s experimental findings suggest that the effective length of text that CLIP can handle optimally is no more than 20 tokens, beyond which it struggles to utilize the additional information effectively. This limitation can result in the loss of crucial information necessary for accurate retrieval. Although models like LongCLIP attempt to address CLIP's token limit by increasing token capacity, alternatives like large language models (e.g., Llama 3) may be better suited for handling lengthy text inputs.

The main strength of multimodal models lies in their capacity to encapsulate data across various modalities within a shared latent space, facilitating comparison and relationship establishment. In contrast, employing unimodal models would compromise this key capability, impeding cross-modality comparisons. Nevertheless, it is feasible to mitigate the limitation of unimodal models in cross-modality comparisons by homogenizing data modalities into a

5. `https://huggingface.co/`

singular format, like text. A potential solution is to convert visual data into text through image descriptions, allowing unimodal models to handle the transformed data exclusively in text format. Using a multimodal generative model such as LLaVA, images are translated into textual descriptions, which can then be compared with existing text-based content. Transforming visual data into text-based representations facilitates the comparison of images and text within a unified latent space, even when using text-only models.

Finally, our study evaluates the potential benefits of domain-specific models, such as PubMedCLIP, over general-purpose models like CLIP for improving multimodal search systems in medical information retrieval.

4 Experimental setup

This section presents the experimental setup, outlining the planned experiments along with their objectives and focus. It also presents the evaluation metrics used to assess the impact of each approach and details the computational resources employed for execution.

4.1 Experiments

To address the outlined research questions, we conducted five experiments, focusing on model variations to investigate how different factors influence retrieval effectiveness.

Exp. 1 **Results fusion effectiveness:** Firstly, we will evaluate the impact of using different results fusion approaches, such as CombSUM, CombMAX, and CombMNZ, by applying them to retrieval outputs from CLIP and comparing their effectiveness.

Exp. 2 **Effect of context length:** Our second experiment will examine whether the context length of a model influences the obtained results. To explore this, we will compare the effectiveness of the CLIP model with the LongCLIP model.

Exp. 3 **Domain-specific model effectiveness:** Our third experiment will investigate whether a domain-specific model can outperform a general-purpose model. This experiment will involve comparing the effectiveness of PubMedCLIP against CLIP.

Exp. 4 **Unimodal vs. Multimodal effectiveness:** Our fourth experiment examines whether a unimodal approach can outperform a multimodal baseline with the dataset at hand. This will involve comparing the outcomes of Llama against CLIP.

Exp. 5 **Dominant data type approach:** The fifth experiment assesses whether selecting a dominant data type (text) and converting visual content to a textual representation can outperform the baseline, which uses the original multimodal data. To investigate this, we will compare the effectiveness of searches using the existing topic sections with those using LLaVa's generated topic image descriptions. This comparison will involve all utilized models, not just the text model Llama.

4.2 Measure

Following established methodologies and metrics from ImageCLEFmed 2013, we report MAP as the primary metric, along with GM-MAP (Geometric Mean, or GMAP), bpref,

and P@10/@30 as complementary metrics. The highest scores in each column are bolded, and statistically significant results are marked in the tables, based on a two-tailed paired permutation test with 100,000 permutations. A Holm-Bonferroni correction was applied at the 0.05 significance level (95% confidence interval) to account for multiple comparisons when evaluating the effectiveness of different dense model-based approaches on the selected dataset. In addition, we report effect sizes (Cohen's d_z), standard errors (SE), and 95% confidence intervals (CI), computed for each comparison relative to its respective baseline.

4.3 Computational resources

Initial experiments were conducted on a server equipped with two NVIDIA GeForce RTX 2080 Ti GPUs, each with 11GB of VRAM, which provided sufficient resources for running smaller models like CLIP. As the complexity of the experiments increased, the computational tasks were migrated to a more advanced computing environment managed by SLURM (Jette and Wickberg, 2023). This setup featured multiple GPUs, including NVIDIA Tesla V100 and NVIDIA A100 models, with 32GB and 80GB of VRAM, respectively. Some steps were run without GPUs and on the less advanced setup to minimize resource usage when possible.

5 Results

The overall results for all experiments are summarized in Table 1, with the scores presented as averages to provide an overview of effectiveness across all topics. Additionally, effect sizes, standard errors, and 95% confidence intervals are reported for all statistically significant comparisons ($p < 0.05$, Holm-Bonferroni corrected) in Table 2. To evaluate each proposed approach, we conducted individual searches for each section of the topic documents against each section of the article documents, testing all possible combinations. Thus, the results tables are organized by topic section (Description, Images) and article section (Title, Abstract, Fulltext, Images, Captions). The scores represent the outcomes of comparing each section from the topic (left-most label) to each section from the article (row-label).

5.1 Results fusion effectiveness

The first experiment examines the impact of different result fusion methods—CombSUM, CombMAX, and CombMNZ—without altering the underlying CLIP model. CombSUM serves as the baseline for comparison, and the effectiveness results for each method are presented in the top section of Table 1. Since topic descriptions and article titles, abstracts, and fulltexts are directly compared, results fusion is unnecessary. Therefore, CombMAX and CombMNZ scores remain unchanged from the baseline and are omitted from the table.

CombMAX consistently outperforms the baseline, particularly when comparing topic images to article images and captions. The medium effect size indicates that the observed improvement is likely to be meaningful in practice, even if not large, indicating better prioritization of relevant documents at the top results. In contrast, CombMNZ produces results nearly identical to CombSUM, with no meaningful improvements observed.

Table 1: Summary of experimental results. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in "Gen. I. Desc." rows), statistically significant differences relative to the description and image baselines are marked with 'd' and 'i', respectively. Highest scores per column are bolded.

		CLIP CombSUM[Exp. 1]					**CLIP CombMNZ**[Exp. 1]					**CLIP CombMAX**[Exp. 1-5]				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.0396**	**0.0074**	**0.1028**	**0.0714**	**0.0543**	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0177	0.0020	0.0799	0.0371	0.0305	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0280	0.0022	0.1013	0.0343	0.0324	-	-	-	-	-	-	-	-	-	-
	Images	0.0096	0.0022	0.0554	0.0143	0.0143	0.0096	**0.0022**	0.0554	0.0143	0.0143	0.0142	0.0026	0.0593	0.0143	0.0229
	Captions	0.0204	0.0018	0.0664	0.0371	0.0295	**0.0204**	0.0018	**0.0664**	**0.0371**	**0.0295**	0.0370*	**0.0031**	0.0941	**0.0629**	**0.0467***
Images	Title	0.0074	0.0015	0.0470	0.0200	0.0124	0.0074	0.0015	0.0470	0.0200	0.0124	0.0164*	0.0027	0.0597	0.0286	0.0210
	Abstract	0.0048	0.0006	0.0420	0.0171	0.0095	0.0048	0.0006	0.0420	0.0171	0.0095	0.0181	0.0011	0.0725	0.0343	0.0276
	Fulltext	0.0042	0.0007	0.0465	0.0086	0.0067	0.0042	0.0007	0.0465	0.0086	0.0067	0.0140	0.0012	0.0674	0.0171	0.0171
	Images	0.0170	0.0018	0.0546	0.0143	0.0105	0.0100*	0.0015	0.0510	0.0086	0.0124	**0.0397***	0.0034	**0.0963**	0.0486*	0.0305*
	Captions	0.0076	0.0015	0.0480	0.0200	0.0143	0.0076	0.0015	0.0480	0.0200	0.0143	0.0212*	0.0028	0.0718	0.0257	0.0267
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0048[di]	0.0002	0.0341[d]	0.0086[d]	0.0086[d]
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0128	0.0003	0.0537	0.0171	0.0095[d]
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0073[di]	0.0002	0.0435[d]	0.0200	0.0095[d]
	Images	-	-	-	-	-	-	-	-	-	-	0.0085[i]	0.0016	0.0599	0.0086[i]	0.0143[i]
	Captions	-	-	-	-	-	-	-	-	-	-	0.0050[di]	0.0007	0.0482[d]	0.0086[d]	0.0067[d]

		LongCLIP CombMAX[Exp. 2]					**PubMedCLIP CombMAX**[Exp. 3]					**Llama CombMAX**[Exp. 4]				
Description	Title	0.0122*	0.0010	0.0671	0.0314	0.0238*	0.0437	0.0047	0.1122	**0.0686**	**0.0562**	0.0304	0.0067	0.0771	0.0714	0.0419
	Abstract	0.0304	**0.0035**	0.0944	**0.0514**	0.0314	0.0504	**0.0069**	**0.1180**	0.0629	0.0457	0.0457*	0.0119	0.0947	0.0771	0.0629*
	Fulltext	**0.0370**	0.0019	**0.1046**	0.0457	**0.0352**	0.0184	0.0016	0.0866	0.0457	0.0314	0.0682*	**0.0141**	0.1152	0.0971*	**0.0771***
	Images	0.0192	0.0035	0.0748	0.0343	0.0210	0.0220	0.0014	0.0651	0.0429	0.0371	-	-	-	-	-
	Captions	0.0209	0.0033	0.0687	0.0486	0.0333	0.0344	0.0064	0.0869	0.0600	0.0524	**0.0686**	0.0123	**0.1171**	**0.1057**	0.0667
Images	Title	0.0184	0.0023	0.0644	0.0371	0.0295	0.0102	0.0007	0.0705	0.0200	0.0143	-	-	-	-	-
	Abstract	0.0189	0.0026	0.0618	0.0200	0.0267	0.0176	0.0016	0.0832	0.0143	0.0143	-	-	-	-	-
	Fulltext	0.0106	0.0008	0.0548	0.0171	0.0133	0.0101	0.0006	0.0655	0.0171	0.0133	-	-	-	-	-
	Images	0.0305	0.0030	0.0913	0.0371	0.0324	**0.0530**	0.0057	0.1011	0.0714	0.0429	-	-	-	-	-
	Captions	0.0114	0.0020	0.0560	0.0057	0.0152	0.0224	0.0041	0.0760	0.0457	0.0314	-	-	-	-	-
Gen. I. Desc.	Title	0.0002[di]	0.0000	0.0245[d]	0.0000[di]	0.0000[di]	0.0087[d]	0.0005	0.0416[d]	0.0114[d]	0.0124[d]	0.0016[d]	0.0002	0.0354[d]	0.0029[d]	0.0029[d]
	Abstract	0.0035[di]	0.0003	0.0432[d]	0.0086[d]	0.0086[d]	0.0150[d]	0.0012	0.0669[d]	0.0257	0.0181[d]	0.0027[d]	0.0003	0.0337[d]	0.0029[d]	0.0057[d]
	Fulltext	0.0021[di]	0.0001	0.0417[d]	0.0029[d]	0.0057[d]	0.0045[d]	0.0004	0.0492[d]	0.0086	0.0114	0.0012[d]	0.0001	0.0185[d]	0.0000[d]	0.0000[d]
	Images	0.0100[di]	0.0015	0.0549	0.0086	0.0124[i]	0.0030[di]	0.0004	0.0250[di]	0.0029[i]	0.0029[di]	-	-	-	-	-
	Captions	0.0026[di]	0.0002	0.0283[d]	0.0057[d]	0.0048[d]	0.0047[di]	0.0008	0.0359[di]	0.0057[di]	0.0057[di]	0.0012[d]	0.0002	0.0352[d]	0.0000[d]	0.0010[d]

Table 2: Effect sizes (Cohen's d_z), standard errors (SE), and 95% confidence intervals (CI) for significant comparisons in Tables 1, 4, and 5. Interpretations of effect sizes follow standard thresholds: small (0.2), medium (0.5), and large (0.8).

		Comparison	**Metric**	d_z	**SE**	**95% CI**	**Interpretation**
Original Qrels	CLIP CombMNZ	Topic Images vs. Article Images	MAP	-0.18	0.17	[-0.53, 0.16]	Small
	CLIP CombMAX	Topic Description vs. Article Captions	MAP	0.40	0.17	[0.05, 0.76]	Small
			P30	0.60	0.18	[0.23, 0.98]	Medium
		Topic Images vs. Article Title	MAP	0.48	0.18	[0.12, 0.84]	Small
		Topic Images vs. Article Images	MAP	0.38	0.17	[0.03, 0.74]	Small
			P10	0.58	0.18	[0.21, 0.95]	Medium
			P30	0.66	0.19	[0.28, 1.03]	Medium
		Topic Images vs. Article Captions	MAP	0.38	0.17	[0.02, 0.74]	Small
	LongCLIP CombMAX	Topic Description vs. Article Title	MAP	-0.51	0.18	[-0.88, -0.14]	Medium
			P30	-0.51	0.18	[-0.88, -0.15]	Medium
	Llama CombMAX	Topic Description vs. Article Abstract	MAP	0.58	0.18	[0.21, 0.95]	Medium
			P30	0.61	0.18	[0.24, 0.99]	Medium
		Topic Description vs. Article Fulltext	MAP	0.44	0.18	[0.08, 0.80]	Small
			P10	0.56	0.18	[0.19, 0.93]	Medium
			P30	0.52	0.18	[0.16, 0.89]	Medium
Subset	LongCLIP CombMAX	Topic Description vs. Article Title	MAP	-0.46	0.18	[-0.82, -0.10]	Small
Expanded Qrels	CLIP CombMNZ	Topic Images vs. Article Images	MAP	-0.43	0.18	[-0.79, -0.07]	Small
			bpref	-0.35	0.17	[-0.71, -0.00]	Small
	CLIP CombMAX	Topic Description vs. Article Captions	bpref	0.48	0.18	[0.12, 0.84]	Small
		Topic Images vs. Article Title	MAP	0.59	0.18	[0.22, 0.96]	Medium
			bpref	0.53	0.18	[0.16, 0.90]	Medium
		Topic Images vs. Article Abstract	MAP	0.43	0.18	[0.08, 0.79]	Small
			bpref	0.49	0.18	[0.13, 0.86]	Small
		Topic Images vs. Article Captions	MAP	0.50	0.18	[0.14, 0.87]	Medium
			bpref	0.55	0.18	[0.18, 0.92]	Medium
	LongCLIP CombMAX	Topic Images vs. Article Captions	MAP	-0.46	0.18	[-0.82, -0.10]	Small
	PubMedCLIP CombMAX	Topic Description vs. Article Abstract	MAP	0.54	0.18	[0.17, 0.90]	Medium
			bpref	0.50	0.18	[0.14, 0.87]	Medium
	Llama CombMAX	Topic Description vs. Article Abstract	MAP	0.44	0.18	[0.08, 0.80]	Small
	PubMedCLIP CombMAX	Topic Description vs. Article Images	bpref	0.36	0.17	[0.00, 0.71]	Small
			P30	0.30	0.17	[-0.05, 0.65]	Small
		Topic Description vs. Article Captions	P30	0.46	0.18	[0.10, 0.82]	Small

An example of CombMAX's advantage is seen in topic 29, where the query includes two head CT scans, and a relevant article (per qrels) contains one matching CT and one unrelated MRI. CombMAX ranked the article 16th by emphasizing the strongest match, while CombSUM ranked it 124th due to the MRI diluting the overall score. This highlights CombMAX's strength when a single strong match determines relevance.

Overall, **CombMAX emerges as the most effective result fusion method in this context**, consistently outperforming CombSUM and CombMNZ across most metrics.

5.2 Effect of context length

The second experiment investigates the impact of context length by comparing CLIP (77 tokens) with LongCLIP (248 tokens), using CombMAX across all runs. The baseline involves CLIP, and the experiment assesses whether longer context lengths improve results by maintaining the same fusion method for comparison. Results are at the bottom of Table 1.

Results indicate that CLIP generally performs better with short texts like titles and captions, aligning with its original training setup. LongCLIP shows mixed results: it slightly improves effectiveness on longer inputs such as abstracts, but often underperforms on shorter texts, with some statistically significant drops. Visual comparisons show varied results across models, likely due to differences in the text–image alignment learned during training.

In summary, while **context length influences retrieval effectiveness, a longer context does not consistently yield better results**. The benefits of longer context length depend on the type and structure of the input data.

5.3 Domain-specific model effectiveness

The third experiment evaluates whether a domain-specific model, PubMedCLIP, outperforms the general-purpose CLIP model in biomedical retrieval tasks, using CombMAX. As shown in the bottom portion of Table 1, the goal is to determine if fine-tuning a model for a specific domain yields better results than a model pre-trained on diverse data.

While statistical significance is limited, PubMedCLIP generally performs better, particularly in retrieving abstracts and article images. Its improvements suggest stronger domain alignment, especially in extracting abstract-level semantics and visual information. However, inconsistent effectiveness across certain queries, reflected in lower GM-MAP, indicates challenges in handling difficult or ambiguous topics.

In summary, the results show that the **domain-specific model (PubMedCLIP) outperforms the general-purpose model (CLIP) for biomedical searches**, achieving higher effectiveness in retrieving relevant cases.

5.4 Unimodal vs. Multimodal effectiveness

The fourth experiment evaluates whether a unimodal text model, Llama, can outperform a multimodal model, CLIP, using the CombMAX fusion method. Since Llama handles only textual input, it is evaluated solely on text-based searches, while CLIP also includes mixed and visual comparisons (see bottom of Table 1) to assess the effectiveness of unimodal compared to multimodal models.

Llama generally outperforms CLIP across most textual inputs, particularly with longer texts like abstracts and fulltexts, reflecting its stronger language modeling capabilities and training on longer contexts. The observed medium effect sizes suggest that these improvements are not only statistically significant but also practically meaningful. A clear example is topic 29, where Llama ranked a relevant article at the top, while CLIP placed it at 309. The key factor is context length: CLIP's 77-token limit only covered the introduction, missing critical content, whereas Llama, with its 8,192-token capacity, processed the entire article, including two case studies and conclusions. This enabled a deeper semantic understanding, crucial to determining the relevance of the article to this case description.

While CLIP performs slightly better on short texts like titles, Llama surprisingly surpasses it on image captions as well, possibly due to CLIP's limited token capacity when handling complex topic descriptions.

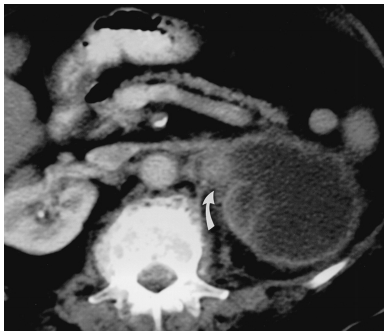
Overall, **Llama consistently outperforms CLIP in nearly all textual sections, as well as in mixed and visual searches**. Llama's superior effectiveness highlights its strength with purely textual content, while CLIP's multimodal capabilities offer no significant advantage in this context. This supports the hypothesis from Section 3.1 that **textual information is prioritized over visual content for this task under evaluation**.

5.5 Dominant data type approach

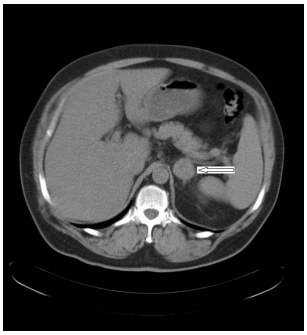
The fifth experiment assesses whether converting visual content into text and prioritizing text can outperform a multimodal approach. This is tested by comparing searches using existing topic sections with those using LLaVa-generated image descriptions (labeled "Gen. I. Desc."). While the focus is on the effectiveness of these generated descriptions, mixed searches that combine them with article images are also analyzed to assess multimodality against unimodality. Instead of relying solely on the text-only Llama model, all models are tested using the CombMAX fusion method. To ensure fair comparisons, each setup is assessed against the corresponding model with statistical relevance tests conducted against the "Topic Description" and "Images" sections. Results show that nearly all metrics are statistically significant for at least one baseline, supporting multiple significant conclusions.

The generation of image descriptions imposed a token limit of 1024 per image to ensure detail and prevent truncation, but actual token counts ranged from 47 to 133, averaging 79 tokens. Descriptions typically began with high-level features, such as color or subfigure count, and then moved on to more specific details within the image. However, the evaluation revealed mixed accuracy due to the model's lack of medical domain fine-tuning, resulting in some errors and inconsistencies. Additionally, a recurring issue was the inconsistency between the topic-generated image descriptions and the image captions, differing in detail, wording, length, and intent. This mismatch, shown by an example in Figure 1, contributed to the observed effectiveness differences.

For CLIP and LongCLIP, visual searches consistently outperformed those using generated descriptions, suggesting that these text surrogates fail to capture the necessary detail or alignment with article content, potentially due to errors in the description generation process and varying levels of detail. PubMedCLIP showed some improvement when comparing generated descriptions to article abstracts, but effectiveness dropped sharply in visual com-



(a) Query image example with LLaVa-generated description: "The image is a black and white medical image of a person's abdomen, likely an X-ray or CT scan. The abdomen is filled with various organs, including the liver, spleen, and pancreas. The liver is located on the left side of the image, while the spleen is situated in the middle, and the pancreas is on the right side. There is a small arrow pointing towards the right side of the image, possibly indicating a specific area of interest or a point of reference. The overall image provides a detailed view of the internal organs within the abdomen.".



(b) Article image example with caption: "CT scan showing an adrenal metastasis to the contralateral gland, 2 years after a right nephrectomy for primary RCC.".

Figure 1: Example of similar abdominal CT scan images, both showing an arrow pointing to a region but with completely different descriptions in terms of wording and medical detail.

parisons. Llama, relying solely on text, also performed worse with generated descriptions than with original topic descriptions, likely due to differences in wording mentioned above.

The experiment assessed whether using text as the dominant data type could outperform a multimodal baseline. While a comparison between the unimodal Llama model and the multimodal CLIP model was obvious, intended to test text-only versus multimodal effectiveness, it was not entirely fair due to differences in models and augmented data. Results showed that Llama with generated image descriptions did not surpass CLIP, though the comparison is inequitable. The experiment suggests that comparing article images with generated captions and topic images could further explore whether visual searches have an advantage over mixed searches, as in the previous experiment the multimodal approach showed no advantage over the unimodal. However, in this experiment, **all textual searches using generated descriptions performed worse than their multimodal baselines, likely due to limitations in the generated data**. Additionally, the misalignment between topic and article captions hindered effectiveness, supporting prior conclusions that topic and article images differ significantly. Despite this effectiveness disparity, CLIP and LongCLIP performed best in visual searches, while PubMedCLIP excelled in comparisons involving topic images and article captions, benefiting from medical domain fine-tuning.

5.6 Comparative analysis against top submissions

To contextualize the effectiveness of our retrieval system, we compare it to the top submissions from the ImageCLEFmed 2013 case-based retrieval task. However, this comparison is not entirely fair due to differences in evaluation methodology. In a typical ImageCLEF

Table 3: Top submissions from ImageCLEFmed 2013 Case-based retrieval task (2013) alongside our highest scores for the same task. Asterisk (*) mark results discussed in Section 6.

	Runid	Retrieval type	MAP	GM-MAP	bpref	P10	P30
Top sub.	SNUMedinfo9 (Choi et al., 2013)	Textual	**0.2429**	**0.1163**	**0.2417**	**0.2657**	**0.1981**
	FCT_CB_MM_rComb (Mourão et al., 2013)	Mixed	**0.1608**	**0.0779**	**0.1426**	**0.1800**	**0.1257**
	FCT_SEGHIST_6x6_LBP (Mourão et al., 2013)	Visual	0.0281	0.0009	0.0335	0.0429	0.0238
Ours original	Llama CombMAX (desc. & cap.)	Textual	0.0686	0.0123	0.1171	0.1057	0.0667
	PubMedCLIP CombMAX (img. & cap.)	Mixed	0.0224	0.0041	0.0760	0.0457	0.0314
	PubMedCLIP CombMAX (img. & img.)	Visual	**0.0530**	**0.0057**	**0.1011**	**0.0714**	**0.0429**
Ours* subset	Llama CombMAX (desc. & cap.)	Textual	0.1210	0.0301	0.1171	0.1629	0.1267
	PubMedCLIP CombMAX (img. & abs.)	Mixed	0.0767	0.0144	0.0832	0.1514	0.1229
	PubMedCLIP CombMAX (img. & img.)	Visual	**0.0967**	**0.0134**	**0.1011**	**0.1257**	**0.1105**
Ours* expanded	Llama CombMAX (desc. & full.)	Textual	0.0419	0.0293	0.0733	0.1657	0.1448
	CLIP CombMAX (img. & title)	Mixed	0.0229	0.0130	0.0425	0.1143	0.0905
	PubMedCLIP CombMAX (img. & img.)	Visual	**0.0282**	**0.0146**	**0.0515**	**0.1257**	**0.0914**

evaluation campaign, the top 30–60 results from each submitted run are merged to create judgment pools of approximately 1,000 cases per topic, which are combined to create pools of approximately 1,000 cases per topic, which are then manually assessed (Kalpathy-Cramer et al., 2015). Since our retrieved documents were not included in this pooling process, only an average of 19.73% of our retrieved articles were judged across all topics, introducing a considerable margin of uncertainty in the effectiveness evaluation.

As shown in Table 3, even our best-performing setup falls significantly behind the top submissions, but we achieved better results in purely visual searches using all the studied multimodal models (CLIP, LongCLIP, PubMedCLIP) compared to top visual submission, which relied on sparse feature extraction methods. This suggests that dense approaches can capture more relevant information. However, in mixed and textual searches, our system significantly underperformed. The top submissions enhanced their text components using an external corpus (MEDLINE) to perform term expansion, contributing to their success. Although we used a somewhat similar approach using the medical-specific model (PubMedCLIP), its inherent token capacity limitations prevented it from handling the large textual sections effectively, likely explaining the discrepancies in retrieval effectiveness.

5.7 Summary

From the five experiments, we can draw several conclusions. The majority of the effect sizes fall within the small and medium ranges. No large effects ($d_z \geq 0.8$) are observed, indicating that while meaningful differences exist, they are generally modest. We begin by analyzing the overall results shared across all experiments and summarizing the key findings.

In the ImageCLEFmed 2013 case-based retrieval task, a total of 709 documents were judged as relevant and 14,319 as non-relevant across 35 query topics, with between 372 and 480 documents judged per topic, covering only about 0.57% of the 75,000 article collection per query topic. This small sample size, due to the pooling technique, may limit the completeness of relevance assessments. Ideally, all documents should be judged to ensure a more accurate evaluation, as unjudged documents are often assumed non-relevant (Clough and Sanderson, 2013), potentially overlooking relevant ones and affecting retrieval effectiveness.

Our overall results are lower than the top submissions of the task. The bpref measure consistently yields better results than MAP across all experiments, indicating that more unjudged documents were retrieved, some of which could be relevant. On average, no more than 19.73% of our retrieved articles were judged across all topics, limiting the evaluation. While bpref accounts for incomplete relevance judgments, it only focuses on the ranking of relevant over non-relevant documents. The highest percentage of judged retrieved articles corresponds with the highest MAP achieved, suggesting that unjudged articles might be relevant. Conversely, the lowest percentage of judged articles resulted in our lowest MAP score (0.0002), yet it did not show the lowest bpref measure. This difference highlights the known property of MAP, which treats all unjudged articles as non-relevant, while bpref handles incomplete judgments by considering only on judged documents.

From all the experiments, we conclude that for the dataset used, CombMAX is the best fusion method out of the ones tested (Exp. 1). Context length affects effectiveness (Exp. 2), showing both advantages and disadvantages based on the input data. Domain-specific models are better suited for their respective domains (Exp. 3). Text-based models can outperform multimodal models when text is the primary information source (Exp. 4). Finally, text searches based on generated descriptions significantly underperform those using the original model on both text and visuals, probably due to limitations in the augmented topic section (Exp. 5).

6 Mitigating incomplete judgments

Many missing judgments in the ground truth may affect reliability of results and conclusions. This can be addressed by adapting the evaluation or expanding the ground truth.

Adapting the evaluation focuses on the subset of the dataset with existing ground truth. This approach, explored in Section 6.1, evaluates system effectiveness within this subset, providing a partial but informative picture. It is a simple, practical method using existing data, but it has limitations. The evaluation assumes that the subset represents the entire collection, which requires careful interpretation as it does not provide a complete evaluation.

Expanding the ground truth can be achieved through manual or semi-supervised techniques. While manual annotation is the most straightforward method, it is often infeasible due to the need for domain experts, as well as its time-intensive and costly nature. In contrast, semi-supervised learning involves using a small labeled dataset to train a model that predicts relevance for unlabeled documents. This method, explored in Section 6.2, can efficiently label large datasets with minimal manual effort, but its accuracy heavily depends on the quality of the initial labeled data and the model's generalization capabilities.

6.1 Retrieval on judged documents

Using trec_eval, we re-ranked our runs and excluded all unjudged documents from the retrieved set, enabling us to calculate metrics solely based on the judged documents, whether relevant or non-relevant. Table 4 presents the overall results for all experiments, considering only the judged documents in the retrieval set. Effect sizes, standard errors, and 95% confidence intervals are also reported for all statistically significant comparisons in Table 2 under "Subset". We analyze these results and compare them with those obtained without excluding unjudged documents, as discussed in Section 5.

Table 4: Summary of experimental results considering only judged documents. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in "Gen. I. Desc." rows), statistically significant differences relative to the description and image baselines are marked with 'd' and 'i', respectively. Highest scores per column are bolded.

		CLIP CombSUM[Exp. 1]					**CLIP CombMNZ**[Exp. 1]					**CLIP CombMAX**[Exp. 1-5]				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.1009**	**0.0207**	**0.1028**	**0.1629**	0.1219	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0672	0.0070	0.0799	0.1343	0.1181	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0968	0.0093	0.1013	0.1371	**0.1305**	-	-	-	-	-	-	-	-	-	-
	Images	0.0509	0.0096	0.0554	0.0943	0.0914	0.0509	**0.0096**	0.0554	0.0943	**0.0914**	0.0565	0.0094	0.0593	0.0886	0.0971
	Captions	0.0592	0.0063	0.0664	0.1029	0.0876	**0.0592**	0.0063	**0.0664**	**0.1029**	0.0876	0.0865	0.0083	0.0941	**0.1371**	**0.1086**
Images	Title	0.0489	0.0092	0.0470	0.0714	0.0733	0.0489	0.0092	0.0470	0.0714	0.0733	0.0662	**0.0128**	0.0597	0.1029	0.0933
	Abstract	0.0389	0.0036	0.0420	0.0686	0.0676	0.0389	0.0036	0.0420	0.0686	0.0676	0.0673	0.0055	0.0725	0.1029	0.0933
	Fulltext	0.0488	0.0053	0.0465	0.0914	0.0895	0.0488	0.0053	0.0465	0.0914	0.0895	0.0717	0.0082	0.0674	0.1057	0.1038
	Images	0.0539	0.0081	0.0546	0.0800	0.0838	0.0465	0.0072	0.0510	0.0743	0.0800	**0.0895**	0.0098	**0.0963**	0.1171	0.1019
	Captions	0.0523	0.0092	0.0480	0.0943	0.0829	0.0523	0.0092	0.0480	0.0943	0.0829	0.0710	0.0124	0.0718	0.0914	0.1029
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0328^{d}	0.0011	0.0341^{d}	0.0743^{d}	0.0486^{di}
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0465	0.0016	0.0537	0.0743^{d}	0.0619^{d}
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0351^{di}	0.0011	0.0435^{d}	0.0743	0.0476^{di}
	Images	-	-	-	-	-	-	-	-	-	-	0.0494^{i}	0.0090	0.0599	0.0857	0.0943
	Captions	-	-	-	-	-	-	-	-	-	-	0.0384^{di}	0.0033	0.0482^{d}	0.0829	0.0781^{d}

		LongCLIP CombMAX[Exp. 2]					**PubMedCLIP CombMAX**[Exp. 3]					**Llama CombMAX**[Exp. 4]				
Description	Title	0.0545^{*}	0.0043	0.0671	0.1400	0.1076	0.1102	0.0127	0.1122	**0.1657**	0.1467	0.0774	0.0225	0.0771	0.1343	0.1152
	Abstract	0.0807	0.0127	0.0944	0.1343	0.1200	**0.1136**	**0.0261**	**0.1180**	0.1571	**0.1562**	0.0971	0.0267	0.0947	0.1571	0.1162
	Fulltext	**0.0855**	0.0062	**0.1046**	**0.1486**	**0.1305**	0.0727	0.0076	0.0866	0.1486	0.1324	0.1200	0.0296	0.1152	**0.1714**	**0.1410**
	Images	0.0697	**0.0138**	0.0748	0.0829	0.0971	0.0578	0.0032	0.0651	0.1057	0.1095	-	-	-	-	-
	Captions	0.0604	0.0119	0.0687	0.1171	0.0981	0.0797	0.0163	0.0869	0.1371	0.1305	**0.1210**	**0.0301**	**0.1171**	0.1629	0.1267
Images	Title	0.0652	0.0120	0.0644	0.1171	0.1019	0.0710	0.0047	0.0705	0.1229	0.1133	-	-	-	-	-
	Abstract	0.0712	0.0135	0.0618	0.1057	0.0943	0.0767	0.0144	0.0832	0.1514	0.1229	-	-	-	-	-
	Fulltext	0.0478	0.0045	0.0548	0.0943	0.0857	0.0579	0.0037	0.0655	0.1229	0.1095	-	-	-	-	-
	Images	0.0853	0.0092	0.0913	0.1143	0.1010	0.0967	0.0134	0.1011	0.1257	0.1105	-	-	-	-	-
	Captions	0.0529	0.0101	0.0560	0.0943	0.0848	0.0755	0.0191	0.0760	0.1257	0.1152	-	-	-	-	-
Gen. I. Desc.	Title	0.0183^{di}	0.0002	0.0245^{d}	0.0486^{d}	0.0210^{di}	0.0447^{d}	0.0030	0.0416^{d}	0.0743^{d}	0.0705^{di}	0.0219^{d}	0.0011	0.0354^{d}	0.0571^{d}	0.0590^{d}
	Abstract	0.0369^{di}	0.0019	0.0432^{d}	0.0686^{d}	0.0667^{d}	0.0665^{d}	0.0080	0.0669^{d}	0.1000	0.0838^{di}	0.0285^{d}	0.0014	0.0337^{d}	0.0514^{d}	0.0533^{d}
	Fulltext	0.0346^{d}	0.0005	0.0417^{d}	0.0714^{d}	0.0476^{d}	0.0370^{d}	0.0025	0.0492^{d}	0.0886	0.0733^{di}	0.0138^{d}	0.0005	0.0185^{d}	0.0286^{d}	0.0438^{d}
	Images	0.0525	0.0067	0.0549	0.0886	0.0943	0.0233^{di}	0.0020	0.0250^{di}	0.0429^{di}	0.0590^{di}	-	-	-	-	-
	Captions	0.0234^{d}	0.0017	0.0283^{d}	0.0543^{d}	0.0543^{d}	0.0321^{di}	0.0053	0.0359^{di}	0.0543^{di}	0.0590^{di}	0.0185^{d}	0.0014	0.0352^{d}	0.0600^{d}	0.0533^{d}

As expected, the bpref values remain unchanged since this metric already excludes unjudged documents, aligning with this new approach. However, the other metrics (MAP, GM-MAP, P10, P30) consistently improve over the previous results. Notably, the MAP reaches 0.1210—nearly twice the highest MAP when unjudged documents are included. Precision also stands out, reaching 0.1714 for P10 and 0.1562 for P30.

Overall, the scores approach those of the best-submitted runs, particularly for mixed retrieval. While our approaches outperform visual retrieval, they remain below the top textual retrieval submission, as seen in the "Ours subset" section of Table 3.

Since we did not have access to the top submitted runs from the ImageCLEF 2013 competition, it was not possible to re-rank or directly evaluate our models against those exact submissions. Instead, by evaluating only on the official judged pool of ImageCLEF 2013, our leave out unjudged results scenario ensures a consistent and fair comparison, as it uses the same set of documents assessed in the original campaign. This approach mitigates biases from unjudged documents and aligns our evaluation methodology with the official judging process as closely as possible, reinforcing the validity of our comparisons.

Importantly, the analysis of retrieval effectiveness considering only judged documents confirms that the conclusions presented in response to the research questions remain consistent and valid. The main trends observed in model effectiveness and fusion strategies hold true under this evaluation strategy, supporting the robustness of our findings.

6.2 Using expanded relevance judgments dataset

We expanded the ImageCLEFmed 2013 case-based retrieval task relevance judgments (qrels) dataset using an MLLM-as-a-Judge approach (Pires et al., 2025), which used Gemini 1.5 Pro to simulate human assessment, increasing the original qrels from 15,028 to 558,653 relevance judgments. Table 5 presents the overall results for all experiments using the expanded dataset, which we analyze and compare to those from the original qrels. We also report effect sizes, standard errors, and 95% confidence intervals for all statistically significant comparisons in Table 2 under "Expanded Qrels".

Overall, the expanded qrels yield 16 statistically significant results, one more than the original qrels. The findings align with previous observations, reinforcing key observations: CombMAX emerges as the most effective result fusion method in this medical context; different context lengths impact effectiveness, but LongCLIP's larger context does not consistently outperform shorter ones, as seen with CLIP; the domain-specific PubMedCLIP surpasses the general-purpose CLIP for biomedical searches; the Llama model consistently outperforms the CLIP model; and all textual searches using generated descriptions underperform compared to their multimodal baselines.

Compared to the results with the original qrels, some MAP values show an increase, but the improvement is not substantial. Notably, Llama continues to achieve the highest MAP when comparing topic descriptions with article captions. However, the top MAP value is still seen with the original qrels. In contrast, GM-MAP, P10, and P30 consistently show clear improvements, while bpref exhibits a substantial decline across all experiments.

Table 5: Summary of experimental results using expanded relevance judgments dataset. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in "Gen. I. Desc." rows), statistically significant differences relative to the description and image baselines are marked with 'd' and 'i', respectively. Highest scores per column are bolded.

		CLIP CombSUM[Exp. 1]					**CLIP CombMNZ**[Exp. 1]					**CLIP CombMAX**[Exp. 1-5]				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.0283**	**0.0188**	**0.0566**	**0.1371**	**0.1152**	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0179	0.0104	0.0416	0.1114	0.0990	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0249	0.0123	0.0466	0.0943	0.0933	-	-	-	-	-	-	-	-	-	-
	Images	0.0137	0.0080	0.0349	0.0800	0.0733	0.0137	0.0080	0.0349	0.0800	0.0733	0.0165	0.0089	0.0375	0.0714	0.0781
	Captions	0.0185	0.0101	0.0372	0.0914	0.0762	**0.0185**	**0.0101**	**0.0372**	**0.0914**	**0.0762**	0.0257	0.0129	**0.0491***	**0.1171**	**0.0933**
Images	Title	0.0150	0.0077	0.0283	0.0886	0.0705	0.0150	0.0077	0.0283	0.0886	0.0705	**0.0229***	**0.0130**	0.0425*	0.1143	0.0905
	Abstract	0.0098	0.0043	0.0217	0.0571	0.0486	0.0098	0.0043	0.0217	0.0571	0.0486	0.0183*	0.0069	0.0336*	0.0800	0.0752
	Fulltext	0.0100	0.0040	0.0203	0.0457	0.0410	0.0100	0.0040	0.0203	0.0457	0.0410	0.0172	0.0053	0.0304	0.0600	0.0571
	Images	0.0125	0.0070	0.0284	0.0686	0.0610	0.0118*	0.0063	0.0271*	0.0686	0.0610	0.0218	0.0108	0.0417	0.0943	0.0829
	Captions	0.0110	0.0060	0.0233	0.0686	0.0552	0.0110	0.0060	0.0233	0.0686	0.0552	0.0212*	0.0099	0.0396*	0.0800	0.0743
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0021	0.0197^{di}	0.0400^{di}	0.0457^{di}
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0012	0.0165^{di}	0.0229^{d}	0.0238^{di}
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0054^{di}	0.0008	0.0124^{di}	0.0400^{d}	0.0219^{di}
	Images	-	-	-	-	-	-	-	-	-	-	0.0102^{di}	0.0060	0.0250^{di}	0.0400^{i}	0.0410^{di}
	Captions	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0047	0.0203^{di}	0.0314^{d}	0.0410^{di}

		LongCLIP CombMAX[Exp. 2]					**PubMedCLIP CombMAX**[Exp. 3]					**Llama CombMAX**[Exp. 4]				
Description	Title	0.0169	0.0062	0.0382	0.0829	0.0743	**0.0416**	**0.0246**	**0.0687**	**0.1600**	**0.1371**	0.0185	0.0124	0.0423	0.1286	0.0895
	Abstract	0.0203	**0.0120**	0.0421	**0.1200**	0.0867	0.0391*	0.0215	0.0622*	0.1257	0.1181	0.0285*	0.0209	0.0557	0.1514	0.1276
	Fulltext	**0.0255**	0.0111	**0.0475**	0.1200	**0.0990**	0.0255	0.0110	0.0469	0.1200	0.0933	**0.0419**	**0.0293**	**0.0733***	**0.1657**	**0.1448***
	Images	0.0174	0.0082	0.0357	0.0914	0.0800	0.0204	0.0064	0.0467	0.1114	0.0981	-	-	-	-	-
	Captions	0.0189	0.0112	0.0416	0.0886	0.0810	0.0292	0.0161	0.0582	0.1314	0.1229	0.0381	0.0221	0.0694	0.1629	0.1343*
Images	Title	0.0219	0.0098	0.0421	0.1057	0.0838	0.0179	0.0051	0.0294	0.0543	0.0514	-	-	-	-	-
	Abstract	0.0189	0.0068	0.0342	0.0686	0.0695	0.0156	0.0057	0.0275	0.0486	0.0610	-	-	-	-	-
	Fulltext	0.0148	0.0046	0.0303	0.0486	0.0581	0.0129	0.0028	0.0255	0.0629	0.0438	-	-	-	-	-
	Images	0.0204	0.0100	0.0425	0.0714	0.0781	0.0282	0.0146	0.0515	0.1257	0.0914	-	-	-	-	-
	Captions	0.0141*	0.0081	0.0308	0.5710	0.0695	0.0220	0.0102	0.0407	0.1086	0.0819	-	-	-	-	-
Gen. I. Desc.	Title	0.0001^{di}	0.0002	0.0024^{di}	0.0086^{di}	0.0048^{di}	0.0137^{d}	0.0042	0.0296^{d}	0.0714^{d}	0.0638^{d}	0.0021^{d}	0.0007	0.0076^{d}	0.0143^{d}	0.0181^{d}
	Abstract	0.0068^{di}	0.0031	0.0165^{di}	0.0400^{d}	0.0333^{d}	0.0130^{d}	0.0042	0.0253^{d}	0.0714	0.0562^{d}	0.0029^{d}	0.0009	0.0094^{d}	0.0171^{d}	0.0190^{d}
	Fulltext	0.0036^{di}	0.0007	0.0110^{di}	0.0171^{d}	0.0257^{di}	0.0065^{d}	0.0026	0.0188^{d}	0.0400^{d}	0.0343^{d}	0.0025^{d}	0.0005	0.0071^{d}	0.0143^{d}	0.0114^{d}
	Images	0.0116^{di}	0.0058	0.0296^{i}	0.0429^{d}	0.0524^{d}	0.0039^{di}	0.0016	0.0104^{di}	0.0314^{di}	0.0248^{di}	-	-	-	-	-
	Captions	0.0029^{di}	0.0013	0.0116^{di}	0.0371^{d}	0.0286^{di}	0.0068^{di}	0.0037	0.0178^{di}	0.0486^{di}	0.0390^{d}	0.0017^{d}	0.0006	0.0068^{d}	0.0114^{d}	0.0133^{d}

The slight variation in MAP values can be attributed to the nature of the expanded judgments dataset. Despite a significant increase in the number of judgments, approximately 99% were labeled as not relevant. Among all the gathered metrics, MAP and GM-MAP are the most affected by dataset imbalance, as they depend on the ranking of relevant documents across the entire retrieved list. In contrast, precision focuses only on a fixed number of top-ranked results (e.g., top 10 and top 30), making it less sensitive to overall dataset distribution. While bpref is designed to mitigate the impact of missing relevance judgments, it is still influenced by dataset imbalance due to the limited number of relevant documents in the ranked list. Since most of the previously excluded unjudged documents are now considered non-relevant, a decline in metric values was expected.

Compared to the top submission of the ImageCLEFmed 2013 competition, we highlight the precision values listed under the "Ours expanded" section of Table 3, which closely approach the top mixed retrieval results. This indicates that the number of relevant documents retrieved in the top 10 and 30 results is similar to the top submissions.

7 Conclusions

To the best of our knowledge, our paper is the first to apply dense models to multimodal medical case retrieval. Our work investigates the effectiveness of different dense-model approaches in improving multimodal ad hoc search, focusing on retrieving articles relevant to medical differential diagnosis. The findings underscore the limitations of the dataset in a dual-modality search scenario, as incorporating visual data did not enhance retrieval effectiveness. The lack of improvement in retrieval effectiveness was likely due to the physician assessors' preference for textual information. Most documents in the judgment pool were retrieved through textual searches (33 submissions), while far fewer submissions focused on visual (5 submissions) or multimodal (4 submissions) tasks. This aligns with the experimental results, which revealed a clear emphasis on text-based retrieval in the retrieved article. Compounding the issue, no more than about 20% of our retrieved articles across all topics were initially judged, leaving a substantial margin of uncertainty in the evaluation process.

We addressed the challenge of highly incomplete relevance judgments by adapting the evaluation by excluding unjudged documents from our retrieval sets, and using an expanded relevance judgment set that covered all missing judgments across experiments. These strategies led to major improvements, with increases in nearly all metrics, especially precision, bringing scores closer to top submissions. Importantly, the overall findings remain consistent with those obtained using the original qrels, reinforcing their robustness.

To answer **RQ1**: *"Which characteristics of dense models have the greatest impact on retrieval effectiveness in multimodal search systems?"*, the results indicate that various dense model characteristics influence retrieval effectiveness. Context length plays a crucial role, with different lengths offering both advantages and disadvantages depending on the input data. Truncated versions of larger texts were used due to the models' token limits, potentially omitting important information. The Llama 3 model, which has the largest token capacity among the models tested, attained the highest MAP in Experiment 4 (Unimodal vs. Multimodal effectiveness), demonstrating the value of larger context lengths. Additionally, domain-specific models significantly improved retrieval effectiveness over general-purpose models.

To answer **RQ2**: *"How does the effectiveness of dense multimodal models compare to traditional search systems in medical case retrieval, and what factors influence their relative effectiveness?"*, the experiments suggest that dense retrieval holds great potential, particularly for semantic similarity searches across different modalities. However, limitations in context length hinder effectiveness, as multimodal models are often trained on shorter inputs, resulting in lower effectiveness than top submissions. Nonetheless, our approaches excelled in visual retrieval, suggesting that a multimodal large language model, especially if fine-tuned for the medical domain, could greatly enhance effectiveness, though at a high computational cost.

Future work could focus on improving both textual and visual data integration to enhance multimodal medical case retrieval. For textual data, handling large inputs more effectively through text-splitting techniques could enable a more thorough analysis, as Llama, despite its large context length, struggles with longer instances. Splitting fulltexts into logical sections and encoding them separately may improve retrieval effectiveness. Additionally, testing a unified description generator, such as LLaVA or a fine-tuned variant for the medical domain, could help resolve inconsistencies observed in the fifth experiment (Dominant data type approach) due to different generators. On the visual side, exploring medical image modality classification to filter out non-relevant images and compound figure separation to isolate relevant subfigures could reduce retrieval noise. As suggested by Garcia Seco de Herrera et al. (2015), these techniques have the potential to improve case-based retrieval but require further investigation and integration. Finally, integrating the top-performing textual and visual search methods, combining sparse and dense models, could further enhance the accuracy and effectiveness of multimodal search systems. Future studies could also extend experiments to additional datasets, reinforcing the findings and broadening applicability. Additionally, topic-level analysis of system behavior remains important future work to better understand model strengths and limitations across different query topics.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable comments and suggestions, which helped improve this paper. This work was financed by Component 5 - Capitalization and Business Innovation, integrated into the Resilience Dimension of the Recovery and Resilience Plan, within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed under Next Generation EU, for the period 2021–2026, within the project HfPT (reference 41). Sérgio Nunes acknowledges the use of computational resources provided by the project "PTicola – Increasing Computationally Language Resources for Portuguese" (reference `https://doi.org/10.54499/CPCA-IAC/AV/594794/2023`).

References

Javed A. Aslam and Mark H. Montague. Models for metasearch. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages

275–284. ACM, 2001. doi: 10.1145/383952.384007.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.*, 38(8):2939–2970, 2022. doi: 10.1007/S00371-021-02166-7.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47, 2014. doi: 10.1613/JAIR.4135.

Sungbin Choi, Jeongeun Lee, and Jinwook Choi. SNUMedinfo at ImageCLEF 2013: Medical retrieval task. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL `https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-ChoiEt2013.pdf`.

Paul D. Clough and Mark Sanderson. Evaluating the performance of information retrieval systems using test collections. *Inf. Res.*, 18(2), 2013. URL `http://www.informationr.net/ir/18-2/paper582.html`.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114.

Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A. Coburn, Keith T. Wilson, Bennett A. Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review. *Progress in Biomedical Engineering*, 5(2):022001, apr 2023. doi: 10.1088/2516-1091/acc2fe.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *CoRR*, abs/2112.13906, 2021. URL `https://arxiv.org/abs/2112.13906`.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=ogjBpZ8uSi`.

Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.*, 22(3):1341–1360, 2021. doi: 10.1109/TITS.2020.2972974.

Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K. Antani, and Henning Müller. Overview of the ImageCLEF 2013 medical tasks. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors,

Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-SecoDeHerreraEt2013b.pdf.

Alba Garcia Seco de Herrera, Roger Schaer, Dimitrios Markonis, and Henning Müller. Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task. *Comput. Medical Imaging Graph.*, 39:46–54, 2015. doi: 10.1016/J.COMPMEDIMAG.2014.04.004.

Alba Garcia Seco de Herrera, Roger Schaer, and Henning Müller. Shangri-la: A medical case-based retrieval tool. *J. Assoc. Inf. Sci. Technol.*, 68(11):2587–2601, 2017. doi: 10.1002/ASI.23858.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783.

MedGIFT Group. medSearch – medical search engine by HES-SO Valais (medSearch 2009). http://fast.hevs.ch:8080/MedSearch/faces/Search.jsp, 2009.

Ruifeng Guo, Jingxuan Wei, Linzhuang Sun, Bihui Yu, Guiyong Chang, Dawei Liu, Sibo Zhang, Zhengbing Yao, Mingjun Xu, and Liping Bu. A survey on advancements in image-text multimodal models: From general techniques to biomedical implementations. *Comput. Biol. Medicine*, 178:108709, 2024. doi: 10.1016/J.COMPBIOMED.2024.108709.

D. Frank Hsu and Isak Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retr.*, 8(3):449–480, 2005. doi: 10.1007/S10791-005-6994-4.

Morris A. Jette and Tim Wickberg. Architecture of the slurm workload manager. In Dalibor Klusácek, Julita Corbalán, and Gonzalo P. Rodrigo, editors, *Job Scheduling Strategies for Parallel Processing - 26th Workshop, JSSPP 2023, St. Petersburg, FL, USA, May 19, 2023, Revised Selected Papers*, volume 14283 of *Lecture Notes in Computer Science*, pages 3–23. Springer, 2023. doi: 10.1007/978-3-031-43943-8_1.

Qiao Jin, Robert Leaman, and Zhiyong Lu. PubMed and beyond: Biomedical literature search in the age of artificial intelligence. *eBioMedicine*, 100:104988, 2024. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2024.104988.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at ImageCLEF 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2014.03.004.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review

of deep learning-based information fusion techniques for multimodal medical image classification. *Comput. Biol. Medicine*, 177:108635, 2024. doi: 10.1016/J.COMPBIOMED.2024.108635.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Zhiyong Lu, Won Kim, and W. John Wilbur. Evaluation of query expansion using MeSH in PubMed. *Inf. Retr.*, 12(1):69–80, 2009. doi: 10.1007/S10791-008-9074-8.

Mark H. Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 538–548. ACM, 2002. doi: 10.1145/584792.584881.

André Mourão and Flávio Martins. NovaMedSearch: A multimodal search engine for medical case-based retrieval. In João Ferreira, João Magalhães, and Pável Calado, editors, *Open research Areas in Information Retrieval, OAIR 2013, Lisbon, Portugal, May 15-17, 2013*, pages 223–224. ACM, 2013. URL http://dl.acm.org/citation.cfm?id=2491798.

André Mourão, Flávio Martins, and João Magalhães. NovaSearch on medical ImageCLEF 2013. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-MouraoEt2013.pdf.

Henning Müller and Jayashree Kalpathy-Cramer. The ImageCLEF medical retrieval task at ICPR 2010 - information fusion. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pages 3284–3287. IEEE Computer Society, 2010. doi: 10.1109/ICPR.2010.803.

Catarina Pires, Sérgio Nunes, and Luís Filipe Teixeira. Expanding relevance judgments for medical case-based retrieval task with multimodal LLMs. *CoRR*, abs/2506.17782, 2025. doi: 10.48550/ARXIV.2506.17782. Presented at the Third Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2025), co-located with SIGIR 2025, Padua, Italy, July 17, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume

139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 105–108. National Institute of Standards and Technology (NIST), 1994. URL http://trec.nist.gov/pubs/trec3/papers/vt.ps.gz.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William R. Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. Clinical information retrieval: A literature review. *J. Heal. Informatics Res.*, 8(2):313–352, 2024. doi: 10.1007/S41666-024-00159-4.

Konstantinos Zagoris, Savvas A. Chatzichristofis, Nikos Papamarkos, and Yiannis S. Boutalis. img(anaktisi): A web content based image retrieval system. In Tomás Skopal and Pavel Zezula, editors, *Second International Workshop on Similarity Search and Applications, SISAP 2009, 29-30 August 2009, Prague, Czech Republic*, pages 154–155. IEEE Computer Society, 2009. doi: 10.1109/SISAP.2009.15.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, volume 15109 of *Lecture Notes in Computer Science*, pages 310–325. Springer, 2024. doi: 10.1007/978-3-031-72983-6_18.

Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.*, 105:104042, 2021. doi: 10.1016/J.IMAVIS.2020.104042.