VisR-Bench: A Visual Retrieval Benchmark for Visually-Rich Documents

Jian Chen[1], **Ruiyi Zhang**[2], **Ming Li**[3], **Shijie Zhou**[1], **Changyou Chen**[1],
[1]University at Buffalo [2]Adobe Research [3]University of Maryland
jchen378@@buffalo.edu, ruizhang@adobe.com

Abstract

Retrieval is essential for multimodal large language models (MLLMs) to handle long contexts and improve factual accuracy. However, existing benchmarks focus on end-to-end answer generation, making retrieval evaluation difficult. To address this, we introduce VisR-Bench, a benchmark for question-driven retrieval in scanned documents. Our queries do not explicitly contain answers, preventing models from relying on keyword matching. Additionally, they avoid ambiguous references to figures or tables by ensuring that each query includes descriptive information necessary to locate the correct content. The dataset spans English and 15 other languages, with English queries enabling fine-grained evaluation across answer modalities (tables, text, figures) and non-English queries focus on multilingual generalization. VisR-Bench provides a comprehensive framework for evaluating retrieval in document understanding.

1 Introduction

The performance of a multimodal retrieval module is critical to ensuring the factual accuracy and efficiency of Retrieval-Augmented Generation (RAG) systems powered by multimodal large language models (MLLMs). It determines the quality of retrieved information from external knowledge bases or long-context before generating a response. Unlike traditional text-based search, aligning natural language queries with multimodal data (e.g., magazines, posters, books) presents unique challenges, including interpreting diverse structured content (e.g., tables, catalogs, figures) and navigating complex document layouts. To address these challenges, several MLLM-based retrieval models have been developed, highlighting the need for a benchmark that systematically evaluates question-driven multimodal retrieval in real-world scenarios.

Existing retrieval datasets fall short in assessing the challenges of MLLM-based RAG systems. An effective benchmark should be question-driven, requiring retrieval models to locate information that is not explicitly stated in the query and perform logical reasoning. For instance, a query such as "When does the first train leave in the morning?" does not explicitly contain the answer; instead, the relevant information is found in a train schedule table, rather than an image of a train. Classic datasets used to train small multimodal encoders primarily focus on text-image similarity rather than QA relevance, making them inadequate for complex retrieval tasks. While VQA datasets contain questions, they often assume the model is provided with the correct input image, and many QA pairs are not designed for retrieval (e.g., "What is the page number of the given page?").

Another key challenge is multilingual retrieval, which remains underexplored. Existing multilingual benchmarks and datasets primarily focus on text-only documents and text-generation tasks, offering limited insights into multimodal retrieval. This highlights the need for benchmarks that assess retrieval performance.

In this paper, we propose VisR-Bench, a question-driven retrieval benchmark designed to evaluate multimodal retrieval performance in visually rich document images. The benchmark encompasses high-quality synthetic QA pairs that are suitable for retrieval tasks. By generating QA pairs for different evidence types including tables, figures, and visual text, our benchmark enables granular performance analysis in multimodal, OCR, and table understanding, addressing the diverse challenges of multimodal retrieval in real-world scenarios. Additionally, our dataset incorporates multilingual documents, allowing for cross-lingual retrieval evaluation. Additionally, our dataset incorporates multilingual documents across 15 languages other than English, exposing language-specific weaknesses in existing retrievers.

2 VisR-Bench

The VisR-Bench dataset is divided into an English split (English only) curated from web crawled data and a multilingual split filtered from the CCpdf dataset (Turski et al., 2023). Figure 1 and Figure 2 presents the document length and number distributions, highlighting greater diversity than previous benchmarks (Tanaka et al., 2023; Islam et al., 2023; Ma et al., 2024b). The blue colors represent the English multimodal split, which can further be categorized into 10 types. Other colors represent the multilingual multimodal split, containing documents in 15 non-English languages.

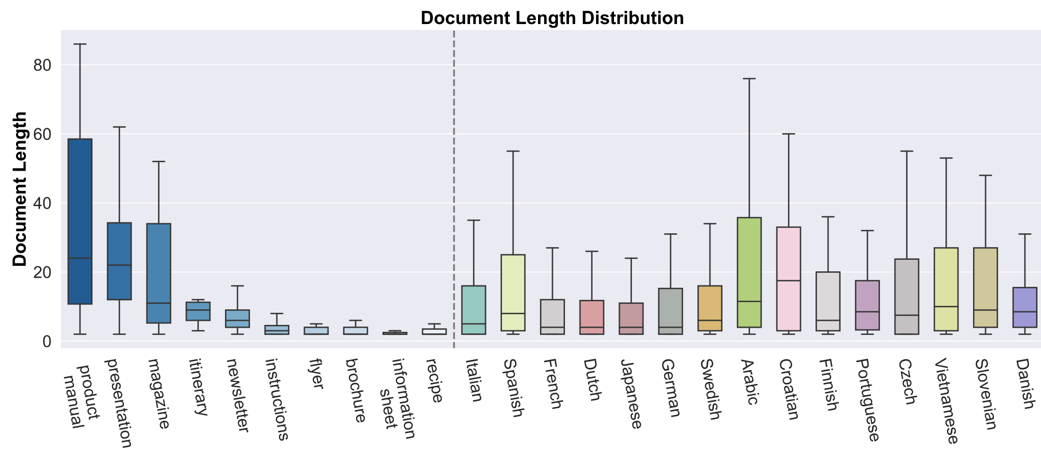

Figure 1: Distribution of average document lengths for the English split (left of the dashed line) and the multilingual split (right of the dashed line).

2.1 English split

To construct the English split, we crawled 4,000 PDF documents and extracted their contents using a document parser[1]. The parser outputs text and tables as Markdown files while saving figures separately as images. To ensure a focus on multimodal content, we retain only English documents that contain both Markdown files and figures and exclude single-page documents, as retrieval is unnecessary for them. After curation, the English split is refined to 387 unique documents. All documents have been validated by human reviewers to ensure the exclusion of harmful content and personally identifiable information (PII). Additionally, we confirm that each document's license and usage terms explicitly permit its use for research purposes.

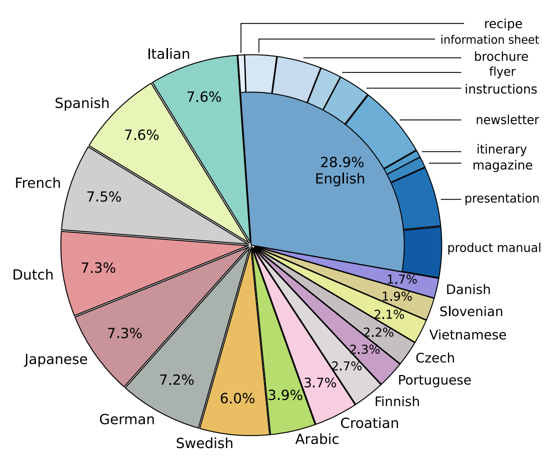

Figure 2: Distribution of language and document types in VisR-Bench.

Figure-related QA To select documents with informative figures, we apply figure classification on the extracted images using the CLIP model `ViT-L/14-336` (Radford et al., 2021). Each figure is classified into one of 19 predefined categories, and we retain 6 relevant types while discarding decorative figures such as logos and banners. We

[1]Adobe Extract API: https://developer.adobe.com/document-services/apis/pdf-extract/

combine figures with their corresponding contexts and use GPT-4o (API version 2024-02-15-preview) to generate QA pairs. For prompt construction, we provide two demonstrations and instruct GPT-4o to generate a new QA pair. To ensure that figures are necessary for answering the questions, we apply a heuristic filtering step: we discard any question that GPT-4o can already answer using only the textual information extracted from the Markdown files. As a result, all remaining questions require both figures and text from the document for accurate answering. This filtering process not only ensures the necessity of figures but also serves as an additional validation step for the correctness of the generated answers. In contrast, most existing benchmarks primarily contain questions that can be answered using extracted text alone, making them less effective for evaluating multimodal retrieval and reasoning.

Text-based QA To generate text-based QA pairs, we first filter pages that contain only text in the extracted Markdown files, excluding those with tables or figures to ensure a sole focus on textual information. We then use GPT-4o to generate QA pairs over the given page. We design a system prompt to enforce key constraints: (1) Questions must simulate a realistic retrieval scenario where a user queries a multi-page document for relevant information. (2) Answers must be explicitly present in the text to prevent hallucination. (3) Questions should not be ambiguous or overly broad, such as asking for the page number or requiring document-level summarization. (4) If a page lacks sufficient content for meaningful questions, the model returns an empty string instead of generating forced or unnatural queries.

Table-related QA Similar to text-based QA, we extract pages that contain tables but no figures to ensure that the generated questions are not influenced by visual elements. This guarantees that the QA pairs focus solely on tabular data and its text context. In addition to the constraints applied to text-based QA, table-related questions are designed to require computation or logical inference rather than simple fact lookup. Instead of directly extracting a single value, the questions encourage tasks such as analyzing trends, making comparisons, identifying rankings, or interpreting correlations within the table data. This ensures that retrieval models must engage in structured reasoning.

2.2 Multilingual Split

Our dataset includes multilingual queries over documents in 15 languages, including Spanish, Italian, German, French, Dutch, Arabic, Croatian, Japanese, Swedish, Vietnamese, Portuguese, Finnish, Czech, Slovenian, and Danish. This subset is designed to evaluate retriever accuracy across a diverse linguistic landscape. The queries are general questions generated by GPT-4o, conditioned on text, tables, and figures, without necessarily incorporating all modalities in each instance.

3 Related Work

Text-based Retrieval Methods Traditional text-based retrieval methods extract text from images using OCR tools (Du et al., 2020; Singh et al., 2021) and apply text-based search techniques. BM25 (Robertson et al., 2009) is a statistical algorithm based on text frequency. Deep learning models such as SBERT (Reimers, 2019) and BGE Models (Chen et al., 2024; Xiao et al., 2023) enables semantically aware search. NV-Embed (Lee et al., 2024), built upon LLMs (Mistral 7B (Jiang et al., 2023)), generates text embedding to enhance retrieval accuracy by capturing richer contextual information. However, these approaches struggle with complex layouts and cannot process visual elements, limiting their performance in real-world applications.

Multimodal Retrieval Methods Multi-modal encoders like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) can be used for image retrieval by similarity in a shared embedding space, but they are optimized for natural images rather than document pages. With the advent of MLLMs, recent approaches customize MLLMs as encoders, leveraging their pre-trained knowledge for improved accuracy. VLM2Vec (Jiang et al., 2024) and GME (Zhang et al., 2024) compute similarity using single-vector embeddings, while ColPali (Faysse et al., 2024), ColPhi (Chen et al., 2025), and ColInternVL2 (Chen et al., 2025) utilize sequences of hidden states and apply sequence interaction scoring (Khattab & Zaharia, 2020) for more effective relevance estimation.

Comapre to Multi-page Datasets Existing multi-page document datasets focus on domain-specific documents, such as SlideVQA (Tanaka et al., 2023), SciMMIR (Wu et al., 2024), and MMVQA (Ding et al., 2024). Wiki-SS (Ma et al., 2024a) emphasizes text-based evidence. DocMatix (Dong et al., 2025) contains noisy and ambiguous queries, and CVQA (Romero et al., 2024) is limited to single natural images, making it unsuitable for document retrieval. Additionally, MMLongBench-Doc (Ma et al., 2024b), MMDocIR (Dong et al., 2025), and M-LongDoc (Chia et al., 2024) are English-only, limiting multilingual applicability.

4 EXPERIMENT

We evaluate 13 retrieval methods on the multimodal and multilingual splits, with results presented in Table 1 and Table A.1. Retrieval methods are categorized into (1) text-based methods, (2) small-encoder models, and (3) large-language model encoders. MLLM-based methods outperform all others, demonstrating their advantage in end-to-end document understanding. VisRAG and VLM2Vec perform poorly, as they are optimized for natural images rather than document understanding. Encoders perform on par with the best text-based methods for figure-based tasks in the multimodal split, highlighting their strength in image encoding, while text-based methods generally outperform encoders in all other tasks. Additionally, figure and table-related tasks prove more challenging than text-based tasks, as evidenced by consistently lower performance across all methods.

	Figure		**Table**		**Text**		**Average**	
Accuracy	top1	top5	top1	top5	top1	top5	top1	top5
Text-based Methods								
BM25 (Chen et al., 2024)	24.27	45.63	38.58	66.43	64.72	89.10	53.18	78.75
SBERT (Reimers, 2019)	25.24	49.27	26.31	52.68	49.96	76.97	36.80	61.94
BGE-large (Xiao et al., 2023)	31.55	56.07	40.36	70.14	57.00	82.68	43.81	68.55
BGE-M3 (Chen et al., 2024)	31.07	56.80	51.11	78.51	67.70	89.89	52.03	74.17
NV-Embed-v2 (Lee et al., 2024)	35.44	65.05	44.04	73.34	61.38	87.46	46.65	71.70
Multimodal Encoders								
CLIP (Radford et al., 2021)	33.90	61.74	24.68	47.59	39.47	70.21	33.54	61.14
SigLip (Zhai et al., 2023)	38.98	69.73	24.73	53.22	39.06	70.97	33.52	64.07
Multimodal Large Language Models								
VisRAG (Yu et al., 2024)	31.96	66.83	19.82	48.53	31.00	61.49	26.72	56.70
VLM2Vec (Jiang et al., 2024)	40.44	76.27	28.51	57.77	39.90	71.69	35.53	66.49
GME (Zhang et al., 2024)	68.04	91.53	61.50	86.38	76.34	95.62	70.28	91.89
Col-InternVL2 (Chen et al., 2025)	68.28	90.31	63.85	86.36	79.19	96.45	72.84	92.31
Col-Phi (Chen et al., 2025)	**68.77**	**93.22**	65.65	**88.51**	81.67	**97.04**	74.98	**93.60**
ColPali-v1.2 (Faysse et al., 2024)	**68.77**	91.77	**66.12**	88.26	**82.63**	96.89	**75.71**	93.36

Table 1: Retrieval accuracy results on VisR-Bench (English split). Bold font indicates the best overall performance for each language.

5 ANALYSIS

Finding 1. Contextualized Late Interaction outperformed Vector Similarity. The superior performance of multi-vector embedding models, ColPali, ColPhi, and ColInternVL2, over the single-vector embedding model GME highlights the advantage of contextualized late interaction, potentially due to finer-grained representation modeling.

Finding 2. LLM-Based Methods and Encoders Are Undertrained. Our evaluation shows that NV-Embed-v2, is outperformed by BM25 on text-based tasks, suggesting a higher potential upper bound for LLM-based model. CLIP and SigLIP struggle on multilingual documents, likely due to insufficient training data in non-English languages, which also limits MLLMs, where they serve as visual encoders.

Finding 3. MLLM Models Benefit from Visual input and Language Modeling Ability. Our evaluation shows that MLLM-based methods consistently outperform text-based models and encoder-based methods, even on text-based tasks. Their ability to leverage both visual layout and language modeling makes them more effective than traditional text-based pipelines.

Finding 4. Arabic Understanding Requires Architectural Modifications All methods perform poorly on Arabic documents, likely due to its right-to-left reading order, requires dynamic design in attention masks and position embeddings.

6 CONCLUSION

We introduce VisR-Bench, a benchmark for question-driven retrieval over multipage documents. Results highlight domain limitations in MLLMs trained on natural vs. document images and the challenges of Arabic retrieval due to its right-to-left reading order.

References

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, and Tong Sun. SV-RAG: LoRA-contextualizing adaptation of MLLMs for long document understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=FDaHjwInXO`.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*, 2024.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3–9, 2024.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*, 2025.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024a.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8802–8812, 2021.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *International Conference on Document Analysis and Recognition*, pp. 348–365. Springer, 2023.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv preprint arXiv:2401.13478*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and N Muennighof. C-pack: packaged resources to advance general chinese embedding. 2023. *arXiv preprint arXiv:2309.07597*, 2023.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

A Multilingual Retrieval Results

	Spanish		Italian		German		French		Dutch	
	top1	top5	top1	top5	top1	top5	top1	top5	top1	top5
Text-based Methods										
BM25	60.25	82.50	59.14	82.02	65.82	86.92	54.07	77.79	59.83	84.88
SBERT	22.77	41.83	21.82	41.12	25.74	48.54	27.43	51.33	27.99	52.25
BGE-large	34.55	60.41	30.27	56.24	39.75	66.82	41.34	67.42	39.14	67.53
BGE-M3	58.16	83.13	52.94	77.96	67.64	88.94	60.68	82.10	63.62	87.73
NV-Embed-v2	42.92	72.71	40.84	66.32	52.23	80.30	49.41	76.13	47.12	78.74
Multimodal Encoders										
CLIP	11.14	29.32	12.39	31.77	19.53	45.69	19.52	44.44	16.22	42.71
SigLIP	13.08	32.36	17.52	40.69	25.69	51.69	24.85	53.15	22.70	50.85
Multimodal Large Language Models										
VisRAG	9.70	28.48	10.69	33.09	14.48	40.22	16.37	42.55	15.22	42.02
VLM2Vec	18.59	44.48	19.42	43.84	26.07	56.10	29.53	60.50	22.51	52.97
GME	60.57	88.08	52.96	79.08	65.97	89.61	66.78	89.55	57.92	85.16
ColInternVL2	58.26	84.57	51.89	77.96	60.35	86.32	64.06	87.17	58.27	84.60
Col-Phi-3-V	65.42	89.00	56.06	81.43	65.02	88.96	67.83	89.65	62.15	88.17
ColPali	71.44	92.62	62.02	85.81	72.96	92.48	72.62	92.09	65.15	89.73

	Arabic		Croatian		Japanese		Swedish		Vietnamese	
Text-based Methods										
BM25	7.43	21.49	52.98	72.71	11.59	38.60	57.44	83.68	48.81	73.01
SBERT	4.02	17.29	17.72	36.67	13.06	41.24	28.26	60.99	17.94	37.07
BGE-large	6.15	19.53	32.67	58.14	31.92	64.97	42.18	74.69	23.94	48.97
BGE-M3	10.55	26.26	59.07	81.46	58.38	84.33	65.25	89.33	44.93	68.82
NV-Embed-v2	5.47	21.73	41.86	68.30	42.17	72.70	53.02	81.40	25.75	60.24
Multimodal Encoders										
CLIP	4.64	18.91	10.46	27.36	14.28	44.86	17.38	48.84	6.67	22.13
SigLIP	5.53	19.56	13.98	33.56	15.62	46.20	26.78	61.66	8.38	25.13
Multimodal Large Language Models										
VisRAG	4.78	19.80	6.38	22.25	21.04	52.37	14.93	49.30	5.53	18.56
VLM2Vec	7.39	24.10	12.31	32.04	19.19	50.02	25.98	62.17	8.22	25.39
GME	15.33	35.72	45.09	72.60	61.11	89.37	59.09	89.79	26.22	51.81
ColInternVL2	5.09	17.50	47.68	73.16	39.65	71.57	61.16	90.51	25.75	54.19
Col-Phi-3-V	8.46	25.95	48.83	74.82	25.28	56.49	64.19	93.08	34.28	65.25
ColPali	14.33	32.59	51.54	76.94	43.85	77.53	65.37	92.11	35.32	66.60

	Portuguese		Finnish		Czech		Slovenian		Danish	
Text-based Methods										
BM25	61.47	79.92	50.11	71.24	66.11	89.34	56.45	81.81	54.38	82.35
SBERT	25.85	50.24	23.34	47.29	26.28	50.00	22.31	48.03	29.56	58.11
BGE-large	38.53	66.08	31.58	57.97	33.97	61.94	35.30	63.89	35.58	71.02
BGE-M3	60.07	82.16	56.90	77.19	65.87	90.22	65.05	88.53	64.42	88.38
NV-Embed-v2	56.98	80.34	34.32	61.63	41.99	70.59	43.91	73.21	52.94	79.48
Multimodal Encoders										
CLIP	16.75	42.17	12.13	36.84	11.86	34.78	13.35	36.29	13.77	45.48
SigLIP	25.30	51.03	17.24	45.84	20.67	47.92	17.03	43.28	23.53	54.38
Multimodal Large Language Models										
VisRAG	12.68	38.29	9.76	34.78	9.46	34.13	8.69	34.50	13.77	46.34
VLM2Vec	23.73	53.46	16.17	44.47	21.07	50.56	15.59	44.09	25.39	58.68
GME	65.29	90.72	38.83	68.80	51.52	82.29	51.34	80.91	54.52	86.94
ColInternVL2	62.32	86.95	46.22	72.85	55.29	85.90	54.75	83.96	60.11	90.10
Col-Phi-3-V	64.99	88.59	49.73	75.82	58.65	88.86	56.81	85.66	61.12	90.67
ColPali	76.03	92.96	42.11	73.07	62.34	91.27	55.82	86.29	62.41	90.10

Table A.1: Retrieval accuracy results for LoCAL-B. Bold font indicates the best model.