# RETRIEVAL OR REPRESENTATION? REASSESSING BENCHMARK GAPS IN MULTILINGUAL AND VISUALLY RICH RAG

**Martin Asenov**
Parexel AI Labs
London, United Kingdom
martin.asenov@parexel.com

**Kenza Benkirane**
Parexel AI Labs
London, United Kingdom
kenza.benkirane@parexel.com

**Dan Goldwater**
Parexel AI Labs
London, United Kingdom
dan.goldwater@parexel.com

**Aneiss Ghodsi**
Parexel AI Labs
San Francisco, United-States
aneiss.ghodsi@parexel.com

## ABSTRACT

Retrieval-augmented generation (RAG) is a common way to ground language models in external documents and up-to-date information. Classical retrieval systems relied on lexical methods such as BM25, which rank documents by term overlap with corpus-level weighting. End-to-end multimodal retrievers trained on large query-document datasets claim substantial improvements over these approaches, especially for multilingual documents with complex visual layouts. We demonstrate that better document representation is the primary driver of benchmark improvements. By systematically varying transcription and preprocessing methods while holding the retrieval mechanism fixed, we demonstrate that BM25 can recover large gaps on multilingual and visual benchmarks. Our findings call for decomposed evaluation benchmarks that separately measure transcription and retrieval capabilities, enabling the field to correctly attribute progress and focus effort where it matters.

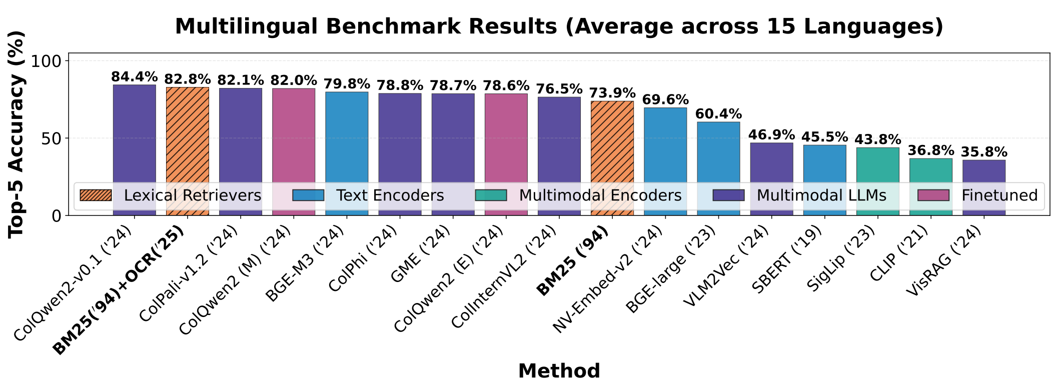



Figure 1: **Multilingual benchmark results across 15 languages.** Average Top-5 retrieval accuracy for different methods. Methods are sorted by performance (highest to lowest). Lexical retrievers (BM25) are shown with diagonal hatching. BM25+OCR indexes text produced by state-of-the-art OCR models and preprocessing techniques per different languages. Release years are shown in parentheses.

# 1 INTRODUCTION

Despite recent progress, retrieval in multilingual and visually rich settings remains challenging for modern systems. Multilingual benchmarks and training corpora are heavily skewed toward high-resource languages, leading to persistent performance gaps and the need for per-language optimizations Chirkova et al. (2024); Ranaldi et al. (2025); Li et al. (2025). Many real-world documents interleave running text with figures, tables, and complex layouts, introducing additional challenges for document retrieval Tanaka et al. (2025). Recent work has increasingly emphasized specialized retrievers, including dense text embeddings Xiao et al. (2023); Chen et al. (2024b), multimodal representations Radford et al. (2021); Zhai et al. (2023); Faysse et al. (2024), or layout-aware models Xu (2020; 2021); Huang (2022), to address the perceived shortcomings of classical text retrieval methods such as BM25 Robertson & Walker (1994). Empirical results on multimodal, visually rich benchmarks such as VisR-Bench Chen et al. (2025) appear to support this trend, showing large performance gaps between sparse text retrievers and modern multimodal approaches.

To examine the impact of optical character recognition (OCR) and text preprocessing, we extend transcription data in VisR-Bench Chen et al. (2025) with three additional OCR models and implement language-specific preprocessing options, including stemming, lemmatization, and morphological analysis. Our results show that improving transcription quality and the resulting text representations leads to significantly better downstream retrieval performance - recovering up to +8.9 Top-5 points for BM25 Robertson & Walker (1994) on average for multilingual datasets by improving transcription quality and normalization. On figure-heavy pages, we observe a distinct failure mode: when figures lack any textual or semantic description, retrieval degrades sharply, whereas even coarse descriptions recover much of the loss, yielding gains of up to +31.1 Top-5 points. This raises a basic question: *are we evaluating retrieval methods, or the pre-processing pipelines?* We show that addressing these sources of error alone allows classical methods such as BM25 to recover a large fraction of the apparent gap.

# 2 RELATED WORK AND PRELIMINARIES

**Multilingual retrieval** performance remains uneven because benchmarks and training data are skewed toward high-resource languages, leaving persistent gaps in low-resource and morphologically rich settings Chirkova et al. (2024); Ranaldi et al. (2025); Li et al. (2025). Recent work therefore emphasizes multilingual dense retrievers and embedding models to improve semantic matching across languages Xiao et al. (2023); Chen et al. (2024b). Any text retriever is fundamentally bounded by the indexed representation, which depends on OCR quality and language-specific preprocessing such as tokenization, stemming, and morphological normalization. Our work complements model-centric studies by showing that much of the observed multilingual gap can be driven by these representation choices rather than retrieval alone.

**Visually rich documents** interleave text with figures and complex layouts, motivating layout-aware encoders and multimodal retrievers that bypass brittle text extraction Tanaka et al. (2025); Xu (2020; 2021); Huang (2022); Radford et al. (2021); Zhai et al. (2023); Faysse et al. (2024). Benchmarks such as VisR-Bench Chen et al. (2025) report large gains for multimodal methods over classical sparse retrieval methods such as BM25 Robertson & Walker (1994). A key caveat is that these comparisons often confound retrieval with upstream transcription quality: if figure content is poorly transcribed, it is effectively missing from the text index. We address this by holding the retriever fixed and varying only OCR/transcription, isolating how improved extraction (and lightweight figure descriptions) can substantially narrow the apparent gap.

# 3 EXPERIMENTS

## 3.1 EXPERIMENTAL SETUP

**Benchmark and metrics** Our goal is to separate retrieval behavior from upstream transcription and normalization choices. We run controlled experiments where we keep the retriever and evaluation protocol fixed while varying only (i) the OCR/transcription used to build the page index, and (ii) language-specific text processing for multilingual retrieval. We evaluate on VisR-Bench Chen et al.

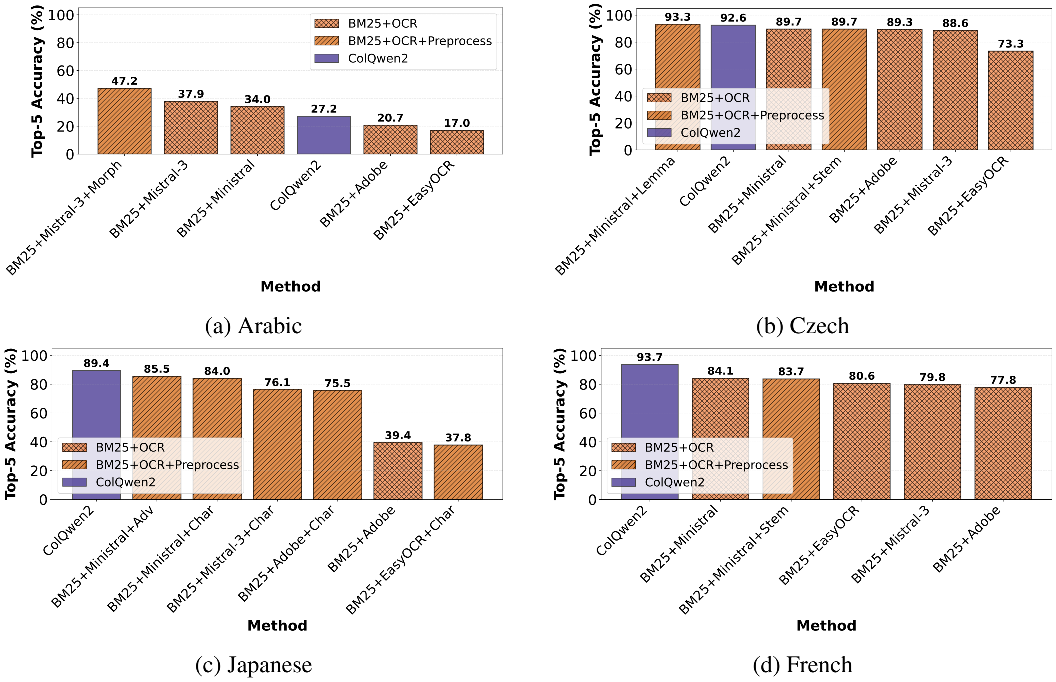



Figure 2: **Language-specific BM25 optimization.** Top-5 retrieval accuracy for BM25 under different OCR/transcription pipelines (Adobe, EasyOCR, Ministral 3B, Mistral OCR 3) and language-specific text processing (lemmatization, stemming, morphological analysis, and segmentation).

(2025), a benchmark for retrieval-augmented question answering over long, visually rich documents. Each example contains a document, a query, and a ground-truth evidence page. The task is *page-level retrieval*: return the evidence page in the Top-$K$ retrieved pages. We report results (i) across 15 languages in Figure 1, and (ii) across visually specific questions in FIgure 4.

**Retrievers** We compare three classes of retrieval methods: (i) sparse lexical retrieval using BM25 Robertson & Walker (1994); (ii) dense text retrievers including SBERT Reimers & Gurevych (2019), BGE-large Xiao et al. (2023), BGE-M3 Chen et al. (2024b), and NV-Embed-v2 Lee et al. (2024); and (iii) multimodal retrievers including CLIP Radford et al. (2021), SigLip Zhai et al. (2023), VisRAG Yu et al. (2024), VLM2Vec Jiang et al. (2024), GME Zhang et al. (2024), and Col* methods Chen et al. (2024a); Faysse et al. (2024). We report baseline results from Chen et al. (2025) for multimodal models, while our evaluation focuses on varying OCR and preprocessing methods. Text-based retrievers index OCR transcriptions of each page, while multimodal methods encode page images directly.

**OCR models** A central variable in our study is transcription quality. We compare the dataset-provided extraction against alternative OCR pipelines:

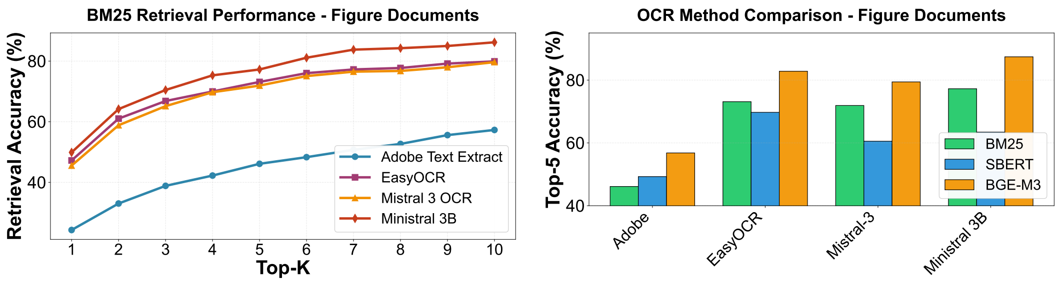
- **Adobe Document Extract**: default parser in the dataset Adobe Developer (2026)
- **EasyOCR**: open-source OCR applied to rendered page images JaidedAI (2020).
- **Mistral OCR 3**: a modern OCR system applied to page images Mistral (2025b).
- **Ministral 3B (VLM transcription)**: a small VLM Mistral (2025a) with the prompt: *"Give me a markdown of what you see in the image. Reply only with the markdown content."*

Adobe Document Extract and EasyOCR perform text-only extraction, while Mistral OCR 3 annotates figures and tables. Ministral 3B is applied per image with the specified prompt.

**Language-Specific Text Processing** For multilingual BM25, we evaluate a set of language-specific preprocessing strategies that directly affect lexical matching. We consider stemming for Romance and Germanic languages using Snowball stemmers Porter (2001); lemmatization for highly inflected languages such as Czech, Slovenian, Croatian, and Finnish using spaCy models Honnibal et al.

(2020); full morphological analysis for Arabic using CAMeL Tools Obeid et al. (2020), which decomposes words into prefixes, stems, and suffixes; and script-aware word segmentation for Japanese with Sudachi Takaoka et al. (2018) and for Vietnamese with pyvi Tran (2021). As a reference, we include a minimal-processing baseline that applies only lowercasing and NLTK tokenization Bird (2006). For each language, we select the best-performing configuration based on Top-5 accuracy and report the chosen setup in Figure 1, with configuration sensitivity illustrated for representative languages in Figure 2.

### 3.2 Results




(a) Figure Documents (b) Figure Documents

Figure 3: **OCR impact on text retrieval methods.** a) BM25 retrieval performance for different Top-K values b) Impact on different retrieval for different combinations of transcription OCR models and text retrieval models.

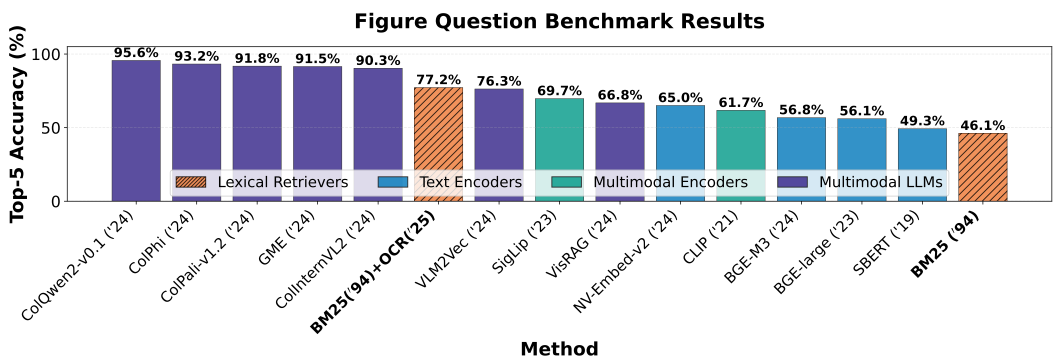



Figure 4: **Figure-heavy focused QA benchmark results.** Average Top-5 retrieval accuracy for different methods. Methods are sorted by performance (highest to lowest). Lexical retrievers (BM25) are shown with diagonal hatching. BM25+OCR uses a small visual language model Mistral (2025a) for transcription. Release years are shown in parentheses.

**Multilingual** Across multilingual settings, we observe that retrieval performance is strongly shaped by transcription and preprocessing choices (Figure 1). Modern OCR substantially improves baseline accuracy, and languages with complex morphology or segmentation (e.g., Japanese, Arabic) benefit disproportionately from appropriate tokenization and morphological processing (Figure 2).

**Visually Rich** For figure-heavy pages, transcribing visual content and adding semantic descriptions recovers most of the performance gap (Figure 4). We observe a two-stage progression: gains from improved transcription fidelity, followed by additional improvements from semantic figure descriptions (Figure 3a). These improvements benefit not only BM25 but also dense text retrievers such as SBERT and BGE (Figure 3b).

**Across Both Settings** Across both multilingual and visually rich retrieval, these results indicate that missing or noisy transcription is a dominant failure mode (Figure 1, Figure 4). Retrieval outcomes are therefore strongly shaped by upstream extraction and processing.

## 4 Conclusion

We revisit the narrative that lexical retrieval underperforms on visually rich and multilingual benchmarks due to inadequate text matching. We show that OCR quality is a critical bottleneck: improving transcription alone recovers most of the performance gap previously attributed to retrieval limitations. These findings call for treating OCR as a first-class component of document retrieval systems.

## References

Adobe Developer. Pdf extract api – overview. https://developer.adobe.com/document-services/docs/overview/pdf-extract-api/, 2026. Accessed: 2026-01-05.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pp. 69–72, 2006.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, and Tong Sun. Lora-contextualizing adaptation of large multimodal models for long document understanding. *arXiv preprint arXiv:2411.01106*, 2024a. doi: 10.48550/arXiv.2411.01106. URL https://arxiv.org/abs/2411.01106.

Jian Chen, Ming Li, Jihyung Kil, Chenguang Wang, Tong Yu, Ryan Rossi, Tianyi Zhou, Changyou Chen, and Ruiyi Zhang. Visr-bench: An empirical study on visual retrieval-augmented generation for multilingual long document understanding. *arXiv preprint arXiv:2508.07493*, 2025.

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*, 2024b. doi: 10.48550/arXiv.2402.03216. URL https://arxiv.org/abs/2402.03216.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. Retrieval-augmented generation in multilingual settings. *arXiv preprint arXiv:2407.01463*, 2024.

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. doi: 10.48550/arXiv.2407.01449. URL https://arxiv.org/abs/2407.01449.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

Yupan et al. Huang. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM MM*, 2022.

JaidedAI. Easyocr. https://github.com/JaidedAI/EasyOCR, 2020.

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. doi: 10.48550/arXiv.2410.05160. URL https://arxiv.org/abs/2410.05160.

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. doi: 10.48550/arXiv.2405.17428. URL https://arxiv.org/abs/2405.17428.

Bo Li, Zhenghua Xu, and Rui Xie. Language drift in multilingual retrieval-augmented generation: Characterization and decoding-time mitigation. *arXiv preprint arXiv:2511.09984*, 2025.

Mistral. Ministral 3 3b. https://docs.mistral.ai/models/ministral-3-3b-25-12, 2025a.

Mistral. Mistral ocr 3. https://mistral.ai/news/mistral-ocr-3, 2025b.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 7022–7032, 2020.

Martin F Porter. Snowball: A language for stemming algorithms, 2001.

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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://arxiv.org/abs/2103.00020.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1410/.

Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pp. 232–241. Springer, 1994.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: A japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24827–24837, 2025.

Viet-Trung Tran. pyvi: Python vietnamese toolkit, 2021. URL https://pypi.org/project/pyvi/. Python package for Vietnamese tokenization and POS tagging.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023. doi: 10.48550/arXiv.2309.07597. URL https://arxiv.org/abs/2309.07597.

Yiheng et al. Xu. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*, 2020.

Yiheng et al. Xu. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, 2021.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. doi: 10.48550/arXiv.2410.10594. URL https://arxiv.org/abs/2410.10594.

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 (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

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. doi: 10.48550/arXiv.2412.16855. URL https://arxiv.org/abs/2412.16855.

# A Usage of LLM models disclosure

Large language models were used for proofreading and rephrasing individual sentences of the manuscript for clarity and conciseness. All claims and results were independently produced and validated by the authors.

# B Multilingual Retrieval: OCR vs. Preprocessing

This section comments on multilingual patterns that are visible in the full ablations and per-language config summaries (Table 3, Table 2) and the full multilingual leaderboard (Table 1), but are not discussed in the main text.

## B.1 When OCR Dominates vs. When Preprocessing Dominates

The complete BM25 ablation in Table 3 reveals two regimes.

**OCR-dominated languages.** For several languages (e.g., Arabic, Japanese, Vietnamese), the choice of OCR model explains most of the performance variance. In these cases, linguistic normalization cannot compensate for missing or corrupted transcription; the dominant gains come from recovering readable text (Table 3).

**Preprocessing-dominated languages.** For highly inflected languages (e.g., Czech, Slovenian, Croatian), OCR choice is often secondary once a reasonable transcription is available. Here, lemmatization or morphology accounts for most of the gains, suggesting that representation quality is limited less by OCR and more by token normalization (Table 3, Table 2).

These regimes help explain why a single global preprocessing pipeline is suboptimal (Table 2).

## B.2 BM25 Sensitivity to Representation Choices

BM25 is highly sensitive to small representation changes: enabling or disabling a single step (e.g., segmentation for Japanese, morphology for Arabic) can change Top-5 accuracy by over 10 points even when OCR is fixed (Table 3). At the same time, the best configurations are typically simple (one or two steps enabled), indicating that most gains come from lightweight, language-specific normalization rather than complex pipelines (Table 2).

## B.3 Multilingual Leaderboard Context

Table 1 provides the full multilingual leaderboard across text-based and multimodal retrievers. The most important takeaway for interpretation is that the *relative* standing of text-based retrievers is strongly affected by representation choices (BM25 vs. BM25*), whereas multimodal methods are invariant to OCR. This makes OCR/preprocessing a key confounder when multilingual comparisons mix modalities (Table 1). It is also worth noting that documents have different page-count distributions across languages, which should be considered when comparing results Chen et al. (2025).

# C Retrieval Across Figure, Table, and Text Documents

This section comments on patterns across Figure/Table/Text subsets and controlled OCR ablations across retrievers (Table 4) that are not discussed in the main text.

## C.1 Dense Retrievers Also Benefit from Better Transcription

Although the main paper emphasizes BM25, the controlled OCR ablations show that dense retrievers also benefit from improved transcription. In particular, changing only the transcription while keeping the retriever fixed yields large gains for BGE-M3 and SBERT on figure-heavy pages (Table 4). This indicates that the representation bottleneck is not specific to lexical matching: dense embedding models are likewise constrained by missing or noisy text.

Table 1: **Multilingual retrieval accuracy (%) on VisR-Bench.** Top-1 and Top-5 accuracy across 15 languages.

| | Spanish | | Italian | | German | | French | | Dutch | | Arabic | | Croatian | | Japanese | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 |
| **Text-based Methods** | | | | | | | | | | | | | | | | |
| BM25 Robertson & Walker (1994) | 60.23 | 82.55 | 59.14 | 82.14 | 65.79 | 86.87 | 54.05 | 77.84 | 59.83 | 84.94 | 7.25 | 20.73 | 52.95 | 72.90 | 11.84 | 39.35 |
| SBERT Reimers & Gurevych (2019) | 22.77 | 41.83 | 21.82 | 41.12 | 25.74 | 48.54 | 27.43 | 51.33 | 27.99 | 52.25 | 4.02 | 17.29 | 17.72 | 36.67 | 13.06 | 41.24 |
| BGE-large Xiao et al. (2023) | 34.55 | 60.41 | 30.27 | 56.24 | 39.75 | 66.82 | 41.34 | 67.42 | 39.14 | 67.53 | 6.15 | 19.53 | 32.67 | 58.14 | 31.92 | 64.97 |
| BGE-M3 Chen et al. (2024b) | 58.16 | 83.13 | 52.94 | 77.96 | 67.64 | 88.94 | 60.68 | 82.10 | 63.62 | 87.73 | 10.55 | 26.26 | **59.07** | **81.46** | 58.38 | 84.33 |
| NV-Embed-v2 Lee et al. (2024) | 42.92 | 72.71 | 40.84 | 66.32 | 52.23 | 80.30 | 49.41 | 76.13 | 47.12 | 78.74 | 5.47 | 21.73 | 41.86 | 68.30 | 42.17 | 72.70 |
| **BM25*** Robertson & Walker (1994) | 61.07 | 84.98 | 60.30 | 84.96 | 66.87 | 89.36 | 58.43 | 84.14 | 62.18 | 87.13 | **25.92** | **47.20** | 58.88 | 77.35 | 49.35 | 85.47 |
| **Multimodal Encoders** | | | | | | | | | | | | | | | | |
| CLIP Radford et al. (2021) | 11.14 | 29.32 | 12.39 | 31.77 | 19.53 | 45.69 | 19.52 | 44.44 | 16.22 | 42.71 | 4.64 | 18.91 | 10.46 | 27.36 | 14.28 | 44.86 |
| SigLip Zhai et al. (2023) | 13.08 | 32.36 | 17.52 | 40.69 | 25.69 | 51.69 | 24.85 | 53.15 | 22.70 | 50.85 | 5.53 | 19.56 | 13.98 | 33.56 | 15.62 | 46.20 |
| **Multimodal Large Language Models** | | | | | | | | | | | | | | | | |
| VisRAG Yu et al. (2024) | 9.70 | 28.48 | 10.69 | 33.09 | 14.48 | 40.22 | 16.37 | 42.55 | 15.22 | 42.02 | 4.78 | 19.80 | 6.38 | 22.25 | 21.04 | 52.37 |
| VLM2Vec Jiang et al. (2024) | 18.59 | 44.48 | 19.42 | 43.84 | 26.07 | 56.10 | 29.53 | 60.50 | 22.51 | 52.97 | 7.39 | 24.10 | 12.31 | 32.04 | 19.19 | 50.02 |
| GME Zhang et al. (2024) | 60.57 | 88.08 | 52.96 | 79.08 | 65.97 | 89.61 | 66.78 | 89.55 | 57.92 | 85.16 | 15.33 | 35.72 | 45.09 | 72.60 | 61.11 | 89.37 |
| ColInternVL2 Chen et al. (2024a) | 58.26 | 84.57 | 51.89 | 77.96 | 60.35 | 86.32 | 64.06 | 87.17 | 58.27 | 84.60 | 5.09 | 17.50 | 47.68 | 73.16 | 39.65 | 71.57 |
| ColPhi Chen et al. (2024a) | 65.42 | 89.00 | 56.06 | 81.43 | 65.02 | 88.96 | 67.83 | 89.65 | 62.15 | 88.17 | 8.46 | 25.95 | 48.83 | 74.82 | 25.28 | 56.49 |
| ColPali-v1.2 Faysse et al. (2024) | 71.44 | 92.62 | 62.02 | 85.81 | 72.96 | 92.48 | 72.62 | 92.09 | 65.15 | 89.73 | 14.33 | 32.59 | 51.54 | 76.94 | 43.85 | 77.53 |
| ColQwen2-v0.1 Faysse et al. (2024) | **75.04** | **94.34** | **65.18** | **88.24** | **78.63** | **95.77** | **77.81** | **93.69** | **70.30** | **92.12** | 12.05 | 27.16 | 55.27 | 79.20 | **65.81** | 89.41 |
| **Finetuned on Multilingual Data** | | | | | | | | | | | | | | | | |
| ColQwen2 (E) | 67.25 | 90.60 | 57.10 | 82.29 | 71.99 | 93.18 | 72.01 | 91.39 | 60.26 | 86.57 | 10.15 | 26.70 | 44.50 | 71.12 | 61.54 | 87.70 |
| ColQwen2 (M) | 69.77 | 92.48 | 59.77 | 85.39 | 72.71 | 92.97 | 72.70 | 92.16 | 64.37 | 89.10 | 11.67 | 28.07 | 50.44 | 76.83 | 63.91 | **89.98** |

| | Swedish | | Vietnamese | | Portuguese | | Finnish | | Czech | | Slovenian | | Danish | | Macro Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 |
| **Text-based Methods** | | | | | | | | | | | | | | | | |
| BM25 Robertson & Walker (1994) | 57.49 | 83.68 | 48.86 | 73.22 | 61.47 | 79.98 | 50.11 | 70.33 | 66.19 | 89.34 | 56.45 | 81.81 | 54.52 | 82.64 | 51.08 | 73.88 |
| SBERT Reimers & Gurevych (2019) | 28.26 | 60.99 | 17.94 | 37.07 | 25.85 | 50.24 | 23.34 | 47.29 | 26.28 | 50.00 | 22.31 | 48.03 | 29.56 | 58.11 | 22.27 | 45.47 |
| BGE-large Xiao et al. (2023) | 42.18 | 74.69 | 23.94 | 48.97 | 38.53 | 66.08 | 31.58 | 57.97 | 33.97 | 61.94 | 35.30 | 63.89 | 35.58 | 71.02 | 33.12 | 60.37 |
| BGE-M3 Chen et al. (2024b) | 65.25 | 89.33 | 44.93 | 68.82 | 60.07 | 82.16 | **56.90** | 77.19 | 65.87 | 90.22 | **65.05** | 88.53 | 64.42 | 88.38 | 56.90 | 79.77 |
| NV-Embed-v2 Lee et al. (2024) | 53.02 | 81.40 | 25.75 | 60.24 | 56.98 | 80.34 | 34.32 | 61.63 | 41.99 | 70.59 | 43.91 | 73.21 | 52.94 | 79.48 | 42.06 | 69.59 |
| **BM25*** Robertson & Walker (1994) | 59.26 | 86.67 | **54.86** | **83.71** | 62.68 | 85.56 | 56.22 | **80.32** | **68.59** | **93.27** | 61.20 | 85.93 | 53.37 | 85.65 | 57.28 | 82.78 |
| **Multimodal Encoders** | | | | | | | | | | | | | | | | |
| CLIP Radford et al. (2021) | 17.38 | 48.84 | 6.67 | 22.13 | 16.75 | 42.17 | 12.13 | 36.84 | 11.86 | 34.78 | 13.35 | 36.29 | 13.77 | 45.48 | 13.34 | 36.77 |
| SigLip Zhai et al. (2023) | 26.78 | 61.66 | 8.38 | 25.13 | 25.30 | 51.03 | 17.24 | 45.84 | 20.67 | 47.92 | 17.03 | 43.28 | 23.53 | 54.38 | 18.53 | 43.82 |
| **Multimodal Large Language Models** | | | | | | | | | | | | | | | | |
| VisRAG Yu et al. (2024) | 14.93 | 49.30 | 5.53 | 18.56 | 12.68 | 38.29 | 9.76 | 34.78 | 9.46 | 34.13 | 8.69 | 34.50 | 13.77 | 46.34 | 11.57 | 35.78 |
| VLM2Vec Jiang et al. (2024) | 25.98 | 62.17 | 8.22 | 25.39 | 23.73 | 53.46 | 16.17 | 44.47 | 21.07 | 50.56 | 15.59 | 44.09 | 25.39 | 58.68 | 19.41 | 46.86 |
| GME Zhang et al. (2024) | 59.09 | 89.79 | 26.22 | 51.81 | 65.29 | 90.72 | 38.83 | 68.80 | 51.52 | 82.29 | 51.34 | 80.91 | 54.52 | 86.94 | 51.50 | 78.70 |
| ColInternVL2 Chen et al. (2024a) | 61.16 | 90.51 | 25.75 | 54.19 | 62.32 | 86.95 | 46.22 | 72.85 | 55.29 | 85.90 | 54.75 | 83.96 | 60.11 | 90.10 | 50.06 | 76.49 |
| ColPhi Chen et al. (2024a) | 64.19 | 93.08 | 34.28 | 65.25 | 64.99 | 88.59 | 49.73 | 75.82 | 58.65 | 88.86 | 56.81 | 85.66 | 61.12 | 90.67 | 52.59 | 78.83 |
| ColPali-v1.2 Faysse et al. (2024) | 65.37 | 92.11 | 35.32 | 66.60 | 76.03 | 92.96 | 42.11 | 73.07 | 62.34 | 91.27 | 55.82 | 86.29 | 62.41 | 90.10 | 56.89 | 82.15 |
| ColQwen2-v0.1 Faysse et al. (2024) | **70.16** | **95.37** | 35.39 | 64.51 | **76.32** | **93.53** | 49.72 | 75.82 | 65.03 | 92.57 | 61.67 | 88.98 | **72.31** | **94.76** | **62.05** | **84.36** |
| **Finetuned on Multilingual Data** | | | | | | | | | | | | | | | | |
| ColQwen2 (E) | 62.30 | 91.37 | 24.50 | 53.29 | 49.44 | 80.19 | 40.88 | 68.43 | 52.87 | 84.12 | 49.44 | 80.19 | 63.92 | 92.22 | 52.54 | 78.62 |
| ColQwen2 (M) | 66.12 | 94.05 | 27.10 | 55.05 | 71.86 | 92.09 | 40.88 | 70.04 | 61.32 | 89.63 | 58.43 | **89.35** | 64.52 | 93.22 | 57.04 | 82.03 |

Table 2: Best configuration per language based on Top-5 accuracy. We report the optimal OCR model and linguistic processing features for each of the 15 languages tested.

| Language | Best OCR | Top-1 (%) | Top-5 (%) | Features |
|---|---|---|---|---|
| Spanish | EasyOCR | 61.07 | 84.98 | Stemming |
| Italian | EasyOCR | 60.30 | 84.96 | None |
| German | Ministral 3B | 66.87 | 89.36 | Stemming |
| French | Ministral 3B | 58.43 | 84.14 | None |
| Dutch | EasyOCR | 62.18 | 87.13 | None |
| Arabic | Mistral OCR 3 | 25.92 | 47.20 | Morphology |
| Croatian | Mistral OCR 3 | 58.88 | 77.35 | Lemmatization |
| Japanese | Ministral 3B | 49.35 | 85.47 | Adv. Tokenizers |
| Swedish | Ministral 3B | 59.26 | 86.67 | Stemming |
| Vietnamese | Ministral 3B | 54.86 | 83.71 | None |
| Portuguese | Ministral 3B | 62.68 | 85.56 | None |
| Finnish | EasyOCR | 56.22 | 80.32 | Stemming |
| Czech | Ministral 3B | 68.59 | 93.27 | Lemmatization |
| Slovenian | EasyOCR | 61.20 | 85.93 | Lemmatization |
| Danish | EasyOCR | 53.37 | 85.65 | Stemming |
| **Average** | — | 57.28 | 82.78 | — |

### C.2 Why Figure-Focused Gains Are Disproportionately Large

The OCR ablations show substantially larger gains on figure-heavy pages than on text-only pages (Table 4). A qualitative inspection suggests two main mechanisms: (i) evidence is often contained in plot labels, legends, and axis annotations that are absent from default OCR, and (ii) even coarse

Table 3: Complete BM25 ablation study across all languages and OCR models on multilingual VisR-Bench. ✓=enabled, ×=disabled. Top-1 / Top-5 accuracy (%).

| | | Adobe Text Extract | | | EasyOCR | | | Ministral 3B | | | Mistral OCR 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Language | Config | MSLS | T-1 | T-5 | MSLS | T-1 | T-5 | MSLS | T-1 | T-5 | MSLS | T-1 | T-5 |
| Spanish | 1 | ×✓×× | 57.65 | 82.00 | ×✓×× | 61.07 | 84.98 | ×✓×× | 56.98 | 82.92 | ×✓×× | 56.69 | 79.82 |
| | 2 | ×××× | 60.23 | 82.55 | ×××× | 58.49 | 81.70 | ×××× | 60.07 | 83.87 | ×××× | 59.03 | 80.18 |
| Italian | 1 | ×✓×× | 56.57 | 80.70 | ×✓×× | 57.54 | 83.41 | ×✓×× | 55.56 | 81.33 | ×✓×× | 55.05 | 79.03 |
| | 2 | ×××× | 59.14 | 82.14 | ×××× | 60.30 | 84.96 | ×××× | 58.74 | 83.35 | ×××× | 57.42 | 80.09 |
| German | 1 | ×✓×× | 66.09 | 87.82 | ×✓×× | 67.47 | 89.19 | ×✓×× | 66.87 | 89.36 | ×✓×× | 66.87 | 87.34 |
| | 2 | ×××× | 65.79 | 86.87 | ×××× | 62.90 | 85.82 | ×××× | 66.39 | 88.06 | ×××× | 66.29 | 85.89 |
| French | 1 | ×✓×× | 52.72 | 78.02 | ×✓×× | 52.89 | 81.43 | ×✓×× | 56.02 | 83.68 | ×✓×× | 54.41 | 79.56 |
| | 2 | ×××× | 54.05 | 77.84 | ×××× | 53.87 | 80.64 | ×××× | 58.43 | 84.14 | ×××× | 56.89 | 79.76 |
| Dutch | 1 | ×✓×× | 59.24 | 84.69 | ×✓×× | 62.30 | 86.73 | ×✓×× | 59.42 | 83.81 | ×✓×× | 58.77 | 82.65 |
| | 2 | ×××× | 59.83 | 84.94 | ×××× | 62.18 | 87.13 | ×××× | 59.96 | 84.10 | ×××× | 59.77 | 82.65 |
| Arabic | 1 | ✓××× | 7.84 | 22.59 | ✓××× | 5.67 | 18.25 | ✓××× | 19.11 | 40.84 | ✓××× | 25.92 | 47.20 |
| | 2 | ×××× | 7.25 | 20.73 | ×××× | 4.78 | 16.98 | ×××× | 15.68 | 34.03 | ×××× | 19.08 | 37.88 |
| Croatian | 1 | ××✓× | 56.77 | 75.57 | ××✓× | 54.47 | 76.49 | ××✓× | 55.47 | 75.79 | ××✓× | 58.88 | 77.35 |
| | 2 | ×✓×× | 52.95 | 72.90 | ×✓×× | 51.76 | 74.45 | ×✓×× | 51.98 | 73.34 | ×✓×× | 55.21 | 75.34 |
| | 3 | ×✓✓× | 56.77 | 75.57 | ×✓✓× | 54.47 | 76.49 | ×✓✓× | 55.47 | 75.79 | ×✓✓× | 58.88 | 77.35 |
| | 4 | ×××× | 52.95 | 72.90 | ×××× | 51.76 | 74.45 | ×××× | 51.98 | 73.34 | ×××× | 55.21 | 75.34 |
| Japanese | 1 | ×××✓ | 48.72 | 80.55 | ×××✓ | 14.24 | 43.39 | ×××✓ | 49.35 | 85.47 | ×××✓ | 46.87 | 77.74 |
| | 2 | ×✓×× | 43.47 | 75.47 | ×✓×× | 10.50 | 37.76 | ×✓×× | 48.89 | 84.00 | ×✓×× | 44.94 | 76.14 |
| | 3 | ×××× | 11.84 | 39.35 | ×××× | 11.38 | 39.56 | ×××× | 11.93 | 39.82 | ×××× | 11.72 | 39.48 |
| Swedish | 1 | ×✓×× | 58.75 | 85.36 | ×✓×× | 54.91 | 84.27 | ×✓×× | 59.26 | 86.67 | ×✓×× | 59.68 | 83.72 |
| | 2 | ×××× | 57.49 | 83.68 | ×××× | 54.32 | 83.30 | ×××× | 58.50 | 85.58 | ×××× | 58.75 | 82.45 |
| Vietnamese | 1 | ×××✓ | 48.24 | 71.77 | ×××✓ | 18.05 | 40.80 | ×××✓ | 56.41 | 83.51 | ×××✓ | 57.60 | 81.33 |
| | 2 | ×××× | 48.86 | 73.22 | ×××× | 21.56 | 48.76 | ×××× | 54.86 | 83.71 | ×××× | 57.34 | 81.33 |
| Portuguese | 1 | ×✓×× | 60.32 | 80.16 | ×✓×× | 54.79 | 81.01 | ×✓×× | 61.04 | 85.13 | ×✓×× | 61.77 | 83.25 |
| | 2 | ×××× | 61.47 | 79.98 | ×××× | 56.19 | 80.34 | ×××× | 62.68 | 85.56 | ×××× | 63.17 | 83.37 |
| Finnish | 1 | ××✓× | 49.12 | 68.57 | ××✓× | 51.33 | 76.20 | ××✓× | 51.03 | 75.13 | ××✓× | 51.49 | 72.23 |
| | 2 | ×✓×× | 54.54 | 73.61 | ×✓×× | 56.22 | 80.32 | ×✓×× | 54.69 | 78.64 | ×✓×× | 56.83 | 75.36 |
| | 3 | ×✓✓× | 49.12 | 68.57 | ×✓✓× | 51.33 | 76.20 | ×✓✓× | 51.03 | 75.13 | ×✓✓× | 51.49 | 72.23 |
| | 4 | ×××× | 50.11 | 70.33 | ×××× | 51.18 | 75.13 | ×××× | 47.98 | 73.00 | ×××× | 51.18 | 72.01 |
| Czech | 1 | ××✓× | 68.91 | 92.15 | ××✓× | 45.03 | 75.88 | ××✓× | 68.59 | 93.27 | ××✓× | 68.99 | 91.27 |
| | 2 | ×✓×× | 66.19 | 89.34 | ×✓×× | 44.63 | 73.32 | ×✓×× | 66.91 | 89.66 | ×✓×× | 67.31 | 88.62 |
| | 3 | ×✓✓× | 68.91 | 92.15 | ×✓✓× | 45.03 | 75.88 | ×✓✓× | 68.59 | 93.27 | ×✓✓× | 68.99 | 91.27 |
| | 4 | ×××× | 66.19 | 89.34 | ×××× | 44.63 | 73.32 | ×××× | 66.91 | 89.66 | ×××× | 67.31 | 88.62 |
| Slovenian | 1 | ××✓× | 60.57 | 84.14 | ××✓× | 61.20 | 85.93 | ××✓× | 62.99 | 84.23 | ××✓× | 59.59 | 80.73 |
| | 2 | ×✓×× | 56.45 | 81.81 | ×✓×× | 53.32 | 82.89 | ×✓×× | 56.54 | 83.60 | ×✓×× | 54.84 | 79.03 |
| | 3 | ×✓✓× | 60.57 | 84.14 | ×✓✓× | 61.20 | 85.93 | ×✓✓× | 62.99 | 84.23 | ×✓✓× | 59.59 | 80.73 |
| | 4 | ×××× | 56.45 | 81.81 | ×××× | 53.32 | 82.89 | ×××× | 56.54 | 83.60 | ×××× | 54.84 | 79.03 |
| Danish | 1 | ×✓×× | 55.81 | 84.07 | ×✓×× | 53.37 | 85.65 | ×✓×× | 57.39 | 85.08 | ×✓×× | 56.53 | 78.77 |
| | 2 | ×××× | 54.52 | 82.64 | ×××× | 52.80 | 84.51 | ×××× | 55.52 | 83.07 | ×××× | 55.81 | 77.33 |

MSLS = Morphology/Stemming/Lemmatization/Segmentation. ✓indicates enabled, ×indicates disabled. In the table we reuse the 'stemming' for Japanese to mark character level tokenization.

VLM-style transcriptions recover enough lexical anchors (numbers, variable names, captions) to improve page discrimination (Table 4).

### C.3 Limits of Representation-Only Improvements

Even with the best OCR and normalization, text-based retrievers remain well below state-of-the-art multimodal systems on visually grounded questions (Table 4). These cases correspond to evidence that is not recoverable as text (e.g., spatial relations, graphical trends, non-textual encodings), highlighting a clear boundary: representation improvements close much of the gap for text-bearing visual content, but not for fundamentally visual reasoning (Table 4).

### C.4 Benchmarking Implication

If changing only OCR yields double-digit gains for a fixed retriever (Table 4), then benchmark gaps should not be interpreted as evidence of superior retrieval models. In such settings, OCR and preprocessing should be treated as benchmark variables rather than hidden implementation details.

Table 4: **Retrieval accuracy (%) on VisR-Bench across document types. A:** Text-based retrieval with default OCR. **B:** Controlled OCR ablation (retriever fixed; only transcription varies). **C:** Multimodal retrievers operating directly on document images. *Macro Avg.* is the unweighted mean over Figure/Table/Text.

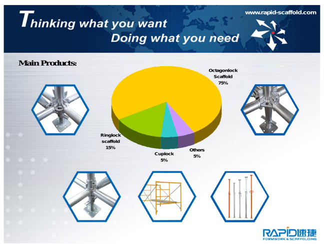
| | | Figure | | Table | | Text | | Macro Avg. | |
|---|---|---|---|---|---|---|---|---|---|
| Retriever | OCR / Transcription | top1 | top5 | top1 | top5 | top1 | top5 | top1 | top5 |
| **A. Text-Based Retrieval (Default OCR)** | | | | | | | | | |
| BM25 Baeza-Yates et al. (1999) | Adobe Text Extract | 24.27 | 46.12 | 38.66 | 66.54 | 64.74 | 89.10 | 42.56 | 67.25 |
| SBERT Reimers & Gurevych (2019) | Adobe Text Extract | 25.24 | 49.27 | 26.31 | 52.68 | 49.96 | 76.97 | 33.84 | 59.64 |
| BGE-large Xiao et al. (2023) | Adobe Text Extract | 31.55 | 56.07 | 40.36 | 70.14 | 57.00 | 82.68 | 42.97 | 69.63 |
| BGE-M3 Chen et al. (2024b) | Adobe Text Extract | 31.07 | 56.80 | 51.11 | 78.51 | 67.68 | 89.89 | 49.96 | 73.95 |
| NV-Embed-v2 Lee et al. (2024) | Adobe Text Extract | 35.44 | 65.05 | 44.04 | 73.34 | 61.38 | 87.46 | 46.95 | 75.28 |
| **B. Controlled OCR Ablation (Retriever Fixed)** | | | | | | | | | |
| BM25 | EasyOCR* | 47.22 | 73.12 | 43.39 | 70.89 | 65.58 | 90.62 | 52.06 | 78.21 |
| | Mistral-3 OCR* | 45.52 | 71.91 | 41.61 | 68.12 | 66.34 | 91.38 | 51.16 | 77.14 |
| | Ministral 3B* | 49.88 | 77.24 | 41.51 | 69.34 | 65.03 | 90.10 | 52.14 | 78.89 |
| SBERT | EasyOCR* | 39.95 | 69.73 | 29.58 | 58.46 | 49.95 | 77.86 | 39.83 | 68.68 |
| | Mistral-3 OCR* | 34.14 | 60.53 | 23.18 | 49.96 | 51.30 | 78.18 | 36.21 | 62.89 |
| | Ministral 3B* | 37.77 | 63.44 | 22.36 | 47.59 | 48.42 | 76.00 | 36.18 | 62.34 |
| BGE-M3 | EasyOCR* | 55.21 | 82.81 | 47.96 | 76.06 | 69.55 | 92.08 | 57.57 | 83.65 |
| | Mistral-3 OCR* | 51.57 | 79.42 | 45.42 | 74.92 | 69.10 | 91.66 | 55.36 | 82.00 |
| | Ministral 3B* | 59.32 | 87.41 | 48.06 | 77.51 | 67.88 | 91.83 | 58.42 | 85.58 |
| **C. Multimodal Retrieval (Reference)** | | | | | | | | | |
| CLIP Radford et al. (2021) | – | 33.90 | 61.74 | 24.68 | 47.59 | 39.47 | 70.21 | 32.68 | 59.85 |
| SigLip Zhai et al. (2023) | – | 38.98 | 69.73 | 24.73 | 53.22 | 39.06 | 70.97 | 34.26 | 64.64 |
| VisRAG Yu et al. (2024) | – | 31.96 | 66.83 | 19.82 | 48.53 | 31.00 | 61.49 | 27.59 | 58.95 |
| VLM2Vec Jiang et al. (2024) | – | 40.44 | 76.27 | 28.51 | 57.77 | 39.90 | 71.69 | 36.28 | 68.58 |
| GME Zhang et al. (2024) | – | 68.04 | 91.53 | 61.50 | 86.38 | 76.34 | 95.62 | 68.63 | 91.18 |
| Col-InternVL2 Chen et al. (2024a) | – | 68.28 | 90.31 | 63.85 | 86.36 | 79.19 | 96.45 | 70.44 | 91.04 |
| Col-Phi Chen et al. (2024a) | – | 68.77 | 93.22 | 65.65 | 88.51 | 81.67 | 97.04 | 72.03 | 92.92 |
| ColPali-v1.2 Faysse et al. (2024) | – | 68.77 | 91.77 | 66.12 | 88.26 | 82.63 | 96.89 | 72.51 | 92.31 |
| ColQwen2-v0.1 Faysse et al. (2024) | – | **74.58** | **95.64** | **67.43** | **88.98** | **83.68** | **97.61** | **75.23** | **94.08** |

* OCR models are used strictly for transcription; retrieval configuration is unchanged. Multimodal retrievers bypass OCR and are shown for contextual reference rather than direct comparison.

## D EXAMPLE ANNOTATIONS

In this section, we present example annotations produced by different text extraction methods, illustrating a progression in OCR quality: from limited extraction focused on figure captions, to broader text coverage, and finally to full image-level content descriptions. We opt for a simple prompt for the Ministral 3B - *"Give me a markdown of what you see in the image. Reply only with the markdown content."*. It would be an interesting area of exploration to optimize for lexical recall and query-agnostic retrieval effectiveness, e.g. prompting the model to provide concise, noun-heavy and non-redundant descriptions.

### D.1 Example 1




**Adobe Text Extract**

```
![Figure](fileoutpart5.png)
```

**EasyOCR**

```
wwrapid-scaffold.com
Thinking what you want Doing what you need
Main Produds:
Octagonlock Scaffold 75%
Ringlock scaffold 15%
Others 5%
Cuplock 5%
RAPIDwII FormwoRK & SCAFFOLDING
```

**Mistral OCR 3**

```
www.rapid-scaffold.com

# Thinking what you want

# Doing what you need

![img-0.jpeg](img-0.jpeg)

![img-1.jpeg](img-1.jpeg)
Main Products:

![img-2.jpeg](img-2.jpeg)

![img-3.jpeg](img-3.jpeg)

![img-4.jpeg](img-4.jpeg)

![img-5.jpeg](img-5.jpeg)

![img-6.jpeg](img-6.jpeg)

RAPID

![Pie Chart](https://example.com/piechart.png
     )

The pie chart shows the distribution of main
     products:
- Octagonlock Scaffold: 75%
- Ringlock Scaffold: 15%
- Cuplock: 5%
- Others: 5%

The images around the pie chart represent
     different types of scaffolding products.
```

**Ministral 3B**

````
```markdown
# Main Products Breakdown

## Logo & Tagline
- ![World Map](placeholder) [www.rapid-
     scaffold.com](http://www.rapid-scaffold.
     com)
- **Tagline:**
  Thinking what you want
  Doing what you need

---

## Product Distribution Chart
- **Octagonlock Scaffold:** 75%
- **Ringlock Scaffold:** 15%
- **Cuplock:** 5%
- **Others:** 5%

---

## Product Images
1. **Octagonlock Scaffold**
   ![Octagonlock Scaffold](placeholder)

2. **Ringlock Scaffold**
   ![Ringlock Scaffold](placeholder)

3. **Cuplock**
   ![Cuplock](placeholder)

4. **Generic Scaffold Frame**
   ![Generic Scaffold Frame](placeholder)

5. **Formwork System**
   ![Formwork System](placeholder)

6. **Vertical Support System**
   ![Vertical Support System](placeholder)

---

## Branding
- **RAPID**
  - Formwork & Scaffolding
  - Logo: [RAPID](placeholder) with Chinese
     characters:
```

*(Note: Replace placeholder images with
     actual descriptions or references if
     needed for visual representation.)*
````

### D.2 Example 2


**Adobe Text Extract**

```
![Figure](fileoutpart7.png)
```

**EasyOCR**

```
SMARTWEBCANADA smart websites, smart technology; smart
     people
Design
Service
'ptimization
Ads/SEM
Marketing
Video
Mobile
blo
E-Mail
Content
Branding
SMM
Introduction
```

**Mistral OCR 3**

```
SMARTWEBCANADA
smart websites, smart technology, smart
     people

![img-0.jpeg](img-0.jpeg)

# Introduction

![Services Hexagon](https://example.com/
     services-hexagon.png)

The image displays a hexagonal diagram
     representing various services offered by
      SmartWebCanada. The hexagons are
     labeled as follows:

- Video
- E-Mail
- Content
- SMM (Social Media Marketing)
- Mobile
- Ads/SEM (Search Engine Marketing)
- Branding

Additionally, there is a circular diagram on
     the left side showing the interconnected
      services of Design, Service,
     Optimization, and Marketing.
```

**Ministral 3B**

~~~~
```markdown
# SmartWebCanada - Hexagonal Diagram Overview

## Core Services
- **Service**
  - Design
  - Optimization
  - Marketing

## Key Components
- **Video**
  ![Video Icon](placeholder) *(Image: Camera
     setup)*

- **Mobile**
  ![Mobile Icon](placeholder) *(Image: Mobile
      device)*

- **Ads/SEM (Search Engine Marketing)**
  ![Ads/SEM Icon](placeholder) *(Image:
     Google search page)*

- **Content**
  ![Content Icon](placeholder) *(Image: Blog
     setup)*

- **E-Mail**
  ![E-Mail Icon](placeholder) *(Image: Email
     inbox)*

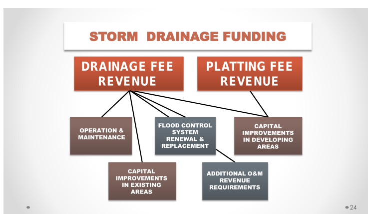
- **SMM (Social Media Marketing)**
  ![SMM Icon](placeholder) *(Image: Social
     media app icons)*

- **Branding**
  ![Branding Icon](placeholder) *(Image:
     Branding materials on a desk)*

## Introduction Section
- **Text:** Introduction
```

*(Note: Placeholder images are described with
      typical icons and devices.)*
~~~~

### D.3 EXAMPLE 3




**Adobe Text Extract**

```
# STORM DRAINAGE FUNDING

![Figure](fileoutpart42.png)
```

**EasyOCR**

```
STORM DRAINAGE FUNDING
DRAINAGE FEE REVENUE
PLATTING FEE REVENUE
FLOOD CONTROL SYSTEM RENEWAL & REPLACEMENT
CAPITAL
OPERATION & MAINTENANCE
IMPROVEMENTS IN DEVELOPING AREAS
CAPITAL
ADDITIONAL O&M REVENUE REQUIREMENTS
IMPROVEMENTS IN EXISTING AREAS
24
```

**Mistral OCR 3**

````
24

# STORM DRAINAGE FUNDING

Document figure (markdown):
```mermaid
flowchart TD
    A[STORM DRAINAGE FUNDING] --> B[DRAINAGE
     FEE REVENUE]
    A --> C[PLATTING FEE REVENUE]
    B --> D[OPERATION & MAINTENANCE]
    B --> E[FLOOD CONTROL SYSTEM RENEWAL &
     REPLACEMENT]
    B --> F[CAPITAL IMPROVEMENTS IN EXISTING
     AREAS]
    C --> G[CAPITAL IMPROVEMENTS IN
     DEVELOPING AREAS]
    C --> H[ADDITIONAL O&M REVENUE
     REQUIREMENTS]
```
````

**Ministral 3B**

````
```markdown
# Storm Drainage Funding

## Revenue Sources
- **Drainage Fee Revenue**
- **Platting Fee Revenue**

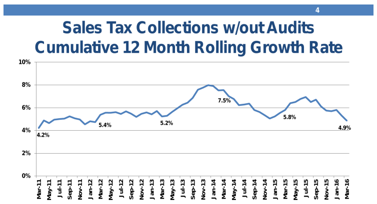
### Allocation Breakdown

#### Drainage Fee Revenue
- **Operation & Maintenance**
  - **Capital Improvements in Existing Areas
    **
- **Flood Control System Renewal &
    Replacement**
  - **Additional O&M Revenue Requirements**

#### Platting Fee Revenue
- **Capital Improvements in Developing Areas
    **
```

*(Note: The diagram does not explicitly show
     direct arrows from Platting Fee Revenue
     to specific categories in the image, so
     only the direct allocations from
     Drainage Fee Revenue are fully
     represented.)*
````

### D.4 EXAMPLE 4




**Adobe Text Extract**

```
Sales Tax Collections w/out Audits Cumulative 12 Month
     Rolling Growth Rate

![Figure](fileoutpart10.png)
```

**EasyOCR**

```
4
Sales Tax Collections wloutAudits
Cumulative 12 Month Rolling Growth Rate 10%
8%
7.5%
6%
5.8%
5.4%
5.2%
4% 4.2%
4.9%
2%
0% 11111118882112171 I!AN 1 L 4
```

**Mistral OCR 3**

````
# Sales Tax Collections w/out Audits
## Cumulative 12 Month Rolling Growth Rate
Document figure (markdown):
```markdown                        | Growth
     Rate
------------------------|------------
Mar-11                    | 4.2%
May-11                    | ~4.5%
Jul-11                    | ~4.7%
Sep-11                    | ~4.6%
Nov-11                    | ~4.8%
Jan-12                    | ~4.7%
Mar-12                    | ~4.9%
May-12                    | ~5.0%
Jul-12                    | ~5.1%
Sep-12                    | 5.4%
Nov-12                    | ~5.2%
Jan-13                    | ~5.3%
Mar-13                    | ~5.4%
May-13                    | ~5.5%
Jul-13                    | ~5.6%
Sep-13                    | ~5.7%
Nov-13                    | ~6.0%
Jan-14                    | ~6.5%
Mar-14                    | ~7.0%
May-14                    | 7.5%
Jul-14                    | ~7.2%
Sep-14                    | ~6.8%
Nov-14                    | ~6.5%
Jan-15                    | ~6.0%
Mar-15                    | ~5.8%
May-15                    | ~5.7%
Jul-15                    | ~5.6%
Sep-15                    | 5.8%
Nov-15                    | ~5.5%
Jan-16                    | ~5.3%
Mar-16                    | 4.9%
```
````

**Ministral 3B**

````
```markdown
# Sales Tax Collections without Audits
## Cumulative 12 Month Rolling Growth Rate

```
![Graph Visualization](Graph representation)

**Timeline:**
- X-Axis: Mar-11 to Mar-16

**Key Data Points:**
- **Mar-11:** ~4.2%
- **Nov-12:** ~5.4%
- **Mar-13:** ~5.2%
- **Nov-13:** ~7.5% (peak)
- **Mar-15:** ~5.8%
- **Mar-16:** ~4.9%

**Trend Line:**
- Fluctuating growth rate with peaks and
     troughs.
- General trend shows variability with a
     notable peak around Nov-13 (~7.5%) and a
      decline towards Mar-16 (~4.9%).
```
````
````