This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

M4-RAG: A Massive-Scale Multilingual Multi-Cultural Multimodal RAG

David Anugraha[1*], Patrick Amadeus Irawan[2], Anshul Singh[3],
En-Shiun Annie Lee[4,5], Genta Indra Winata[6]
[1]Stanford University [2]MBZUAI [3]Indian Institute of Science [4]Ontario Tech University
[5]University of Toronto [6]Capital One

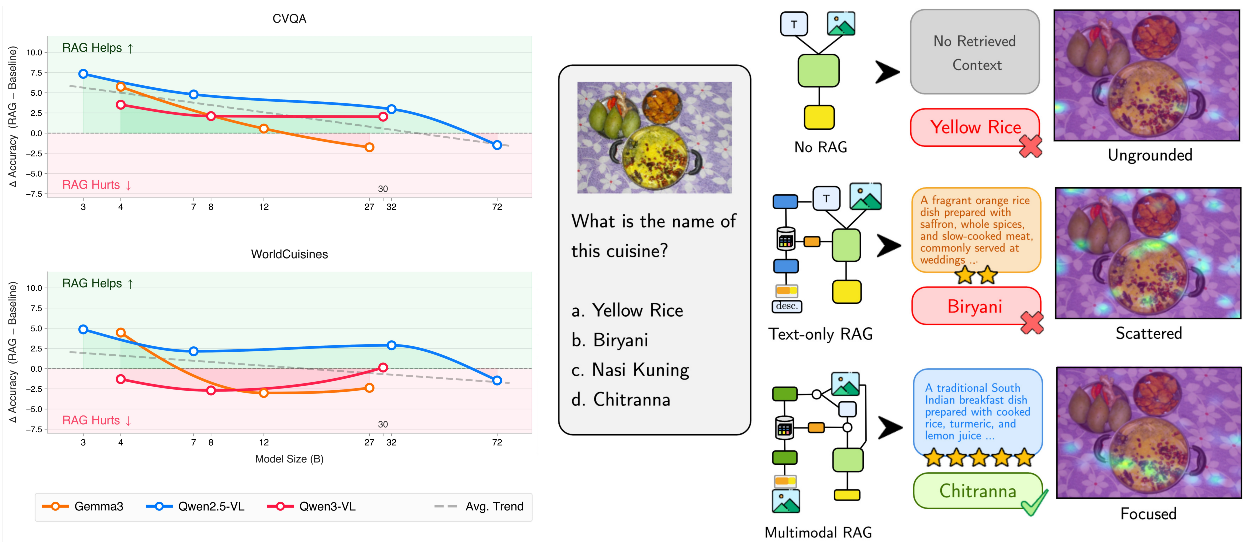

Figure 1. M4-RAG evaluates state-of-the-art Vision Language Models (VLMs) under culturally grounded retrieval settings. **Left:** Summarized results on CVQA (top) and WORLDCUISINES (bottom) across model scales with and without RAG, showing that the retrieved context can either improve or degrade performance depending on the scale. **Right:** Examples showing how the quality of retrieved context (number of ★) affects model behavior and directs attention toward semantically relevant cues. When asked to identify the cuisine in the image, the non-RAG and text-only RAG systems produce incorrect guesses such as "Yellow Rice" or "Biryani". With multimodal RAG, the system retrieves culturally specific evidence (e.g., Lemon Rice, India, Breakfast), guiding the VLM to the correct answer "Chitranna."

Abstract

Vision-language models (VLMs) have achieved strong performance in visual question answering (VQA), yet they remain constrained by static training data. Retrieval-Augmented Generation (RAG) mitigates this limitation by enabling access to up-to-date, culturally grounded, and multilingual information; however, multilingual multimodal RAG remains largely underexplored. We introduce M4-RAG, *a massive-scale benchmark spanning 42 languages, 56 regional dialects and registers, and 189 countries, comprising over 80,000 culturally diverse image-question pairs for evaluating retrieval-augmented VQA across languages and modalities. To balance realism with reproducibility, we build a controlled retrieval environment containing millions of carefully curated multilingual documents relevant to the query domains, approximating real-world retrieval conditions while ensuring consistent experimentation. Our systematic evaluation reveals that although RAG consistently benefits smaller VLMs, it fails to scale to larger models and often even degrades their performance, exposing a critical mismatch between model size and current retrieval effectiveness. Our cross-lingual evaluations also reveal significant performance degradation when prompts or retrieved context are provided in non-English languages. The code, datasets, and evaluation protocols for* M4-RAG *are available as open-source at* https://github.com/davidanugraha/M4-RAG.

*Corresponding author: david.anugraha@stanford.edu

1. Introduction

Recent advances in large language models (LLMs) and vision-language models (VLMs) have demonstrated strong capabilities in reasoning, summarization, and question answering [8, 19, 20, 53, 64]. However, their reliance on static training corpora often leads to outdated or incomplete knowledge, limiting factual accuracy and cross-domain coverage. Retrieval-Augmented Generation (RAG) addresses this limitation by augmenting model outputs with information retrieved from external knowledge sources [36, 59].

RAG has evolved along two complementary directions: multilingual and multimodal RAG. Multilingual RAG [15, 48, 71] enables cross-lingual information access, allowing queries and retrieved documents to appear in different languages, while multimodal RAG [1, 22, 39] incorporates visual or structured inputs such as images, tables, or videos into retrieval and generation pipelines. Despite rapid progress in both areas, their intersection—multilingual multimodal RAG—remains largely unexplored, as shown in Table 1. This gap is particularly important because real-world knowledge access is inherently both multilingual and multimodal [55]. Cultural knowledge exemplifies this challenge since it is inherently long-tail, region-specific, and not reliably encoded in model parameters even for large models [45, 50, 60], making it an ideal and well-motivated testbed for multilingual multimodal RAG. Consequently, the alignment between cross-lingual retrieval and multimodal representations, the ability of multilingual models to ground information across modalities, and the adequacy of evaluation metrics in capturing these complex dependencies are important, despite remaining underexplored.

To address these challenges, we present M4-RAG, a comprehensive evaluation framework for multilingual, multicultural, and multimodal RAG. Our evaluation spans multiple languages and modalities, covering both text-text and text-image retrieval scenarios. Figure 1 illustrates how multimodal retrieval can guide model attention toward relevant visual regions to correctly identify the dish as "Chitranna," whereas in the first two settings, it fails to help the VLMs to focus on the correct objects. Through systematic experiments across diverse model families and language configurations, we show that current RAG systems are less effective on larger VLMs and degrade substantially when queries and retrieved contexts differ in language or modality, highlighting limitations in cross-lingual alignment, retrieval robustness, and multimodal reasoning.

Our key contributions are summarized as follows:

- We introduce the first massive-scale evaluation framework for multilingual multimodal RAG, spanning 42 languages with 56 regional dialects and registers across two cultural VQA datasets: CVQA [50] and WORLDCUISINES [60]. Table 1 compares our framework with existing multilingual and multimodal datasets.[1]
- We conduct a systematic study of retrieval strategies for VLM-based RAG. We find that naive text-based retrieval can degrade performance, while multimodal retrieval provides more reliable gains but does not consistently scale with model size. Furthermore, retrieval relevance correlates with performance but does not guarantee successful evidence integration, particularly for larger models that are less likely to incorporate corrective evidence.
- We perform cross-lingual evaluation across 42 languages and find that current VLMs exhibit significant performance degradation when prompts or retrieved context are provided in non-English languages.

We hope this work provides a foundation for developing RAG systems capable of reasoning seamlessly across languages, modalities, and cultures.

2. M4-RAG

2.1. Tasks and Objectives

We propose M4-RAG, an evaluation framework for multilingual multimodal RAG that prioritizes end-to-end task performance while enabling systematic investigation of when and why retrieval systems help or hinder generation quality. Unlike prior work that evaluates retrieval and generation in isolation, we assess their interaction in realistic multilingual multimodal settings, where questions, images, and knowledge sources may span diverse languages.

Formally, given a VQA instance consisting of a question q in language ℓ_q, an associated image I, and a ground-truth response r^*, along with a multilingual document corpus $\mathcal{C}$ containing relevant factual and cultural knowledge, the system must produce a response r. The RAG pipeline operates in two stages. First, a retriever R_θ selects the top-k most relevant passages from the corpus:

$$R_\theta(q, I, \mathcal{C}) = D_k = \{d_1, d_2, \ldots, d_k\}, \tag{1}$$

where each passage d_i may be in any language ℓ_d, reflecting real-world retrieval scenarios where relevant information is not restricted to the query language. Then, a VLM $\mathcal{M}$ generates a response conditioned on the question, image, and retrieved context:

$$\hat{a} = \mathcal{M}(q, I, D_k), \tag{2}$$

which is evaluated against a^* using task-specific accuracy metrics.

2.2. Evaluation Benchmark Source

We source our VQA pairs from two existing large-scale, culturally-rich datasets, CVQA [50] and WORLDCUISINES [60]. These benchmarks form the foundation of

[1] The dataset is available at https://huggingface.co/datasets/davidanugraha/M4-RAG, and the codebase is released at https://github.com/davidanugraha/M4-RAG.

Dataset	# Size	# Languages	# Dialects	Modality	Domains	Retrieval Source	License
TyDiQA [16]	204k	11	N/A	Text	Multiple†	Wikipedia	N/A
MLQA [35]	46k	7	N/A	Text	Multiple†	Wikipedia	CC-BY-SA 3.0
XOR QA [7]	40k	7	N/A	Text	Multiple†	Wikipedia	CC BY-SA 4.0
MKQA [41]	260k	26	N/A	Text	Multiple†	Wikidata	CC-BY-SA 3.0
Mintaka [51]	20k	9	N/A	Text	Movies, Music, Sports Books, Geography, Politics Video Games, History	Wikidata	CC-BY-4.0
MIRACL [71]	79k	18	N/A	Text	Multiple†	Wikipedia	N/A
AfriQA [47]	12k	10	N/A	Text	Multiple†	Wikipedia	CC-BY-SA 4.0
MIRAGE-Bench [54]	50k	7	N/A	Text	Multiple†	Wikipedia	N/A
ViQuAE [34]	3.6k	1	N/A	Text, Image	Multiple†	Wikipedia	N/A
Encyclopedic VQA [44]	221k	1	N/A	Text, Image	Fine-grained species, Landmarks	Wikipedia	N/A
Xue et al. [63]	500	1	N/A	Text, Image	Genome, Urban	N/A	N/A
UniFashion [73]	260k	1	N/A	Text, Image	Fashion	N/A	N/A
MRAG-Bench [27]	1,353	1	N/A	Text, Image	Multiple†	Wikipedia, ImageNet [21], Flowers102 [46], StanfordCars [32]	N/A
Chart-MRAG Bench [66]	4,738	1	N/A	Text, Image	Family, Race, Politics, Religion, Economy, International Affairs, Internet, Scientific Research	pewresearch	N/A
M4-RAG	80k	42	56	Text, Image	Vehicles, Food, People Sports, Plants & Animals, Objects Brands, Geography, Tradition, Pop Culture	Wikidata, Wikipedia	CC-BY-SA 4.0

Table 1. Comparison of multilingual and multimodal RAG datasets. M4-RAG offers broader linguistic coverage, spanning 42 languages, and explicitly incorporates regional dialects to provide a more fine-grained view of dialectal representation. This enables more precise analysis of cultural and linguistic variation. In addition, our benchmark is released under a permissive open-source license to facilitate reuse and further research. †Details for these entries are not specified in the original papers.

our evaluation by providing high-quality, human-annotated examples that are both multilingual and deeply grounded in cultural context. Cultural knowledge is particularly well-suited for this evaluation since it is inherently long-tail and region-specific, making it unlikely to be reliably encoded in model parameters even for large models, and thus a natural testbed for retrieval augmentation. In total, these datasets cover 42 languages and 56 regional dialects, with further details provided in the Supplementary Materials.

CVQA. CVQA is a multilingual dataset with more than 10,000 VQA pairs spanning 10 diverse cultural categories across 30 countries and 31 languages. We include CVQA to complement the domain-specific nature of WORLDCUISINES, thereby broadening the evaluation landscape with a wider variety of cultural domains and knowledge sources. This diversity is essential for assessing whether RAG systems can retrieve and reason over a broad spectrum of cultural information, rather than performing well only within a single domain.

WORLDCUISINES. WORLDCUISINES is a massive-scale benchmark containing 60k VQA pairs that are parallel across 30 languages and dialects, centered on global cuisine. We select WORLDCUISINES due to its extensive multilingual parallelism that enables controlled analysis of cross-lingual retrieval behavior under consistent semantic content. The dataset also includes intentionally challenging scenarios, such as adversarial prompts where the provided context is misleading, which offers a valuable stress test to examine whether RAG can help generation models to re-anchor their responses in factual evidence instead.

2.3. Knowledge Base Creation

To systematically evaluate retrieval quality, we construct a new, large-scale multilingual knowledge corpus for each evaluation benchmark. These corpora are built from Wikipedia snapshots dated April 2025 to ensure broad thematic coverage and temporal alignment with the creation timelines of the associated datasets. Another reason to use Wikipedia is for its open licensing and redistribution. This alignment helps preserve contextual fidelity, as cultural information and entity descriptions evolve over time [18].

For each VQA instance, we construct a set of multilingual queries that capture complementary types of evidence. These include a question-only query, an answer-only query, and culturally enriched queries that expand an answer with domain-relevant terms (for example, using "Japanese cuisine" for a sushi-related item). These query types are used in combination to maximize corpus coverage, retrieving the top 25 articles independently in English and in the corresponding target language, to ensure that the non-English passages reflect culturally accurate terminology rather than direct query translations. Articles are parsed into sections using Wikipedia's heading structure, preserving semantic coherence within each retrieval unit, then cleaned to remove non-content elements such as scripts, tables, and navigation elements, and deduplicated across queries and languages, yielding 223,468 articles for WORLDCUISINES and 306,794 articles for CVQA.

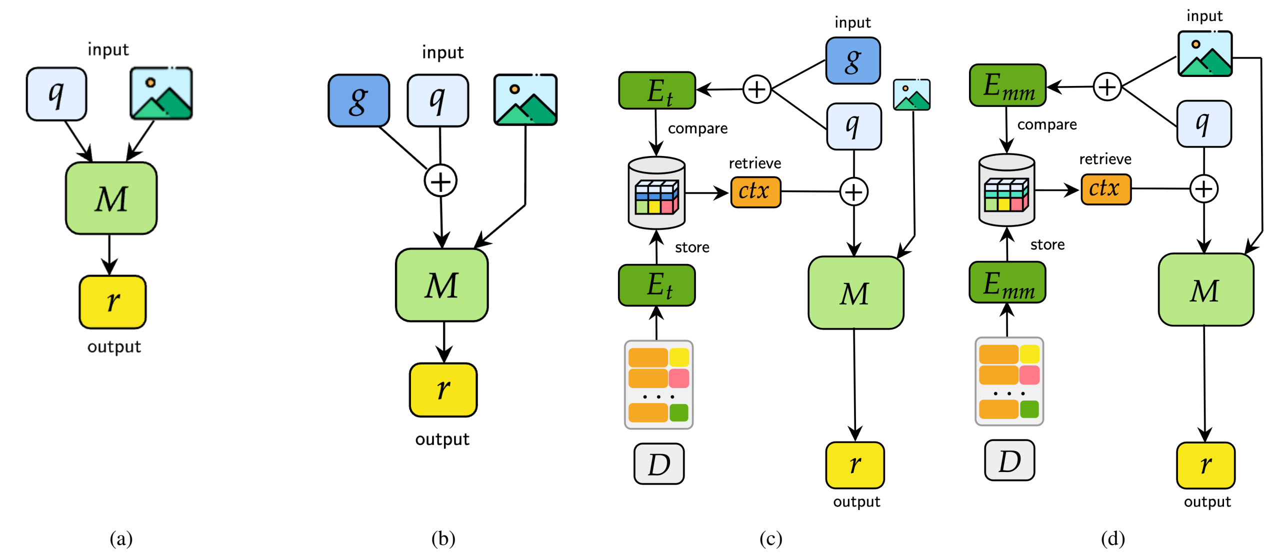

Figure 2. The overall evaluation framework of M4-RAG comprises four configurations: **(a)** a No-RAG baseline, where the VLM (M) directly takes the question and image as input and predicts a response answer; **(b)** a No-RAG setup augmented with oracle context, which is concatenated with the question and image to probe the upper bound of how much perfectly relevant knowledge can help; **(c)** a text-based RAG configuration, where a text encoder (E_t) encodes the query, compares it against an indexed document collection, retrieves the top textual context, and feeds this retrieved text together with the original inputs; and **(d)** a multimodal RAG configuration, where documents are stored with embeddings from a text encoder and retrieval can leverage both textual and visual signals with an image encoder (E_{mm}), yielding richer multimodal context. Across **(c)** and **(d)**, the retrieved context is treated as an additional conditioning signal that steers the model toward culturally relevant knowledge while keeping the backbone VLM architecture unchanged.

3. Experimental Setup

3.1. Retrieval Settings

To assess the impact of retrieval, we evaluate all VLMs under four main configurations, resulting in six experimental variants per model due to different retrieval strategies. For all RAG-based methods, we retrieve the top-k passages with $k = 5$.

1. **Baseline (No Retrieval):** The VLM receives only the question q and image I, without any external context. This serves as a simple baseline to quantify the model's performance without any retrieval assistance.
2. **Oracle Context:** The VLM is provided with the oracle context, representing an upper bound on performance. For WORLDCUISINES, this corresponds to the human-labeled food description from the knowledge base. For CVQA, since no ground-truth evidence passages are provided, we simulate oracle context using a caption generated by `Qwen2.5-VL-72B-Instruct`, conditioned on the image, question, and human-annotated ground-truth answer. This design ensures the caption is tightly grounded in verified information. To validate caption quality, four annotators evaluated 200 randomly sampled image-caption pairs on a 1-5 Likert scale for relevance to the corresponding question, with all samples receiving a score of 5 with full inter-annotator agreement. Further details are provided in the Supplementary Materials.
3. **Text-Based RAG:** Retrieval is performed using textual queries derived from the input, using `E5` [57] as the multilingual dense retrieval. This approach evaluates text-only retrieval by representing visual content through generated captions [25, 40, 72]. Therefore, we consider two settings:
   - *Oracle-Query RAG:* The VLM uses the oracle context as the query to retrieve passages. This provides a reliable textual query and serves as a strong reference point for text-based retrieval.
   - *Caption-Query RAG:* A caption is first generated from image I and question q using `Qwen2.5-VL-72B-Instruct`. The VLM then combines the question q and generated caption to retrieve passages, simulating scenario from multiple past works where the image is converted into text for retrieval [25, 40].
4. **Multimodal RAG:** The question q and image I are used jointly to retrieve passages, leveraging both textual and visual information holistically. We test two multimodal embedding models: `mmE5` (11B) [11] and `B3` (7B) from VLM2Vec [30].

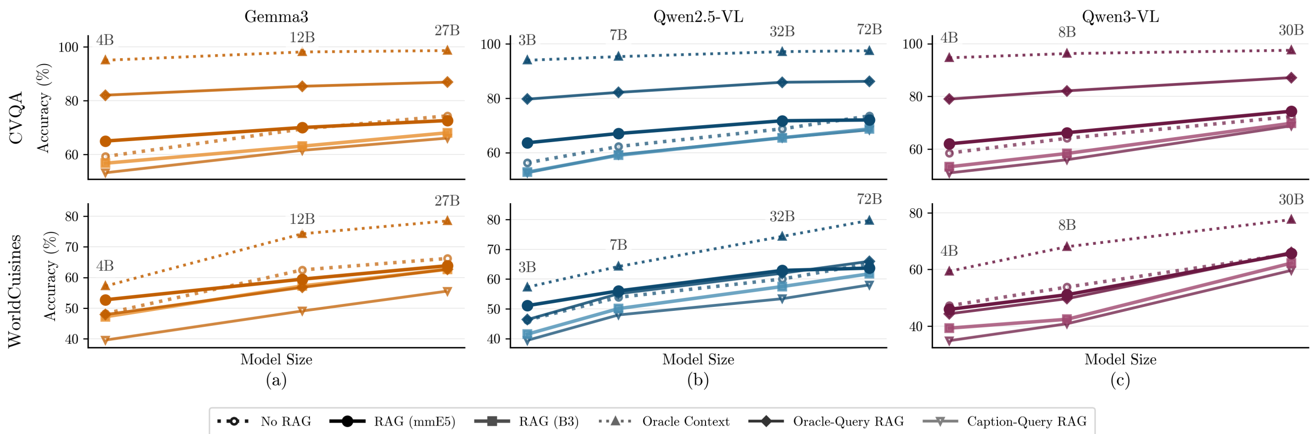

Figure 3. Overall VQA performance on CVQA and WORLDCUISINES across different model families and sizes, with different retrieval configurations. Each column corresponds to a VLM family (`Qwen2.5-VL`, `Gemma3`, `Qwen3-VL`), and each panel plots accuracy as a function of model size. Across all families and scales, adding retrieval (solid lines) consistently improves over the No-RAG baseline (dotted black), with the multimodal RAG variants approaching the oracle context upper bound. Gains are especially pronounced on the more culturally nuanced WORLDCUISINES benchmark, where smaller models with RAG can match or exceed much larger non-RAG models, illustrating that external knowledge is more beneficial than pure parameter scaling in this setting. Among RAG settings, `mmE5`-based retrieval generally outperforms `B3` and caption-query retrieval, highlighting the importance of a strong multimodal encoder and joint use of image and query signals to surface culturally relevant evidence.

3.2. Vision Language Models

We evaluate end-to-end VQA performance across four prominent open-source multilingual VLM families, each available at multiple scales: `Gemma3` [53] at 4B, 12B, and 27B; `Qwen2.5-VL` [9] at 3B, 7B, 32B, and 72B; `Qwen3-VL` with reasoning [65] at 4B, 8B, and 30B-A3B; and `Pangea` [69] at 7B.

3.3. Multilingual VQA

We analyze the impact of language on the VQA task in a cross-lingual experimental setting. We measure how the model's performance changes when the language of its instructional prompts and provided context varies. To do this, we created multilingual versions of our prompts and the oracle contexts for both CVQA and WORLDCUISINES. We use `Gemini-2.5-Flash` to produce high-quality translation of two key components. To ensure their fidelity, all translations were subsequently reviewed and validated by annotators.

- **Multilingual Prompts:** The entire instruction template, including system messages and formatting cues, was translated from English into each of the target languages. This creates the `Multilingual Prompts` setting to see whether models achieve better cultural grounding and task performance when instructions are provided in the native language of the query. We measure this by analyzing the performance change from the English prompt baseline.
- **Multilingual Oracle Context:** Similarly, we created the `Oracle Multilingual Context` setting to investigate whether models perform better on a cultural VQA task when the evidence is provided in the culture's language. For this oracle setup, we translated the oracle English context into each target language. This allows us to isolate the model's ability to perform cultural reasoning when all information is presented in the target language. By comparing this to the English context baseline, we can directly quantify whether models benefit from language that is aligned with the VQA's cultural context.

3.4. Evaluation Metrics

For VLM generations, we use macro-averaged accuracy for all datasets by comparing the multiple choice answer. For annotations we use VLM-as-a-judge using reasoning rubric based since it improves reasoning and more interpretable [5, 6, 33]. The rubrics and prompts for evaluation can be found in the Supplementary Materials.

4. Results and Analysis

4.1. Overall Performance

Figure 3 presents the overall trends across all experiments. `Gemma3 27B` performs the best in both CVQA and WORLDCUISINES for baseline, with accuracy of 74.34% and 66.20%, respectively. As expected, providing the oracle context consistently yields the highest performance across all models and datasets, serving as an upper bound for the quality of contextual information. In this setting, `Gemma3 27B` performs the best in CVQA, while `Qwen-2.5-VL`

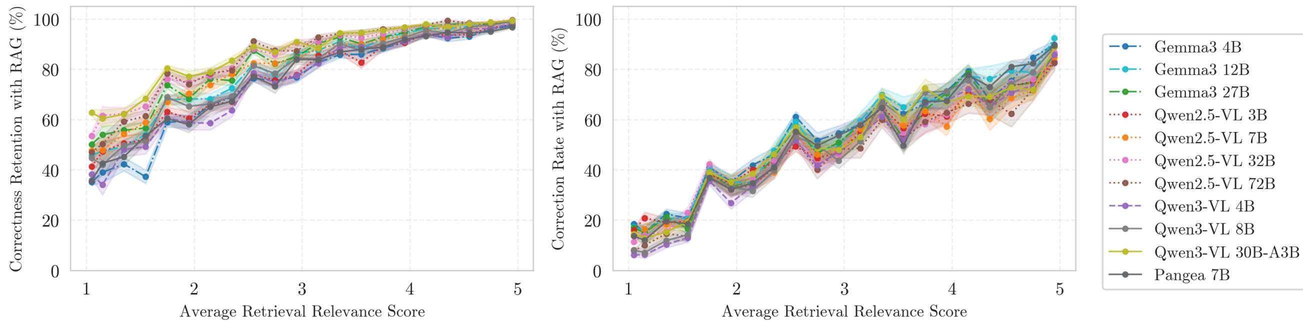

Figure 4. The effect of retrieval quality on RAG performance for various models on the CVQA dataset, using `mmE5` for multimodal retrieval. **Left:** The "Correctness Retention" rate measures the percentage of responses that were correct without RAG and remained correct with RAG. **Right:** The "Correction Rate" measures the percentage of responses that were incorrect without RAG but were successfully corrected by RAG.

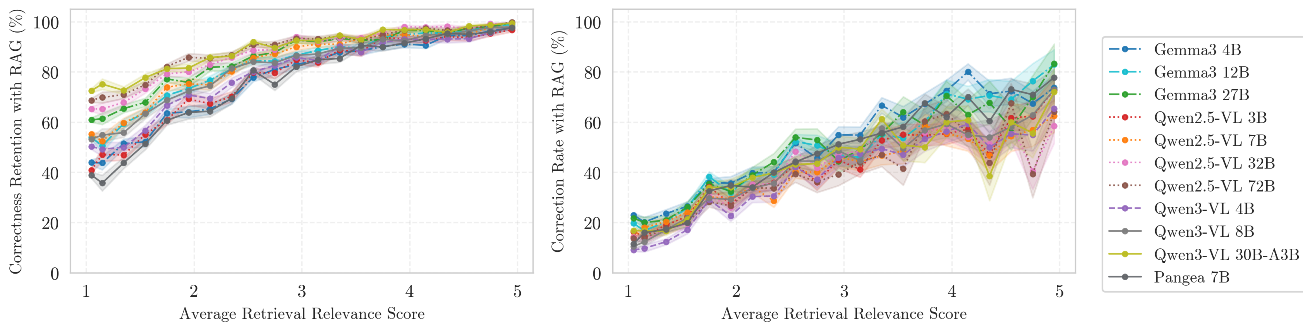

Figure 5. The effect of retrieval quality on RAG performance for various models on the CVQA dataset, using `B3` for multimodal retrieval. **Left:** The "Correctness Retention" rate measures the percentage of responses that were correct without RAG and remained correct with RAG. **Right:** The "Correction Rate" measures the percentage of responses that were incorrect without RAG but were successfully corrected by RAG.

`72B` performs the best when given the oracle context in WORLDCUISINES.

Among retrieval strategies, text-based retrieval performs the worst, even worse than the baselines across model sizes and datasets, indicating that naively converting the image to text can introduce noise that harms VLM performance. In contrast, multimodal retrieval consistently outperforms text-based retrieval, although the `B3` embedding model shows comparatively lower gains. We also observe that reasoning VLMs consistently outperform non-reasoning models of comparable or larger size under RAG settings, suggesting that reasoning capability helps models better integrate retrieved context.

4.2. Model Scaling

Figure 3 illustrates how performance scales with model size across all families. While accuracy generally improves with scale, the benefit of retrieval does not follow the same trend. Text-based RAG performs the worst overall and scales poorly with model size, indicating that this retrieval approach is unlikely to provide meaningful benefits even for larger models.

For multimodal retrieval, `mmE5` and `B3` provide initial gains over the baseline, but these gains do not scale consistently. For larger models, the baseline eventually matches or surpasses multimodal RAG performance, suggesting that models struggle to effectively leverage retrieved context at scale. Reasoning models are more robust to this effect, maintaining retrieval gains longer as scale increases, but the overall trend holds where the slope of improvement diminishes with scale, and larger models show reduced reliance on external context regardless of retrieval strategy. This suggests a fundamental tension between model scale and retrieval utility, where stronger parametric knowledge increasingly competes with rather than complements externally retrieved evidence.

4.3. When Does RAG Succeed and Fail?

In order to understand the effect of retrieval quality on RAG performance, we analyze the quality of the retrieved con-

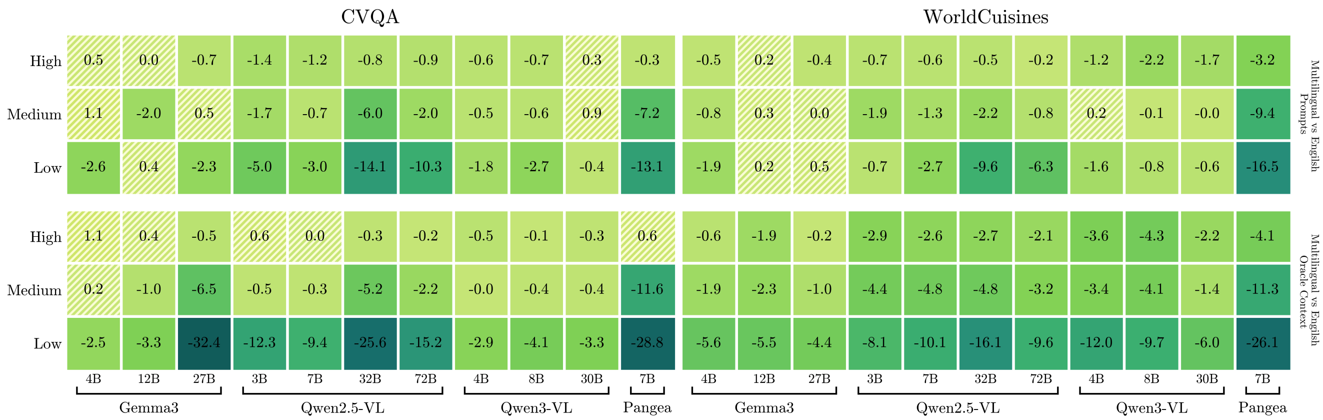

Figure 6. Performance deltas (multilingual − English) in M4-RAG across languages grouped by vitality (high-, medium-, and low-resource). The top rows show the effect of switching from English to multilingual prompts, while the bottom rows show the effect of switching from English to multilingual oracle context. Negative values (darker green) indicate that the multilingual condition performs worse than English, while values near zero or positive (yellowish-green, hatched) indicate stability or gains.

text from both multimodal embeddings on both CVQA and WorldCuisines using VLM-as-a-judge, to see how relevant the retrieved context is with respect to the image, query, and the actual ground truth answer. We use two complementary metrics to support our analysis: *correctness retention*, which is the percentage of initially correct answers that remain correct after RAG context is provided, and *correction rate*, which is the percentage of initially incorrect answers successfully remediated by RAG context. Here, we will only show the plots for CVQA since WorldCuisines also provides the same linear trend, in which figures are shown in the supplementary materials instead.

Figure 4 and Figure 5 illustrate the impact of retrieval quality on the efficacy of RAG systems using `mmE5` and `B3`, respectively, applied to the CVQA dataset. Both metrics show a positive correlation between the average retrieval relevance score with the performance of all tested models. This demonstrates that the relevance of the retrieved context is aligned with the RAG system's overall success. However, their behavior differs in important ways. Correctness retention degrades sharply under poor retrieval, dropping to 40–60% at scores below 2.0, confirming that irrelevant context actively misleads models into abandoning correct answers. As relevance improves, retention converges tightly across all models toward near-perfect rates (95–100%), suggesting that high-quality retrieval reliably reinforces correct parametric knowledge regardless of model family or size.

The correction rate tells a different story. While high-relevance context enables models to fix 80–90% of original errors, the correction rate never saturates to the same degree as retention, and model spread remains wide even at the highest relevance scores. This asymmetry indicates that leveraging retrieved evidence to overturn a wrong answer is fundamentally harder than preserving a correct one, and that current VLMs still struggle to reliably integrate externally retrieved evidence even when it is entirely correct. This gap is more pronounced with `B3` than `mmE5`, where wider inter-model variance suggests that weaker retrievers amplify differences in context integration ability across models.

Beyond the main trend, both plots reveal an important scaling effect. Larger models, exemplified by `Qwen2.5-VL 72B` and `Gemma3 27B`, exhibit greater reliance on their parametric knowledge. In the left plot, large models form the upper boundary of correctness retention (i.e., they more often preserve correct baseline answers), while in the right plot they frequently form the lower boundary of correction rate (i.e., they are less likely to change an incorrect baseline when given high-quality retrieved evidence). On the other hand, smaller model would take the counterparts, respectively on both multimodal embedding strategies. Taken together, these patterns indicate that model scale increases inertial priors, that is stronger internal beliefs that are less readily updated by external context. In other words, larger models show reduced context integration (or lower context susceptibility) as they are less prone to be misled by poor retrieval, but at the same time, also less likely to adopt corrective information supplied by good retrieval. This phenomenon suggests a potential point of diminishing returns for RAG, where beyond a certain scale, improvements in model capacity do not straightforwardly translate to better utilization of retrieved evidence.

4.4. Multilingual Performance Gaps

Our cross-lingual experiments reveal a strong English-centric bias in current VLMs. As shown in Figure 6, shifting from English to multilingual prompts consistently degrades performance across all resource levels, though the effect is relatively mild for high-resource languages (mostly within

-1% to -2%) and becomes more severe for low-resource languages. This indicates that models best interpret task instructions in English regardless of the cultural context of the query.

The multilingual gap is far more pronounced even when the oracle context is provided in the target language rather than English. Contrary to the intuition that culturally aligned context should help, performance deteriorates dramatically, with drops as large as -32.4% for `Qwen2.5-VL 32B` on CVQA and -28.8% on `Pangea` on CVQA for low-resource languages. Note that `Pangea` is a model explicitly trained on multilingual and multicultural Wikipedia data, yet it is still among the most severely affected, suggesting that exposure to multilingual training data does not straightforwardly confer robustness to non-English retrieved context at inference time. This asymmetry between prompt switching and context switching indicates that models can tolerate non-English instructions to some extent, but fail much more severely when the retrieved evidence itself is in a non-English language, suggesting that cross-lingual evidence integration is a deeper bottleneck than instruction following.

This trend is not uniform across model families. The `Qwen` family exhibits a sharper collapse in low-resource settings compared to `Gemma`, showing that model scale alone does not resolve this bias. `Pangea` also shows some of the largest drops. An interesting observation is that smaller models show a lesser performance drop overall because they tend to code-switch to English even when prompted in a target language, whereas larger models attempt to respond fully in the target language and fail more dramatically as a result.

5. Related Work

Culture has been studied across multiple dimensions, including social norms [23], country-specific variation [26, 45, 50], general world knowledge [71], and food-related knowledge [2, 26, 38, 42, 60, 67]. Understanding these facets is crucial for building AI systems that can reason appropriately across diverse cultural contexts. Early work on multicultural RAG has largely relied on machine-translated benchmarks [17, 36]. While this approach enables broader language coverage, it often fails to fully leverage high-quality human-curated resources and can introduce translation artifacts that obscure culturally specific nuances.

Several multilingual text-only retrieval datasets have been proposed, including MINERS [59], Mintaka [51], MIRACL [71], MKQA [41], and MLQA [35]. These datasets span diverse domains such as books, geography, politics, and general knowledge, providing broad language coverage and enabling progress in multilingual natural language understanding. However, their focus on textual information limits their applicability to tasks that require multimodal reasoning [29, 56]. Tasks such as visual question answering [4, 61, 62], image-grounded reasoning [52, 58, 61], and culturally contextualized visual understanding remain largely unsupported. This gap highlights the need for benchmarks that jointly integrate multilingual and multimodal information [13]. Recent efforts toward evaluating retrieval-augmented multimodal systems include MRAG-Bench [28], MRAMG-Bench [68], MIRAGE-Bench [70], BordIRLines [37], and BERGEN [49], but these benchmarks focus primarily on English settings.

At the same time, vision-language models (VLMs) have rapidly expanded their multilingual capabilities. Large proprietary or large-scale models such as Qwen2.5-VL [64], Qwen3-VL [8], Gemma3 [53], and PaLI / PaLI-X [12] demonstrate strong cross-modal reasoning across multiple languages. In parallel, a growing ecosystem of open multilingual VLMs has emerged, including InternVL [14], mBLIP [24], PALO [43], Maya [3], Aya Vision [19], and PaliGemma [10]. Despite these advances, systematic evaluation of multilingual VLMs on knowledge-intensive, culturally grounded multimodal tasks remains limited, highlighting the need for benchmarks that assess multilingual multimodal reasoning in realistic retrieval-augmented settings.

To address these gaps, we introduce M4-RAG that combines multilingual and multimodal inputs, enabling end-to-end evaluation across languages with comprehensive experimental analysis.

6. Conclusion

In this work, we present M4-RAG, a massive-scale multilingual and multimodal benchmark spanning 42 languages, 56 regional dialects and registers, and over 80,000 culturally grounded image-question pairs for evaluating RAG in realistic settings. We conduct comprehensive experiments across 11 models and 3 retrieval configurations, including cross-lingual settings, to characterize the behavior of multilingual multimodal RAG systems. Our experiments reveal that while RAG reliably boosts smaller VLMs, larger models exhibit diminished correction rates and stronger reliance on parametric knowledge, and that multilingual context degrades performance even in models explicitly trained on multilingual data. These findings suggest that the fundamental challenge is not whether to retrieve, but how to enable effective integration of retrieved information, i.e. since larger models retain correct answers but fail to leverage retrieval for error correction, this may indicate misalignment between retrievers and foundation models. We therefore advocate for model-aware retrieval strategies that optimize for integration utility rather than query relevance alone, through directions such as joint retriever-VLM post-training or test-time adaptation. We hope M4-RAG serves as a foundation for future work toward RAG systems that reason robustly across languages, modalities, and cultural contexts.

References

[1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025. 2

[2] Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. Indifoodvqa: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176, 2024. 8

[3] Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*, 2024. 8

[4] Stanislaw Antol, Aishwarya Agarwal, Jiasen Lu, et al. Vqa: Visual question answering. In *Proceedings of ICCV*, 2015. 8

[5] David Anugraha, Shou-Yi Hung, Zilu Tang, Annie En-Shiun Lee, Derry Tanti Wijaya, and Genta Indra Winata. mr3: Multilingual rubric-agnostic reward reasoning models. *arXiv preprint arXiv:2510.01146*, 2025. 5

[6] David Anugraha, Zilu Tang, Lester James V Miranda, Hanyang Zhao, Mohammad Rifqi Farhansyah, Garry Kuwanto, Derry Wijaya, and Genta Indra Winata. R3: Robust rubric-agnostic reward models. *arXiv preprint arXiv:2505.13388*, 2025. 5

[7] Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 547–564, 2021. 3

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 8

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 8

[11] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*, 2025. 4

[12] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[13] Ziyi Chen, Shuming Mohan, et al. X-vnli: Evaluating cross-lingual visio-linguistic intelligence. In *Proceedings of EMNLP*, 2023. 8

[14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 8

[15] Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, 2024. 2

[16] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. 3

[17] Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, 2024. 8

[18] Nicole Creanza, Oren Kolodny, and Marcus W Feldman. Cultural evolutionary theory: How culture evolves and why it matters. *Proceedings of the National Academy of Sciences*, 114(30):7782–7789, 2017. 3

[19] Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025. 2, 8

[20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 2

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[22] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *ICLR*, 2025. 2

[23] Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, 2023. 8

[24] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mblip: Efficient bootstrapping of multilingual vision-llms. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, 2024. 8

[25] Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. *arXiv preprint arXiv:2510.14605*, 2025. 4

[26] Tianyi Hu, Maria Maistro, and Daniel Hershcovich. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, 2024. 8

[27] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024. 3

[28] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. MRAG-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[29] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of CVPR*, 2019. 8

[30] 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. 4

[31] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020. 1, 8

[32] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3

[33] Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pages 11286–11315, 2024. 5

[34] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3108–3120, 2022. 3

[35] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7315–7330, 2020. 3, 8

[36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 2, 8

[37] Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. Bordirlines: A dataset for evaluating cross-lingual retrieval augmented generation. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 1–13, 2024. 8

[38] Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, 2024. 8

[39] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024. 2

[40] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022. 4

[41] Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021. 3, 8

[42] Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5674–5685, 2023. 8

[43] Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*, 2024. 8

[44] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124, 2023. 3

[45] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024. 2, 8

[46] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008*

Sixth Indian conference on computer vision, graphics & image processing, pages 722–729. IEEE, 2008. 3

[47] Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, et al. Afriqa: Cross-lingual open-retrieval question answering for african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, 2023. 3

[48] Jeonghyun Park and Hwanhee Lee. Investigating language preference of multilingual rag systems. *arXiv preprint arXiv:2502.11175*, 2025. 2

[49] David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. Bergen: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, 2024. 8

[50] 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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 11479–11505, 2024. 2, 8

[51] Priyanka Sen, Alham Fikri Aji, and Amir Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, 2022. 3, 8

[52] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Yonatan Bisk, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of ACL*, 2019. 8

[53] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2, 5, 8

[54] Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad. Mirage-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 274–298, 2025. 3

[55] Ashish V Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, 2022. 2

[56] Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, et al. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32967–32982, 2025. 8

[57] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024. 4

[58] Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. Learning visual grounding from generative vision and language models. In *Proceedings of CVPR*, 2025. 8

[59] Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. Miners: Multilingual language models as semantic retrievers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2742–2766, 2024. 2, 8

[60] Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, 2025. 2, 8

[61] Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*, 2025. 8

[62] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[63] Junxiao Xue, Quan Deng, Fei Yu, Yanhao Wang, Jun Wang, and Yuehua Li. Enhanced multimodal rag-llm for accurate visual question answering. *arXiv preprint arXiv:2412.20927*, 2024. 3

[64] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2, 8

[65] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5

[66] Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*, 2025. 3

[67] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Foodlmm: A versatile food assistant using large multi-modal model. *IEEE Transactions on Multimedia*, 2025. 8

[68] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. Mramg-bench: a comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3616–3626, 2025. 8

[69] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024. 5

[70] Weichen Zhang, Yiyou Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. Mirage-bench: Llm agent is hallucinating and where to find them. *arXiv preprint arXiv:2507.21017*, 2025. 8

[71] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11: 1114–1131, 2023. 2, 3, 8

[72] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*, 2023. 4

[73] Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. Unifashion: A unified vision-language model for multimodal fashion retrieval and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1490–1507, 2024. 3