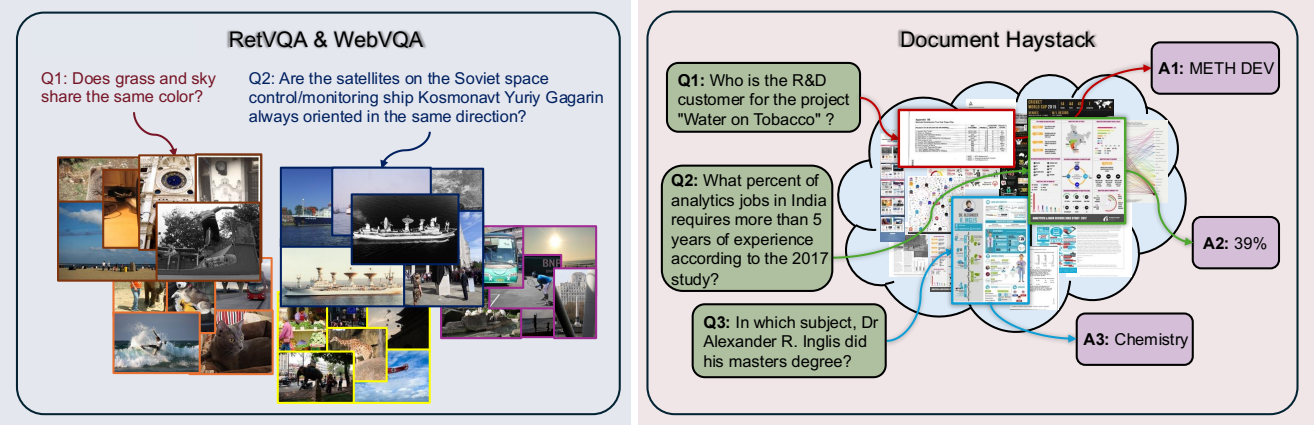
Document Haystacks: Vision-Language Reasoning Over Piles of 1000+ Documents

Jun Chen[1]*, Dannong Xu[2]*, Junjie Fei[1]*, Chun-Mei Feng[3], Mohamed Elhoseiny[1]
[1]King Abdullah University of Science and Technology
[2]The University of Sydney, [3]IHPC, A*STAR
{jun.chen,junjie.fei,mohamed.elhoseiny}@kaust.edu.sa
daxu8019@uni.sydney.edu.au, fengcm.ai@gmail.com



(a) Previous Benchmarks: Each question paired with a limited image set

(b) Our Benchmarks: All questions mapped to an extensive document collection

Figure 1. **Comparison between previous and proposed benchmarks.** Given a question as input, all benchmarks aim to retrieve relevant images from an image pool to correctly answer the question. Unlike prior benchmarks like RetVQA [32] and WebVQA [7], which structure their datasets by pairing each question with a limited set of images (typically ≤ 30), our benchmarks, DocHaystack and InfoHaystack, map each question to a substantially larger document collection, scaling up to 1,000 visual documents. This expanded scope more accurately represents large-scale document retrieval scenarios and offers a greater challenge in retrieval accuracy and visual question answering.

Abstract

Large multimodal models (LMMs) have achieved impressive progress in vision-language understanding, yet they face limitations in real-world applications requiring complex reasoning over a large number of images. Existing benchmarks for multi-image question-answering are limited in scope, each question is paired with only up to 30 images, which does not fully capture the demands of large-scale retrieval tasks encountered in the real-world usages. To reduce these gaps, we introduce two document haystack benchmarks, dubbed DocHaystack and InfoHaystack, designed to evaluate LMM performance on large-scale visual document retrieval and understanding. Additionally, we propose V-RAG, a novel, vision-centric retrieval-augmented generation (RAG) framework that leverages a suite of multimodal vision encoders, each optimized for specific strengths, and a dedicated question-document relevance module. V-RAG sets a new standard, with a 9% and 11% improvement in Recall@1 on the challenging DocHaystack-1000 and InfoHaystack-1000 benchmarks, respectively, compared to the previous best baseline models. Additionally, integrating V-RAG with LMMs enables them to efficiently operate across thousands of images, yielding significant improvements on our DocHaystack and

* Equal contribution

InfoHaystack benchmarks. Our code and datasets are available at https://github.com/Vision-CAIR/dochaystacks

1. Introduction

Large Multimodal Models (LMMs) [1, 20, 30, 41] have made remarkable progress in the vision-language understanding. However, these models still face challenges when tasked with reasoning over extensive collections of images or documents [43], limiting their effectiveness in real-world applications, such as visual search or querying over large sets of images or documents, like those stored on personal devices or in photo albums. However, there lacks such proper benchmarks to evaluate these capabilities. To address this gap, we introduce the DocHaystack and InfoHaystack benchmarks, designed to evaluate LMMs on large-scale image retrieval and understanding capabilities, pushing the boundaries of LMM performance in complex, real-world scenarios.

The existing multi-image retrieving and reasoning benchmarks are primarily constructed on a small scale, as highlighted in works such as [32, 37]. Each question in these benchmarks is paired with only up to 30 images as illustrated in Figure 1 (a). However, this limited scope does not align well with real-world scenarios, which often require retrieval and reasoning across hundreds or thousands of images or documents. In contrast, our established benchmarks, depicted in Figure 1 (b), allow for querying questions from a large-scale collection of up to 1,000 documents, necessitating that models retrieve and reason over an extensive set of documents for each question. This scale better simulates practical applications and their demands.

The main challenge in constructing such benchmarks is collecting specific questions while ensuring there are no ambiguous answers across a large set of images. Existing datasets, such as those in DocVQA and InfographicVQA [27, 28], contain numerous "general" questions, like "What is the table number?", where answers could be derived from multiple images, leading to non-unique responses. To address this, we implemented a rigorous data filtering pipeline. First, we employed both a large language model (LLM) and human annotators to systematically filter out "general" questions based on carefully defined criteria. Additionally, we used the LLM to exclude questions relying on generic knowledge, such as "What is the capital of Missouri?", which can be answered without image context. This approach ensures that the questions in the benchmark can only be answered through specific visual cues from the provided images, maintaining the benchmark's integrity for evaluating image-based understanding.

To enable the current LMMs effectively reason over a large number of images, we propose a vision-centric retrieval-augmented generation (RAG) framework, named V-RAG. V-RAG combines multiple multimodal vision encoders, leveraging each encoder's unique strengths to enhance retrieval accuracy. Additionally, it incorporates an LMM-filter module to assess the relevance of each document to the query, refining the retrieval process by ensuring that only relevant documents are prioritized. This integrated approach allows V-RAG to navigate extensive document collections efficiently. Experimental results demonstrate that V-RAG achieves 9% and 11% improvement in Recall@1 on the DocHaystack-1000 and InfoHaystack-1000 compared to previous best text-to-image retrieval methods. Additionally, we found that integrating V-RAG brings GPT-4o over a 55% acc improvement on DocHaystack-200 and a 34% acc improvement on InfoHaystack-200, indicating the effectiveness of our V-RAG.

Our contributions are as follows:

- We introduce the Document Haystack benchmarks, including DocHaystack-100/200/1000 and InfoHaystack-100/200/1000, with the most challenging setup consisting of 1,000 documents for each inquiry. These benchmarks advance document retrieval and reasoning tasks by requiring models to navigate and reason across extensive document collections, surpassing prior benchmarks limited to smaller retrieval tasks.
- We propose a vision-centric retrieval-augmented generation framework, V-RAG, which enhances the retrieval capabilities of LMMs. V-RAG achieves substantial improvements over previous best text-to-image retrieval methods by 9% and 11% on DocHaystack-1000 and InfoHaystack-1000, respectively.

2. Related Works

VQA benchmarks. VQA play a critical role in assessing a model's ability to understand and reason across visual contexts [6, 11, 12]. Traditional VQA datasets typically measure a model's comprehension of object attributes [15, 19], spatial relationships [15], as well as its understanding of documents [27, 28], charts [26], mathematics [23, 40, 46], and open knowledges [25, 35]. Additionally, these benchmarks explore models' knowledge across varied fields, including science and the arts [22, 44]. This broad array of benchmarks has greatly advanced vision-language models by cultivating diverse visual comprehension skills, particularly for modern foundation models in vision-language understanding [1, 3, 8, 20, 22, 30, 31, 39, 47]. Notably, these benchmarks have primarily focused on question answering within single image or document. In contrast, our benchmark shifts the focus towards retrieval and comprehensive understanding across a large collection of visual documents, presenting new challenges and expanding the scope of visual question answering.

Several previous efforts have tackled the challenge of vi-

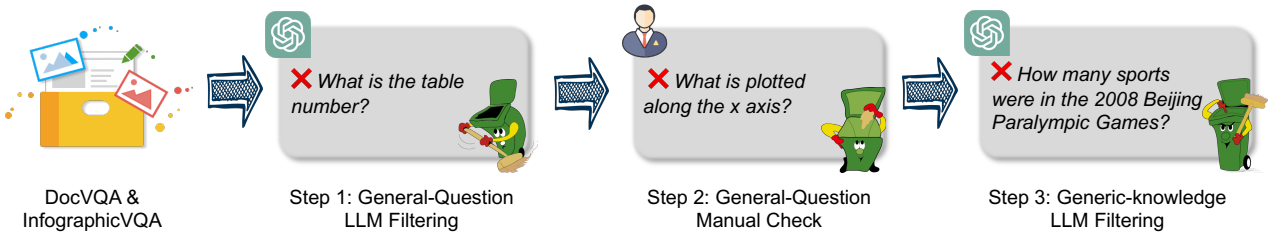


Figure 2. **Data Curation Pipeline.** Our benchmarks are curated based on the DocVQA and InfographicVQA datasets, following a three-step filtering process to obtain document-specific question-answer pairs. In Step 1, we filter out general questions (e.g., "What is the table number?"), as these could be answered by multiple documents and lack specificity. Step 2 involves a manual review by human annotators to further remove general questions. In Step 3, we eliminate generic-knowledge questions (e.g., "How many sports were in the 2008 Beijing Paralympic Games?") that can be answered directly by large language models without requiring image input."

sual question answering and reasoning across multiple images [5, 7, 32, 37, 38, 42]. For instance, datasets such as MultimodalQA [37] and ISVQA [5] require models to have multi-image reasoning abilities. Meanwhile, WebQA [7] and RetVQA [32] involve an additional step where models must first retrieve relevant images from a limited image pool before answering visual questions based on these results. However, these benchmarks are generally constrained to relatively small image pools, where each question is paired with an image set containing up to 30 images. In contrast, our proposed benchmarks, DocHaystack and InfoHaystack, significantly expand this scope by requiring models to retrieve and reason from a much larger set of up to 1,000 documents, presenting a notably greater challenge in retrieval and multi-image reasoning.

Large multimodal models (LMMs). LMMs have achieved substantial advancements in understanding and reasoning across single or multiple images [1, 8, 20, 30, 41, 47]. These models have significantly enhanced vision-language understanding across numerous dimensions and applications [12, 23, 44, 46]. LMMs benefit primarily from large-scale image-text alignment and extensive language modeling, which emerge them with advanced understanding and reasoning abilities. However, despite these breakthroughs, LMMs still encounter challenges when handling large-scale image or document sets [43]. This difficulty is due to the inherent complexity of processing such complex data. To address this, retrieval-based approaches have been developed to extend the capacity of vision-language models, augmenting their ability to process and reason over a larger number of images.

Retrieval-augmented generation (RAG). RAG integrates retrieval systems [4, 11, 16, 33, 45], with generative models, enhancing them with additional knowledge. While RAG has been extensively explored in language domains [2, 13, 17, 24], its application in vision-language contexts is also advancing. In vision-language RAG, models like MuRAG [9] leverage image-text memory to retrieve top-k neighbors by comparing inner-product similarities. RetVQA [32] uses an image-question relevance encoder, combining BERT [10] and Faster R-CNN [34] to filter relevant images, while MIRAGE [43] employs a CLIP-based encoder [33] to train a retriever. These frameworks extend model capabilities, enabling retrieval and reasoning across hundreds or thousands of images. In contrast, we propose V-RAG, a vision-centric RAG framework that integrates multiple vision encoders to more effectively capture image features, and introduces a LMM-based question-document relevance comparison module. Our results demonstrate that V-RAG surpasses existing methods on our DocHaystack and InfoHaystack benchmarks, setting a new standard for large-scale visual retrieval and reasoning.

3. DocHaystack and InfoHaystack Benchmarks

To support effective retrieval and reasoning across extensive document collections, we present two new benchmarks—DocHaystack and InfoHaystack—designed to ensure each question yields a unique, document-specific answer. Derived from DocVQA [28] and InfographicVQA [27], these benchmarks address the challenge of answer ambiguity by selectively curating questions that can only be answered by a single document within a large dataset.

Benchmark construction pipeline. There exists many general questions in the existing benchmarks and lead to multiple answers for different document context. For example, general questions like "`What is the table number?`" may apply to various documents and yield multiple valid answers, while a targeted question like "`Who is the reviewer for the article titled 'An antithyroid factor in milk'?`" is likely to produce a unique answer, as only a single document or a limited set of documents would contain that information. Therefore, our benchmark construction follows a structured three-step filtering pipeline, illustrated in Figure