



VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents

Supplementary Material

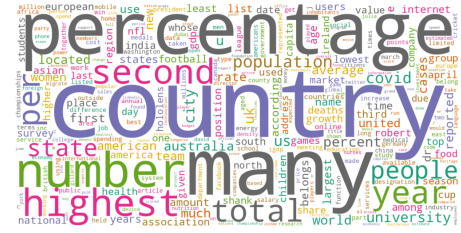
| Statistics | Number |
|---------------------------------|----------------|
| Total Images | 231,339 |
| Total Questions | 43,474 |
| - Single-Hop Questions | 33,244 (76.5%) |
| - Multi-Hop Questions | 10,230 (23.5%) |
| - Extractive Answer | 19,797 (45.5%) |
| - Abstractive Answer | 23,677 (54.5%) |
| QA Source Datasets | 9 |
| - Existing DocumentVQA Datasets | 7 |
| - Existing TableQA Datasets | 1 |
| - Our Newly Created Datasets | 1 |
| Maximum Question Length | 58 |
| Maximum Answer Length | 130 |
| Average Question Length | 13.7 |
| Average Answer Length | 3.7 |

Table A. Main statistics in OpenDocVQA.

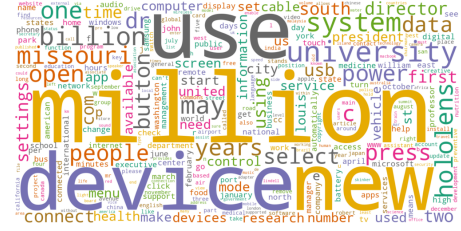
A. OpenDocVQA Details

Dataset Statistics. The main statistics of OpenDocVQA are presented in Table A. There are two types of questions: single-hop (45.5%) and multi-hop (23.5%). Answers to questions are categorized as extractive (45.5%) and abstractive (54.5%) types. OpenDocVQA consists of nine open-domain DocumentVQA datasets, including a newly created MHDocVQA dataset to address multi-hop questions over multiple documents, and collected and filtered QA datasets as follows.

- **DocVQA** [42] includes industry document images collected from the UCSF Industry Document Library.
- **InfoVQA** [43] includes infographics downloaded from the Internet for the search query “infographics”.
- **VisualMRC** [56] is a visual machine reading comprehension on webpage screenshot images.
- **ChartQA** [41] is a chart understanding dataset with human-written and machine-generated questions focusing on visual and logical reasoning.
- **OpenWikiTable** [27] is an open-domain question answering over tables. We took screenshot images of the tables, converting them into images with complex text layouts to handle visually-rich table data.
- **DUDE** [28] is a multi-page, multi-domain, and multi-industry QA dataset that requires processing long documents and understanding different types of documents.
- **MPMQA** [68] requires comprehending multimodal content in an entire product manual and answering questions.



(a) Word cloud of questions.



(b) Word cloud of answers.

Figure A. Word cloud distributions of question and answer texts.

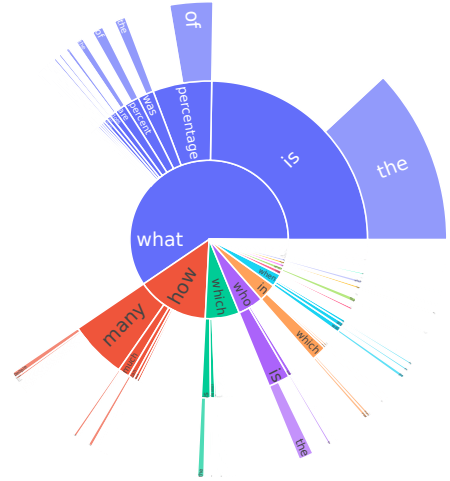


Figure B. Distribution of first three words of the question.

- **SlideVQA** [57] requires multi-hop reasoning over multiple slide images containing various text formats, layouts, and visual content such as plots and charts.

Figure A presents word clouds of the most frequently appeared words in the question and answer texts, illustrating that OpenDocVQA covers a wide range of topics and words. This observation is further supported by Figure B, which is a sunburst of the first three words of the questions.

Filtering DocumentVQA datasets. We applied the following five heuristic rules to automatically filter out likely

Multi-hop Question Generation Prompt

```
EXAMPLE1:
question1: In which country is the GWP smallest?
answer1: Denmark
question2: What is the staple diet of Denmark?
answer2: Fish, cheese
combined question: What is the staple diet of the country where the GWP is the smallest?

EXAMPLE2:
question1: To which League does Chicago Cubs belong?
answer1: mlb
question2: What is the average MLB team value?
answer2: $1.5b
combined question: What is the average the league where Chicago Cubs belongs to team value?

EXAMPLE3
question1: Which is the capital city of Germany?
answer1: Berlin
question2: What year did Berlin host the OKFestival?
answer2: It's 2014.
combined question: What year did the capital city of Germany host the OKFestival?

Based on the above 3 examples, provide a combined question for the following case,
such that the answer to the combined question is the same as the answer2:
question1: {single-hop question}
answer1: {single-hop answer}
question2: {single-hop question}
answer2: {single-hop answer}
combined question:
```

Table B. Multi-hop question generation prompt. “{single-hop question}” and “{single-hop answer}” are placeholders of two single-hop questions.

Multi-hop Question Filtering Prompt

```
question1: {single-hop question}
answer1: {single-hop answer}
question2: {single-hop question}
answer2: {single-hop answer}

Based on the questions and answers above, please answer the following question shortly.
If the answer is not identified, the answer is 'None': {multi-hop question}
```

Table C. Multi-hop question filtering prompt. “{single-hop question}” and “{single-hop answer}” are placeholders of two single-hop questions. “{multi-hop question}” denotes the generated multi-hop questions.

context-dependent questions:

- The question has one or more demonstrative pronouns, including “this”, “these”, and “those”.
- The question has one or more personal pronouns, including “she”, “he”, “her”, “his”, and “him”.
- The question has one or more specific keywords, including “the document” and “mention”.
- The question does not contain entities except for numbers.
- The question is shorter than six words.

Any samples matching at least one of these rules were removed from our dataset. After applying the rules, we

manually reviewed all the questions to ensure context-independence, guided by the instruction: “*When you see the question without a given document, can you find a unique document in the corpus to provide a unique answer?*”. To validate our review, we randomly sampled 50 questions with their gold and top-5 retrieved documents (from VDocRetriever) and found no ambiguous cases, confirming the high quality of our process.

Prompts for creating multi-hop questions. Table B shows the prompt for combining two single-hop questions

| Dataset | Task Description |
|---------------|---|
| DocVQA | You have to find an industry document that answers my question. |
| InfoVQA | Given a question, retrieve an infographic to answer the question. |
| VisualMRC | I’m looking for a screenshot image that answers the question. |
| ChartQA | Given a user query, retrieve a chart image that answers the query. |
| OpenWikiTable | Given a user query, retrieve a table image for answering the question. |
| DUDE | You need to retrieve evidence from a PDF page to address the question. |
| MPMQA | I want to know the answer to the question. Can you find evidence from manual pages? |
| SlideVQA | Given a question, retrieve a slide image to answer the question. |
| MHDocVQA | Given a multihop-question, retrieve multiple pages that can help answer the question. |

Table D. Instructions in the visual document retrieval task.

| Model | Model Checkpoint |
|-------------|--------------------------------------|
| Contriever | facebook/contriever-msmarco |
| E5 | intfloat/e5-base-v2 |
| GTE | thenlper/gte-base |
| E5-Mistral | intfloat/e5-mistral-7b-instruct |
| NV-Embed-v2 | nvidia/NV-Embed-v2 |
| CLIP | openai/clip-vit-large-patch14-336 |
| DSE | Tevatron/dse-phi3-docmatix-v1 |
| VisRAG-Ret | openbmb/VisRAG-Ret |
| Phi3V | microsoft/Phi-3-vision-128k-instruct |
| Idefics3 | HuggingFaceM4/Idefics3-8B-Llama3 |

Table E. Model checkpoints stored on HuggingFace.

| Hyperparameters | Value |
|----------------------------|--------|
| Learning Rate | 1e-4 |
| Gradient Accumulation | 4 |
| Adam W β_1 | 0.9 |
| Adam W β_2 | 0.999 |
| LoRA Attention Dimension r | 8 |
| LoRA Scaling Alpha | 64 |
| LoRA Dropout | 0.1 |
| LoRA Target | *.proj |
| BF16 | True |

Table F. Hyperparameters used for pre-training and fine-tuning.

to generate multi-hop questions. Moreover, Table C shows the prompt for filtering the generated multi-hop questions.

B. Experimental Details

Instruction templates. Following a standard LLM-based retrieval training and evaluation strategy [60], we applied natural language instruction templates to the original question for the visual document retrieval task:

Instruct: {task description} \n Query: {question},

where “{task description}” is a placeholder for a one-sentence task description as shown in Table D. Note that the instruction format was applied to only LLM-based retrievers, including E5-Mistral [60], NV-Embed-v2 [30],

| Max Image Resolution | Retrieval nDCG@5 | Encoding Time | QA ANLS | QA Generation Time |
|----------------------|------------------|---------------|---------|--------------------|
| 336×336 | 28.7 | 85.0 | 37.2 | 394.5 |
| 672×672 | 72.8 | 106.4 | 42.7 | 490.9 |
| 1344×1344 | 72.9 | 204.4 | 56.2 | 789.7 |

Table G. Impact of image resolution on InfoVQA under the single-pool setting. Average time (ms) to encode a single document or generate a single answer is measured on a single A100 GPU.

DSE [37], Phi3 [1], and VDocRetriever. Our preliminary experiments observed that using the instruction during both training and evaluation improved the performance of LLM-based retrievers. However, applying the same instruction format to non-LLM-based retrievers, such as Contriever [22], resulted in a performance decline due to lacking instruction-following capabilities. Furthermore, we appended an instruction regarding the desired output format for the DocumentVQA task:

\n Answer briefly.

Model checkpoints Table E shows model initialization checkpoints stored on HuggingFace ¹.

Model hyperparameters Table F lists hyperparameters in pre-training and fine-tuning used for our models.

C. Additional Experimental Analysis

How does image resolution impact performance? Table G shows that increasing image resolution improved the model’s capability to understand and encode the document; however, it also significantly increased the inference time for both retrieval and QA tasks. Moreover, the performance in the QA task exhibited greater sensitivity to image resolution compared to the retrieval task, indicating that the QA task demands more detailed visual understanding.

¹<https://huggingface.co>

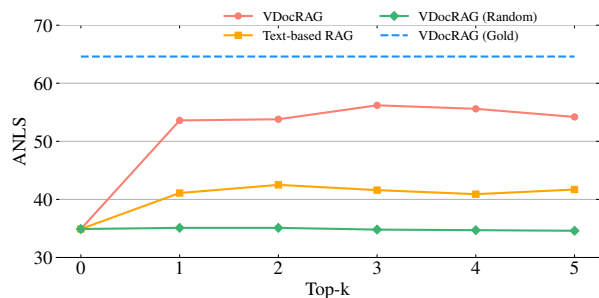


Figure C. QA performance with various top-k on InfoVQA under the single-pool setting. () denotes document sources.

How many retrieved documents to augment? Figure C shows that incorporating three documents yielded the best results in VDocRAG. While adding a few documents may include helpful contexts, adding more low-ranked or randomly sampled documents introduces noise and deteriorates generation due to the imperfections of retrievers.

Additional qualitative results. Figure D shows qualitative results of VDocRAG compared to text-based RAG. VDocRAG demonstrates significant performance advantages in understanding layouts and visual content, such as tables, charts, figures, and diagrams. These findings highlight the critical role of representing documents as images to improve the performance of the RAG framework.

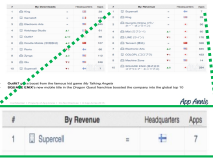
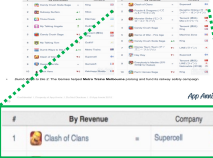
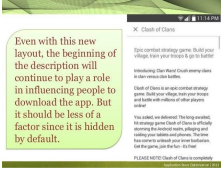

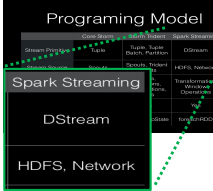
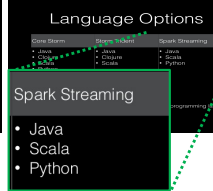
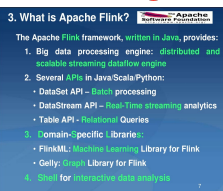
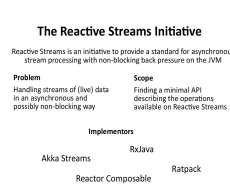
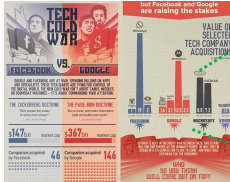





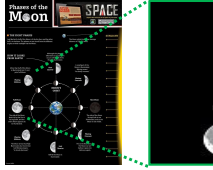
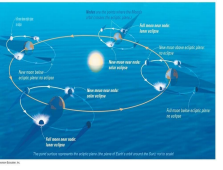
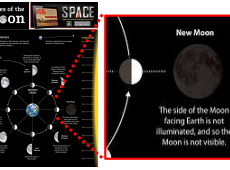
| | VDocRetriever | Text-based Retriever |
|---|---|--|
| <p>How many apps does the company which makes Clash of Clans make?</p> <p>Ground-truth: 7</p> <p>Text-based RAG: 61 ✗</p> <p>VDocRAG: 7 ✓</p> | <p>Top1 ✓</p>  <p>Top2 ✓</p>  | <p>Top1 ✗</p>  <p>Top2 ✗</p>  |
| <p>What is the Stream Source for the API which uses Java, Scala, and Python?</p> <p>Ground-truth: HDFS, Network</p> <p>Text-based RAG: Fink ✗</p> <p>VDocRAG: HDFS, Network ✓</p> | <p>Top1 ✓</p>  <p>Top2 ✓</p>  | <p>Top1 ✗</p>  <p>Top2 ✗</p>  |
| <p>Which is Microsoft's biggest acquisition to date?</p> <p>Ground-truth: Skype</p> <p>Text-based RAG: Oculus ✗</p> <p>VDocRAG: Skype ✓</p> | <p>Top1 ✓</p>  | <p>Top1 ✗</p>  <p>Top2 ✗</p>  |
| <p>How many layers are used in the gloves for the DPE suit?</p> <p>Ground-truth: Three</p> <p>Text-based RAG: Two ✗</p> <p>VDocRAG: Three ✓</p> | <p>Top1 ✓</p>  | <p>Top1 ✗</p>  <p>Top2 ✗</p>  |
| <p>What is the phase before full moon?</p> <p>Ground-truth: Waxing Gibbous</p> <p>Text-based RAG: New Moon ✗</p> <p>VDocRAG: Waxing Gibbous ✓</p> | <p>Top1 ✓</p>  | <p>Top1 ✗</p>  <p>Top2 ✓</p>  |

Figure D. Additional qualitative results of VDocRAG compared to Text-based RAG.