Towards Natural Language-Based Document Image Retrieval: New Dataset and Benchmark

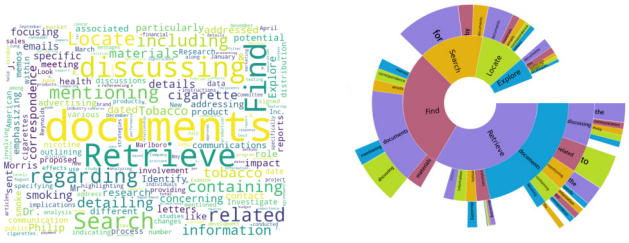
Supplementary Material

In the supplementary materials, we provide more details on dataset construction, along with the experimental training and testing processes. Additionally, we discuss further experimental findings and future work, concluding with an overview of the Licensing, Hosting, and Maintenance Plan, as well as the Datasheet.

1. Dataset Details

1.1. Query Analysis

We construct a word cloud in Fig. 1a that removes stop words. The word cloud displays the most frequently occurring words and reveals the primary intention and key topics of retrieval. Fig. 1b presents a sunburst diagram depicting the distribution of the first four words in the queries. The diagram reveals that queries frequently begin with words such as "retrieve", "find", and "search", which indicates the intention of information retrieval. The outer ring of the sunburst provides even more granular details, likely representing specific topics or types of information being sought, such as "documents", "emails", and "memos".



(a) (b)

Figure 1. Query Analysis. (a) Word cloud of queries after removing stop words. (b) Distribution of first four words in queries in the NL-DIR dataset.

To extract more descriptive words, we process the queries using spacy's NER module to extract adjectives and nouns. The query length distribution, shown in Fig.2, is concentrated between 3 and 9 words.

1.2. Collection Details of the NL-DIR Dataset

Initially, we attempt to build the Natural Language-Based Document Image Retrieval (NL-DIR) dataset by utilizing existing information retrieval datasets by rendering the documents into images. However, these images often fail to reflect the distribution of real-world documents. Therefore, we decide to construct the dataset by generating the corresponding queries.

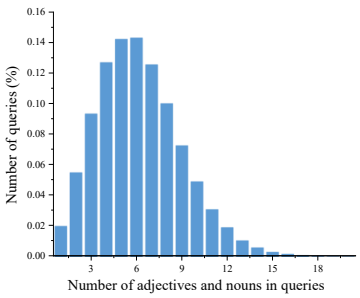


Figure 2. Query length distribution after NER processing.

We conduct extensive research and find datasets with a large scale of document images and relatively good OCR results. "Relatively good OCR results" refers to commercial OCR systems compared to the open-source Tesseract OCR. Finally, some images from DocVQA [10] and OCR-IDL [2] are sampled to build the dataset. This allows us to obtain real-world document images with relatively good layout and content information as mentioned in the main paper. For OCR-IDL [2], we collect and use the first page of its PDF files as a default choice, which contains richer semantic content. The subsequent construction process can be found in the main paper. In the following part, we also provide the prompts used in the query generation and filtering processes, as well as the standards for manual verification.

1.3. Scoring Models and Manual Verification

The scoring models used to pre-score the ten generated queries include a large language model (ChatGPT [11]), a multimodal large vision-language model (Qwen-VL-Plus [1]), and two contrastive models (CLIP [12] and BLIP [8]). The large language model (LLM) enables more effective analysis of the query content when combined with the OCR text, while the large vision-language model (LVLM) incorporates some visual elements from the document images for scoring. The final two models, which have undergone extensive pre-training on cross-modal image-text alignment, provide a preliminary score based on the degree of similarity between the text query and the document image. We collect and assign the aforementioned scores to the corresponding queries for each image.

By designing and providing a visualization interface as Fig. 3, we display each image, the queries, and their scores for human verification. During the human verification process, we first remove damaged document images and inappropriate queries. The reserved pairs are filtered based on

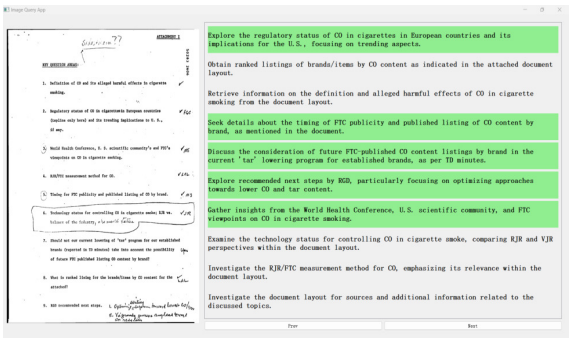


Figure 3. Visualization for manual verification.

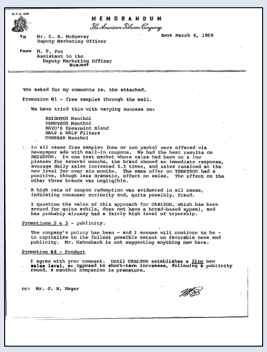
the above scores and query quality, ensuring that queries are as strongly related to the current image as possible. Then the annotators are asked to filter out ambiguous queries and images as much as possible.

Specifically, for document images, we will remove those with significant quality degradation or very low information content. For queries, we apply filtering rules to exclude those containing specific characters, such as queries that include "UCSF" or the original document source information. When both document images and queries are involved, we use the filtering methods mentioned in the body of the text to filter them accordingly.

To alleviate data bias, we strive to ensure consistent query quality across different document categories during manual filtering. Finally, through these two processes, we obtain a high-quality, fine-grained NL-DIR dataset.

1.4. Visualized Examples of NL-DIR

This section presents examples and analysis of the five most common types of document images and their corresponding query statements. As shown in Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8, the vocabulary related to document images is labeled with different colors in each query, showing the diversity of generated queries.



Queries:

1.Retrieve documents discussing the effectiveness of free sample promotions via mail for various tobacco brands, emphasizing results and consumer response.

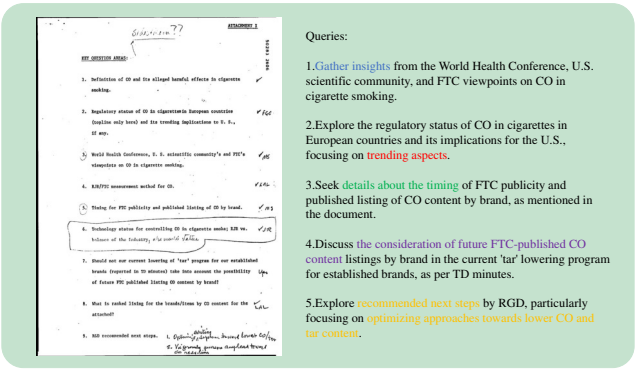
2.Retrieve documents discussing the appropriateness of introducing a menthol companion product in relation to the current sales status of CARLTON.

3.Search for documents discussing the role of consumer curiosity and potential fraud in response to free sample promotions.

4.Search for documents evaluating the value of offering free samples through mail-in coupons for established brands versus new products.

5.Find documents analyzing the impact of different promotional strategies on sales, particularly focusing on the success rates of newspaper ad campaigns with mail-in coupons.

Figure 4. Letter: when the textual information is abundant and there is relatively little structured information, the corresponding five queries mainly focus on understanding the content of the entire image.



Queries:

1.Gather insights from the World Health Conference, U.S. scientific community, and FTC viewpoints on CO in cigarette smoking.

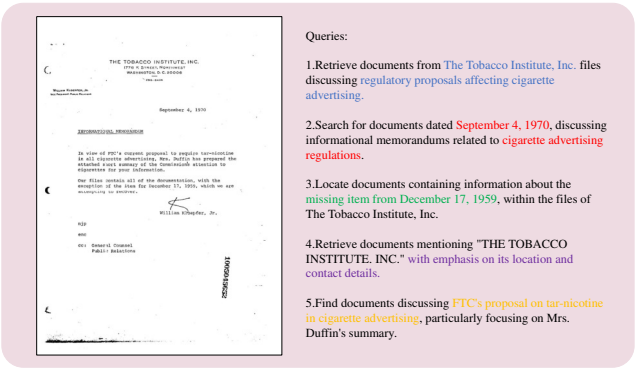
2.Explore the regulatory status of CO in cigarettes in European countries and its implications for the U.S., focusing on trending aspects.

3.Seek details about the timing of FTC publicity and published listing of CO content by brand, as mentioned in the document.

4.Discuss the consideration of future FTC-published CO content listings by brand in the current 'tar' lowering program for established brands, as per TD minutes.

5.Explore recommended next steps by RGD, particularly focusing on optimizing approaches towards lower CO and tar content.

Figure 5. Report: if the document image contains entries, the query is likely to be finely granular in its focus, with varying focus for each entry.



Queries:

1.Retrieve documents from The Tobacco Institute, Inc. files discussing regulatory proposals affecting cigarette advertising.

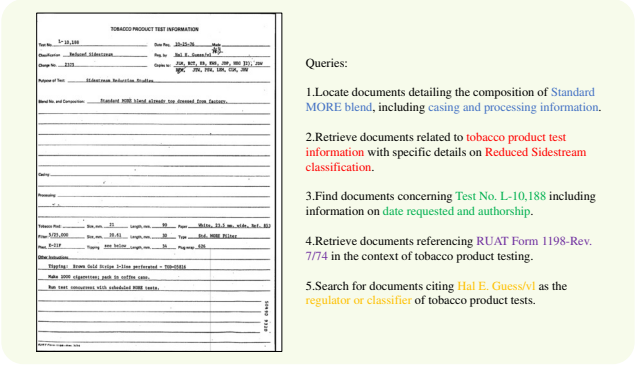
2.Search for documents dated September 4, 1970, discussing informational memorandums related to cigarette advertising regulations.

3.Locate documents containing information about the missing item from December 17, 1959, within the files of The Tobacco Institute, Inc.

4.Retrieve documents mentioning "THE TOBACCO INSTITUTE, INC." with emphasis on its location and contact details.

5.Find documents discussing FTC's proposal on tar-nicotine in cigarette advertising, particularly focusing on Mrs. Duffin's summary.

Figure 6. Memo: it can be observed that the query will mine and search for some unique information within the document.



Queries:

1.Locate documents detailing the composition of Standard MORE blend, including casing and processing information.

2.Retrieve documents related to tobacco product test information with specific details on Reduced Sidestream classification.

3.Find documents concerning Test No. L-10,188 including information on date requested and authorship.

4.Retrieve documents referencing RUAT Form 1198-Rev. 7/74 in the context of tobacco product testing.

5.Search for documents citing Hal E. Guess/vl as the regulator or classifier of tobacco product tests.

Figure 7. Form: in the context of form images, greater emphasis is placed on the querying of the names and contents of different fields within the form.

2. Experimental Details

2.1. Recall and Re-Ranking Setting

Recall stage. In the zero-shot setting, for Contrastive VLMs, we directly extract their original visual represen-

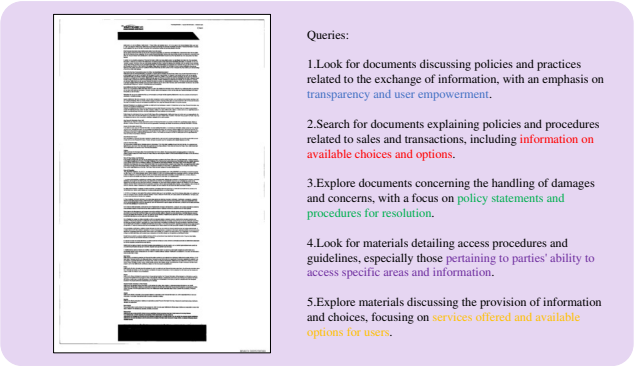


Figure 8. Document: when the document image is rich in textual information, the query will be constructed to analyze the document as a whole and summarise its content.

tations. In contrast, for Generative VDU models, we take the final output from each visual encoder and perform mean pooling to obtain the visual representation. Both representations are stored in the FAISS vector library [1]. During the extraction process for Generative VDU models, we also attempt to use the entire VLM to extract representations from the final output of the language layer as visual features. However, the retrieval performance is similar to that of directly extracting the visual encoder's representations, but the process take significantly longer. Therefore, we do not discuss this approach further. The dot product is then utilized to query the document image representation in the vector library, ultimately yielding zero-shot results.

In the fine-tuning setting, we use LoRA [5] to fine-tune the text encoders (i.e., CLIP [12] and BLIP [8]), with parameters set to $r = 8$ and lora alpha $= 16$. After fine-tuning, we align these encoders with various VDU models. Linear layers are employed to fine-tune the mean pooling features, consisting of two layers with a residual connection that maps the original feature dimensions to 512. We also attempt to fine-tune both visual and text encoders simultaneously using LoRA; however, this approach not only significantly increase training time but also lead to a decrease in retrieval efficiency.

For SigLIP [13], we directly utilize its original model structure and apply LoRA to fine-tune both the text and visual encoders with parameters set as $r = 32$, lora alpha $= 32$, weight decay $= 1 \times 10^{-4}$, warmup steps $= 2.5\%$, lr $= 5 \times 10^{-5}$. With a batch size of 32, fine-tuning for 10 epochs yields the best recall retrieval result.

Re-ranking stage. We conduct a comparison with contemporaneous models, such as DSE [9] and ColPali [3], which utilize large visual-language models to encode queries and images. Considering factors like encoding time, storage space, retrieval efficiency, and training costs, we test these models in zero-shot setting.

During the fine-tuning of the re-ranker, we primarily focus on the models that perform well in the recall stage. We use several models that have original cross-attention modules, specifically BLIP-ITM [8] and Pix2Struct-base [7], or incorporate additional cross-attention for fine-tuning. For the pre-trained BLIP-ITM model, we fine-tune its language module directly. In the case of Pix2Struct-base, we add an additional ITM head and fine-tune the language module accordingly. For models with additional cross-attention, we enable interactions between the original fine-grained features to improve re-ranking results. The fine-tuning parameters for these models are set as follows: $r = 32$, lora alpha $= 32$, lr $= 1 \times 10^{-3}$ (lr $= 1 \times 10^{-4}$ for LoRA), weight decay $= 1 \times 10^{-4}$. The optimizer follows a cosine decay schedule with $T_{\text{max}} = 10$ and $\eta_{\text{min}} = 1 \times 10^{-5}$, and we use a batch size of 8.

In the future, we will release the dataset with its construction code, evaluation code, model code, and weights to facilitate reproducibility for researchers.

2.2. Case Analysis

To better observe the results of fine-grained interactions during the re-ranking stage, we use the attention scores from the query and key in the cross-attention module to visualize the interactions between the query and the images. We aggregate the attention scores for each token in the query and superimpose the heatmap on top of the original image, allowing us to identify the regions in the image that are most relevant to the query.

The query corresponding to the image below is: "Retrieve documents from B. P. Horrigan regarding SALEM Lights 100 tar level developments." As seen in Fig. 9, the areas related to "SALEM Lights 100 tar level developments" are prominently displayed. This indicates that the re-ranking stage has a certain degree of fine-grained matching and scoring capability, allowing for a more effective re-ranking of the original results.

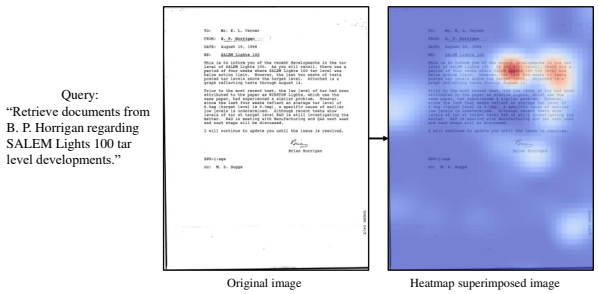


Figure 9. Visualization of cross-attention in the re-ranking stage.

[1] https://github.com/facebookresearch/faiss

3. Additional Experimental Analysis

To assess the generalization of models trained on our dataset to other tasks, we evaluate their performance on a downstream document classification task. As shown in Tab.1, we compare the zero-shot classification results of the original and trained SigLIP models on the RVL-CDIP and Tobacco3482 datasets, demonstrating the effectiveness of our approach in improving document representation learning.

Table 1. Zero-shot comparison of original and trained Model.

Dataset	Original	Trained
RVL-CDIP[4]	7.43	10.74
Tobacco3482[6]	44.57	55.92

Recent LVLMs, such as InternVL2-2B and Qwen2-VL-2B, have demonstrated strong document understanding capabilities. We leverage these models to generate content summaries for document images using the prompt: "Please describe the document image." The generated descriptions serve as retrieval queries, which we then encode using the BGE model for document retrieval. As shown in Tab.2, the retrieval performance of these caption-based queries is comparable to that of OCR-IR. However, similar to OCR-IR, generating content summaries requires significant computational resources and time.

Table 2. Comparison with models as image-captioners (CAP-IR).

Metric	InternVL2-2B	Qwen2-VL-2B	OCR-IR	Ours
Recall@1	52.83	47.31	52.72	81.03
Recall@10	71.63	68.07	72.16	94.17
MRR@10	53.90	59.02	58.85	85.68

To gain deeper insights into the retrieval performance across different document categories, we further analyze the re-ranked retrieval results. As shown in Tab.3, we report the retrieval performance and the number of queries for five representative document categories.

Table 3. Retrieval performance on five representative categories.

Category	Letter	Report	Memo	Form	Document
Query_nums	3675	1760	1520	1180	1175
Recall@1	83.10	82.67	87.50	73.64	89.62
Recall@10	93.88	95.80	94.61	94.66	97.96
MRR@10	86.86	87.37	89.92	81.33	92.98

4. Future Work

This study presents a preliminary exploration of document image retrieval, offering valuable insights into dataset construction and model optimization. However, as research progresses, several key directions warrant further investigation and improvement.

First, large-scale training data and the powerful representational capacity of LVLMs have enabled state-of-the-art retrieval performance in the recall stage. However, these models often incur significant computational and memory costs, raising concerns about efficiency. Currently, there is a lack of alignment models specifically designed for high-resolution document images and rich textual content. Effective cross-modal representation alignment facilitates the mapping of image and text information into a shared vector space, thereby enhancing fine-grained understanding and retrieval performance. This can help bridge the gap between cross-modal document image retrieval and purely text-based retrieval using OCR. Future research should focus on designing more efficient and compact models optimized for high-resolution document images and their textual content while improving image-text alignment. Furthermore, with advancements in generative models, the integration of cross-attention mechanisms with generative understanding models holds great potential. However, significant room remains for experimentation and improvements in the re-ranking stage. Despite the progress made in document image retrieval, a critical future direction lies in tightly integrating fine-grained generative understanding capabilities with the practical demands of document image retrieval.

Second, as document retrieval technology evolves, real-world applications often require retrieving multi-page documents. This necessitates models capable of processing and understanding multi-page document images while capturing long-range contextual dependencies. Additionally, there is a growing need for fine-grained paragraph-level retrieval. Currently, retrieval units in this study are typically single-page documents, and the models lack precise paragraph-level localization, which can impact retrieval accuracy in certain scenarios. Future research should explore long-document modeling for multi-page documents and precise paragraph-level localization. This is not only crucial for improving retrieval accuracy but also provides broader applications in document analysis and search systems.

In summary, future advancements in document image retrieval will focus on overcoming computational and memory efficiency bottlenecks, enhancing the ability to capture long-document information, and improving paragraph-level retrieval precision. As technology advances and application scenarios expand, document image retrieval is expected to play an increasingly vital role in improving information access efficiency and enhancing user experience.

5. Prompt Design

In this section, we provide the prompts used in the query generation and filtering processes, as shown in Table Tab. 4.

6. Licensing, Hosting and Maintenance Plan

Author Statement. We bear all responsibilities for the licensing, distribution, and maintenance of our dataset.

License. NL-DIR is under CC-BY-NC-SA 4.0 license.

Hosting. NL-DIR can be viewed and downloaded on huggingface at https://huggingface.co/datasets/nianbing/NL-DIR. Prior to the publication of the article, we typically present a selection of illustrative samples, after which we will release the entire dataset. We assure its long-term preservation for future reference and use. The annotations for retrieval queries are provided in the JSON file format, while the raw pictures are available in the PNG format.

We do not hold any copyright for the document images; the copyrights belong to the UCSF Industry Documents Library and the document authors. For user convenience, we provide a download method for these document images, provided users agree that the data is only used for research purposes and not for commercial purposes. If copyright holders request the deletion or modification of certain images, we will hide or delete key information in the images to minimize the impact on the query. If the retention of images is not allowed, we will retain the query data and provide metadata for the corresponding images.

The Croissant metadata record is stored in https://huggingface.co/datasets/nianbing/NL-DIR-sample/blob/main/croissant.json.

Metadata. Metadata can be found at https://huggingface.co/datasets/nianbing/NL-DIR.

7. Datasheet

7.1. Motivation

For what purpose was the dataset created?

Answer: NL-DIR establishes a fine-grained semantic retrieval dataset and benchmark for document images in real-world scenarios, which evaluates the retrieval performance of existing contrastive vision-language models (VLMs) and generative visual document understanding (VDU) models. NL-DIR provides an evaluation of existing models for document image understanding and cross-modal dense representation. As far as I know, NL-DIR is the first comprehensive benchmark for fine-grained document image semantic retrieval.

7.2. Composition

What do the instances that comprise the dataset represent? (e.g., documents, photos, people, countries)

Answer: Each instance represents a document image and five fine-grained semantic queries in our dataset. The document image is a PDF screenshot collected from UCSF Industry Documents Library [2] in PNG format. The query is generated by LLM and then stored in a JSON file after being scored and manually filtered by a scoring model.

How many instances are there in total (of each type, if appropriate)?

Answer: We collected a total of 41,795 document images, each corresponding to five queries. The specific dataset statistics can be found in the main paper.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Answer: We collect over 60K document images from the Industry Documents Library. The corresponding layout text information is extracted from the annotations of DocVQA [10] and OCR-IDL [2], which use Microsoft OCR and Amazon Textract respectively as OCR engines.

Is there a label or target associated with each instance?

Answer: Yes, for each document image, we generate and filter five queries.

Is any information missing from individual instances?

Answer: All instances are complete.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Answer: Some instances may have similar images or queries, but when filtering, we try to ensure a strong correlation between queries and images as much as possible.

Are there recommended data splits (e.g., training, development/validation, testing)?

Answer: Yes, we have done a reasonable split of the NL-DIR dataset, which is reflected in the already split JSON file, we will make all JSON files public after the publication.

Are there any errors, sources of noise, or redundancies in the dataset?

Answer: No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Answer: All data will be publicly accessible in the dataset repository. Our annotations will be stored in JSON format.

Does the dataset contain data that might be considered confidential?

Answer: No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Answer: No.

7.3. Collection Process

The data collection process is described in the main paper and supplementary materials.

[2] https://www.industrydocuments.ucsf.edu

7.4. Uses

Has the dataset been used for any tasks already?

Answer: Yes, NL-DIR has been used to evaluate the cross-modal retrieval capabilities of as many as 9 different models.

What (other) tasks could the dataset be used for?

Answer: NL-DIR is mainly used for the evaluation of the cross-modal retrieval capability of document-related visual and language models.

Is there a repository that links to any or all papers or systems that use the dataset?

Answer: No.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Answer: The document images we collected are all from IDL, and the corresponding OCR information is from DocVQA and OCR-IDL. The query generation and filtering methods have been provided in the main paper. However, we will do our best to maintain the dataset if the copyright holder requests the removal of certain data in the future.

Are there tasks for which the dataset should not be used?

Answer: No

7.5. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Answer: Yes. The benchmark is publicly available on the Internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

Answer: The benchmark is available on Huggingface at https://huggingface.co/datasets/nianbing/NL-DIR.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Answer: CC-BY-NC-SA 4.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

Answer: No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

Answer: No.

7.6. Maintenance

Who will be supporting/hosting/maintaining the dataset?

Answer: The authors will be supporting, hosting, and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Answer: Please contact the one of the authors (guohao2022@iie.ac.cn, qinxugong@njust.edu.cn).

Is there an erratum?

Answer: No. We will make announcements if there are any.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Answer: Yes. We will post a new update in https://huggingface.co/datasets/nianbing/NL-DIR if there is any.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period and then deleted)?

Answer: People's information may appear in the reference images. People may contact us to exclude specific data instances if they appear in the reference images.

Will older versions of the dataset continue to be supported/hosted/maintained?

Answer: Yes. Old versions will also be hosted in https://huggingface.co/datasets/nianbing/NL-DIR-sample

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Answer: Yes, according to our dataset construction method, if the data is compliant and reasonable, expanding the dataset is allowed.

Table 4. The prompts used in the query generation and filtering processes.

Model		Template
Generate	**ChatGPT [11]**	You are an expert who can use image's OCR to generate a query for retrieval. More specifically, you will obtain the following content: 1. Instruction: A statement used to describe specific task details. 2. Layout-aware OCR Document: Text extracted from an image and arranged according to to the layout when it appears in the image to maintain the relative position between the texts appearing in the image. You need to understand the document layout with the help of spaces and line breaks in the document. NOW YOU TURN: Instruction: You need to generate ten different layout-related queries that cover all the different aspects of the entire document as much as possible. These queries are used for retrieving layout-aware documents based on the above conditions and the following. A query cannot be a simple and detailed description, but should express the purpose of the search. Layout-aware OCR Document : **{document}** Queries:
Score	**ChatGPT [11]**	Here are ten queries used to retrieve image documents, and we would like to request your feedback on the quality of the queries. Please rate the quality of the ten given queries based on the content of the Layout -aware OCR Document. Each query receives a score of 0 to 10, with higher scores indicating higher quality. Layout-aware OCR Document: **{document}** Queries: **{queries}** Please provide a comprehensive explanation of your evaluation to avoid any potential biases. Output format: Scores: Reasons:
	Qwen-VL-Plus [1]	You are an expert in using images and their OCR text to score queries for retrieval. Here are ten queries used to retrieve image documents, and we would like to request your feedback on the quality of the queries. Please rate the quality of the ten given queries based on the content of the Layout -aware OCR Document and the document image. The document image, from which you can obtain some visual elements that are not included in the Layout-aware OCR Document. Each query receives a score of 0 to 10, with higher scores indicating higher quality. Layout-aware OCR Document: **{document}** Queries: **{queries}** Please provide a comprehensive explanation of your evaluation to avoid any potential biases. Output format: Scores: Reasons:

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 7

[2] Ali Furkan Biten, Rubèn Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-idl: Ocr annotations for industry document library dataset. In *European Conference on Computer Vision*, pages 241–252. Springer, 2022. 1, 5

[3] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 3

[4] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 4

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[6] Jayant Kumar, Peng Ye, and David Doermann. Learning document structure for retrieval and classification. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 1558–1561. IEEE, 2012. 4

[7] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 3

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3

[9] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024. 3

[10] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1, 5

[11] OpenAI. Chatgpt. *https://chat.openai.com/*, 2022. 1, 7

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[13] 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*, pages 11975–11986, 2023. 3