IV. International Conference on Electrical, Computer and Energy Technologies (ICECET 2024)
25-27 July 2024, Sydney-Australia

RAG Beyond Text: Enhancing Image Retrieval in RAG Systems

Sukanya Bag
Data Science and Insights
Genpact India Private Limited
Bengaluru, India
sukanya.bag1@genpact.com

Ayushman Gupta
Data Science and Insights
Genpact India Private Limited
Bengaluru, India
ayushman.gupta@genpact.com

Rajat Kaushik
Data Science and Insights
Genpact India Private Limited
Bengaluru, India
rajat.kaushik1@genpact.com

Chirag Jain
Data Science and Insights
Genpact India Private Limited
Bengaluru, India
chirag.jain4@genpact.com

***Abstract*-This paper presents a novel methodology for the extraction and retrieval of images in RAG (Retrieval Augmented Generation) powered Question Answering Conversational Systems that circumvents the limitations of Optical Character Recognition and Large Language Model (OCR-LLM) powered traditional image retrieval approaches. We are leveraging the positional information of images in a vast array of multi-modal (text/image) documents for ingesting image information alongside text, followed by advanced retrieval and prompt engineering techniques to develop an RAG system that maintains the integrity of textual and visual data correlation in responses to queries pertaining to both text and images in QnA solutions and is adept at retrieving both OCR-compatible and OCR-incompatible images. We have successfully incorporated this approach over a variety of multimodal documents ranging from research papers, application documentations, surveys to guides and manuals containing text, images and even tables with images and managed to achieve SoTA (State of The Art) performance over simple to complex queries asked on the mentioned documents.**

Furthermore, our approach performed explicitly better in cases where Vision Models like GPT-4 Vision fails to accurately retrieve images which are OCR incompatible and pertains to highly customized scientific devices or diagrams and in cases where the image's visual representation is not semantically aligned with textual information but is important to be retrieved for completeness in the response.

Index Terms—Document Question Answering, GPT4, Large Language Models, Langchain, Mu-RAG (Multi-modal Retrieval Augmented Generation), RAG (Retrieval Augmented Generation)

I. INTRODUCTION

The advent of LLMs and LLM powered RAG systems has fostered a growing need for sophisticated image extraction and retrieval in such systems as well. Traditional mechanisms often rely heavily on Optical Character Recognition (OCR) integrated with Language Learning Models (LLMs) to interpret and contextualize images within documents. However, this integration poses significant challenges, including bottlenecks in processing speed and accuracy issues stemming from the OCR component. These challenges become even more pronounced when dealing with images that are not OCR-compatible like flowcharts, diagrams, scientific devices, or manuals leading to a loss of information and discontinuity between text and visual elements which is extremely crucial to address in responses generated by LLMs.

To address these limitations, there has been an increasing emphasis on developing alternative strategies that can bypass the dependency on OCR while enhancing the performance and reliability of image responses in RAG systems. Such an advancement is essential for a wide array of applications, from legal and medical document management to academic research and content archiving, where the accurate retrieval and contextualization of images are of utmost importance.

In this paper, we introduce an innovative solution that transcends the traditional OCR-LLM framework, offering an efficient and accurate method for multiple image retrieval aligned with textual responses. Furthermore, we address the critical need to maintain the continuity between textual and visual elements within documents for allowing the language model to provide the user with step-by-step assistance with textual information aligned with respective relevant image/(s). Traditional methods often disrupt this continuity, leading to a disjointed representation of information. Our methodology ensures that the spatial arrangement of images in relation to the text is preserved, thereby upholding the document's structural integrity and enable highly accurate QnA.

This paper presents two approaches to image retrieval in Retrieval Augmented Generation Systems that enhance the accuracy and efficiency of Question Answering:

A. Traditional OCR-LLM Based Image Retrieval

Image Content Extraction with OCR + Captioning with LLM - This initial method combines Optical Character Recognition capabilities for extracting textual content in images followed by an LLM which is utilized to create a meaningful 2-6 (not limited to) worded caption from the raw OCR extracted text representing the respective image. This information along with the image's metadata is paired and loaded into a vector store as embeddings. While this traditional approach does provide a foundational solution, but it is limited to the OCR compatibility of the images, meaning that it can only retrieve relevant images if those images have textual information in them which are semantically aligned with the text.

B. Image Localization Tag (ILT) Based Image Retrieval

Bearing in mind the drawbacks and limitations of the traditional OCR-LLM based approach for image retrieval, we devise a new approach to retrieve images, irrespective of the content and kind of image constraint - a relevant image without any textual info can also be important! Hanging on the same thought for a moment, another point of view comes in - a relevant image having textual info will not always semantically align with the textual content / questions asked to the RAG system. Completeness in the answer is a very crucial parameter to be taken into consideration while developing QnA solutions leveraging RAG.

Hence, our Image Localization Tag (ILT) approach focuses more on injecting the image's information in the respective position of the image in a document so that it will always maintain two things very crucial for enhancing image retrieval alongside text -

1) Text-Image Continuity dictating the original document's structure/content.
2) Establishing an acquired semantic correlation between text and images based on spatial proximity of images alongside text.

2024 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 979-8-3503-9591-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECET61485.2024.10698598

II. Literature Review

Previous works in RAG-based QA systems have focused on semantic congruence between text and images. Lewis et al. [1] introduced the concept of RAG (*Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*) to enhance text generation with retrieved documents. However, the scope for non-textual content was limited. Subsequent research expanded on this by integrating images, but often faced challenges in retrieving contextually apt visuals when semantic alignment was absent.

Zhou et al. [2] proposed an image captioning approach which leverages news articles as additional context to describe the image. This method focuses on news image captioning, which narrows its scope and applicability compared to our versatile approach. Zhou et al. [3] in their Style-Aware image captioning method proposed captioning content based on relevant style. This method neglect OCR compatibility and broader applicability across various multi-modal documents. Zhuolin Yang et al. [4] similarly proposed Retrievalaugmented Visual Language Model which uses Flamingo model and utilizes external database for relevant knowledge retrieval for few-shot image caption generation. As the model is trained on generic image data, it has no OCR capabilities and would fail for domain specific use-cases. Other works [5–8] have also shown excellent capabilities of Retrieval-augmented based approach for image captioning. However, they are suitable for general use-cases. When applied over domain specific use-cases like research papers, user manuals, guides etc., we observed that they fail to generate suitable captions.

Traditional systems often struggled with semantic limitations, relying solely on textual cues for image retrieval. Not only that, it has been seen in past works that inconsistency became a very significant bottleneck in relevancy of the context retrieved, that is the sheer inability to fetch images along with text, maintaining the text image continuity as dictated in the original knowledge base, which is extremely crucial to the context but lacking direct semantic connection. Moreover, recent as well as previous research on language and large language model shows that despite the advancements in RAG systems, an over-reliance on language model or Large Language Models (LLM) reasoning has been identified as a bottleneck, leading to sub-optimal results in the acquisition of necessary text/visual content. One of the works to support this statement is Khatun and Brown et al. [9] who find that subtle changes in prompt wording change a model's response.

III. Methodology

A. Traditional OCR-LLM Based Image Retrieval

We begin by systematically extracting images from a set of documents. For each image, we record its index within the document, the page number on which it appears, and the name of the document itself. These images are then stored in a blob storage system, ensuring that they are catalogued and retrievable for further processing. Each stored image undergoes Optical Character Recognition (OCR) to extract the embedded textual content. The OCR process is pivotal as it converts visual information into machine encoded text, which serves as the basis for further interpretation and analysis by the LLM. The raw text obtained from the OCR is input into a Large Language Model (LLM) for caption generation. In our implementation, we utilize the GPT-3.5 Turbo 4k model, a state-of-the-art LM known for its ability to produce concise and coherent text outputs. The model processes the OCR-extracted text to generate captions that succinctly represent the content of the images.

The LLM-generated captions are then paired with their respective image filenames, forming key-value pairs. This step facilitates the organization of image data and its corresponding textual description, which is crucial for efficient retrieval. These pairs are stored in a structured text file format and subsequently ingested into a vector store. The vector store houses embeddings of the captions, which we refer to as our 'image vector store.' Fig. 1 shows the workflow pipeline of document ingestion which includes image OCR and caption generation step and text and image vector store generation.

Upon receiving a user query, the system initiates a retrieval process within the text vector store. The objective is to fetch text chunks that are most relevant to the user's question. This retrieval is guided by the contextual information encapsulated in the user's query, ensuring that the search within the vector store is focused and precise. While the textual retrieval is underway, the system concurrently conducts a Similarity Search within the image vector store. The aim is to identify the image caption that best aligns with the contextual cues obtained from the text vector store. The Cosine Similarity Search algorithm computes the degree of relevance between the LLM-generated captions and the user query, surfacing the most pertinent image-caption pair. Fig. 2 shows the QnA workflow pipeline using OCR-LLM approach. Fig. 4 shows more detailed workflow of the QnA pipeline using OCR-LLM approach.

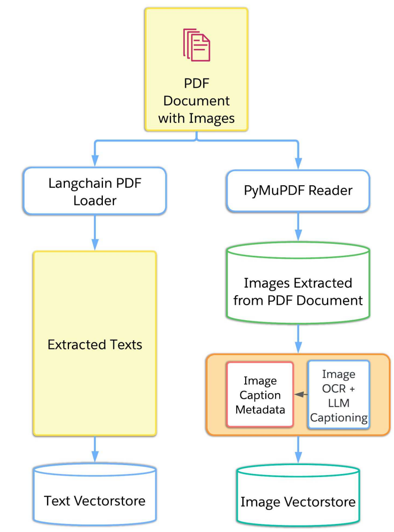


Fig. 1. Ingestion pipeline using OCR-LLM approach.

1) Solution Bottlenecks:

- Dependence on OCR and Captioning Quality: The system's performance relies heavily on OCR quality and captioning. Non-text images, like scientific figures or flowcharts, pose a challenge, as OCR's inability to find text leads to irrelevant captions. This is exemplified by Fig. 3, where a scientific image lacks text, resulting in an unrelated caption and unsuccessful image retrieval.
- Semantic Similarity and Retrieval Issues: Retrieval depends on semantic similarity between user queries and captions, which can cause inaccuracies. For example, a low similarity score for the query "How to switch role on AWS?" may prevent the retrieval of important images, as seen when the score is only 0.56.
- Limitations in Retrieving Multiple Images: The system's top k selection for similarity search struggles with retrieving several relevant images for one query. A high k may include irrelevant images, while a low k might miss or omit necessary ones. The default setting of $k = 1$ focuses on accuracy for individual images but fails to support multi-image queries.
- Separate Image and Text Responses: Images and textual responses are provided separately, disrupting spatial alignment.

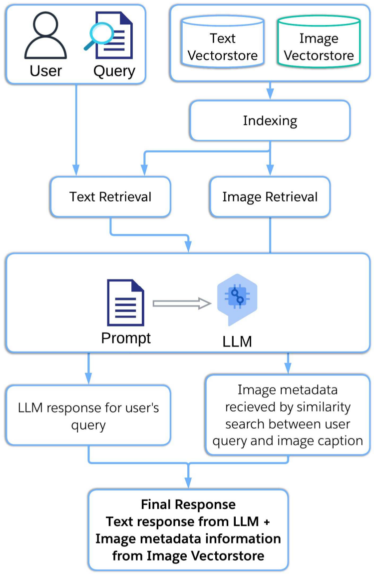


Fig. 2. QnA pipeline using OCR-LLM approach.

Question: Show me examples of how specialized attention heads in a Transformer recover protein structure and function, based solely on language model pre-training.

Response:
(a) Attention in head 12-4, which targets amino acid pairs that are close in physical space (see inset subsequence 117D-157I) but lie apart in the sequence. Example is a de novo designed TIMbarrel (5BVL) with characteristic symmetry.

(b) Attention in head 7-1, which targets binding sites, a key functional component of proteins.
Example is HIV-1 protease (7HVP). The primary location receiving attention is 27G, a binding site for protease inhibitor small-molecule drugs.

Fig. 3. Example response using OCR-LLM approach.

and the flow of information, which could hinder the user's understanding of the content.

- Cost Inefficiency and API Demands: The solution's high-cost stems from numerous LLM calls for captioning. For a corpus with 10,000 documents and 100 images each, about 1 million OCR and LLM calls are needed, with additional API calls for Q&A increasing costs substantially.

B. Image Localization Tag (ILT) for Image Retrieval

Our study introduces a sophisticated technique for document image retrieval that combines spatial and semantic alignment of images with associated text. This approach overcomes the limitations of Optical Character Recognition (OCR) and Large Language Models (LLM) captioning in analyzing multi-modal documents. We employ bounding box computations for each image to pinpoint its location and size within the document. These bounding boxes create a vital connection between each visual component and its textual context. The PyMuPDF library is utilized to extract images and their bounding box information, which is then used to insert an Image Localization Tag (ILT) at the image's position in a modified version of the document.

1) *ILT Generation:* An ILT comprises a unique image identifier, image SHA1 hash ID, and the image file extension. It is noteworthy that for our experiments we have used truncated version of SHA1 Hash ID. The motivation behind this approach was that, for images that are present within tables, the ILTs tend to overlap with the table contents due to the length of SHA1 Hash IDs. To tackle this issue, we decided to shorten the Hash ID. We shrunk the Hash ID to 8 digits in our case with [$H \bmod 10^n$]. Here H is decimal (base 10) representation of the SHA1 Hash ID of image and n is the number of digits the Hash ID is to be shrunk down to. Although for our experimentation we incorporated image object identifier with modified image SHA1 Hash ID and file extension to create our ILTs, it must be noted that ILTs are also highly flexible to one's use-case and needs, and not to be limited to the metadata we incorporated in our ILTs. For example, using different value for n for truncating the image Hash IDs. Images are stored externally by their hash ID and extension for efficient retrieval in response to LLM queries, with an example ILT looking like <image: filename(23523473.png)>. The ILTs, hence rightfully serves as a sophisticated images' contextual placeholder or a visual context marker within the document, encapsulating both the spatial coordinates and the semantic essence of the image. This ensures that each image is not only anchored in its original location but is also inherently connected to the relevant textual information.

2) *ILT Integration:* In the subsequent step, we embed the ILT within the document at the precise location specified by the bounding box information that we get with the help of PyMuPDF library. This embedding is performed with great attention to the original layout, preserving the exact region specified by the bounding box to avoid any misalignment issues. The modified document now contains a rich interplay of text and ILTs, mirroring the original structure of the document while enhancing it for advanced text retrieval capabilities. It is noteworthy to be mentioned that the original document is preserved on our storage bucket which we can display for the response's source for reference.

The document, enhanced with text and ILTs, is incorporated into our vector store, maintaining its layout, and meaning while facilitating efficient multi-modal retrieval. Fig. 5 shows the overall workflow of processing the document and integrating into the vector store. To extract segments pertinent to user queries, we employ the MMR retriever, which selects text chunks based on their cosine similarity to the query while minimizing redundancy with previously chosen chunks.

3) *LLM Prompting with Chain of Thought:* Our Chain Of Thought (CoT) Prompt Tuning technique refines the document retrieval process by creating targeted prompts that guide the LLM to consider Image Localization Tags (ILTs) during its response generation. This ensures that the LLM's output maintains fidelity to the document's layout and the images' contextual relevance. When the LLM retrieves contexts containing ILTs, these prompts are crucial for preserving the original structure and meaning of the document. Following the LLM's output, which includes the pertinent ILTs, we engage in a post-processing step. This involves identifying ILTs in the LLM response, extracting associated image data, and then substituting the ILTs with the actual images stored externally. The result is a comprehensive response that accurately reflects the placement and relevance of images as per the original document structure. Fig. 6 shows the overall workflow of QnA pipeline from taking in users' query to processing final response with relevant image references from LLM response. Fig. 8 shows a detailed workflow of QnA pipeline using Image Localization Tag approach.

Our methodology, which emphasizes the computation and integration of bounding box regions, not only optimizes image retrieval within documents but also guarantees contextually rich and precise responses from the LLM. Despite occasional shortcomings when images are located far from their relevant context, our approach excels at accurately retrieving images that are accompanied by textual figure descriptions. As a result, this method significantly enhances the system's performance and lays the groundwork for RAG-powered question-answering systems to generate more coherent and contextually aligned multi-modal responses.

4) *Overcoming the bottlenecks of traditional OCR-LLM Approach:*

- Cost Efficient - Saves the cost for up to 1M standalone captioning calls to LLM.
- Highly Accurate Image Retrieval - Can retrieve images of

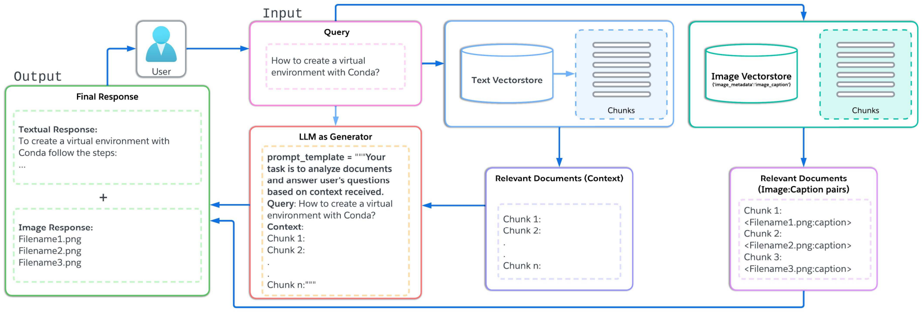


Fig. 4. OCR-LLM QnA Pipeline Workflow Demo

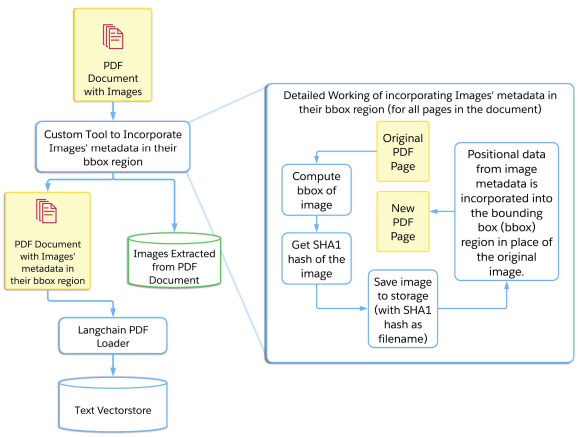


Fig. 5. Ingestion pipeline using Image Localization Tag based approach.

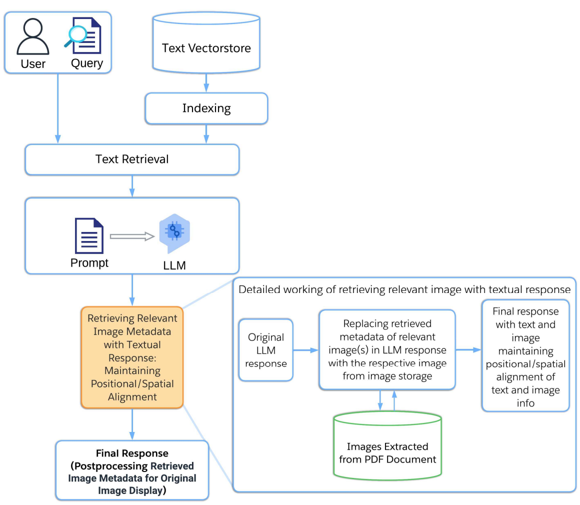


Fig. 6. QnA pipeline using Image Localization Tag based approach.

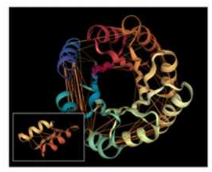
any kind – ranging from a vast array of images like images found on internet of basic natural things to biomedical images, flowcharts or logic diagrams, scientific instruments to software/application snapshots.

- Text Image Continuity - Facilitates multiple image retrieval and image retrieval maintaining spatial alignment of text and images as dictated by the original document's structure.
- Ability to retrieve OCR-incompatible images - Highly controlled solution which does not depend on LLM's reasoning to retrieve a particular image or not, as in many cases there are images which are necessary to be retrieved but are not semantically aligned with the textual information related to it.

Fig. 7. shows an example of a case where images retrieved are OCR-incompatible.

Question: Show me examples of how specialized attention heads in a Transformer recover protein structure and function, based solely on language model pre-training.

Response:
(a) Attention in head 12-4, which targets amino acid pairs that are close in physical space (see inset subsequence 117D-157I) but lie apart in the sequence. Example is a de novo designed TIMbarrel (5BVL) with characteristic symmetry.



(b) Attention in head 7-1, which targets binding sites, a key functional component of proteins.
Example is HIV-1 protease (7HVP). The primary location receiving attention is 27G, a binding site for protease inhibitor small-molecule drugs.

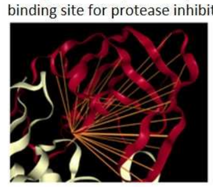


Fig. 7. Example response using Image Localization Tag approach.

IV. Dataset and Experiments

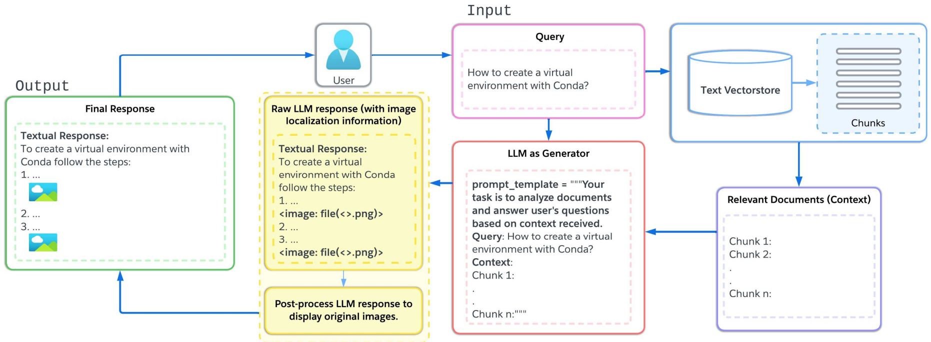
A. Dataset

To develop our RAG powered question-answering system for documents containing both text and images, we created a diverse dataset from a variety of sources. This collection includes a range of materials such as application guides, user manuals, programming instructions, literature reviews, and research articles. We gathered part of this dataset from the publicly available "Technically-oriented PDF Collection" on GitHub, which features a wide range of technical documents. Additionally, we included academic papers from the "CVPR 2019 Papers" dataset found on Kaggle [10].

Beyond these sources, we specifically chose a set of Standard Operational Procedures (SOPs) on topics like programming languages and application usage, as well as technical manuals from the internet. These documents were carefully selected to represent the kind of practical information that professionals might seek in their daily work, with a focus on areas like cloud computing and various software applications.

Our comprehensive dataset contains 100 documents with a total of 200+ questions on which we tested these documents. The dataset is categorized as follows:

- Handpicked Cloud/Programming/Various Application Documentation: 60
- User/Service Manuals (SOPs): 10
- Programming Guides: 10
- Literature Surveys: 10



- Research Papers: 10

Fig. 8. Image Localization Tag QnA Pipeline Workflow Demo

While training developers is one example of how this dataset might be used, its design is flexible enough to train any employee within an organization. By including an organization's private SOPs, our dataset can help streamline the training and on-boarding process, making it easier for new hires to get up to speed quickly. Our dataset has been put together for research purposes and is available to interested parties upon request. We have taken care to ensure that it adheres to all legal and ethical guidelines around the sharing and use of copyrighted content.

B. Experiments

In our preliminary experiments, we employed the OCR-LLM methodology on documents containing OCR-compatible images, primarily consisting of user manuals and programming guides. To assess the results, we utilized a collection of 20-30 queries. The validation of these responses was facilitated by RAGAS, a library specifically designed for evaluating RAG responses. However, since there is no existing library for evaluating image responses within RAG pipelines, we conducted manual evaluations of the image responses with the assistance of Subject Matter Experts (SMEs). The textual response outcomes displayed consistency across all queries. Given that the OCR-LLM approach is constrained by providing only a single response per query, the majority of the image responses were accurate. Nevertheless, for queries lacking a relevant image response, the pipeline tended to produce at least one pertinent image, even if it was unrelated to the query. This occurrence can be attributed to the static k value employed in the Langchain similarity search.

Subsequently, we replaced the OCR-LLM component of our pipeline with GPT-4 Vision to generate image captions. As GPT4 Vision is proficient in interpreting various types of images and has demonstrated excellent performance in explaining images containing text, we opted to use it for caption generation of images within the documents. GPT-4 Vision performed exceptionally well for images with text, including graphs, diagrams, screenshots, and other OCR-compatible images. However, it occasionally struggled to produce satisfactory captions for domain-specific images, such as scientific instruments and intricate tools. Table I shows a detailed comparison between captions generated by GPT-4 Vision and their captions taken from the document. For generic image contents GPT4 Vision was able to generate accurate captions but it failed when image contents are domain or product specific, which is our focus of improvement.

For our final approach, we maintained the same set of queries and continued to use RAGAS for evaluating textual responses, while manual evaluation was conducted for image responses. This method resulted in improved accuracy in image responses. With the help of Image Localization tags and Chain of Thought prompting, we were able to achieve accurate image response, surpassing the performance of GPT-4 Vision. Although our approach performs well in retrieving complex non-OCR images, it must be noted that this approach is dependent on the position of image with respect to its relevant textual information. For images that are further away from its textual context, our approach works well if the images have description (figure information) that is semantically aligned with parent textual context.

V. Results

Based on the comparison between OCR-LLM and Image Localization Tag approaches across various document types, from Table II it is evident that the Image Localization Tag approach consistently outperforms the OCR-LLM approach. Across research papers, manuals, programming documentations, and guides/surveys, the Image Localization Tag approach consistently achieves higher accuracy. Specifically, it scores a 91% for research papers, 94% for programming guides, and a commendable 95% for manuals and guides/surveys, whereas the OCR-LLM approach scores range from 60% to 70%, suggesting that the Image Localization Tag approach offers superior performance in accurately localizing and helping in extraction of information from documents across various domains, making it a more effective choice compared to the OCR-LLM approach.

VI. System Performance and Usability Analysis

The research leverages Langchain primarily for text generation within the system. While experiments have been conducted using a proprietary deployment of Azure OpenAI, the architecture is not limited to this; it is compatible with various open-source large language models (LLMs) as long as Langchain is employed. The system's complexity is concentrated in the document ingestion phase, wherein PDFs undergo processing. Here, a modified version of each document is created, and images are stored in an external repository. The computational load is directly proportional to the document's length and image content, with more extensive documents increasing system latency. Nonetheless, since the ingestion and querying components operate independently, user interactions, which are limited to the querying interface, remain largely unaffected by the ingestion process's computational demands.

From the perspective of user experience, the system facilitates user interaction exclusively through the querying interface, without

TABLE I. Comparison of image caption generated by GPT-4 Vision with original caption from document.

Image	Caption from Document	GPT 4 Vision Generated Caption

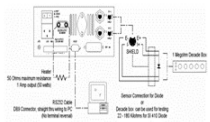
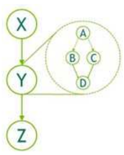
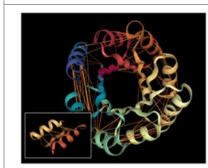
	NI USB Data Acquisition System.	National Instruments data acquisition device
	Model 9700 Temperature Controller rear panel connections	Schematic of electronic component testing setup
	SCM10 rear panel connector pins	Industrial device with various connectivity ports
	Diagram of child graph	Abstract diagram of interconnected nodes and cycles
	Complex protein structure	Colorful abstract 3D knot structure illustration

TABLE II. Performance comparison between OCR-LLM and Image Localization Tag Approaches

Document Type	SME evaluation (OCR-LLM)	SME evaluation (ILT)
Research Papers	65%	91%
Manuals	60%	95%
Programming Guides	70%	94%
Guides / Surveys	65%	95%

involving them in document uploading. Unlike existing RetrievalAugmented Generation (RAG) systems that provide only textual responses, our image-based RAG system incorporates visual elements into its responses. This is particularly beneficial for documents where images are integral to comprehension, such as manuals, guides, and tutorials. For example, in a medical instrument manual where each step is accompanied by critical images, traditional RAG systems would only generate text-based instructions, which can be less user-friendly as it may require users to revisit the document. Our system enhances user comprehension by sequentially presenting relevant images alongside each step, thereby reducing, or eliminating the need to consult the original document.

VII. FUTURE SCOPE

Our objective is to advance this methodology by integrating an effective and precise mechanism for retrieving complex diagrams characterized by a combination of shapes, images, and descriptive texts. Several documents feature such diagrammatic illustrations that are split into distinct elements of shapes, images, and textual annotations. Our existing system is unable to effectively capture these multifaceted diagrams, an area we plan to enhance in subsequent phases of our research. Our current solution, as discussed, fails to fetch images that are further from its relevant context if it has no relevant description along with it. We aim to improve this this in further advancements of our research.

VIII. CONCLUSION

Our research introduces a new image retrieval method for RAG systems that overcomes the challenges of traditional OCR-LLM approaches, improving multi-modal document comprehension and accurately retrieving images, essential for technical documents and research papers. Through thorough testing, we've shown that our ILT-based approach enhances text-image correlation and retains document structure, producing contextually relevant RAG system responses. Our method outshines current techniques, especially with intricate scientific imagery and non-text-aligned visuals. This work advances image retrieval for multi-modal documents, offering a more effective and economical solution, and paves the way for future enhancements in RAG systems, including better handling of complex images and expanded ILT metadata use.

ACKNOWLEDGMENT

This project was supported by Genpact India Pvt. Ltd.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. t. Yih, T. Rocktäschel, et al. Retrievalaugmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[2] M. Zhou, G. Luo, A. Rohrbach, and Z. Yu. Focus! relevant and sufficient context selection for news image captioning. *arXiv preprint arXiv:2212.00843*, 2022.

[3] Y. Zhou and G. Long. Style-aware contrastive learning for multi-style image captioning. *arXiv preprint arXiv:2301.11367*, 2023.

[4] Z. Yang, W. Ping, Z. Liu, V. Korthikanti, W. Nie, D.-A. Huang, L. Fan, Z. Yu, S. Lan, B. Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023.

[5] Z. Shi, H. Liu, M. R. Min, C. Malon, L. E. Li, and X. Zhu. Retrieval, analogy, and composition: A framework for compositional generalization in image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1990–2000, 2021.

[6] S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, pages 1–7, 2022.

[7] R. Ramos, D. Elliott, and B. Martins. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*, 2023.

[8] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.

[9] A. Khatun and D. G. Brown. Reliability check: An analysis of gpt3's response to sensitive topics and prompt wording. *arXiv preprint arXiv:2306.06199*, 2023.

[10] https://www.kaggle.com/datasets/paultimothymooney/cvpr-2019-papers