EcoDoc: A Cost-Efficient Multimodal Document Processing System for Enterprises Using LLMs

Ravi K. Rajendran*, **Biplob Debnath***, **Murugan Sankaradas and Srimat T. Chakradhar**
Department of Integrated Systems
NEC Laboratories America Inc., Princeton, NJ
{rarajendran,biplob,murugs,chak}@nec-labs.com

Abstract

Enterprises are increasingly adopting Generative AI applications to extract insights from large volumes of multimodal documents in domains such as finance, law, healthcare, and industry. These documents contain structured and unstructured data (images, charts, handwritten texts, etc.) requiring robust AI systems for effective retrieval and comprehension. Recent advancements in Retrieval-Augmented Generation (RAG) frameworks and Vision-Language Models (VLMs) have improved retrieval performance on multimodal documents by processing pages as images. However, large-scale deployment remains challenging due to the high cost of LLM API usage and the slower inference speed of image-based processing of pages compared to text-based processing. To address these challenges, we propose EcoDoc, a cost-effective multimodal document processing system that dynamically selects the processing modalities for each page as an image or text based on page characteristics and query intent. Our experimental evaluation on TAT-DQA and DocVQA benchmarks shows that EcoDoc reduces average query processing latency by up to 2.29× and cost by up to 10×, without compromising accuracy.

1 Introduction

Enterprises are increasingly leveraging Generative AI applications to process and extract insights from vast collections of documents across domains such as finance, legal, healthcare, and industry. These documents contain a mixture of structured (tables, forms) and unstructured (free text, scanned, typewritten, handwritten notes) data, requiring robust AI systems for retrieval, comprehension, and response generation. Recent advancements in Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) frameworks have enabled enterprises to integrate domain-specific retrieval with large-scale generative models, improving contextual relevance in AI-driven document understanding. However, efficiently handling multimodal enterprise documents, those with both textual and visual elements, remains a significant challenge in large-scale deployments due to computational and cost constraints.

Metrics	Text-based (Traditional)	Page Image (VLM)	Text + Page Image (Both)	EcoDoc (ours)
Ingestion time	Slow	Fast	Slowest	Fast
Query processing time	Fast	Slower	Slowest	Fast
Cost (LLM API usage)	Low	Higher	Highest	Low
Accuracy	Low	High	Highest	Highest

Table 1: Comparison of document processing methods on various metrics in the pipeline.

Traditional document processing pipelines primarily relied on text-based retrieval, where documents were parsed through Optical Character Recognition (OCR), and the images were passed through captioning models generating image descriptions as text and stored in retrievable text chunks for downstream processing. While effective for text-heavy documents, this approach struggles with visually complex documents, where critical information is embedded in tables, charts, and layout-specific structures. More recently, Vision-Language Model (VLM)-based indexing and retrievers such as ColPali (Faysse et al., 2025) and VisRAG (Yu et al., 2025) has emerged as a promising alternative, allowing for direct processing of page images without explicit text extraction. This prevents information loss that occurs during OCR-based parsing and enables a richer, more holistic document representation. With VLM-based page embedding techniques, enterprises can now index multimodal documents more efficiently, ensuring both faster retrieval and higher fidelity in captured information (Faysse et al., 2025).

Despite advancements in indexing techniques, the inference phase continues to be a major bot-

*Equal Contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1530–1537

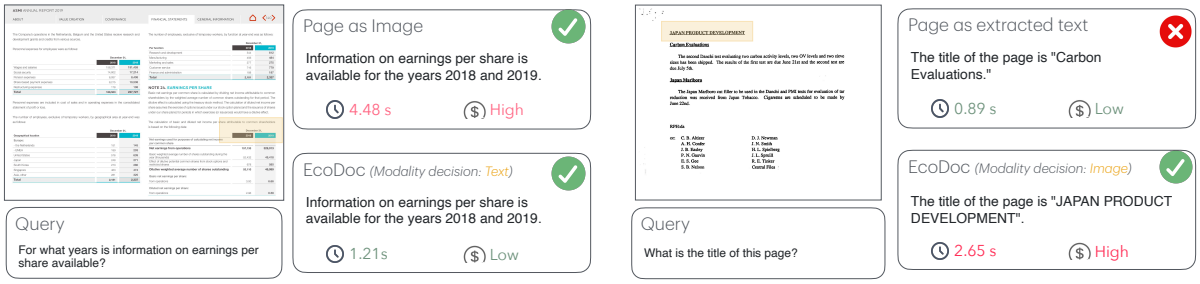

Figure 1: Illustration of EcoDoc's effectiveness in choosing the right representation for inference. Relevant context for the query is highlighted . (a) A query from TAT-DQA (Zhu et al., 2022) containing both tabular and textual data, where EcoDoc opted for cost-efficient text representation over image. (b) A query from DocVQA (Mathew et al., 2021) on a typewritten document, where despite OCR successfully extracting text, the query required position-aware processing, leading EcoDoc to process the page as an image for improved accuracy.

tleneck in large-scale deployments. For example, in a product catalog query scenario, we observed that performing inference using a Vision-Language Model (VLM) incurs approximately 40% higher computational costs and results in twice the latency compared to text-based inference. However, accuracy remains a critical factor for many enterprise applications. To address this, some enterprises (Anthropic, 2024) adopt a dual representation strategy, where each document page is processed as both text and image during inference. While this approach enhances accuracy, it substantially increases computational costs and latency. Table 1 presents a comparative analysis of cost, latency, and accuracy trade-offs across different multimodal document processing approaches.

To optimize the inference phase, this paper introduces EcoDoc, a system that dynamically selects the most efficient representation of a document page for processing through Large Language Models (LLMs). Based on the context of pages relevant to a given query, EcoDoc adaptively chooses between image or text. This adaptive strategy maintains the accuracy benefits while significantly enhancing inference speed and reducing computational costs. As a result, it enables scalable, cost-effective, and accurate document processing for large-scale enterprise applications.

Figure 1 illustrates EcoDoc's effectiveness in selecting the optimal representation during inference. In Figure 1(a), EcoDoc determines that text-based processing is sufficient, enabling lower-cost inference while maintaining accuracy comparable to the image-based approach. On the other hand, in Figure 1(b), EcoDoc selectively opts for image-based processing despite its higher computational cost, ensuring a more accurate response when necessary. This adaptive selection strategy optimizes both efficiency and accuracy based on the specific requirements of the query.

In summary, our contributions in this paper are as follows:

- We propose EcoDoc, a multimodal document processing system designed to optimize cost and latency for large-scale enterprise deployments.
- EcoDoc introduces a dynamic modality selector that intelligently chooses between processing each page as an image or text based on the query and the content of the retrieved pages.
- We evaluate EcoDoc on two benchmarks - DocVQA (Mathew et al., 2021) and TAT-DQA (Zhu et al., 2022), highlighting significant cost savings (up to 10×) and query processing time improvement (up to 2.29×) while maintaining comparable accuracy.

2 EcoDoc System

In this section, we introduce EcoDoc, as shown in Figure 2. EcoDoc extends the Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020) to handle multimodal enterprise documents. It operates in two phases: the indexing phase and the question-answering phase.

2.1 Indexing (Data Ingestion)

In the indexing phase, documents undergo an offline preprocessing step to optimize retrieval efficiency during inference. Rather than applying a

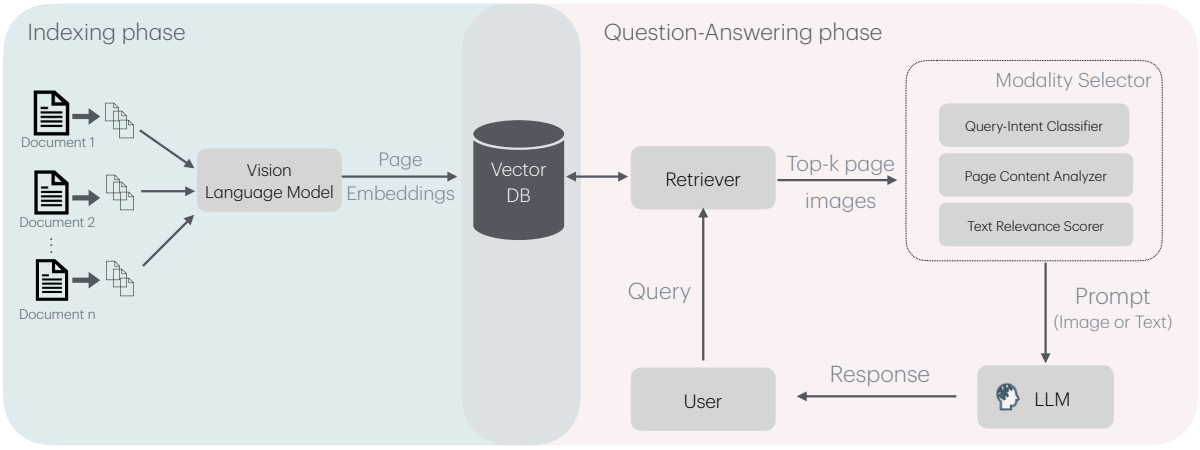

Figure 2: EcoDoc system workflow. The left side shows the document ingestion pipeline, while the right side illustrates the query-answering pipeline. The modality selector dynamically selects the appropriate page representation (i.e., image or text) to be processed by the LLM for generating the answer.

multi-stage pipeline involving Optical Character Recognition (OCR), layout analysis, image captioning, and text chunking for every page, EcoDoc adopts a simplified yet effective approach by converting each page into an image, as proposed in ColPali (Faysse et al., 2025). This image representation is then processed through a vision-language model (VLM), which generates dense embeddings that capture both textual and visual semantics in a unified representation. These embeddings are stored in a vector database. By leveraging page-level embeddings, EcoDoc avoids the computational overhead of extracting and processing individual text and visual elements.

2.2 Question Answering

The question-answering process begins with the retrieval stage, where a retriever selects the most relevant pages from a vector database in response to a given textual query. EcoDoc performs a similarity search over the precomputed page embeddings to identify the top-k pages that are most relevant to the query. These top-k pages then serve as input to the next phase, where answers are generated using large language models (LLMs).

2.2.1 Challenges

Traditionally, page images are passed into the LLM for answer generation. However, processing images directly is considerably more time-consuming and expensive than using their textual counterparts. Our observations indicate that leveraging page images instead of their text versions results in a 2$\times$ increase in average query processing latency and a 40% rise in average cost. This disparity arises because visual data demands greater computational resources and processing power, making image-based queries less efficient.

Notably, not every query necessitates processing the page image. Some queries can be adequately answered using only the text extracted from the page image, while others require the richer context provided by the visual representation. For instance, questions related to visual structure, layout, or non-textual elements of a page may benefit from image-based processing. Conversely, queries centered on textual content can often be resolved more efficiently using the text version alone.

By dynamically selecting the appropriate representation - either the page image or its textual version for each query, we can significantly reduce both the processing time and cost associated with answer generation. The challenge, however, lies in determining when to rely on the image and when to use the text. To address this, EcoDoc employs a sophisticated hybrid approach. It analyzes the content of the retrieved pages, interprets the intent behind the user query, and strategically decides whether to process the image or the text. This strategic selection optimizes resource usage while ensuring accurate and efficient answers.

Now, we describe how EcoDoc addresses these problems by analyzing the contents of the retrieved pages, the intent of the user query, and finally taking a hybrid approach in the following section.

2.2.2 Selecting Right Modality

Given a query, EcoDoc first determines the most suitable representation - image or text, for generating the answer. Although it could rely on LLMs to make this decision directly, invoking the LLM for every query would significantly increase both cost and processing time. To address this challenge, EcoDoc employs a more efficient, precomputed approach by generating two distinct lists of potential questions: one comprising questions best answered using images and the other containing questions that are more effectively addressed through textual representations.

EcoDoc utilizes a *Query-Intent Classifier* that leverages the following prompt to pre-generate a set of representative queries using an LLM.

Prompt for generating list of queries

You are an expert in document understanding. Your task is to generate representative user queries that would be issued to a document question-answering system. For each query, classify the preferred modality required to answer it accurately:

• `"text"`: The query can be answered reliably using only OCR-extracted plain text from the document.

• `"pageimage"`: The query requires visual cues such as layout, spatial relationships, formatting, tables, handwritten elements, or other non-textual features.

Generate a list of 10 diverse queries for each modality. For each query, provide a short explanation of why the specified modality is required.

Expected JSON Output Format:

```
{
  "text_samples": [
    { "query": "...", "reason": "..." },
    ...
  ],
  "pageimage_samples": [
    { "query": "...", "reason": "..." },
    ...
  ]
}
```

Table 2 shows a set of sample queries generated by the LLM. Questions that involve understanding the visual layout, spatial relationships, or visual characteristics of the page often require processing the image representation. In contrast, queries that focus on retrieving specific facts or textual information can typically be answered more efficiently using the text version. For example, questions like "Is there a signature at the bottom?" or "What color is the chart?" rely on visual cues from the image, whereas queries such as "List all items in the table" or "What is the invoice number?" can be addressed directly from the text data.

Sample Queries	
Text-based Inference	**Image-based Inference**
1. What is the invoice number?	1. Is there a signature at the bottom of the page?
2. What is the date of the document?	2. What color is the chart in the top-right corner?
3. Who is the sender of the letter?	3. How many tables are present in the document?
4. List all line items in the invoice.	4. Is there a company logo on the first page?
5. What is the total amount due?	5. What is the title at the top of the document?
6. What is the shipping address?	6. Which section is in bold and underlined?
7. What is the name of the customer?	7. Is there a table with three columns on the page?
8. What are the terms and conditions?	8. Does the document include any handwritten notes?
9. What is the product description listed?	9. What is the label directly above the chart?
10. Who signed the contract?	10. Is the footer visible on the page?

Table 2: Sample query set generated by the LLM for the Query-Intent Classifier.

To classify a new query, EcoDoc computes query text embedding and compares it against the embeddings of the pre-generated questions in both lists. The similarity between the query embedding and each question embedding is computed. For each modality class, the similarity scores across all associated questions are averaged. The corresponding class - image or text, with the highest averaged similarity is then assigned to the query, guiding the decision on whether to process the each of the retrieved pages as an image or text.

If the decision is to use the image representation, the page is directly fed as an image to the LLM to generate an answer. On the other hand, if text-based processing is selected, additional steps are taken to ensure that the extracted text is relevant to the query. To facilitate this, EcoDoc employs a *Page Content Analyzer* that utilizes a layout detector to determine the presence of textual content on the page. If text is detected, the page is processed using an OCR engine to extract the text. The extracted content is then evaluated for relevance to the original query using semantic similarity, computed by the *Text Relevance Scorer*. If the similarity score exceeds a predefined threshold (empirically set to 0.45), the OCR-extracted text is used to generate the answer. Otherwise, the page is processed as an image to ensure that any important visual context is not overlooked.

This hybrid decision-making process enables EcoDoc to balance computational efficiency, cost, and answer accuracy. Visually rich or non-textual pages are processed as images to retain critical context, while pages with relevant, structured text are handled via faster, more cost-effective text-based inference. This adaptive strategy reduces the reliance on expensive image processing while improving the relevance and quality of the answers generated.

3 Performance Evaluation

To assess the performance of EcoDoc, we report the system's efficiency - measured by latency and cost and the accuracy of the generated responses. Response accuracy is evaluated through manual inspection of the generated results.

3.1 Datasets

To evaluate the effectiveness of EcoDoc, we benchmark against two widely used datasets: DocVQA (Mathew et al., 2021) and TAT-DQA (Zhu et al., 2022). These datasets are ideal for evaluating document-and-query-dependent modality selection, as they represent a diverse mixture of textual and visual elements, including images, charts, tables and handwritten texts. TAT-DQA, with its emphasis on financial documents, contains structured text-heavy and tabular data, while the DocVQA, focused on industrial documents, includes more visually rich scanned, typewritten and handwritten texts, offering a balanced evaluation set across different document types.

3.2 Experiment Setup

In the experimental setup, we utilize the ColPali (Faysse et al., 2025) framework provided by Byaldi[1] for indexing. The dataset consists of pages stored as images, which are used to create the embeddings. Since the indexed data only stores compact page embeddings rather than full document images, the system maps the retrieved embeddings back to their corresponding original documents and pages based on the mapping established during data ingestion. EcoDoc's retriever module ensures that the exact source pages are fetched for further processing. Only the top k retrieved pages are passed to the response generation phase and in our experiments, we evaluate top-1 and top-4 retrieval results. For generating responses, we specifically use GPT-4o (OpenAI, 2024), leveraging its capabilities to process the retrieved context and produce accurate answers. We use processing document pages as images as the baseline for comparison.

3.3 Results

In this work, our primary focus is on optimizing inference cost rather than enhancing retrieval accuracy. To ensure a fair evaluation of our proposed techniques, we report accuracy and inference cost metrics only for queries where the top-k retrieved pages contain the necessary context required to generate a correct response using LLMs. By narrowing our evaluation to these cases, we can better isolate the impact of inference cost optimization without conflating it with potential retrieval errors.

Method	DocVQA		TAT-DQA	
	k=1	k=4	k=1	k=4
Baseline	0.52	0.73	0.66	0.70
EcoDoc	0.52	0.73	0.65	0.69

Table 3: Query response accuracy

3.3.1 Query Response Accuracy

To evaluate response accuracy, we compare EcoDoc's adaptive inference strategy against a baseline on the DocVQA and TAT-DQA benchmarks across varying retrieval depths ($k = 1$ and $k = 4$). As shown in Table 3, EcoDoc achieves accuracy on par with the baseline while significantly reducing reliance on image-based processing. On DocVQA, EcoDoc matches the baseline performance with accuracy scores of 0.52 and 0.73 for $k = 1$ and $k = 4$, respectively. On TAT-DQA, EcoDoc attains scores of 0.65 and 0.69, closely approximating the baseline's 0.66 and 0.70. These results indicate that EcoDoc incurs only a marginal 1% reduction in accuracy on TAT-DQA, demonstrating its effectiveness in maintaining high answer quality while optimizing processing efficiency.

3.3.2 Inference cost

To evaluate the inference efficiency, we measure and report the latency and LLM API usage costs for both the baseline and EcoDoc. Figure 3 presents the average response time, showing that while the baseline approach (processing pages as images) achieves high accuracy, it also incurs the highest latency due to the computational need for image-based processing. Similarly, Figure 4 shows the normalized compute cost per query, where EcoDoc demonstrates significantly lower processing costs by efficiently prioritizing text-based inference.

In TAT-DQA, EcoDoc reduced latency by 1.35× and lowered costs by 10× compared to the baseline. In DocVQA, EcoDoc achieved a 2.29× reduction in latency, while cost savings reached 4.17×. The high cost savings in TAT-DQA can be attributed to the higher proportion of text-based processing, which is cheaper. However, the higher latency is due to the complexity of the queries, which re-

[1] https://github.com/AnswerDotAI/byaldi

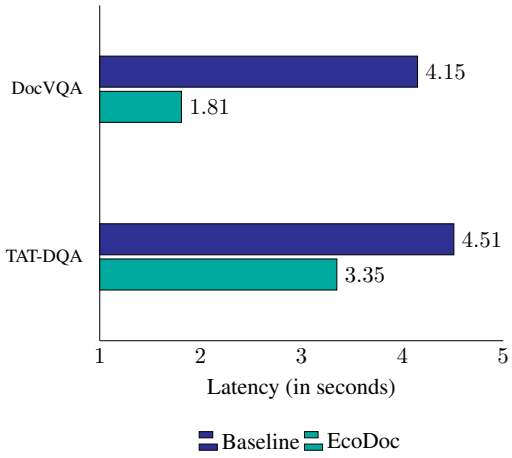

Figure 3: Latency comparison on TAT-DAQ and DocVQA datasets.

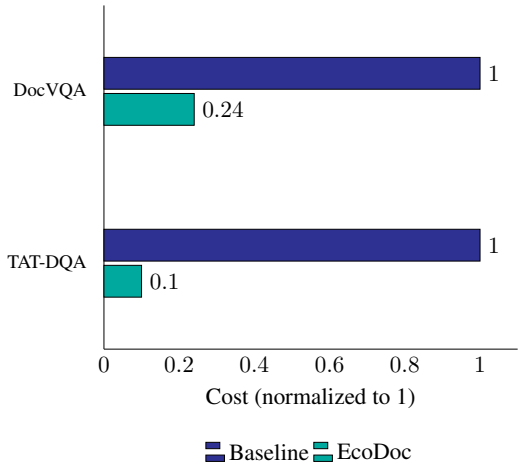

Figure 4: Relative cost comparison on TAT-DAQ and DocVQA datasets.

quire more reasoning and produce longer outputs, thereby increasing inference time. Conversely, in DocVQA, the relatively lower cost savings stem from increased reliance on image-based processing. Nevertheless, the queries in DocVQA require more concise information retrieval, contributing to faster inference. These improvements are driven by EcoDoc's dynamic modality selection, which prioritizes text processing when sufficient and selectively applies image-based inference only when necessary, optimizing both cost and latency.

3.4 EcoDoc Deployment

We describe a deployment use case where EcoDoc is utilized to analyze an extensive product catalog encompassing shipping and packing supplies, as well as other industrial supplies and bulk business goods. The catalog contains thousands of products, each accompanied by a brief description, weight, dimensions, product images, and pricing information. As shown in Figure 5, a user poses the query "What can I use to ship my guitar?". EcoDoc processes the query and retrieves two relevant products from the catalog, suitable for shipping both small and large guitars. Additionally, EcoDoc presents brief information about each recommended product, including its description and specifications. To enhance user confidence and ensure transparency, EcoDoc also displays corresponding product images, allowing the user to visually verify the items and confirm their suitability for shipping needs. Overall, using EcoDoc reduced deployment costs by 70% and processed queries twice as fast compared to the baseline, where each page was always processed as an image.

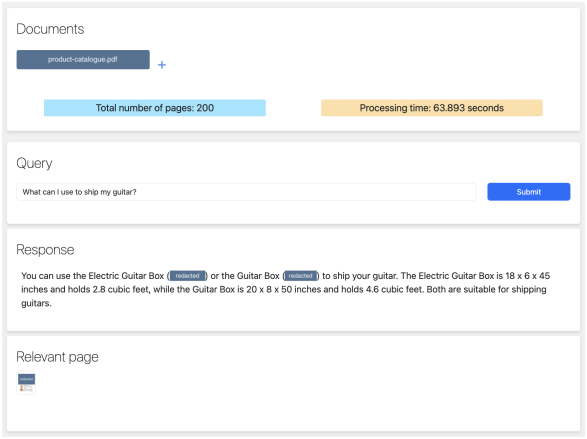

Figure 5: Deployment use case of EcoDoc analyzing a product catalog to identify suitable shipping supplies for a guitar, presenting relevant products with descriptions, specifications, and images to assist the user in making informed decisions.

4 Related Work

Multimodal document processing has garnered significant attention due to its potential in handling diverse tasks across text and image modalities. Multimodal retrieval encompasses tasks such as identifying texts that respond to queries related to specific images (Hu et al., 2023a; Luo et al., 2023), retrieving text-image pairs for question answering (Chang et al., 2022), and finding images that match textual descriptions (Han et al., 2017). To address the diverse nature of these tasks, UniIR Wei et al. (2023) proposed a universal multimodal retrieval model capable of handling a wide range of retrieval scenarios across modalities.

The integration of retrieved multimodal information has proven beneficial for applications like

in-context learning (Tan et al., 2024; Liu et al., 2023) and knowledge incorporation (Hu et al., 2023b; Luo et al., 2021), with use cases spanning from answer generation to image synthesis (Sharifymoghaddam et al., 2024). However, much of the existing research relies on curated academic datasets, where modalities are neatly separated, preprocessed, and aligned (e.g., images with corresponding captions). This structured setup does not fully align with real-world retrieval-augmented generation (RAG) scenarios, where documents often present unstructured and interleaved modalities.

Recent advancements aim to mitigate these challenges by developing models that encode entire document images directly for retrieval tasks. For instance, DSE (Ma et al., 2024), ColPali (Faysse et al., 2025) and VisRAG (Yu et al., 2025) simplify the RAG pipeline by treating documents as images, reducing preprocessing complexity and streamlining retrieval. Nevertheless, these methods introduce new challenges, such as increased query processing times and higher costs associated with large language model (LLM) API usage.

In light of these limitations, EcoDoc proposes a dynamic strategy that intelligently determines when to input image data or text data into the LLM. By evaluating query-specific factors such as content complexity and multimodal context, EcoDoc optimizes the decision-making process to reduce LLM API usage cost and processing overhead. This strategy not only enhances system efficiency but also strikes a balance between leveraging visual and textual information, ensuring improved performance and cost-effectiveness in multimodal document processing.

5 Conclusion

In this work, we introduced EcoDoc, a cost-efficient system for multimodal document processing that optimizes inference by leveraging document structure and query intent. By incorporating text-to-visual density analysis, query-to-page-text semantic similarity, and query intent classification, EcoDoc significantly reduces latency and cost while preserving high accuracy during inference. EcoDoc effectively balances cost and performance, surpassing systems that process multimodal documents solely as images. Through evaluations on datasets from diverse domains, we showed that EcoDoc achieves substantial efficiency improvements without sacrificing response quality.

References

Anthropic. 2024. Build with claude: Pdf support.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of CVPR*, pages 16495–16504.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of ICCV*, pages 1463–1471.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023a. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023b. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of CVPR*, pages 23369–23379.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2023. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *arXiv preprint arXiv:2312.01714*.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of ACL*, pages 8573–8589.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of EMNLP*, pages 6417–6431.

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

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2024. Unirag: Universal retrieval augmentation for multi-modal large language models. *arXiv preprint arXiv:2405.10311*.

Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z Li. 2024. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning. *arXiv preprint arXiv:2405.20834*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *Preprint*, arXiv:2311.17136.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4857–4866. ACM.