Doc-React: Multi-page Heterogeneous Document Question-answering

Junda Wu[1], **Yu Xia**[1], **Tong Yu**[2], **Xiang Chen**[2], **Sai Sree Harsha**[2],
Akash V. Maharaj[2], **Ruiyi Zhang**[2], **Victor Bursztyn**[2], **Sungchul Kim**[2],
Ryan Rossi[2], **Julian McAuley**[1], **Yunyao Li**[2], **Ritwik Sinha**[2]
[1]University of California San Diego [2]Adobe Inc.

Abstract

Answering questions over multi-page, multimodal documents, including text and figures, is a critical challenge for applications that require answers to integrate information across multiple modalities and contextual dependencies. Existing methods, such as single-turn retrieval-augmented generation (RAG), struggle to retrieve fine-grained and contextually relevant information from large, heterogeneous documents, leading to suboptimal performance. Inspired by iterative frameworks like ReAct, which refine retrieval through feedback, we propose Doc-React, an adaptive iterative framework that balances information gain and uncertainty reduction at each step. Doc-React leverages InfoNCE-guided retrieval to approximate mutual information, enabling dynamic sub-query generation and refinement. A large language model (LLM) serves as both a judge and generator, providing structured feedback to iteratively improve retrieval. By combining mutual information optimization with entropy-aware selection, Doc-React systematically captures relevant multimodal content, achieving strong performance on complex QA tasks.

1 Introduction

Answering queries over multi-page, multimodal documents with textual and visual information is a critical challenge (Ma et al., 2024a; Tanaka et al., 2023). These documents, which are typically long, often have diverse layouts and interleaved content (Nguyen et al., 2024), requiring reasoning across modalities. Specifically, as illustrated by the use case in Figure 1, the input is a user query and multi-page documents, either provided by the user or from a database. The goal is to produce an accurate and concise text answer by analyzing the relevant content across pages.

Existing methods (Yu et al., 2024; Yang et al., 2023; Faysse et al., 2024; Yao et al., 2022; Wu et al., 2024d) retrieve multimodal content and feed it into single-turn RAG models. While effective in simpler scenarios, these methods fall short in multi-page settings, missing fine-grained, contextually relevant information. As illustrated in Figure 1, when applying (Yu et al., 2024; Faysse et al., 2024), the query Q does not explicitly mention the region with “252M mobile broadband subscriptions.” Simply retrieving figures semantically similar to Q fails to identify that the region refers to North America and further locate the figure containing North America’s “active social network users” (described in S_2). This highlights the need for an iterative approach to dynamically refine queries, discover relevant figures, and gather necessary information. Inspired by iterative methods (Yao et al., 2022), we adopt a dynamic approach to refine queries iteratively, addressing the challenges of multi-page document QA.

We propose Doc-React, an adaptive iterative framework for multi-page document retrieval and acting. Doc-React refines retrieval and query formulation in distinct steps to maximize normalized information gain, balancing information gain and uncertainty reduction with InfoNCE-guided retrieval. It addresses residual gaps through sub-queries refined by an LLM, which provides feedback as both judge and generator. By maximizing step-wise information gain, Doc-React iteratively captures fine-grained, contextually relevant information, adapting queries and improving retrieval efficiency. In summary, our contributions are:

- We propose Doc-React, an adaptive iterative framework that dynamically refines multimodal retrieval and QA performance through LLM-guided feedback.
- We introduce a method to balance information gain and uncertainty reduction, dynamically adapting sub-queries and retrieval prompts.
- We conduct extensive experiments showing

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 67–78
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

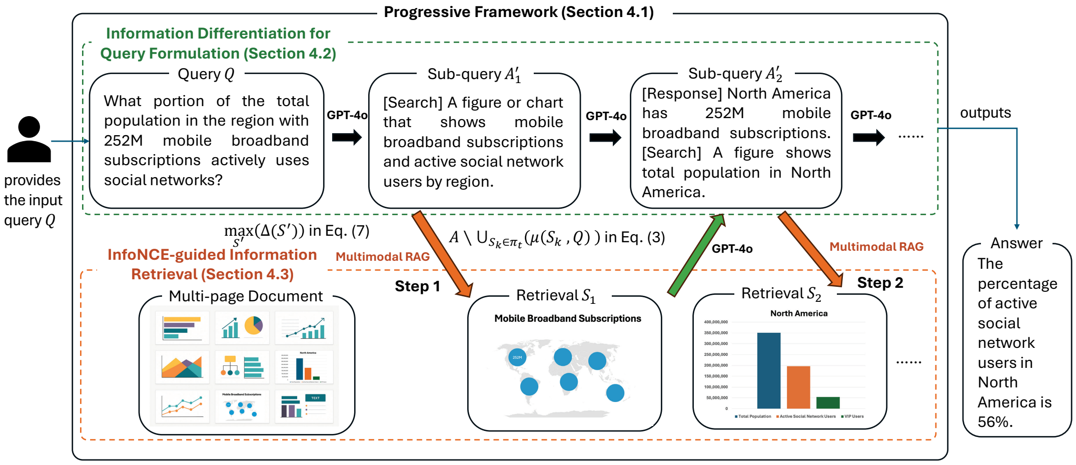
Progressive Framework (Section 4.1)
Information Differentiation for Query Formulation (Section 4.2)
provides the input query Q
Query Q
What portion of the total population in the region with 252M mobile broadband subscriptions actively uses social networks?
GPT-4o
Sub-query A_1'
[Search] A figure or chart that shows mobile broadband subscriptions and active social network users by region.
GPT-4o
Sub-query A_2'
[Response] North America has 252M mobile broadband subscriptions. [Search] A figure shows total population in North America.
GPT-4o
……
outputs
$\max_{S'}(\Delta(S'))$ in Eq. (7)
$A \setminus \cup_{S_k \in \pi_t}(\mu(S_k, Q))$ in Eq. (3)
InfoNCE-guided Information Retrieval (Section 4.3)
Multimodal RAG
Step 1
GPT-4o
Multimodal RAG
Step 2
Multi-page Document
Retrieval S_1
Mobile Broadband Subscriptions
252M
Retrieval S_2
North America
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
……
Answer
The percentage of active social network users in North America is 56%.

Figure 1: Doc-React applied to the multi-page document QA task. The framework processes a user query as input and operates on multi-page documents. It iteratively refines information retrieval and query formulation to maximize information gain and reduce uncertainty, ultimately generating an accurate and contextually relevant answer.

significant improvements over baselines for complex multi-page QA tasks.

2 Related Work

2.1 Iterative and Adaptive Retrieval

Iterative approaches, such as ReAct (Yao et al., 2022), Chain-of-thought (Wei et al., 2022; Wu et al., 2024c,a) and MM-React (Yang et al., 2023), leverage dynamic feedback mechanisms to refine queries and align context (Shinn et al., 2024; Huang et al., 2023; Zhang et al., 2024), effectively addressing challenges in noisy or incomplete retrieval scenarios (Zhou et al., 2024). These methods demonstrate the importance of iterative frameworks (Hu et al., 2024b; Yang et al., 2024) in achieving more accurate and context-aware retrieval. Similarly, mutual information optimization techniques, inspired by InfoNCE (Oord et al., 2018), rely on iterative strategies to balance information gain and uncertainty reduction, further highlighting the relevance of iterative methods for handling complex retrieval tasks. In our framework, InfoNCE-guided mutual information optimization is not only theoretically motivated but also serves as a practical mechanism for adaptive multi-step retrieval (detailed in Section 4).

2.2 Multimodal Document Retrieval

Advancements in multimodal learning enable systems to jointly process text and visual information (Yan et al., 2024; Wu et al., 2025, 2024b; Yao et al., 2024; Liu et al., 2024a). Models like LayoutLM (Xu et al., 2020b,a) leverage spatial embeddings for document layout understanding, excelling at tasks like form and invoice parsing. These models are limited to single-page documents (Liu et al., 2024b) and cannot handle reasoning over multi-page (Ma et al., 2024a), and heterogeneous layouts (Tanaka et al., 2023).

2.3 Retrieval-Augmented Generation (RAG)

RAG (Lewis et al., 2020) and FiD (Izacard and Grave, 2020) combine retrieval and generation for open-domain QA. While effective, their reliance on static pipelines limits adaptability to complex document layouts. Retrieval methods like BM25 (Robertson and Walker, 1994) and dense models such as ColBERT (Khattab and Zaharia, 2020) efficiently match content but fail to capture the cross-modal dependencies for multimodal documents. MuRAR (Zhu et al., 2024) generates multimodal responses to convey rich cross-modal information.

3 Formulation: Multi-page Doc-QA

We introduce a novel formulation for the multi-page document question-answering (Doc-QA) task, where the input consists of heterogeneous documents that interleave images and texts in unstructured, free-form layouts. This practical task involves disentangling multimodal information embedded within the images of document pages to accurately answer a given query. The objective of a multi-page Doc-QA is to extract sufficient multimodal information from the relevant pages of a document and use it to generate a correct answer. Our novel formulation is different from the formulation of existing works (Mathew et al., 2021; Chang et al., 2022; Masry et al., 2022; Liu et al., 2024b) that we consider question-answering with a long sequence of document page images due to heterogeneous layouts of various types of documents. Different from tool-chain based agentic framework (Yang et al., 2023; Ma et al., 2024b; Kumar et al., 2024), our formulation focuses on balancing between information gain and generation uncertainty when multiple images are in the prompts.

Given a N-page document $D = \{p_1, p_2, \ldots, p_N\}$ and a query Q, an underlying policy is required to identify multiple groups of document pages $\pi = \{S_i | S_i \subseteq D\}$, where each group of document pages S_i provides relevant information to a specific aspect of the query Q. The multimodal information necessary to derive the answer A must then be extracted from each subset S_i, conditioned on the query Q, using a multimodal large language model (MLLM) denoted as $\mu(S_i, Q)$. The final output is a text-only response, generated based on the extracted multimodal content. To quantify the informativeness of this process, we measure the mutual information $I(\mu(S_i, Q); A)$ as the amount of information required to determine the answer A.

We further formalize this task as an information optimization problem, which seeks a balance between reducing uncertainty and ensuring information completeness (Alemi et al., 2016):

$$\begin{aligned} \min_{I} \quad & \sum_{S_i \in \pi} H(\mu(S_i, Q)), \quad S_i \subseteq D, \quad \forall S_i \in \pi, \\ \text{s.t.} \quad & I\Big(\bigcup_{S_i \in \pi} \mu(S_i, Q); A\Big) \geq I(\mu(D, Q); A), \end{aligned} \tag{1}$$

where $H(\cdot)$ denotes the entropy, capturing uncertainty in the extracted information. A key practical challenge is the increasing estimation error as the number of input images grows. Empirically, we can observe that there exist subsets $S_1, S_2 \subseteq D$ such that $H(\mu(S_1, S_2, Q)) > H(\mu(S_1, Q)) + H(\mu(S_2, Q))$. This implies that increasing the size of the input set does not always improve answer accuracy, while the increased input size also retrains the usage of MLLMs, due to their limited perception of multimodal context.

4 Method: Doc-React

The formulation's major challenge is incrementally identifying and aggregating multimodal information from a multi-page document corpus (Hu et al., 2024a; Cho et al., 2024). Inspired by information-entropy optimization (Alemi et al., 2016; Still, 2009), we propose **Doc-React**, which greedily maximizes the entropy-normalized information gain for each round of retrieval and action.

Algorithm 1 Doc-React: Adaptive Retrieval with Normalized MI Gain Maximization

```
Require: Document $D = \{p_1, p_2, \ldots, p_N\}$, Query $Q$, MLLM $\mu$, Retrieval model $f(\cdot, \cdot)$, Differentiation prompt $P_{\text{diff}}$, Maximum iteration $T$, Candidate pool $\Omega$.
Initialize $\pi_0 \leftarrow \emptyset$.
for $t = 0$ to $T - 1$ do
    Sub-query Formulation with Residual Information Approximation:
    Formulate sub-queries according to Eq.(5);
    Candidate Evaluation:
    for each $S' \in \Omega$ do
        Estimate the lower bound of the mutual information gain according to Eq.(6);
        Compute entropy $H(\mu(S', Q))$ from MLLM's generation distributions;
        Compute normalized gain $\Delta(S')$ according to Eq.(4);
    end for
    Select Best Subset:
    $S^*_{t+1} \leftarrow \arg\max_{S' \in \Omega'} \Delta(S')$
    Update:
    $\pi_{t+1} \leftarrow \pi_t \cup \{S^*_{t+1}\}$
    if stopping criteria met then
        break
    end if
end for
return Final selected subsets $\pi_T$, and final retrieval results $\mu(\{S_k\}_{k \in \pi_T}, Q)$.
```

4.1 Progressive Framework Refining Retrieval and Query Formulation

At time step t, we maintain a selection of document page subsets π_t, where each subset's $S_k \in \pi_t$ multimodal information is extracted by the MLLM $\mu(S_k, Q)$. To solve the information optimization problem in Eq.(1), we propose to maximize the *normalized information gain* for the next step $t+1$:

$$\Delta(S_{t+1}) = \frac{I\left(\mu(S_{t+1}, Q); A \mid \bigcup_{S_k \in \pi_t} \mu(S_k, Q)\right)}{H(\mu(S_{t+1}, Q))}, \tag{2}$$

where A represents the ground-truth answer to the query. However, different from conventional goal-oriented optimization tasks (Sutton, 2018; Levine et al., 2016), which assume a good representation of the goal, our task poses a challenge of lacking such *oracle information*.

4.2 Information Differentiation

For query formulation, we propose an information differentiation process to approximate each step's maximum information gain,

$$A'_t = A \setminus \bigcup_{S_k \in \pi_t} \mu(S_k, Q), \tag{3}$$

where A'_t denotes the information residual at time step t. Takes Eq.(3) into Eq.(2), we can derive the approximation of the original information gain,

$$\Delta(S_{t+1}) = \frac{I(\mu(S_{t+1}, Q); A'_t)}{H(\mu(S_{t+1}, Q))}, \tag{4}$$

where the residual information is practically approximated by the MLLM

$$\tilde{A}'_t = \mu(\{S_k\}_{k \in \pi_t}, Q, P_{\text{diff}}), \tag{5}$$

which aggregates retrieved information $\{S_k\}_{k\in\pi_t}$ and the query Q, quantifying the information gap.

Theorem 1 (**Information Gain Derivation via Information Differentiation**). *With the information differentiation enabled as A'_t (in Eq.(3)), the information gain $\Delta(S_{t+1})$ in Eq.(2) can be derived via the information differentiation in Eq.(4).*

We prove the equivalency between Eq.(2) and Eq.(4) in Appendix D. Then, this gap is addressed by leveraging the LLM-as-a-judge (Zheng et al., 2023; Rosset et al., 2024; Gu et al., 2024) to evaluate and decompose complex queries (Xia et al., 2024). Specifically, the gap is formulated into sub-queries by following the in-context examples (in Appendix E).

4.3 InfoNCE-guided Information Retrieval

To further track the information gain $I(\mu(S_{t+1}, Q); A'_t)$, we introduce existing multimodal retrieval models (Faysse et al., 2024; Yu et al., 2024) as prior, which can efficiently provide such information quantification through a retrieval process. Given the multimodal retrieval model $f(\cdot, \cdot)$ which measures the distance between multimodal inputs, we can derive the variational lower bound for the mutual information estimation following InfoNCE (Oord et al., 2018),

$$\begin{aligned} I(\mu(S_{t+1}, Q); A'_t) &\geq -\mathcal{L}_{\text{InfoNCE}}(S_{t+1}) \\ &= \mathbb{E}\left[\log \frac{\exp(f(S_{t+1}, [Q, A'_t]))}{\frac{1}{|\Omega|}\sum_{S'\in\Omega}\exp(f(S', [Q, A'_t]))}\right], \end{aligned} \tag{6}$$

where $\Omega = \{S' \mid \forall S' \subseteq D\}$, while in practice we select from a random subset Ω' to reduce the computation complexity. In addition, the estimation uncertainty $H(\mu(S_{t+1}, Q))$ can be calculated based on the MLLM's generation probabilities.

Dataset		MMLongBench		SlideVQA	
Method	Model	F1	Acc	F1	Acc
VisRAG	GPT-4o	29.02	27.78	48.37	46.93
	InternVL	11.43	11.03	18.13	17.28
	MiniCPM	8.28	8.17	12.62	11.83
ColPali	GPT-4o	32.17	30.76	53.42	52.33
	InternVL	8.86	8.57	16.34	15.23
	MiniCPM	7.84	7.74	13.34	12.48
DocReact	GPT-4o	**38.07**	**38.29**	**54.87**	**55.04**

Table 1: Comparison with multimodal retrieval-augmented generation baselines.

Dataset		MMLongBench		SlideVQA	
Method	Model	F1	Acc	F1	Acc
Standard (i.e., w/o CoT)	GPT-4o	**42.00**	**41.71**	53.58	53.41
	InternVL	4.38	4.31	19.00	18.20
	MiniCPM	5.59	5.55	13.85	13.03
CoT	GPT-4o	41.17	40.77	52.12	51.71
	InternVL	10.79	10.47	17.45	16.74
	MiniCPM	6.37	6.13	12.45	11.73
DocReact	GPT-4o	38.07	38.29	**54.87**	**55.04**

Table 2: Comparisons with multi-image multimodal LLM baselines.

Therefore, using Eq.(6) in Eq.(2), we optimize step-wise normalized information gain $\Delta(S_{t+1})$, by maximizing its lower-bound estimation,

$$\begin{aligned} S^*_{t+1} &= \arg\max_{S'\in\Omega} \Delta(S'), \\ \pi_{t+1} &\leftarrow \pi_t \cup \{S^*_{t+1}\}, \end{aligned} \tag{7}$$

for which, we can achieve the near-optimal solution to the original problem in Eq.(1). We illustrate our method in Algorithm 1.

5 Experiments

For our comparison experiments, we use two lines of baselines, including **Multi-image MLLMs:** GPT4-o, InternVL (Chen et al., 2024), and MiniCPM (Yao et al., 2024), suitable for multi-page inputs; and **Multimodal RAG:** ColPali (Faysse et al., 2024) and VisRAG-Ret (Yu et al., 2024). To evaluate baselines and DocReact on multi-page document question-answering, we apply our method on two datasets, **SlideVQA** (Tanaka et al., 2023) and **VisualWebBench** (Liu et al., 2024b). For details, please refer to Appendix C.

Method	MMLongBench		SlideVQA	
	F1	Acc	F1	Acc
ColPali	32.17	30.76	53.42	52.33
DocReact (w/o ColPali)	37.22	37.37	54.12	54.29
DocReact (w/ ColPali)	**38.07**	**38.29**	**54.87**	**55.04**

Table 3: Ablation study comparing Doc-React with and without ColPali retrieval. Results are reported on the MMLongBench and SlideVQA benchmarks.

Comparison with Multimodal RAG We evaluate DocReact's effectiveness against multimodal RAG methods in Table 1. While multimodal RAG methods mitigate input information overload, they struggle with complex queries requiring progressive multi-image reasoning and retrieval. In contrast, DocReact addresses these challenges through step-wise information differentiation and retrieval, achieving strong performance on both datasets.

Comparison with Multi-image MLLMs We evaluate DocReact's efficiency in processing multi-page document inputs compared to multi-image MLLMs. As shown in Table 2, GPT-4o performs well on MMLongBench, where the limited number of candidate images allows it to process the full context effectively. However, *simply relying on GPT-4o (i.e., the baseline Standard) is not scalable for real-world scenarios with a larger number of candidate images*, where DocReact's iterative approach becomes essential for efficiently retrieving fine-grained, contextually relevant information. While CoT reasoning offers limited improvement on GPT-4o due to inefficiencies in handling multimodal inputs, DocReact achieves competitive performance on MMLongBench and surpasses GPT-4o on the SlideVQA dataset, showing its scalability and effectiveness in handling multi-page QA tasks.

Case Study We validate DocReact's advantage by a case study (Appendix A). In this example, the query Q does not explicitly mention the region with "252M mobile broadband subscriptions." Advanced retrieval methods such as ColPali (Faysse et al., 2024) fail to infer that the region refers to North America and subsequently miss the figure containing relevant information on active social network users. In contrast, DocReact dynamically refines its search process. Through intermediate reasoning steps (Thought 2 and Thought 3), it identifies the missing region, retrieves the relevant figure, and gets the information (Observation 3), ultimately yielding the accurate result of 56% (Action 4).

Ablation Study on the Effect of Retriever Instead of relying solely on the strong retrieval approach ColPali (Faysse et al., 2024; Yao et al., 2024; Chen et al., 2024), we conducted additional experiments to compare **DocReact (w/ ColPali)** and **DocReact (w/o ColPali)**. In the latter setting, we replaced ColPali with a weaker retrieval method as indicated in Table 3. This comparison allows us to better understand whether the observed improvements are attributable to our framework itself, rather than being solely due to the strong retrieval backbone. Our experiments demonstrate that even without leveraging the strong ColPali method, our approach still outperforms directly using ColPali.

6 Conclusion

We address the challenge of answering questions over multi-page, multimodal documents by introducing Doc-React, an adaptive iterative framework for retrieval and acting. By balancing information gain and uncertainty reduction at each step, the framework leverages InfoNCE-guided retrieval and a large language model (LLM) for dynamic query refinement. Through systematic optimization and feedback, Doc-React effectively captures fine-grained and contextually relevant information, and achieves strong performance.

7 Limitations

Our approach maximizes the normalized mutual information gain, which is achieved through an iterative framework that approximates this objective (as discussed in the theoretical analysis in Appendix D). While the framework effectively refines retrieval and query formulation, the approximation process may lead to slight inefficiencies in specific cases. Future work could focus on refining the optimization process to further streamline the approximation.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024b. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Di Huang, Ziyuan Nan, Xing Hu, Pengwei Jin, Shaohui Peng, Yuanbo Wen, Rui Zhang, Zidong Du, Qi Guo, Yewen Pu, et al. 2023. Anpl: towards natural programming with interactive decomposition. *Advances in Neural Information Processing Systems*, 36:69404–69440.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Somnath Kumar, Yash Gadhia, Tanuja Ganu, and Akshay Nambi. 2024. Mmctagent: Multi-modal critical thinking agent framework for complex visual reasoning. *arXiv preprint arXiv:2405.18358*.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024b. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024a. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, et al. 2024b. Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action. *arXiv preprint arXiv:2412.05479*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, et al. 2024. Gui agents: A survey. *arXiv preprint arXiv:2412.13501*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents. *arXiv preprint arXiv:2402.17896*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Susanne Still. 2009. Information-theoretic approach to interactive learning. *Europhysics Letters*, 85(2):28005.

Richard S Sutton. 2018. Reinforcement learning: An introduction. *A Bradford Book*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, et al. 2024a. Ocean: Offline chain-of-thought evaluation and alignment in large language models. *arXiv preprint arXiv:2410.23703*.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian J McAuley. 2024b. Commit: Coordinated instruction tuning for multimodal large language models. *CoRR*.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, et al. 2025. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024c. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024d. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A Rossi, Haoliang Wang, and Julian McAuley. 2024. Knowledge-aware query expansion with large language models for textual and relational retrieval. *arXiv preprint arXiv:2410.13765*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. 2024. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*.

Yujing Yang, Boqi Chen, Kua Chen, Gunter Mussbacher, and Dániel Varró. 2024. Multi-step iterative automated domain modeling with large language models. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, pages 587–595.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mmreact: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xianren Zhang, Xianfeng Tang, Hui Liu, Zongyu Wu, Qi He, Dongwon Lee, and Suhang Wang. 2024. Divide-verify-refine: Aligning llm responses with complex instructions. *arXiv preprint arXiv:2410.12207*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *arXiv preprint arXiv:2410.23856*.

Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. 2024. Murar: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering. *arXiv preprint arXiv:2408.08521*.

A Case Study

In this section, we illustrate a case study about the comparison between DocReact and CoT baseline. The desired answer, according to the ground truth, is 56%. We show in the use case that DocReact can successfully enable information verbal differentiation by identifying the missing information in the current observation (highlighted in green font). In addition, DocReact can further formulate sub-queries based on the information differentiation results, which can be used to retrieve more document image potentially containing such information (highlighted in orange font).

Question: What percentage of the TOTAL POPULATION in the region where there are 252M mobile broadband subscriptions are ACTIVE SOCIAL NETWORK USERS?

(i) The result by our DocReact

Thought 1: I need to find the region where there are 252M mobile broadband subscriptions and then determine the percentage of the total population in that region who are active social network users.

Action 1: Search[A figure or chart that shows mobile broadband subscriptions and active social network users by region.]

Observation 1: 6,572,950,124 mobile subscriptions

Thought 2: The observation does not provide the necessary information about the region with 252M mobile broadband subscriptions. I need to search again for a figure or chart that shows mobile broadband subscriptions and active social network users by region.

Action 2: Search[A figure or chart that shows mobile broadband subscriptions and active social network users by region, specifically looking for 252M mobile broadband subscriptions.]

Observation 2: North America has 252M mobile broadband subscriptions

Thought 3: I need to search for a figure or chart that shows active social network users in North America.

Action 3: Search[A figure or chart that shows active social network users in North America.]

Observation 3: 56% North America's share of active social network

Action 4: Finish[56%]

(ii) The result by the baseline Standard

The image does not provide a direct percentage of active social network users in the region with 252M mobile broadband subscriptions. However, it does show that there are 173,871,400 active social network users in the region with 252M mobile broadband subscriptions. To calculate the percentage, we would need the total population of that region, which is not provided in the image.

(iii) The result by the baseline CoT

The image shows a bar graph with different regions and their respective mobile broadband subscriptions and active social network users. To find the percentage of active social network users, we need to look at the bar corresponding to the region with 252M mobile broadband subscriptions. The bar indicates that the percentage of active social network users in that region is 17%. Therefore, the answer is 17%.

B Implementation Discussions

While our approach may involve more API calls than baseline methods due to its iterative nature, it is important to note that iterative frameworks are both necessary and common for addressing complex retrieval tasks. Prior works, such as ReAct (Yao et al., 2022), have demonstrated the effectiveness of dynamic feedback mechanisms for refining queries in noisy or incomplete scenarios. Similarly, mutual information optimization methods inspired by InfoNCE (Oord et al., 2018) rely on iterative strategies to balance information gain and uncertainty reduction effectively.

Doc-React builds on these principles by leveraging an iterative query refinement framework to progressively retrieve the most relevant information. This approach significantly reduces the number of input tokens processed per API call by avoiding redundant or irrelevant content. Consequently, while iterative, Doc-React achieves a lower overall computational cost in terms of processed tokens compared to baseline methods, which often rely on processing larger, less refined multimodal contexts in a single pass. The iterative design is not only practical but also essential for tackling the complexity of multi-page, multimodal document QA tasks. Given the markedly inferior performance of open-sourced MLLMs (e.g., MiniCPM and InternVL) on such challenging tasks, we opt to build DocReact on the stronger baseline GPT-4o to enable a more rigorous comparison. In practice, we limit the process to a maximum of 4 iterations of retrieval and action steps. The MLLM has the flexibility to terminate early by directly outputting the final answer. However, if the MLLM does not voluntarily end the interaction, the system enforces termination after the final iteration. At this point, the answer is generated using aggregated information from all previous rounds.

C Experiment Details

C.1 Dataset Details

We evaluate our method on the following datasets:

- **SlideVQA** (Tanaka et al., 2023): Designed for structured documents like presentation slides, focusing on reasoning with visual elements such as diagrams and charts.
- **MMLongBench-Doc** (Ma et al., 2024a): Evaluates tasks requiring long-context reasoning over multimodal documents.

C.2 Baseline Details

We include the following baselines in our comparisons:

- **Multi-image MLLMs:** GPT4-o, InternVL (Chen et al., 2024), and MiniCPM (Yao et al., 2024).
- **Multimodal RAG:** ColPali (Faysse et al., 2024) and VisRAG-Ret (Yu et al., 2024).

C.3 Implementation Details

- **Framework:** Implemented using the Hugging Face Transformers library for integration with retrieval and generative models.
- **Query Refinement:** Iterative query refinement is driven by LLM feedback and InfoNCE-based optimization.
- **Hyperparameters:** We set generation temperature as 0 for deterministic decoding. We set the token limit of 100 for generation and the maximum number of images input as 50 for multi-image MLLM baselines.
- **Evaluation Metrics:** We follow the evaluation metrics introduced by MMLongBench-Doc (Ma et al., 2024a) where Acc and F1 are used to meature QA performance.
- **Hardware:** Experiments are conducted on 4 NVIDIA A6000 GPUs with multi-GPU support.

D Theoretical Analysis

This section provides more details and the proof of Theorem 1.

Proof. Denote the accumulated information as

$$Z_t = \bigcup_{S_k \in \pi_t} \mu(S_k, Q).$$

We can assume the information differentiation by

$$A = g(Z) + A'_t, \tag{8}$$

where $g(Z)$ is the part of A explained by Z. By definition of conditional mutual information, the information gain term in Eq.(2) can be derived

$$\begin{aligned} I(\mu(S_{t+1}, Q); A \mid Z) = H(A \mid Z) \\ - H(A \mid \mu(S_{t+1}, Q), Z). \end{aligned}$$

Using the decomposition assumption (in Eq.(8)) of information differentiation, we show that

$$\begin{aligned} H(A \mid Z) = H(A'_t \mid Z), \\ H(A \mid \mu(S_{t+1}, Z) = H(A'_t \mid \mu(S_{t+1}, Z). \end{aligned}$$

In addition, since the differentiated information A'_t is assumed to be independent from the currently collected information Z_t, we can further reduce the results as

$$\begin{aligned} & H(A'_t \mid Z) = H(A'_t), \\ & H(A'_t \mid \mu(S_{t+1}, Z) = H(A'_t \mid \mu(S_{t+1}). \end{aligned}$$

Therefore, we could conclude with the equivalency

$$\begin{aligned} I(\mu(S_{t+1}; A \mid Z) = H(A'_t) - H(A'_t \mid \mu(S_{t+1})) \\ = I(\mu(S_{t+1}); A'_t). \end{aligned}$$

□

E In-context Examples

We illustrate the three in-context learning examples we used to align MLLMs' behaviors in generating thoughts of information differentiation (**Thought** t), actions of multimodal search sub-queries (**Action** t), and MLLM received feedback back through MLLM information retrieval (**Observation** t).

Example 1:

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny extends into, then find the elevation range of the area.
Action 1: Search[A figure of the regional map of Colorado orogeny, where eastern sector is displayed.]
Observation 1: Based on the give images, we can observe as follows. The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas. This took place from 1780 to 1650 million years ago (Mya), during the Paleoproterozoic (Statherian Period). It is recorded in the Colorado orogen, a >500-km-wide belt of oceanic arc rock that extends southward into New Mexico. The Colorado orogeny was likely part of the larger Yavapai progeny.
Thought 2: It does not mention the eastern sector of the Colorado orogeny. I need to look up eastern sector.
Action 2: Lookup[Extract the eastern sector information from the given figure.]
Observation 2: Based on the give images, we can observe as follows. The Colorado orogen, formerly called the Colorado province, is a >500-km-wide belt of oceanic arc rock (1.78–1.65 Ga) that extends southward into New Mexico and composes a major part of the Proterozoic provinces of southwestern United States. This transcontinental collisional event occurred during the Paleoproterozoic (Statherian Period). The Wyoming sector of the Colorado orogeny was formerly called the Medicine Bow orogeny. The eastern sector extends into the High Plains and is called the Central Plains orogeny. The boundary between the Colorado orogeny and the Wyoming craton is the Cheyenne belt, a 5-km-wide mylonitic shear zone that verges northward. The Cheyenne belt transects and cuts off the south edge of the older Trans-Hudson progeny.
Thought 3: The eastern sector of Colorado orogeny extends into the High Plains, so High Plains is the area. I need to search High Plains and find its elevation range.
Action 3: Search[A figure of High Plains, where its elevation range is annotated.]
Observation 3: Based on the give images, we can observe as follows. The High Plains are a subregion of the Great Plains, mainly in the Western United States, but also partly in the Midwest states of Nebraska, Kansas, and South Dakota, generally encompassing the western part of the Great Plains before the region reaches the Rocky Mountains. The High Plains are located in eastern Montana, southeastern Wyoming, southwestern South Dakota, western Nebraska, eastern Colorado, western Kansas, eastern New Mexico, western Oklahoma, and to just south of the Texas Panhandle. The southern region of the Western High Plains ecology region contains the geological formation known as Llano Estacado which can be seen from a short distance or on satellite maps. From east to west, the High Plains rise in elevation from around 1,800 to 7,000 ft (550 to 2,130 m).
Thought 4: High Plains rise in elevation from around 1,800 to 7,000 ft, so the answer is 1,800 to 7,000 ft.
Action 4: Finish[1,800 to 7,000 ft]

Example 2:

Question: Which magazine was started first Arthur's Magazine or First for Women?
Thought 1: I need to search Arthur's Magazine and First for Women, and find which was started first.
Action 1: Search[A wiki document page about Arthur's Magazine.]
Observation 1: Based on the give images, we can observe as follows. Arthur's Magazine (1844–1846) was an American literary periodical published in Philadelphia in the 19th century. Edited by Timothy Shay Arthur, it featured work by Edgar A. Poe, J.H. Ingraham, Sarah Josepha Hale, Thomas G. Spear, and others. In May 1846 it was merged into Godey's Lady's Book.
Thought 2: Arthur's Magazine was started in 1844. I need to search First for Women next.
Action 2: Search[A webpage introduction about First for Women.]
Observation 2: Based on the give images, we can observe as follows. First for Women is a woman's magazine published by Bauer Media Group in the USA. The magazine was started in 1989. It is based in Englewood Cliffs, New Jersey. In 2011 the circulation of the magazine was 1,310,696 copies.
Thought 3: First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.
Action 3: Finish[Arthur's Magazine]

Example 3:

Question: Guitars for Wounded Warriors is an album that was recorded in the village in which New York county?
Thought 1: I need to search Guitars for Wounded Warriors, then find in which New York county it was recorded.
Action 1: Search[A document page about Guitars for Wounded Warriors.]
Observation 1: Based on the give images, we can observe as follows. Guitars for Wounded Warriors is the sixth solo studio album by guitarist Brian Tarquin, released in August 2014 by Cleopatra Records. In 2016, it received a Best Compilation Album nomination by the Independent Music Awards. All tracks were recorded at Tarquin's Jungle Room Studios in New Paltz (village), New York. Being moved by the lack of support for returning veterans through his life Tarquin decided to compose a heavy instrumental rock album as a way to show his appreciation to all veterans. So once again he enlisted top notch players to guest with him on the album, Billy Sheehan on Black Hawk, Ron "Bumblefoot" Thal and Reb Beach on Taliban Terror, Steve Morse on Freedom, Chris Poland on Charlie Surfs and Hunting, Chuck Loeb on Escape Kabul, Hal Lindes on Sand & Blood, Gary Hoey on 5 Klicks To Hell and Baghdad, Alex De Rosso Dokken on Broken Arrow, and The Flyin' Ryan Brothers on Surgical Strike. The entire album was engineered, produced and composed by Tarquin especially for each one of the guest guitarists. Partial proceeds are donated to the Fisher House Foundation from sales.
Thought 2: Guitars for Wounded Warriors was recorded in New Paltz (village), New York. I need to search New Paltz (village), New York and find which county it is in.
Action 2: Search[A figure of map that indicates the location of New Paltz (village), New York.]
Observation 2: Based on the give images, we can observe as follows. New Paltz is a village in Ulster County located in the U.S. state of New York. It is approximately 80 miles (130 km) north of New York City and 70 miles (110 km) south of Albany. The population was 7,324 at the 2020 census.
Thought 3: New Paltz (village), New York is in Ulster County, so the answer is Ulster County.
Action 3: Finish[Ulster County]