Hi3Doc: Hierarchical Tri-Level Representations for Multimodal Long-Document Understanding

Supplementary Material

Table 4. Overall accuracy of general-purpose LVLMs and our Hi3Doc framework. Hi3Doc, when combined with lightweight GPT-mini backbones, consistently outperforms strong commercial LVLMs, demonstrating the competitiveness of retrieval-plus-reasoning pipelines for long-document understanding.

LVLM	MMLongBench-Doc	LongDocURL	SlideVQA	DUDE
GPT-4o (up to 120 pages)	0.428	0.343	0.713	0.281
GPT-5mini (up tp 50 pages)	0.281	0.313	0.707	0.359
Gemini-2.5-Pro (up tp 120 pages)	0.438	0.568	0.795	0.447
Hi3Doc (GPT-4mini)	0.524	0.653	0.802	0.508
Hi3Doc (GPT-5mini)	**0.561**	**0.661**	**0.835**	**0.570**

6. Benchmark Details

We evaluate our method on four representative multimodal long-document benchmarks: MMLongBench-Doc [9], LongDocURL [5], SlideVQA [13], and DUDE [15]. As summarized in Table 5, the queries have a comprehensive coverage of text (TXT), table (TAB), chart (CHA), layout (LAY), and image (IMG) according to the ground-truth evidence modalities, and their supporting evidence can be distributed within a single page, across multiple pages, or be absent (unanswerable). The underlying documents span common long-document formats such as PDF, PPTX, and PNG, and cover diverse real-world domains, including academic, financial, governmental, instructional, and workplace documents, slide decks, as well as web pages from large-scale book, document, and multimedia archives. DUDE and SlideVQA contain relatively short documents (average lengths 5.57 and 20.0, respectively) and mainly test fine-grained modeling and understanding of multimodal elements. In contrast, MMLongBench-Doc and LongDocURL consist of much longer documents (average lengths 86.5 and 47.5), placing greater emphasis on a model's ability to retrieve and aggregate evidence across ultra-long, multi-page contexts.

7. Hi3-HPC Implementation Details

To perform hierarchical semantic modeling of long documents, we first extract fine-grained element information at the page level and then aggregate it across pages at the block level, yielding page-block representations with clear structure and coherent semantics.

As shown in Fig. 7, we configure the LLM as a multimodal document analyzer and instruct it, via prompts, to identify key element types on each page (text paragraphs, images, tables, charts, diagrams, etc.), to describe the topic of each element in one or two sentences of natural language, and to provide a high-level page summary for the entire page. The output is represented as a combination of "page summary + element list," which explicitly preserves element info and provides fine-grained cues for subsequent cross-page aggregation based on element semantics. We then perform the page-block summary stage. As illustrated in Fig. 8, we feed the single-page summaries of a sequence of consecutive pages (e.g., groups of 20 pages) into the LLM and ask it to group consecutive pages discussing the same topic into several page blocks according to the semantic relations between the single-page summaries and their element info. For each block, the LLM generates a higher-level chunk summary and outputs the corresponding page range as well as the pages where key elements appear.

Through this two-stage procedure, we retain element-level semantic labels at the page level and obtain cross-page chunk representations with coherent topics and clear boundaries at the block level, providing structured semantic priors for subsequent cross-page retrieval and multimodal evidence localization.

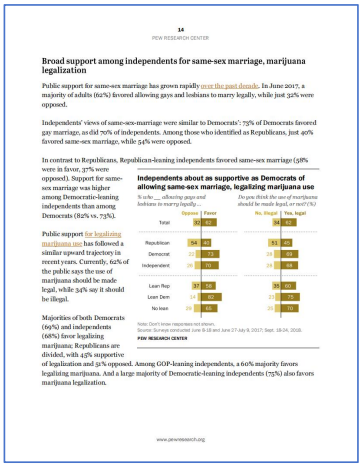
8. VQA Performance Analysis

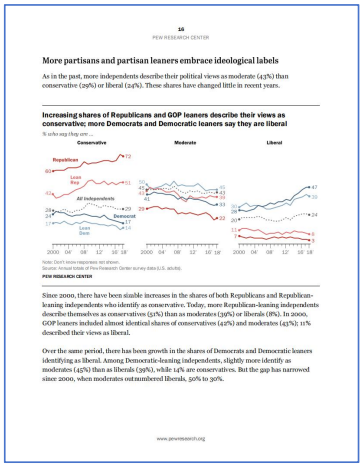
8.1. Compared to General LVLM

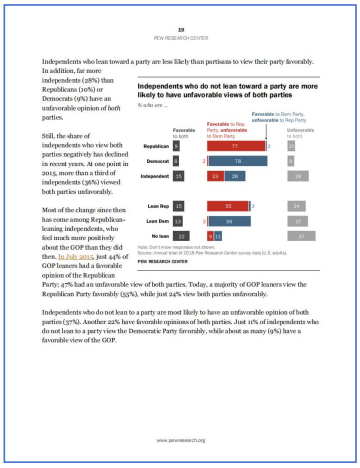
We further compare Hi3Doc with recent commercial LVLMs that directly take document images as input, including GPT-4o [10], GPT-5mini [12], and Gemini-2.5-Pro [8]. As summarized in Table 4, these LVLMs are evaluated in an end-to-end setting with their native page limits (50 or 120 pages), whereas Hi3Doc employs the same backbone understanding models (GPT-4mini and GPT-5mini) but is equipped with our retrieval-centric long-document pipeline. Across all four benchmarks, Hi3Doc consistently outperforms the pure LVLM baselines, even with the smaller GPT-4mini backbone, indicating that the benefits mainly come from our hierarchical retrieval and evidence aggregation rather than from the scale of the backbone model. These results demonstrate that simply scaling up general-purpose LVLMs is insufficient for robust long-document VQA: end-to-end models remain constrained by context length and struggle to reason over dispersed evidence in

Single-page Summary Prompt Template

You are an expert multimodal document analyst. For the given document page:
1. Identify each key element (especially **focus on following element type:text paragraph, image, chart, table, diagram**). Ignore any minor or irrelevant elements that do not contribute to understanding the page content.
2. For each element, briefly describe the topic or content **in one or two sentence**.
3. At the end, provide a one-sentence summary of what the entire page is about.
Output format(plain text only, no JSON):
<element1 type>: <short description of the topic>
<element2 type>: <short description of the topic>
...
So the page talks about <overall page summary>.







Summary of Page 14

text paragraph: Public support for same-sex marriage has grown rapidly over the past decade, with 62% of adults favoring it in June 2017.
chart: A chart showing the percentage of people who oppose and favor allowing gays and lesbians to marry legally, as well as the percentage who think the use of marijuana should be made legal or not.
text paragraph: Independents' views of same-sex marriage were similar to Democrats', with 73% of Democrats and 70% of independents favoring gay marriage.
Page Summary : So the page talks about public opinion on same-sex marriage and marijuana legalization across different political affiliations.

Summary of Page 16

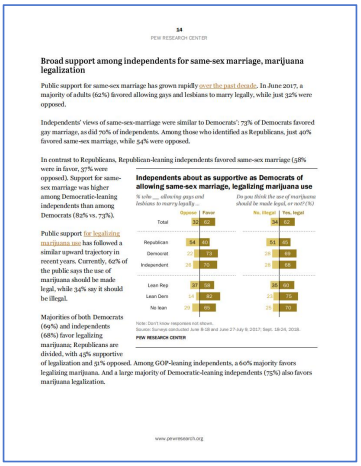
text paragraph: More partisans and partisan leaners embrace ideological labels, with increasing shares of Republicans and GOP leaners describing their views as conservative, while more Democrats and Democratic leaners say they are liberal.
chart: Line graph showing trends in the percentage of Republicans, Democrats, and Independents who identify as conservative, moderate, or liberal from 2000 to 2018.
Page Summary : So the page talks about the changing political ideologies among Republicans, Democrats, and Independents over time.

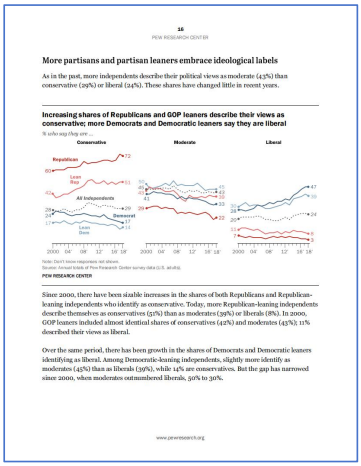
Figure 7. Single-page summary prompt template. The LLM is prompted to act as a multimodal document analyst, identify key elements on each page (e.g., text paragraphs, tables, charts, images), generate a brief description for each element, and produce a one-sentence summary of the overall page content.

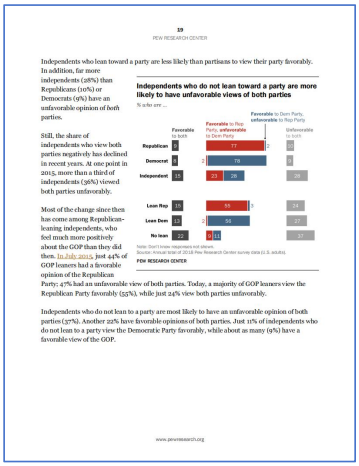
Page-block Summary Prompt Template

You are an expert document analyst.You will be given a list of information for consecutive document pages from a document.Your task is to group consecutive pages into coherent page ranges, where each range covers pages that discuss the same topic.
Different ranges must focus on distinct topics.

Output Format:
[start page number, end page number]: summary, element information
Note:
the summary should be short and concise in one or two sentence.
the element information should focus on visually-rich element like table, figure, chart, diagram, you should tell me which page contains what element.







Chunk Summary

{"[14:15]": "Public opinion on same-sex marriage and marijuana legalization, including differences among independents and leaners. Element information: Page 14 has a chart tracking support/opposition for same-sex marriage and legalization of marijuana; Page 15 gives subgroup figures for non-leaning independents (textual with summary percentages).",

"[16:17]": "Trends in ideological self-identification (conservative/moderate/liberal) among Republicans, Democrats, and independents, and the stable moderate tilt of non-leaning independents. Element information: Page 16 features a line graph showing ideology trends (2000–2018); Page 17 summarizes the distribution of ideology among non-leaning independents (text).",

"[18:21]": "Party favorability and the rise of partisan antipathy over time, showing growing very unfavorable views of the opposing party among leaners and partisans. Element information: Page 18 includes a line graph of favorability patterns to both parties (1994–2018); Page 19 presents a bar chart of favorable/unfavorable ratings by group (Republican, Democrat, independents, leaners, no-lean); Page 20 references trend charts of unfavorable views over time; Page 21 gives numerical trend highlights (textual summary).",}

Figure 8. Page-block summary prompt template. Given a sequence of single-page summaries, the LLM groups consecutive pages discussing the same topic into coherent page blocks, and for each block outputs a higher-level chunk summary together with the corresponding page range and key element information.

Table 5. Overview of evidence distribution, evidence modality, document length, and domains for the multimodal long-document QA benchmarks used in our experiments.

Benchmark	Evidence Distribution			Evidence Modality					Average Document Length	Topic/Domains
	Single	Cross	Una	TXT	LAY	CHA	TAB	IMG		
MMLongBench-Doc	✔	✔	✔	✔	✔	✔	✔	✔	86.5	Academic, financial, governmental, and instructional documents
LongDocURL	✔	✔	✔	✔	✔		✔	✔	47.5	Research, instructional, and workplace documents
SlideVQA	✔	✔		✔	✔	✔	✔	✔	20.0	Slidedecks
DUDE	✔	✔		✔	✔	✔	✔	✔	5.57	Web pages from large-scale book, document, and multimedia archives

Table 6. Statistics on MMLongBench-Doc based on evidence modality and evidence distribution. Hi3Doc achieves the best performance across different cross-modal QA settings, and significantly outperforms other methods in the challenging case where the answer is distributed across multiple pages. The best results are marked in bold, and the second-best results are underlined.

Method	Model	Evidence Modalities					Evidence Locality			Overall	
		TXT	LAY	CHA	TAB	IMG	SIN	MUL	UNA	ACC	F1 Score
OCR+LLM	Mixtral-Instruct-V0.1 [1]	0.342	0.213	0.195	0.213	0.192	0.277	0.219	0.324	0.269	0.247
	GPT-4o [10]	0.411	0.234	0.285	0.381	0.224	0.354	0.293	0.186	0.301	0.305
LVLM	InternVL-Chat-V1.5 [3]	0.140	0.162	0.071	0.101	0.166	0.149	0.122	0.175	0.146	0.130
	GPT-4o [10]	0.463	0.460	0.453	0.500	0.441	0.545	0.415	0.202	0.428	0.449
RAG-based	M3DocRAG [4]	0.300	0.235	0.189	0.201	0.208	0.324	0.148	0.580	0.210	0.226
	MMRAG-DocQA [7]	0.459	0.344	0.449	0.511	0.375	0.535	0.368	**0.762**	0.523	0.460
	Colpali [6]+GPT4.1mini [11]	0.454	0.441	0.487	0.448	0.426	0.601	0.307	0.422	0.464	0.462
	Colpali [6]+GPT5mini [12]	0.524	0.468	0.492	0.489	0.497	0.639	0.358	0.488	0.513	0.519
	Hi3Doc+GPT4.1mini [11]	0.491	0.479	0.531	0.557	0.480	0.620	0.442	0.443	0.524	0.518
	Hi3Doc+GPT5mini [12]	**0.573**	**0.596**	**0.560**	**0.643**	**0.551**	**0.671**	**0.504**	0.418	**0.561**	**0.569**

very long documents. In contrast, Hi3Doc's retrieval-aware design effectively decomposes long documents into semantically coherent chunks and presents focused evidence to the backbone model, allowing a lightweight GPT-mini family model to exceed the performance of much larger commercial LVLMs. This confirms that a specialized retrieval-plus-reasoning pipeline remains highly competitive, even in the presence of strong off-the-shelf LVLMs.

8.2. Detailed VQA Results on MMLongBench-Doc

Table 6 reports the detailed breakdown of model performance on MMLongBench-Doc across evidence modalities and evidence locality, complementing the visual summaries in Fig. 3 and Fig. 4 of the main paper. For each model, we report accuracy on Text (TXT), Layout (LAY), Chart (CHA), Table (TAB), and Image (IMG) evidence, as well as on Single-page (SIN), Cross-page (MUL), and Unanswerable (UNA) cases, together with overall accuracy (ACC) and F1 score. Consistent with the observations in Figure 3, Hi3Doc+GPT5mini achieves the best or near-best accuracy on all five modalities, with particularly strong gains on layout- and chart-intensive queries where structural understanding is crucial.

The locality columns in Table 6 provide a more fine-grained view of the evidence distribution analysis in Fig. 4. Hi3Doc+GPT5mini achieves the highest accuracy on both single-page (SIN) and cross-page (MUL) questions, and Hi3Doc+GPT4.1mini consistently ranks second, confirming that Hi3Doc effectively supports retrieval and reasoning over long documents, especially when relevant evidence is dispersed across multiple pages. In contrast, conventional RAG baselines (M3DocRAG, MMRAG-DocQA, and Colpali-based variants) are noticeably weaker on cross-page cases. On unanswerable questions (UNA), Hi3Doc-based models lag behind some baselines, and we further discuss this limitation and potential improvements in Section 4.4.2.

8.3. Detailed VQA Results on LongDocURL

Table 7 reports the detailed results on LongDocURL under different retrieval methods and task settings. For both ChatGPT-4.1mini and ChatGPT-5mini, Hi3Doc consistently achieves the best overall accuracy (0.654 and 0.661, respectively), outperforming all four RAG baselines (Colpali, VisRAG, VDocRAG, and V-RAG). Across

Table 7. Statistics on LongDocURL based on evidence modality, evidence distribution and task type. Hi3Doc achieves the best performance across different cross-modal QA settings, and significantly outperforms other methods in the challenging case where the answer is distributed across multiple pages. The best results are marked in bold, and the second-best results are underlined.

Retrieval Method	Understanding Model	Evidence Modality				Evidence Distribution		Task Type			Overall Accuracy
		TXT	TAB	FIG	LAY	SIN	MUL	Uude	Reas	Loca	
Colpali [6]	ChatGPT-4.1mini [11]	0.656	0.639	0.661	0.525	0.692	0.552	0.669	0.551	0.564	0.616
VisRAG [17]		0.637	0.618	0.642	0.508	0.663	0.542	0.652	0.550	0.531	0.599
VDocRAG [14]		0.403	0.435	0.485	0.319	0.447	0.361	0.424	0.401	0.362	0.402
V-RAG [2]		0.608	0.599	0.624	0.504	0.639	0.529	0.628	0.551	0.514	0.581
Hi3Doc		**0.683**	**0.668**	**0.689**	**0.571**	**0.698**	**0.614**	**0.698**	**0.645**	**0.578**	**0.654**
Retrieval Method	Understanding Model	Evidence Modality				Evidence Distribution		Task Type			Overall Accuracy
		TXT	TAB	FIG	LAY	SIN	MUL	Uude	Reas	Loca	
Colpali [6]	ChatGPT-5mini [12]	0.657	0.609	0.631	0.507	0.667	0.551	0.662	0.558	0.532	0.606
VisRAG [17]		0.637	0.604	0.615	0.510	0.655	0.541	0.647	0.558	0.521	0.594
VDocRAG [14]		0.403	0.431	0.459	0.316	0.432	0.369	0.423	0.418	0.346	0.399
V-RAG [2]		0.602	0.585	0.593	0.504	0.613	0.534	0.618	0.546	0.502	0.572
Hi3Doc		**0.700**	**0.668**	**0.669**	**0.567**	**0.701**	**0.626**	**0.717**	**0.675**	**0.554**	**0.661**

evidence modalities, Hi3Doc attains the highest accuracy on TXT, TAB, FIG, and LAY, with particularly noticeable gains on figure and layout questions where document structure and cross-modal cues are important. When examining evidence distribution, Hi3Doc yields the strongest performance on both single-page (SIN) and cross-page (MUL) cases, indicating that its hierarchical chunking and retrieval strategy remains effective even when relevant evidence is scattered over multiple pages.

The task-type breakdown ("Uude", "Reas", and "Loca" corresponding to understanding, reasoning, and locating questions, respectively) further confirms the robustness of Hi3Doc on LongDocURL. Under all three types, Hi3Doc delivers the best accuracy in almost all cases; for ChatGPT-5mini, it reaches 0.717 on understanding and 0.675 on reasoning, clearly surpassing the strongest baseline Colpali. While the improvements on locating-oriented queries are relatively smaller, Hi3Doc still remains competitive and does not sacrifice performance on these tasks. Overall, the LongDocURL results demonstrate that Hi3Doc generalizes well beyond MMLongBench-Doc, providing consistent gains across modalities, evidence distributions, and task types when paired with different LLM backbones.

9. Evidence Retrieval Efficiency

We further analyze the evidence retrieval capability of Hi3Doc under a limited context budget on MMLongBench-Doc and LongDocURL. Figures 9 and 10 plot page-level recall and precision of Colpali at different Top-K values, together with the average recall and precision of Hi3Doc shown as horizontal lines. On both benchmarks, Colpali's

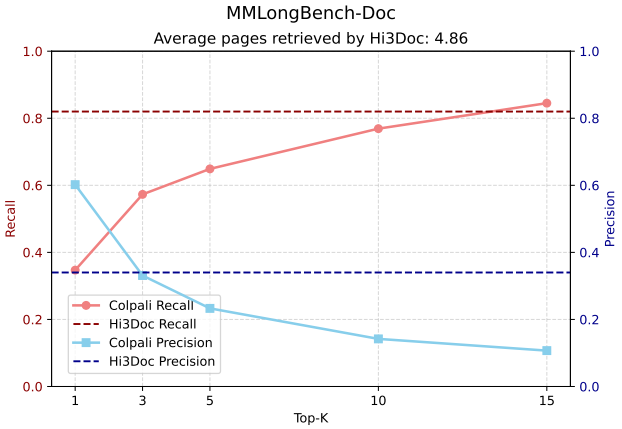


Figure 9. Page-level recall and precision of Colpali under different Top-K values on MMLongBench-Doc, compared with Hi3Doc. Solid curves denote Colpali's recall and precision, and dashed horizontal lines denote Hi3Doc's. Colpali only reaches a similar recall level when Top-K is increased to 10–15, which substantially reduces precision and requires many more pages than Hi3Doc.

recall increases monotonically with larger Top-K, while its precision continuously drops. In contrast, the number of pages retrieved by Hi3Doc is fixed by the pipeline design and does not depend on the Top-K parameter. Experiments show that, to reach a recall level comparable to Hi3Doc, Colpali must increase Top-K to 10–15, i.e., retrieve 2–3 times more pages than Hi3Doc. This substantially enlarges the input context for the downstream LVLM and simultaneously degrades precision, leading to much sparser relevant evidence. In contrast, Hi3Doc attains the same or

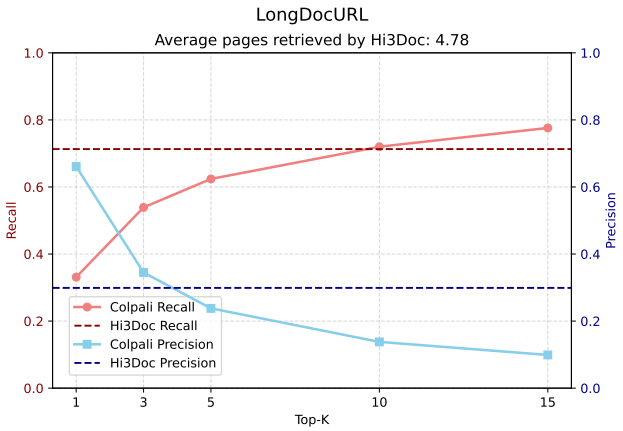


Figure 10. Page-level recall and precision on LongDocURL. Solid curves show Colpali at different Top-K, dashed lines show Hi3Doc with an average of 4.78 retrieved pages; Colpali needs $K \approx 10$ to match Hi3Doc's recall.

even higher recall with a fixed and relatively small number of retrieved pages, indicating that under the same or an even tighter context budget, Hi3Doc is more effective at selecting question-relevant evidence and providing more focused retrieval results for long-document VQA.

References

[1] Mistral AI. Mixtral-instruct-v0.1. https://mistral.ai/news/mixtral-of-experts/, 2024. Released by Mistral AI. Accessed: 2025-11-13. 7, 4

[2] Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24817–24826, 2025. 5

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, Jifeng Dai, ..., and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. Version describing the InternVL 2.5 series. 5, 4

[4] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024. 1, 2, 4

[5] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, 2025. 5, 1

[6] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 5, 4

[7] Ziyu Gong, Yihua Huang, and Chengcheng Mai. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering with hierarchical index and multi-granularity retrieval. *arXiv preprint arXiv:2508.00579*, 2025. 1, 2, 4

[8] Google DeepMind. Gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed: 2025-11-21. 1

[9] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024. 4, 1

[10] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2025-11-21. 1, 4

[11] OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025. Accessed: 2025-05-20. 5, 4

[12] OpenAI. Introducing GPT-5. https://openai.com, 2025. Accessed: 2025-08-07. 5, 1, 4

[13] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13636–13645, 2023. 5, 1

[14] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24827–24837, 2025. 5

[15] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 5, 1

[16] Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. Molorag: Bootstrapping document understanding via multi-modal logic-aware retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14056, 2025. 5

[17] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 5