	DocHaystack-100			DocHaystack-200			DocHaystack-1000		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
BM25 (OCR)	63.30	75.23	79.82	65.14	71.56	75.23	56.88	66.06	69.72
Jina-CLIP [18]	16.51	31.19	41.28	9.17	24.77	30.28	3.67	7.34	12.84
Nomic-Embed-Vision [29]	16.51	24.77	28.44	13.76	21.10	25.69	1.83	2.75	6.42
CLIP [33]	46.79	65.14	69.72	44.04	58.72	65.14	23.85	41.28	45.87
SigLIP [45]	51.38	67.89	76.15	47.71	63.30	70.64	33.03	49.54	57.80
OpenCLIP [16]	58.72	75.23	79.82	56.88	70.64	75.23	34.86	49.54	57.80
V-RAG (ours)	**81.65**	**88.99**	**88.99**	**77.98**	**84.40**	**84.40**	**66.06**	**77.98**	**78.90**
	InfoHaystack-100			**InfoHaystack-200**			**InfoHaystack-1000**		
	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
BM25 (OCR)	56.77	65.81	70.97	51.61	65.16	69.03	38.71	51.61	58.06
Jina-CLIP	43.23	51.61	58.06	36.77	46.45	51.61	23.87	33.55	37.42
Nomic-Embed-Vision	34.84	50.32	56.77	30.97	43.23	48.39	20.65	30.97	35.48
CLIP	69.68	78.71	85.81	65.16	77.42	81.94	45.81	64.52	70.32
SigLIP	58.06	71.61	80.00	55.48	67.74	76.77	39.35	55.48	61.94
OpenCLIP	72.26	85.16	**92.90**	66.45	81.94	**89.03**	53.55	65.81	72.90
V-RAG (ours)	**79.35**	**90.97**	**92.90**	**74.84**	**88.39**	88.39	**64.52**	**74.19**	**78.06**

Table 2. **Retrieval Results.** We compare our V-RAG model with other text-to-image and text-to-text (using OCR) retrieval methods across both benchmarks. V-RAG consistently outperforms baseline models on Recall@1, Recall@3, and Recall@5 metrics. Notably, V-RAG leverages an ensemble of text-to-image models along with a large multimodal model in a two-stage filtering approach. Top-performing values in each column are highlighted in **bold**.

Model	DocHaystack			InfoHaystack		
	100	200	1000	100	200	1000
LLaVA-OV [20]	-	-	-	-	-	-
GPT-4o [30]	27.52	23.85	-	23.87	20.00	-
Gemini [1]	50.46	48.62	-	29.03	21.94	-
Qwen2-VL [41]	41.28	12.84	-	20.00	14.19	-
MIRAGE [43]	3.67	3.67	2.75	7.74	7.10	6.45
LLaVA-OV+V-RAG	69.72	65.14	55.05	43.22	41.94	36.77
GPT-4o+V-RAG	81.65	72.48	66.97	65.16	63.23	56.77
Gemini+V-RAG	73.39	65.14	58.72	57.42	57.42	47.10
Qwen2-VL+V-RAG	82.57	74.31	66.06	65.81	65.81	**60.00**
Qwen2-VL-f.t.+V-RAG	**86.24**	**79.82**	**73.39**	**67.10**	**67.74**	**60.00**

Table 3. **The VQA results for the DocHaystack and InfoHaystack.** We evaluate with many closed-source and open-source multimodal model, and also integrating them with our V-RAG retrieval framework. - denotes that those models can not be inferred due to their token context constraints. To enable GPT-4o and Qwen2-VL to process hundreds of images, we employ low-resolution mode and adjust image size for compatibility.

peak learning rate of 1e-4 over a single epoch. Additionally, we leverage LoRA [14] with a rank of 8 to efficiently adapt the model's parameters during training.

5.2. Main Experimental Results

We evaluated a range of open-source and closed-source vision-language models for VQA tasks. We also evaluate several text-to-image and text-to-text (with OCR) retrieval models to evaluate their retrieval capabilities on our benchmarks. More detailed performance analysis are described in the following sections.

Retrieving results. The retrieval results in Table 2 demonstrate the superiority of our proposed V-RAG framework over several baseline methods across both DocHaystack and InfoHaystack benchmarks. V-RAG consistently achieves the highest Recall@1, Recall@3, and Recall@5 scores on most categories, indicating its robust retrieval capabilities. Notably, V-RAG outperforms text-based retrieving models such as BM25 and also the text-to-image retrieval models like jina-clip, CLIP, SigLIP, and OpenCLIP by substantial margins, especially on the DocHaystack-100 subset, where it reaches Recall@1 of 81.65% and Recall@5 of 88.99%. This pattern continues for larger datasets (DocHaystack-1000), where V-RAG remains competitive, achieving Recall@1 of 66.06%. It achieves the top performance across all recall metrics on DocHaystack. For InfoHaystack benchmarks, V-RAG also outperforms other models, particularly on InfoHaystack-100 and InfoHaystack-200, where it receives Recall@1 of 74.84% and 64.52%, higher than previous best by 8% and 11%, respectively. This consistent performance advantage highlights the effectiveness of V-RAG's ensemble of multiple vision encoders, allowing it to capture more granular details and improve retrieval accuracy over large multimodal models.

Visual question answering (VQA) results. The table presents VQA results for the DocHaystack and InfoHaystack benchmarks across varying dataset sizes (100, 200, 1000) using different multimodal models, both independently and in combination with the V-RAG framework. The results show that Qwen2-VL fine-tuned with V-RAG (Qwen2-VL-f.t.+V-RAG) achieves the highest scores across most benchmarks, with particularly notable performance on DocHaystack-100 (86.24) and InfoHaystack-100 (67.10), indicating superior retrieval and VQA capabilities in these scenarios. When V-RAG is added to other models, substantial improvements are observed, demonstrating the framework's efficacy in enhancing retrieval accuracy. For instance, GPT-4o's performance increases sig-

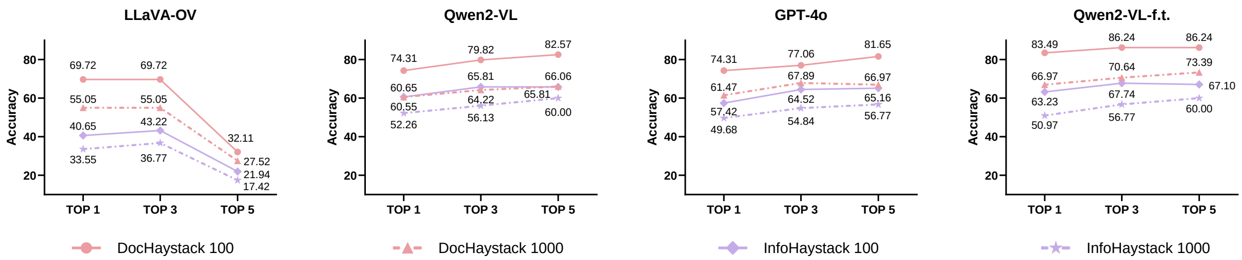

Figure 5. **Top-k selection ablation analysis for LMM-VQA.** We demonstrate the results for LLaVA, Qwen2-VL, GPT-4o and also the finetuned Qwen2-VL model on the DocHaystack-100/1000 and InfoHaystack-100/1000 benchmarks. All the models are integrated with our V-RAG framework. We show the VQA accuracy performance for each ablation.

CLIP	SigLIP	OpenCLIP	VLM-filter	DocHaystack-1000			InfoHaystack-1000		
				R@1	R@3	R@5	R@1	R@3	R@5
✓	✗	✗	✗	23.85	41.28	45.87	45.81	64.52	70.32
✗	✓	✗	✗	33.03	49.54	57.80	39.35	55.48	61.94
✗	✗	✓	✗	34.86	49.54	57.80	53.55	65.81	72.90
✓	✓	✗	✗	40.37	59.63	62.39	59.35	67.74	74.19
✓	✓	✓	✗	42.20	66.06	77.48	56.13	70.97	**78.06**
✓	✓	✓	✓	**66.06**	**77.98**	**78.90**	**64.52**	**74.19**	**78.06**

Table 4. **Ablation study on the V-RAG framework components.** We quantify the impact of each module for the Recall@1, Recall@3 and Recall@5 retrieval performance on the DocHaystack-1000 and InfoHaystack-1000 for our V-RAG framework.

nificantly with V-RAG, particularly for DocHaystack-100 and -200. The analysis highlights that V-RAG integration generally boosts performance across models, with Qwen2-VL-f.t.+V-RAG standing out as the top performer on both benchmarks, especially for the larger 1000-document tasks where retrieval accuracy is more challenging. This suggests that V-RAG's vision-centric, retrieval-augmented approach is highly effective for large-scale multimodal document understanding.

The table also shows that the DocHaystack-1000 and InfoHaystack-1000 present significant challenges for current LMMs. The drop in performance for larger document sets, with top accuracy only reaching 73.39% for DocHaystack-1000 and 60.00% for InfoHaystack-1000, underscores the difficulty our benchmarks.

5.3. Ablation Studies

Ablation study on Top-k Selection. This figure presents the top-k selection ablation analysis for LMM-VQA across four models: LLaVA-OV, Qwen2-VL, GPT-4o, and the fine-tuned Qwen2-VL (Qwen2-VL-f.t.), evaluated on the DocHaystack-100/1000 and InfoHaystack-100/1000 benchmarks. The analysis reports VQA accuracy as a function of top-k selection (Top 1, Top 3, and Top 5). Overall, accuracy tends to improve with larger k-values, suggesting that offering more retrieval options positively impacts model performance. However, for LLaVA-OV, there is a marked decrease in performance at top-5, indicating that this model struggles to process multiple images at this scale.

Ablation study on the V-RAG framework components. The ablation study in Table 4 highlights the contributions of each component in the V-RAG framework on the DocHaystack-1000 and InfoHaystack-1000 benchmarks. Using CLIP alone yields low performance (e.g., Recall@1 of 23.85% on DocHaystack-1000 and 45.81% on InfoHaystack-1000), indicating its limited retrieval capability on its own. Adding SigLIP and OpenCLIP incrementally improves results.

The highest performance is achieved when all three encoders are combined with the VLM-filter module, leading to Recall@1 scores of 66.06% on DocHaystack-1000 and 64.52% on InfoHaystack-1000. This setup also achieves the top Recall@1, Recall@3 and Recall@5 values, demonstrating that the VLM-filter is essential for refining the ensemble outputs and significantly improving retrieval accuracy. These results confirm that each module contributes to V-RAG's overall effectiveness.

6. Conclusion

In this work, we introduced the DocHaystack and InfoHaystack benchmarks to evaluate LMMs for retrieving and reasoning across large-scale documents. Our benchmarks providing a more rigorous and realistic assessment of large multimodal models in real-world, large-scale retrieval scenarios. To tackle these challenges, we proposed V-RAG, a vision-centric retrieval-augmented generation framework that significantly enhances retrieval precision and overall VQA performance. V-RAG achieves this through an ensemble of vision encoders and a specialized relevance filtering module, enabling improved accuracy across diverse visual inputs. Experimental results indicate that integrating V-RAG enables both open-source and closed-source LMMs to achieve superior performance in large-scale image retrieval and complex reasoning tasks.

References

[1] Google AI. Gemini: Google's multimodal ai model. *Google AI Research*, 2024. https://fireflies.ai/blog/gemini-vs-gpt-4. 2, 3, 7

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. 3

[3] Yifan Bai, Zhen Zhang, Yifan Zhang, Yuxuan Li, Yifan Zhang, Yifan Zhang, Yifan Zhang, Yifan Zhang, Yifan Zhang, Yifan Zhang, et al. Qwen-vl: A frontier vision-language model with larger-scale vision pre-training and aligned cross-modal instruction tuning. *arXiv preprint arXiv:2310.06726*, 2023. 2, 6

[4] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. Spotlight Presentation. 3

[5] Aayush Bansal, Karan Sikka, Gaurav Sharma, and Rama Chellappa. Visual question answering on image sets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2020. 3

[6] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024. 2

[7] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14178–14188, 2022. 1, 3

[8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.15339*, 2023. 2, 3

[9] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5558–5570. Association for Computational Linguistics, 2022. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. 3

[11] Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman Khan, Wangmeng Zuo, Xinxing Xu, Rick Siow Mong Goh, and Yong Liu. Vqa4cir: Boosting composed image retrieval with visual question answering. *arXiv preprint arXiv:2312.12273*, 2023. 2, 3

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[13] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938. PMLR, 2020. 3

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 2

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. 2021. If you use this software, please cite it as below. 3, 6, 7

[17] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics, 2020. 3

[18] Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. Jina clip: Your clip model is also your text retriever, 2024. 6, 7

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3, 6, 7