Which View Works Best? Evaluating Representations for Scientific Document Retrieval

Anonymous Author(s)

Overview Scientific papers mix text, tables, and figures, making document retrieval hard when evidence is spread across modalities. Prior work often isolates passages (SciFact) or figures (ArXivQA [Li et al., 2024]) rather than full documents, and many multimodal systems encode PDF pages as images—capturing layout but losing structured text. Retrieval models split into **text retrievers** (ColBERT, E5) strong at semantic matching yet blind to visuals, and **multimodal retrievers** (ColPali [et al., 2024], VisRAG [et al., 2025]) that leverage images but weakly integrate with text. We study which *representation* of a paper works best.

Datasets and Approach **ArXivDocQA:** 251 queries, 2,263 arXiv papers from full LaTeX sources. Queries referencing local figures are decontextualized with GPT-4-1 (e.g., "What do delay fibers do in Fig. b?" → "What is the role of delay fibers in photonic circuits?"). **SciFactDoc:** 482 queries, 263 papers derived from SciFact [et al., 2020] claims; claims are similarly decontextualized.
Representations. We test *text-only* (raw LaTeX, chunked), *figures-only* (extracted images), and *PDF-as-pages* (rendered page images). **Models.** Text retrievers: ColBERT, E5-4K, GTE-Qwen2-7B. Multimodal retrievers: ColPali, VisRAG. **Metric.** MRR@10.

Dataset	Model	Text	Figures	PDF
ArXivDocQA	E5	0.223	–	–
	ColBERT	0.174	–	–
	GTE-Qwen2-7B	0.321	–	–
	VisRAG	0.190	0.626	0.454
	ColPali	0.657	**0.720**	0.657
SciFactDoc	E5	0.233	–	–
	ColBERT	0.192	–	–
	GTE-Qwen2-7B	**0.314**	–	–
	VisRAG	0.127	0.101	0.238
	ColPali	0.263	0.121	0.215

Table 1: Retrieval results (MRR@10). Best values per dataset in bold.

Results Across both datasets, no representation is universally optimal: ArXivDocQA benefits from figure views, SciFactDoc from text. PDF-as-pages never achieves the best results, despite capturing layout. ColPali remains competitive even in text-only settings, suggesting cross-modal generalization.

References

D. W. et al. Fact or fiction: Verifying scientific claims. In *EMNLP*, pages 7534–7550, 2020.

M. F. et al. Colpali: Efficient document retrieval with vision language models, 2024.

S. Y. et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *ICLR 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=zG459X3Xge`.

L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *ACL*, pages 14369–14387, 2024.