Context-Enriched Figure Indexing for Scientific Multi-Modal RAG: An Industry Case Study at Scale

Dan Li*
Elsevier
Amsterdam, The Netherlands
d.li1@elsevier.com

Yongkang Li*†
University of Amsterdam
Amsterdam, The Netherlands
y.li7@uva.nl

Simon Huesgen
University of Amsterdam
Amsterdam, The Netherlands
simonhuesgen@gmx.de

Hosein Azarbonyad
Elsevier
Amsterdam, The Netherlands
h.azarbonyad@elsevier.com

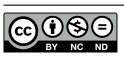
Abstract

Scientific figures encode key experimental evidence and domain knowledge, yet remain underutilized in retrieval-augmented generation (RAG). Directly indexing visual embeddings often misses fine-grained scientific semantics and misaligns with text-centric retrieval pipelines, while author-provided captions are frequently sparse and insufficiently contextualized. In this paper, we report on the design, implementation, and evaluation of a context-enriched figure indexing system built for a large-scale scientific digital library. Rather than introducing specialized multi-modal indexing infrastructure, our system converts figures into context-enriched captions—generated by fusing visual content with local article context via vision–language models—and indexes them alongside text in a unified embedding space using standard text retrievers. We evaluate this design on SciCap and a production corpus of approximately 113,000 figures drawn from Elsevier's ScienceDirect, comparing open-source (MiniCPM) and proprietary (GPT-4o) captioning models. Enriched captions consistently outperform direct image embeddings and original author captions for figure retrieval and downstream question answering. Although GPT-4o produces higher-quality captions under intrinsic evaluation, MiniCPM achieves comparable retrieval utility at no API cost and runs on a single commodity GPU—a trade-off that favors open-source models for large-scale indexing. We share the practical lessons, cost–performance trade-offs, and remaining open challenges from building this system, supporting context-enriched captioning as a scalable and easily adoptable foundation for scientific multi-modal RAG in production.[1]

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Information systems** → **Language models**; Retrieval models and ranking; *Question answering*; Information extraction.

*Both authors contributed equally to this research.
†Corresponding author.
[1]Code is available at https://github.com/dli1/multimodalrag



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
SIGIR-AP 2026, December 13–15, 2026, Singapore, Singapore
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3033-7/2026/12
https://doi.org/10.1145/3848000.3850195

Keywords

Scientific Figure Retrieval, Multi-modal Retrieval-augmented Generation, Context-enriched Captioning

ACM Reference Format:
Dan Li, Yongkang Li, Simon Huesgen, and Hosein Azarbonyad. 2026. Context-Enriched Figure Indexing for Scientific Multi-Modal RAG: An Industry Case Study at Scale. In *Proceedings of the 2026 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2026), December 13–15, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3848000.3850195

1 Introduction

Large-scale scientific digital libraries increasingly deploy retrieval-augmented generation (RAG) systems to power research assistants over millions of peer-reviewed articles [27, 36]. These production systems, however, face a fundamental gap: scientific knowledge is inherently multi-modal [9, 15, 16], yet mainstream RAG pipelines operate almost exclusively on text, leaving the figures that encode key experimental evidence—result plots, architecture diagrams, and workflow illustrations—effectively invisible to retrieval [1]. In practice, a researcher searching a scientific database for, e.g., "a plot showing the recoil energy spectrum fitted to the DAMA modulation model" will not find the relevant figure if its author-provided caption never mentions these terms, even when the figure is precisely what they need. Closing this gap is essential for scientific QA assistants in production, but it must be done under real deployment constraints: indexing pipelines that process millions of documents, strict cost and throughput budgets, and existing text-based retrieval infrastructure that cannot be rebuilt from scratch.

Making figures retrievable through textual descriptions is a long-standing goal: prior to the LLM era, content-analysis pipelines combined computer vision and pattern recognition to generate textual summaries of charts and infographics for search and accessibility [3–5, 22]. Modern multi-modal RAG pipelines, however, predominantly adopt one of two strategies, each with practical limitations at production scale. The first is **direct image embedding**, where a vision encoder (e.g., CLIP [29]) converts the raw image into a vector [1, 44]. These encoders, pre-trained mostly on natural image–caption pairs, struggle with the fine-grained logical structures and dense textual elements (e.g., axis labels, flowchart symbols, architectural hierarchies) characteristic of scientific figures [32, 39], and—equally important in practice—they require

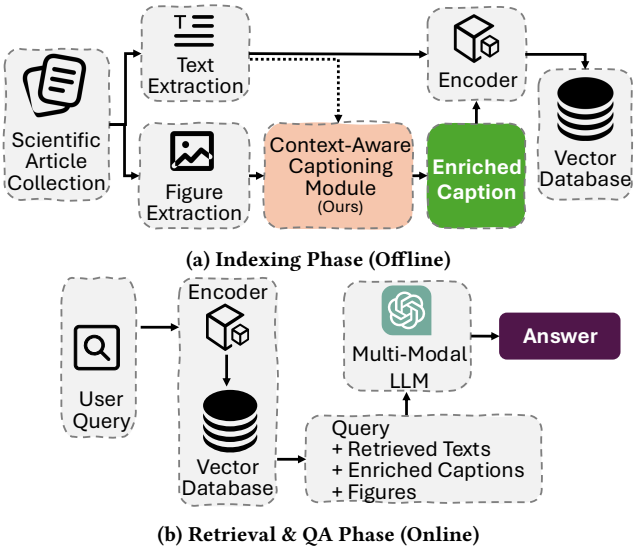


Figure 1: Overview of the proposed multi-modal RAG pipeline. (a) Indexing Phase: Scientific articles are processed to extract text chunks and figures; figures are converted to enriched captions via our context-aware captioning module before encoding. (b) Retrieval & QA Phase: User queries retrieve relevant texts, captions, and figures, which are jointly processed by a multi-modal LLM to generate answers.

maintaining a separate multi-modal index alongside the text index. The second strategy relies on **textual surrogates**, such as indexing the author-provided caption. These captions are frequently too sparse and generic for precise retrieval [30, 35, 40], and summaries of surrounding text often omit purely visual details [9, 10, 42].

In this paper, we report on the design, implementation, and evaluation of a **context-enriched figure indexing** system built for ScienceDirect, Elsevier's full-text scientific database. Rather than pursuing cross-modal alignment or introducing new retrieval infrastructure, our design converts each figure into a single, information-rich textual representation: we use vision–language models (VLMs) to fuse the figure's visual content with its local article context (captions, in-text mentions, and article metadata), producing an *enriched caption* that captures both what the figure shows and what it means within the article's narrative. Enriched captions are indexed alongside regular text chunks in a unified embedding space using standard text retrievers. This design was driven by deployment pragmatics: it requires no specialized multi-modal indexing infrastructure, integrates directly into existing lexical or dense retrieval stacks, and shifts all multi-modal computation to an offline, one-time preprocessing step whose cost can be controlled by the choice of captioning model.

We evaluate this design on the public SciCap dataset [9] and on a production corpus of approximately 113,000 figures drawn from 10,000 peer-reviewed computer science articles, comparing open-source (MiniCPM) and proprietary (GPT-4o) captioning models across intrinsic caption quality, figure retrieval, and end-to-end question answering. The primary contributions of this work are summarized as follows:

- **An industry case study of figure-aware RAG:** We describe the design and trade-offs of a context-aware figure indexing pipeline for a large-scale scientific digital library, including captioning prompts, context-selection strategies, and configuration choices we found effective, so practitioners can adopt the approach directly.
- **Large-scale empirical validation:** On both SciCap and the production corpus, enriched captions consistently outperform direct image embeddings and original author captions for figure retrieval and downstream QA, with the production corpus providing evidence under realistic figure complexity and scale.
- **A cost–performance analysis of open-source versus proprietary VLMs:** We find that although GPT-4o produces higher-quality captions under intrinsic evaluation, an 8B-scale open-source model (MiniCPM) achieves comparable retrieval utility at no per-call API charges and on a single commodity GPU—a decisive trade-off for indexing pipelines at the scale of millions of figures.

2 Related Work

Our work lies at the intersection of retrieval-augmented generation, scientific figure understanding, and multi-modal retrieval.

2.1 Retrieval-Augmented Generation (RAG)

RAG grounds LLM responses in external knowledge by retrieving relevant corpus units and augmenting generation with retrieved context [14, 34]. In scientific domains, work has focused mainly on textual chunking and retrieval [33], from BM25 [31] to neural retrievers such as DPR [11], BGE [41], and ColBERT [12]. Recent studies also investigate the robustness and generalization of dense retrievers under varied and adversarial conditions [18–21]. Yet these methods remain text-centric and struggle when scientific knowledge is encoded in figures, tables, and diagrams [44].

2.2 Scientific Figure Understanding: From Content Analysis to VLMs

Generating textual descriptions of figures for search, summarization, and accessibility long predates the LLM era. Early content-analysis pipelines combined computer vision, OCR, and pattern recognition to interpret charts and infographics [5, 6, 25], and several works targeted *retrieval* directly—generating figure summaries to augment document search [3], retrieving infographics via structure and message content [22], mining scholarly figures at scale [4], and searching visualizations by style and structure [8]. Notably, these systems often aimed at *high-level, message-oriented* descriptions (e.g., identifying an upward trend) rather than enumerating visual elements—an ambition our enriched captions share. Their reliance on bespoke, figure-type-specific engineering, however, limited coverage and made them costly to maintain in production. Modern VLMs replace such specialized pipelines with general-purpose visual understanding: SciCap [9] established foundations for descriptive captioning, SciCap+ [42] and SciGraphQA [16] highlighted the value of external knowledge, and ChartQA [26] and CharXiv [39] pushed chart understanding further. Across both eras, a recurring limitation is treating figures in isolation: a figure's meaning is often defined by its textual references in the article, yet most

pipelines optimize for *descriptive fidelity* ("what is in the image") rather than *functional utility* ("how the image serves retrieval and reasoning"). Our work revisits the retrieval-oriented ambition of the early content-analysis line, using VLMs as a general-purpose replacement for bespoke pipelines, evaluated at production scale.

2.3 Multi-Modal RAG in Scientific Domains

Advances in vision–language models (e.g., Flamingo [2], LLaVA [24], MiniCPM-V [43]) have enabled multi-modal RAG, with benchmarks such as SPIQA [28] and SciMMIR [40] targeting scientific documents. Within modern RAG pipelines, two strategies dominate for incorporating figures: **direct visual embedding**, where encoders like CLIP [29] project images into a shared vector space—suffering from modality misalignment on fine-grained scientific semantics and requiring a separate multi-modal index in production—and **textual surrogates** such as author captions or OCR text [23], which discard visual information and are often sparse. A parallel line builds richer multi-modal RAG architectures: ViDoRAG [38] employs iterative multi-agent reasoning over visual documents, BayesRAG [17] probabilistically corroborates cross-modal evidence at retrieval time, and RAG-Anything [7] unifies heterogeneous content via multimodal knowledge-graph indexing. These systems pursue answer quality through added machinery—agentic iteration and evidence fusion on the serving path, or specialized graph-based indexing infrastructure—whereas our design shifts all multi-modal computation to a one-time offline captioning step and keeps both the index and the serving path a standard text-retrieval stack. We view these directions as complementary, and an empirical comparison under matched production constraints remains future work (see Limitations).

Our work thus treats figure captioning as an *indexing optimization* problem: we generate context-aware enriched captions that fuse visual content with document-level semantics, producing retrieval-optimized representations compatible with existing text-based infrastructure, which proved decisive for deployment in our setting.

3 Methodology

This section describes the design of our context-aware figure indexing pipeline. The design is grounded in a practical observation: effective figure retrieval requires captions that capture not only visual content but also document-level semantics. The pipeline is configurable along three dimensions: textual context selection (§3.2.1), prompting strategy (§3.2.2), and captioning model (§3.2.3). We first present the system overview, then detail each component.

3.1 System Overview

The core objective of our pipeline is to transform scientific figures into retrieval-optimized textual surrogates. Rather than treating a figure as an isolated visual object, we fuse its visual content with the surrounding textual context from the article, generating an *enriched caption* that bridges raw pixels and domain-specific semantics.

This design offers two practical advantages. First, converting figures to text ensures *retrieval compatibility*—figures and text chunks share the same vector space, and the online serving path remains a standard text-retrieval stack. We note that this does not eliminate multi-modal computation altogether: every figure must be processed by a VLM once at indexing time. The design instead *relocates* this computation to an offline, one-time preprocessing step, where throughput requirements are relaxed and cost can be controlled through the choice of captioning model (§3.2.3). Second, by incorporating article context, the resulting captions capture not only "what the figure shows" but also "what the figure means" within its scientific narrative.

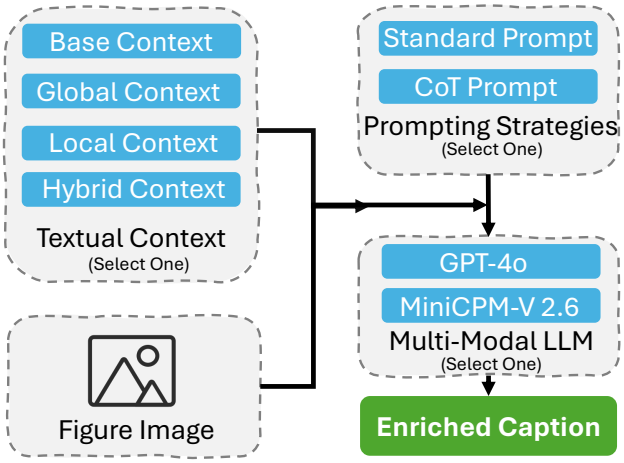


Figure 2: Our proposed enriched context-aware captioning method. We explore different configurations for Textual Context (choosing from Base, Global, Local, or Hybrid) and Prompting Strategies (comparing Standard and CoT) to optimize generation.

As illustrated in Figure 1, our pipeline consists of two phases: (a) an offline *indexing phase* that extracts text and figures from articles, generates enriched captions, and encodes them into a vector database; and (b) an online *retrieval and QA phase* that retrieves relevant content and generates answers via a multi-modal LLM. The following sections detail our captioning methodology (§3.2), the central design component of the pipeline.

3.2 Captioning Method

Our pipeline generates enriched captions by combining figure images with article context as input to a multi-modal LLM, configured along the three dimensions shown in Figure 2.

3.2.1 Textual Context. Textual context refers to the supplementary linguistic information provided alongside the figure image to guide the caption generation process. To systematically evaluate the impact of different information sources, we categorize textual context into four variants:

- (1) **Base Context:** The most straightforward setting, which utilizes the original author-provided caption.
- (2) **Global Context:** Article-level scientific paper information, including the paper's title, abstract, and keywords, which provides the high-level research background.
- (3) **Local Context:** Figure-specific textual information from the surrounding discourse, including (i) *figure mentions*—sentences that

explicitly reference the figure, extracted via regular expressions—and (ii) the *first-appearance section*, i.e., the section in which the figure is first introduced.

• (4) **Hybrid Context:** A fixed combination of the original caption, figure mentions (from Local Context), and article keywords (from Global Context). No per-figure selection is performed; the combination was chosen to balance breadth and specificity.

3.2.2 Prompting Strategies. We design a family of captioning prompts to control how the multi-modal LLM transforms a figure and its textual context into an enriched caption. Concretely, we consider two prompting strategies that progressively increase structure and controllability, while remaining compatible with downstream indexing in our retrieval pipeline.

• (1) **Standard Prompt**: Our starting point is a single, structured instruction template that conditions caption generation on a selected textual context variant (Section 3.2.1). As shown in Figure 3, the template requests a high-level *overview*, a *detailed description* of salient visual elements (e.g., charts, symbols, numeric labels, legends), an explanation of *relationships and patterns*, and domain *terminology and keywords* for retrieval, together with constraints on accuracy and output format. This prompt proved effective in our deployment and serves as our default configuration; we do not claim it is optimal, and its transferability to future VLMs is an open question.

• (2) **Chain-of-Thought Prompt**: A reasoning-oriented variant that decomposes the task into explicit intermediate steps, aimed at reducing omissions in complex scientific figures. It separates (i) *visual foundation*—analyzing the image alone to identify visualization type, axes, labels, and legends; (ii) *contextual meaning*—extracting the study aim, methods, and key claims from the provided article context; (iii) *aligned synthesis*, which integrates the two and resolves conflicts between them; followed by (iv) caption composition and (v) a self-validation pass. The output format is identical to the standard prompt.

Faithfulness considerations. Injected context may lead the model to assert claims not visually grounded in the figure. Both prompts counter this explicitly—the standard prompt restricts content to "information visible in the figure or clearly stated in the context," and the CoT prompt adds a self-validation pass—though we do not quantify hallucination rates here (see Section 7).

3.2.3 Captioning Model. The final dimension of our system is the choice of multi-modal LLM for caption generation. We prioritize open-source models for their cost-effectiveness and reproducibility, while also benchmarking against proprietary alternatives. We evaluate two representative models[2]:

• (1) **MiniCPM-V 2.6:** A lightweight open-source multi-modal LLM[3] that integrates OCR functionality with effective visual-textual synthesis. Crucially, it is deployable on a single GPU (e.g., NVIDIA A10, 24GB), making it a practical solution for large-scale deployment without reliance on resource-intensive infrastructure.

• (2) **GPT-4o:** A state-of-the-art proprietary multi-modal model included as a strong reference point, to assess whether open-source alternatives can achieve competitive results for scientific figure captioning.

[2]We also experimented with other models (e.g., Qwen2-VL-7B, LLaVA-1.5-7B, Claude-3.5); comparisons are provided in Section 5.2.

[3]https://huggingface.co/openbmb/MiniCPM-V-2_6

All models are used off-the-shelf without task-specific fine-tuning. This is a deliberate deployment decision: avoiding fine-tuning keeps the pipeline simple to maintain and immediately applicable as newer VLMs are released, though task-specific adaptation could plausibly improve caption quality further (see Section 7).

4 Experimental Setup

We design experiments to evaluate our context-aware captioning system across two dimensions: intrinsic caption quality and extrinsic RAG performance. This section describes our research questions, datasets, evaluation protocols, and implementation details.

4.1 Research Questions

Our experiments are guided by three research questions:

- **RQ1 (Enriched Caption Generation):** How can visual content and textual context be combined to generate accurate and comprehensive enriched captions?
- **RQ2 (Figure Retrieval Performance):** How do enriched captions compare to raw author captions and direct image embeddings in downstream retrieval tasks?
- **RQ3 (Multi-Modal Answer Generation):** How do different figure representations affect downstream answer quality in an end-to-end RAG pipeline?

4.2 Datasets

We evaluate on two corpora to ensure diversity and domain coverage: (1) **SciCap data**: We sampled from the SciCap dataset [9], covering all the 20 scientific domains. From each domain, we randomly selected 500 to 550 images along with their metadata, including Article ID, title, abstract, and figure mentions. For articles lacking explicit keywords, we generated them using *BERTopic* to ensure consistent contextual inputs. (2) **Production data**: To validate our approach on real-world distributions, we sampled 10,000 computer science articles from ScienceDirect, Elsevier's full-text scientific database comprising millions of peer-reviewed publications[4].

Crucially, access to the structured XML source files—a distinctive advantage of working directly with the publisher's production infrastructure—allowed us to extract high-fidelity metadata (e.g., figure mentions, hierarchical section headers) that is often unavailable or noisy in PDF-only datasets. This production corpus contains approximately 113,000 figure images after sampling, with figures exhibiting high variance in complexity, ranging from sparse diagrams to dense, multi-layered visualizations.

4.3 Evaluation Protocol

We evaluate both intrinsic caption quality and extrinsic RAG performance.

Caption Evaluation. We evaluate generated captions given the figure image and its extracted context (title, abstract, keywords, mentions, and original caption). Since no ground truth exists for enriched captions, we adopt LLM-as-a-judge evaluation with human validation. We define five rubrics on a 4-point scale (Table 1): Technical Correctness (Cor), Completeness (Com), Conciseness

[4]Accessed under a research agreement with Elsevier; licensing prevents redistributing the raw articles, but our preprocessing pipeline is released with the code.

Standard Captioning

Analyze the scientific figure and any provided context to create a concise, retrieval-optimized caption. Follow these guidelines:
Overview: Summarize the figure's main subject, purpose and experimental focus.
Visual Details: Identify and describe important visual elements (e.g. charts, diagrams, symbols, numeric labels, legends).
Relationship and Patterns: Explain interactions, trends, or hierarchies evident in the figure.
Terminology and Keywords: Use relevant domain-specific terms (with plain-language equivalents if needed), then list essential keywords for retrieval.
Relevance and Accuracy: Include only information visible in the figure or clearly stated in the context. Avoid speculation, repetition or irrelevant details.
Formatting: Present the description coherently, in a paragraph of 300 to 400 words. Omit statements about missing elements unless explicitly relevant.

Chain-of-Thought (CoT) Captioning

VISUAL FOUNDATION (Image Analysis Only): Identify visualization type, axes, labels, legends, data representations.
CONTEXTUAL MEANING: Extract study aim, experimental focus, methods, key claims.
ALIGNED SYNTHESIS (Intelligent Combination): Integrate visual+text; resolve conflicts; highlight context-only insights.
CAPTION GENERATION (Composition Guidelines): Compose 300-400 word coherent paragraph; start with purpose; integrate visuals+context; include domain terminology and keywords; maintain objective tone.
SELF-VALIDATION (Quality Control): Verify evidence, completeness, keyword coverage; remove speculation and redundancy.

Figure 3: Overview of two captioning prompts (Standard and CoT), displayed in compact form to accommodate length constraints.

Table 1: Rubrics for figure caption evaluation, scored from 1 to 4 with definitions and endpoint anchors.

Rubric	Definition (1 = lowest, 4 = highest)
Technical Correctness	Accuracy of terminology, values, and conceptual relationships (1 = factual errors; 4 = technically precise and complete).
Completeness	Coverage of salient visual and conceptual elements (1 = misses >50%; 4 = full, detailed coverage).
Conciseness and Focus	Redundancy vs. essential insight (1 = obscures main ideas; 4 = lean and informative).
Context Integration	Use of methodological information from the article (1 = absent; 4 = methods used to enhance analysis).
Visual Integration	Analytical use of visuals beyond listing (1 = superficial listing; 4 = synthesizes visual data for insight).

(Foc), Context Integration (Cin), and Visual Integration (Vin). The rubrics separately assess factual accuracy, information coverage, contextual grounding, and use of visual evidence; Foc penalizes irrelevant or redundant expansion rather than caption length itself. We use fine-grained criteria following [13]. GPT-5-chat evaluates generated captions using both the figure image and textual context, producing scores with short justifications. We note that the judge shares a vendor with one of the captioning models (GPT-4o); since the observed ranking favors GPT-4o on intrinsic quality, any such bias would work against the conclusion we ultimately draw regarding open-source viability.

Human validation. Two of the authors independently rated a small random sample of figure–caption pairs from the production corpus using the same five rubrics, blind to which configuration produced each caption; reported scores are averaged across raters. We did not compute formal inter-annotator agreement statistics, and the sample is small relative to the LLM-scored set; the human ratings should therefore be read as a consistency check on the LLM judge rather than as independent validation.

RAG Evaluation. As no standard benchmark exists for joint figure-text retrieval and QA, we construct a synthetic test set of 450 queries following the methodology of SPIQA [28]. Using Claude Sonnet 4, we generate query-answer pairs balanced across three categories (150 each): *text-only*, *image-only*, and *cross-modal*. Each query undergoes multi-call consensus verification and manual validation. Formally, each test sample is a tuple $\{Q, Q_{type}, E_{gt}, A_{gt}, S_{id}\}$, where Q is the query, Q_{type} indicates the modality, E_{gt} denotes ground-truth evidence, A_{gt} is the reference answer, and S_{id} is the source article. We note that synthetic queries are not a substitute for observed user information needs; our benchmark is best read as a controlled comparison across figure representations under matched conditions, rather than as evidence of end-user impact.

Metrics. For retrieval, we report Recall@k ($k \in \{1, 5\}$) and Mean Reciprocal Rank (MRR). For answer quality, we use both lexical metrics (BLEU, ROUGE-L) and LLM-as-a-judge scores following [30]: Answer Correctness (A.Cor), Answer Relevancy (A.Rel), Text Faithfulness (T.Fai), and Image Faithfulness (I.Fai). These metrics assess downstream answer quality rather than caption quality itself.

4.4 Models

Text Retrieval Models. We consider both lexical and dense retrieval backbones. As a lexical baseline, we use BM25 [31]. For dense retrieval, we evaluate three representative text embedding models: Snowflake Arctic Embed[5], BGE [41], and all-MiniLM[6]. Unless otherwise stated, we use Snowflake Arctic Embed as the default text encoder/retriever in our pipeline, as it yields the strongest retrieval performance in our setting (Figure 5). We nevertheless report results with BM25, BGE, and all-MiniLM to assess robustness across retrieval paradigms.

Visual Embedding Model. For experiments that require direct visual indexing (Section 5.2), we compute image embeddings using CLIP (ViT-B/32)[7].

Vision–Language Models for Captioning. For enriched caption generation, we evaluate both open-source and proprietary vision–language models. The open-source models include LLaVA-1.5 [24],

[5] https://huggingface.co/Snowflake/snowflake-arctic-embed-m
[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[7] https://huggingface.co/openai/clip-vit-base-patch32

Qwen2-VL [37], and MiniCPM-V 2.6 [43], while GPT-4o is used as a representative proprietary model. This model set enables a controlled comparison of caption quality and retrieval utility under different deployment constraints.

4.5 Implementation Details

Unless otherwise specified, we follow the default model choices described in Section 4.4. All open-source model inference is performed on a single NVIDIA A10 GPU. For proprietary captioning results, we query GPT-4o via its API under the same captioning prompts and contextual inputs. All reported results come from a single generation run per configuration; we do not report variance across repeated runs, and small differences between models should be interpreted with this in mind.

5 Experimental Results

We present results organized by our three research questions: caption generation quality (§5.1), retrieval performance (§5.2), and answer generation (§5.3). Our key findings are: (1) textual context selection drives caption quality far more than prompting strategy, with hybrid context performing best; (2) enriched captions outperform visual embeddings and original captions in retrieval; (3) open-source models achieve competitive performance at much lower cost.

5.1 Caption Generation Quality (RQ1)

In this subsection, we investigate **RQ1**: How can visual content and textual context be combined to generate accurate and comprehensive enriched captions.

Impact of Textual Context. As shown in Table 2 (Part A), hybrid context achieves the best average score on both datasets (2.47 on SciCap, 3.53 on the production corpus), combining the author caption, article keywords, and figure mentions. The gains are concentrated in context and visual integration (CIN, VIN), while conciseness (FOC) is consistently lower than with base context alone—enrichment adds retrievable content at the cost of brevity, a trade-off we accept for indexing purposes. Local context (figure mentions) proves more valuable than global context (titles and abstracts), which scores lowest overall (1.95 AVG on SciCap).

Impact of Captioning Model. In Table 2 (Part B), GPT-4o leads on intrinsic caption quality on both datasets, with the gap considerably wider on SciCap (3.07 vs. 2.47 AVG) than on the production corpus (3.73 vs. 3.53 AVG). MiniCPM-V 2.6 is therefore not the stronger captioner in absolute terms. Its appeal is that this quality gap does not translate proportionally into retrieval utility (§5.2), while it incurs no API charges and runs on a single commodity GPU. For an indexing pipeline operating over millions of figures, this makes it the more practical choice despite the lower intrinsic scores.

Impact of Prompting Strategy. Part C of Table 2 compares prompting strategies. CoT prompting produces small and mixed changes: context and visual integration improve (CIN 3.28 → 3.32; VIN 3.73 → 3.81), consistent with explicit reasoning steps helping the model attend to both context and visual evidence, while conciseness declines (FOC 3.14 → 3.06). The net effect on the average score is marginal (3.53 → 3.55). In our setting, the choice of textual context (Part A) matters considerably more than the prompting strategy.

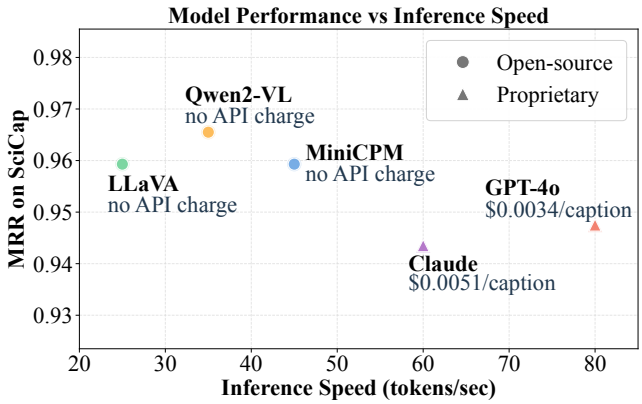


Figure 4: Retrieval performance (MRR on SciCap) against inference speed and per-caption API cost for five captioning models using enriched captions. Retrieval utility varies little across models, while per-caption cost differs substantially. *Throughput for proprietary models is measured end-to-end through their APIs and includes network latency; it is not directly comparable to local inference speed.

Human Evaluation. Table 3 reports human ratings collected on the production corpus, following the same three-part design as the LLM-based evaluation. The orderings agree with the LLM judge on all three comparisons: hybrid context scores highest and global context lowest (Part A, 3.34 vs. 2.26); GPT-4o remains slightly ahead of MiniCPM on intrinsic caption quality (Part B, 3.44 vs. 3.34); and CoT prompting yields only a modest average gain over the standard prompt (Part C, 3.34 → 3.40). Two divergences are worth noting. First, human averages are lower than LLM scores in every comparable condition, by 0.15 to 1.03 points, with the largest gaps on the weakest configurations—the LLM judge appears to compress the low end of the scale, which suggests such scores are better suited to relative comparison than to absolute quality estimation. Second, the two protocols disagree on the effect of CoT prompting on conciseness (FOC): human raters record a clear improvement (2.80 → 3.60) where the LLM judge records a slight decline (3.14 → 3.06), though both agree that completeness does not improve. Given the small sample and the absence of formal agreement statistics, we treat these ratings as a consistency check on the LLM-as-a-judge protocol rather than as independent validation.

Summary. Overall, these results demonstrate that combining visual content with carefully selected local textual context is critical for generating enriched captions that are accurate, context-aware, and suitable for downstream retrieval.

5.2 Retrieval Performance (RQ2)

In this subsection, we next address **RQ2**: How do enriched captions compare to raw author captions and direct image embeddings in downstream retrieval tasks? Table 4 illustrates the retrieval performance across different figure representations. We present two primary findings:

First, enriched captions substantially outperform direct visual embeddings (0.152 → 0.959 MRR on SciCap; 0.092 → 0.468 on

Table 2: Comparative analysis of caption quality across different contexts and models. *** denotes significance over the corresponding baseline ($p < 0.001$, paired t-test); AVG is excluded from significance testing.

Dataset	Method / Model	COR	COM	FOC	CIN	VIN	AVG
Part A: Impact of Textual Context (using MiniCPM)							
SciCap	Base context	2.44	2.88	1.87	2.40	2.46	2.41
	Global context	1.91	2.28	1.32	2.23	1.99	1.95
	Local context	2.20	2.65	1.72	2.25	2.30	2.22
	Hybrid context (Best)	**2.46*****	2.84	1.77	**2.72*****	**2.56*****	**2.47**
Production	Base context	3.83	3.36	3.38	3.17	3.52	3.45
	Global context	3.76	3.22	2.96	3.15	2.38	3.29
	Local context	3.83	3.47	3.12	3.19	3.55	3.43
	Hybrid context (Best)	**3.90*****	**3.62*****	3.14	**3.28*****	**3.73*****	**3.53**
Part B: Comparison of Captioning Models (using Hybrid Context)							
SciCap	MiniCPM	2.46	2.84	1.77	2.72	2.56	2.47
	GPT-4o	**3.34*****	**3.67*****	3.07	**3.14*****	**3.22*****	**3.07**
Production	MiniCPM	3.90	3.62	3.14	3.28	3.73	3.53
	GPT-4o	**4.00*****	**3.92*****	**3.17*****	**3.58*****	**4.00*****	**3.73**
Part C: Impact of Prompt Strategies (using MiniCPM + Hybrid Context)							
Production	Standard Prompt	3.90	3.62	3.14	3.28	3.73	3.53
	CoT Prompt	**3.94**	**3.63**	3.06	**3.32**	**3.81**	**3.55**

Table 3: Human evaluation of caption quality across contexts, models, and prompting strategies on the production corpus. Best average score per part in bold.

Method	COR	COM	FOC	CIN	VIN	AVG
Part A: Impact of Textual Context (using MiniCPM)						
Base context	2.80	2.50	2.20	2.40	3.20	2.62
Global context	2.40	2.00	2.10	2.10	2.70	2.26
Local context	2.70	2.80	2.60	2.70	3.10	2.78
Hybrid context	3.60	3.60	2.80	3.40	3.30	**3.34**
Part B: Comparison of Captioning Models (using Hybrid Context)						
MiniCPM	3.60	3.60	2.80	3.40	3.30	3.34
GPT-4o	3.70	3.50	3.20	3.40	3.40	**3.44**
Part C: Impact of Prompt Strategies (using MiniCPM + Hybrid Context)						
Standard Prompt	3.60	3.60	2.80	3.40	3.30	3.34
CoT Prompt	3.60	3.20	3.60	3.30	3.30	**3.40**

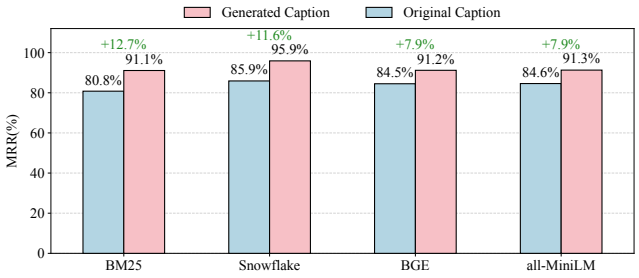


Figure 5: Retrieval performance across various search paradigms on SciCap. Our Enriched Caption (EC) consistently outperforms the original author-written captions (OC) across lexical and dense retrievers.

the production corpus). CLIP-based encoders resolve fine-grained details such as axis labels and domain-specific symbols poorly; converting the figure to text makes these entities—visually latent but named in the surrounding article—explicitly indexable.

Second, enriched captions also outperform author captions (0.859 → 0.959; 0.323 → 0.468), indicating that the gain comes from contextual fusion rather than from text indexing alone: original captions and figure mentions supply precise semantic anchors, while article keywords broaden the surface the retriever can match against.

Figure 4 relates retrieval performance to inference speed and per-caption token cost[8]. Retrieval utility is essentially flat across the five captioning models (MRR between 0.944 and 0.966) despite substantial differences in intrinsic caption quality and price, with the two proprietary models sitting at the lower end of this narrow band while charging \$0.0034–\$0.0051 per caption. This comparison runs on SciCap, where headroom is limited—author captions alone reach 0.859 MRR—so flatness here could in principle be a ceiling artifact. The production corpus, which we treat as our primary evidence, gives the same picture for the two models evaluated there (0.468 vs. 0.448 MRR, Table 4): caption quality and retrieval utility remain only loosely coupled. At indexing scale, the dominant consideration is therefore cost rather than caption quality. One caveat applies to the throughput axis: figures for proprietary models are measured end-to-end through their APIs and include network latency and server-side batching, whereas open-source figures reflect local inference on a single A10 GPU; the two are not directly comparable.

Figure 5 demonstrates the robustness of enriched captions across search paradigms. Unlike original captions, which often lack the keyword density required for lexical search (BM25), EC consistently

[8]Cost per caption is estimated from empirical averages in our experiments: $\text{Cost}_{\text{GPT-4o}} = (170 \times 2.5 \times 10^{-6}) + (227 \times 1.3 \times 10 \times 10^{-6}) \approx \$0.0034$; $\text{Cost}_{\text{Claude-3.5}} = (170 \times 3 \times 10^{-6}) + (235 \times 1.3 \times 15 \times 10^{-6}) \approx \$0.0051$, assuming a 1.3 token/word ratio. Open-source models are reported at \$0 in the sense of no per-call API charges; local inference still consumes hardware, energy, and operational effort that we do not quantify.

Table 4: Retrieval performance comparison across different figure representations. Enriched Captions demonstrate superior retrieval effectiveness compared to pure Visual Embeddings (VE) and sparse Original Captions (OC). *** denotes significance over VE, OC, and EC (GPT-4o) ($p < 0.001$, paired t-test).

Encoding type	Representation	SciCap (S)			Production (P)		
		MRR	R@1	R@5	MRR	R@1	R@5
Visual	VE (Openclip-1)	0.152	0.106	0.211	0.092	0.067	0.133
Textual	OC (Original Author Caption)	0.859	0.840	0.886	0.323	0.240	0.477
Textual	EC (GPT-4o Enriched Caption)	0.959	0.949	0.971	0.448	0.325	0.622
Textual	**EC (MiniCPM Enriched Caption - Ours)**	**0.959**	**0.959*****	**0.971**	**0.468*****	**0.347*****	**0.665*****

improves performance in both vector-based and keyword-based retrieval, and the improvement is stable. This stability stems from the "text anchoring" effect. By populating the index with both visual descriptions and contextual keywords, EC increases the searchable surface area of the figure, ensuring that it remains retrievable even under varied query formulations or retrieval algorithms.

Summary. Enriched captions outperform direct visual indexing while remaining robust across retrieval methods.

5.3 Answer Generation Quality (RQ3)

Finally, we investigate **RQ3**: how do different figure representations affect downstream answer quality in an end-to-end RAG pipeline? In all conditions the top-ranked figure is passed to the answer generator as visual input, so the representation under comparison determines both which figure is retrieved and what textual grounding accompanies it.

We employ GPT-4o as the answer generator, conditioning on all retrieved textual chunks while providing the top-ranked figure as visual input. Tables 5 and 6 report both reference-based metrics and LLM-as-a-judge evaluations.

Both enriched caption variants improve over visual embeddings and original author captions on every metric. The largest gain is in image faithfulness (I.FAI, 3.30 for original captions to 3.92–3.95 for enriched captions). In the deployed configuration the enriched caption both selects the figure and grounds the generator, so these end-to-end numbers—the quantity deployment cares about—do not attribute the gain between the two roles; an ablation holding retrieval fixed would be needed to separate them. Differences between the two captioning models, by contrast, are marginal: MiniCPM captions yield slightly higher BLEU, ROUGE-L, and I.FAI, GPT-4o captions score marginally higher on text faithfulness (4.37 vs. 4.35), and answer correctness and relevancy are identical. These lexical gaps are small in absolute terms (BLEU 0.0991 vs. 0.1010) and we do not report significance for these comparisons, so the appropriate reading is parity between the two captioning models rather than superiority of either.

Summary. Enriched captions bridge retrieval and generation, and a lightweight open-source captioner grounds downstream answers as effectively as a proprietary one.

6 Additional Analysis and Discussion

In this section, we first introduce a conceptual framing—the semantic spectrum — to characterize different captioning strategies, then illustrate our approach through a case study.

Table 5: Answer generation performance comparison on the Production data.

Modality	Representation	BLEU	ROUGE-L
Visual	VE (Openclip-1)	0.0715	0.2494
Textual	OC (Original Caption)	0.0925	0.2783
Textual	EC (GPT-4o EC)	0.0991	0.2884
Textual	**EC (MiniCPM - Ours)**	**0.1010**	**0.2914**

Table 6: LLM-as-a-judge evaluation of answer quality on the Production dataset.

Modality	Representation	A.COR	A.REL	T.FAI	I.FAI
Visual	VE (Openclip-1)	2.98	4.27	3.49	2.71
Textual	OC (Original Caption)	3.49	4.51	4.35	3.30
Textual	EC (GPT-4o)	3.62	**4.54**	**4.37**	3.92
Textual	**EC (MiniCPM - Ours)**	**3.62**	**4.54**	4.35	**3.95**

6.1 The Semantic Spectrum of Figure Captioning

A framing we found useful in practice is that the optimal representation of a scientific figure is not a fixed linguistic translation of its visual pixels, but rather a tunable position on a *semantic spectrum* (see Figure 7). We propose that the role of a caption in RAG should lie somewhere between the **descriptive pole** and the **indexing pole**. We hypothesize that the optimization target for captions can shift along this spectrum based on the application scenario:

• **The Descriptive Pole (Visual Reconstruction):** At this end, the goal is exhaustive visual detail (e.g., "A blue line graph with five data points"). This is essential for *Visual Question Answering (VQA)* where the model must "see" the raw evidence.

• **The Contextual Mid-point (Semantic Grounding):** Our enriched caption resides here. By infusing the caption with paper mentions (e.g., "The 50 GeV energy peak of the Crab Pulsar"), we bridge the gap between abstract visual shapes and domain-specific entities. This is the operating point explored in our scientific RAG setting.

• **The Indexing Pole (Intent Anticipation):** At the far end, the caption acts as a functional tag or an "abstract token." It may include metadata not present in the image. For retrieval, this serves as a query-anticipatory bridge, helping retrieve the figure for high-level methodology queries even when the figure itself contains only numerical data.

Case Study: What context enrichment adds to an author caption.

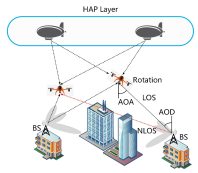
Target Figure	Original Caption	Paper Context	Enriched Caption (Ours)
	Fig. 9. An illustration of 6D localization and HAP-assisted localization approaches for UAVs.	**Title:** An Overview on Integrated Localization and Communication Towards 6G **Abstract:** ...6G should provide ubiquitous connectivity...; accurate location aids beam alignment and network optimization. **Mentions:** [... AOA/AOD estimated in uplink/-downlink 6D localization; quasistatic floating aerial BSs HAP-assisted localization ...]	(figure overall) The image is a schematic diagram illustrating two localization techniques for UAVs in future networks: 6D localization and HAP-assisted localization... In 6D localization, both Azimuth of Arrival (AOA) and Azimuth of Departure (AOD) (mentions caught)... For HAP-assisted localization, the quasistatic floating aerial BSs... (retrieval terminology) Terminology includes 6D localization, AOA, AOD, uplink, downlink...

Figure 6: Context-aware caption generation. Domain-specific entities (green) are extracted from the paper context but absent from the original caption, enriching the figure's semantic representation.

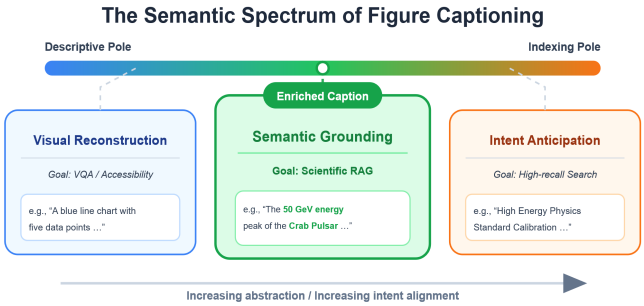


Figure 7: The semantic spectrum of figure captioning. Our enriched caption occupies the mid-point, balancing visual fidelity with retrieval utility for scientific RAG tasks.

Traditional image captioning targets the descriptive pole, mapping visual elements one-to-one into text. For retrieval, a caption instead serves as a semantic anchor that expands the "reachable" territory of potential user queries, including through visually latent or context-derived information. Caption quality therefore depends on both evidential fidelity and retrievability: contextual information may improve reachability, but should remain supported by the article and not contradict visual evidence. Where the optimal balance lies for other multimodal RAG tasks is left to future work.

6.2 Case Study

Figure 6 illustrates the caption generation process on a representative figure. The original author caption ("An illustration of 6D localization and HAP-assisted localization approaches for UAVs") provides minimal retrievable content: it names neither the technical quantities depicted nor the terminology a user would plausibly search for. Our enriched caption recovers domain-specific entities—AOA, AOD, and the *HAP layer*—that appear as labels in the diagram and are elaborated in the surrounding mentions, but are absent from the author caption. This illustrates a mechanism contributing to the retrieval gains in §5.2: enrichment expands the figure's searchable surface area with terms that appear in queries but not in author captions. The example also shows the two provenances an enriched caption mixes: statements grounded in the figure itself, and statements derived from article context. Our prompts permit both by design, so an enriched caption is faithful to the article rather than to the figure alone; whether context-derived statements are also consistent with the figure is what we leave unmeasured.

7 Conclusions & Limitations

We reported on the design and evaluation of a context-enriched figure indexing system for a large-scale scientific digital library. Converting figures into enriched captions that fuse visual content with local article context makes them retrievable through a standard text stack, outperforming author captions and CLIP-based embeddings and yielding higher end-to-end QA scores.

Two lessons stand out. Local context matters more than global: figure mentions contribute most of the gain, while titles and abstracts can add noise. Caption quality also couples only loosely to retrieval utility—GPT-4o writes better captions, yet an 8B open-source model retrieves comparably without per-call API charges, making it practical at indexing scale.

Several limitations remain. Test queries are LLM-generated rather than observed from users and share article provenance with the enriched captions, which are generated once offline and cannot adapt to specific query intent. Because enriched captions include figure mentions and keywords absent from author captions, this asymmetric provenance may inflate gains relative to independent user queries. Quality is scored mainly by an LLM judge sharing a vendor with GPT-4o, checked against a small human sample without agreement statistics; results are single-run and lack confidence intervals, so small differences indicate parity rather than ordering. Recall-oriented metrics leave false positives unmeasured, while near-ceiling SciCap results offer limited discrimination. We also do not quantify visually ungrounded claims, leaving the safeguards of §3.2.2 unvalidated; we use off-the-shelf captioning models, lack comparisons against richer multi-modal RAG architectures under matched cost and latency, and evaluate on a single-publisher corpus.

Acknowledgments

This work was partially supported by the China Scholarship Council (202308440220). We also gratefully acknowledge support from Elsevier. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Abootorabi, M.M., Zobeiri, A., Dehghani, M., Mohammadkhani, M., Mohammadi, B., Ghahroodi, O., Baghshah, M.S., Asgari, E.: Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 16776–16809. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.861

[2] Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html

[3] Bhatia, S., Mitra, P.: Summarizing figures, tables, and algorithms in scientific publications to augment search results. ACM Transactions on Information Systems **30**(1), 1–24 (feb 2012). https://doi.org/10.1145/2094072.2094075

[4] Choudhury, S.R., Wang, S., Giles, C.L.: Scalable algorithms for scholarly figure mining and semantics. In: Proceedings of the International Workshop on Semantic Big Data. pp. 1–6. Association for Computing Machinery (ACM) (jun 2016). https://doi.org/10.1145/2928294.2928305

[5] Demir, S., Carberry, S., McCoy, K.F.: Summarizing information graphics textually. Computational Linguistics **38**(3), 527–574 (sep 2012). https://doi.org/10.1162/coli_a_00091

[6] Ferres, L., Verkhogliad, P., Lindgaard, G., Boucher, L., Chretien, A., Lachance, M.: Improving accessibility to statistical graphs: the igraph-lite system. In: Pontelli, E., Trewin, S. (eds.) Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2007, Tempe, Arizona, USA, October 15-17, 2007. pp. 67–74. ACM (2007). https://doi.org/10.1145/1296843.1296857

[7] Guo, Z., Ren, X., Xu, L., Zhang, J., Huang, C.: Rag-anything: All-in-one rag framework (2025), https://arxiv.org/abs/2510.12323

[8] Hoque, E., Agrawala, M.: Searching the visual style and structure of D3 visualizations. IEEE Trans. Vis. Comput. Graph. **26**(1), 1236–1245 (2020). https://doi.org/10.1109/TVCG.2019.2934431

[9] Hsu, T., Giles, C.L., Huang, T.K.: Scicap: Generating captions for scientific figures. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021. pp. 3258–3264. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.277

[10] Huang, C., Hsu, T., Rossi, R.A., Nenkova, A., Kim, S., Chan, G.Y., Koh, E., Giles, C.L., Huang, T.K.: Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In: Keet, C.M., Lee, H., Zarrieß, S. (eds.) Proceedings of the 16th International Natural Language Generation Conference, INLG 2023, Prague, Czechia, September 11 - 15, 2023. pp. 80–92. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.INLG-MAIN.6

[11] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550

[12] Khattab, O., Zaharia, M.: 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. pp. 39–48 (2020)

[13] Kim, S., Suk, J., Cho, J.Y., Longpre, S., Kim, C., Yoon, D., Son, G., Choi, Y., Shafayat, S., Baek, J., Park, S.H., Hwang, H., Jo, J., Cho, H., Shin, H., Lee, S., Oh, H., Lee, N., Ho, N., Joo, S.J., Ko, M., Lee, Y., Chae, H., Shin, J., Jang, J., Ye, S., Lin, B.Y., Welleck, S., Neubig, G., Lee, M., Lee, K., Seo, M.: The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025. pp. 5877–5919. Association for Computational Linguistics (2025). https://doi.org/10.18653/V1/2025.NAACL-LONG.303

[14] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[15] Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., Liu, Q.: Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models (2024), https://arxiv.org/abs/2403.00231

[16] Li, S., Tajbakhsh, N.: Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs (2023), https://arxiv.org/abs/2308.03349

[17] Li, X., Wang, Y., Luo, H., Liu, S., Liang, J., Fu, Y., Weihuang, Yu, J., Zhu, J.: Bayesrag: Probabilistic mutual evidence corroboration for multimodal retrieval-augmented generation (2026), https://arxiv.org/abs/2601.07329

[18] Li, Y.: Understanding and enhancing robustness in dense information retrieval. In: Campos, R., Jatowt, A., Lan, Y., Aliannejadi, M., Bauer, C., MacAvaney, S., Anand, A., Ren, Z., Verberne, S., Bai, N., Mansoury, M. (eds.) Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part III. Lecture Notes in Computer Science, vol. 16485, pp. 599–607. Springer (2026). https://doi.org/10.1007/978-3-032-21324-2_51

[19] Li, Y., Eustratiadis, P., Fan, Y., Kanoulas, E.: On the robustness of llm-based dense retrievers: A systematic analysis of generalizability and stability (2026), https://arxiv.org/abs/2604.16576

[20] Li, Y., Eustratiadis, P., Kanoulas, E.: Reproducing hotflip for corpus poisoning attacks in dense retrieval. In: Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 15575, pp. 95–111. Springer (2025). https://doi.org/10.1007/978-3-031-88717-8_8

[21] Li, Y., Eustratiadis, P., Lupart, S., Kanoulas, E.: Unsupervised corpus poisoning attacks in continuous space for dense retrieval. In: Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S. (eds.) Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025. pp. 2452–2462. ACM (2025). https://doi.org/10.1145/3726302.3730110

[22] Li, Z., Carberry, S., Fang, H., McCoy, K.F., Peterson, K., Stagitis, M.: A novel methodology for retrieving infographics utilizing structure and message content. Data & Knowledge Engineering **100**, 191–210 (nov 2015). https://doi.org/10.1016/j.datak.2015.05.005

[23] Lin, W., Byrne, B.: Retrieval augmented visual question answering with outside knowledge. arXiv preprint arXiv:2210.03809 (2022)

[24] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

[25] Mahmood, A., Bajwa, I.S., Qazi, K.: An automated approach for interpretation of statistical graphics. In: 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics. vol. 2, pp. 376–379. Institute of Electrical and Electronics Engineers (IEEE) (aug 2014). https://doi.org/10.1109/ihmsc.2014.192

[26] Masry, A., Long, D.X., Tan, J.Q., Joty, S.R., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 2263–2279. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.FINDINGS-ACL.177, https://doi.org/10.18653/v1/2022.findings-acl.177

[27] Oche, A.J., Folashade, A.G., Ghosal, T., Biswas, A.: A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. CoRR **abs/2507.18910** (2025). https://doi.org/10.48550/ARXIV.2507.18910, https://doi.org/10.48550/arXiv.2507.18910

[28] Pramanick, S., Chellappa, R., Venugopalan, S.: SPIQA: A dataset for multimodal question answering on scientific papers. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024). https://doi.org/10.52202/079017-3773, http://papers.nips.cc/paper_files/paper/2024/hash/d74033a247989e8f6f3bf9e0c9629fb5-Abstract-Datasets_and_Benchmarks_Track.html

[29] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html

[30] Riedler, M., Langer, S.: Beyond text: Optimizing RAG with multimodal inputs for industrial applications. CoRR **abs/2410.21943** (2024). https://doi.org/10.48550/ARXIV.2410.21943, https://doi.org/10.48550/arXiv.2410.21943

[31] Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond, vol. 4. Now Publishers Inc (2009)

[32] Salaberria, A., Azkune, G., de Lacalle, O.L., Soroa, A., Agirre, E.: Image captioning for effective use of language models in knowledge-based visual question answering. Expert Syst. Appl. **212**, 118669 (2023). https://doi.org/10.1016/J.ESWA.2022.118669

[33] Shi, Z., Chen, Y., Li, H., Sun, W., Ni, S., Lyu, Y., Fan, R.Z., Jin, B., Weng, Y., Zhu, M., Xie, Q., Guo, X., Yang, Q., Wu, J., Zhao, J., Tang, X., Ma, X., Wang, C., Mao, J., Ai, Q., Huang, J.T., Wang, W., Zhang, Y., Yang, Y., Tu, Z., Ren, Z.: Deep research: A systematic survey (2025), https://arxiv.org/abs/2512.02038

[34] Su, W., Dong, Q., Ai, Q., Liu, Y.: Dynamic and parametric retrieval augmented generation. In: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. p. 453–458. SIGIR-AP 2025, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3767695.3769672

[35] Tang, B.J., Boggust, A.W., Satyanarayan, A.: Vistext: A benchmark for semantically rich chart captioning. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 7268–7298. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.ACL-LONG.401

[36] Wang, J.J., Wang, V.X.: Leveraging large language models to democratize access to costly financial datasets for academic research. CoRR **abs/2412.02065** (2024). https://doi.org/10.48550/ARXIV.2412.02065, https://doi.org/10.48550/arXiv.2412.02065

[37] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR **abs/2409.12191** (2024). https://doi.org/10.48550/ARXIV.2409.12191, https://doi.org/10.48550/arXiv.2409.12191

[38] Wang, Q., Ding, R., Chen, Z., Wu, W., Wang, S., Xie, P., Zhao, F.: ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9113–9134. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.464

[39] Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., Chevalier, A., Arora, S., Chen, D.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/cdf6f8e9fd9aeaf79b6024caec24f15b-Abstract-Datasets_and_Benchmarks_Track.html

[40] Wu, S., Li, Y., Zhu, K., Zhang, G., Liang, Y., Ma, K., Xiao, C., Zhang, H., Yang, B., Chen, W., Huang, W., Moubayed, N.A., Fu, J., Lin, C.: Scimmir: Benchmarking scientific multi-modal information retrieval. In: Ku, L., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 12560–12574. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.FINDINGS-ACL.746

[41] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.Y.: C-pack: Packed resources for general chinese embeddings. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 641–649. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657878

[42] Yang, Z., Dabre, R., Tanaka, H., Okazaki, N.: Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. In: Veyseh, A.P.B., Dernoncourt, F., Nguyen, T.H., Lai, V.D. (eds.) Proceedings of the Workshop on Scientific Document Understanding co-located with 37th AAAI Conference on Artificial Inteligence (AAAI 2023), Remote, February 14, 2023. CEUR Workshop Proceedings, vol. 3656. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3656/paper13.pdf

[43] Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)

[44] Zhao, R., Chen, H., Wang, W., Jiao, F., Long, D.X., Qin, C., Ding, B., Guo, X., Li, M., Li, X., Joty, S.: Retrieving multimodal information for augmented generation: A survey. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023. pp. 4736–4756. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.314