ColPali and ViDoRe focus on page-level, visually rich document retrieval, so the multimodal RAG tests should stress: (1) retrieval quality in vision space, (2) alignment between retrieved pages and generated answers, and (3) robustness across layouts, languages, and modalities.

Below are question *types* and concrete example prompts you can use to probe a ColPali+VLM RAG stack. You can adapt them to your own corpus (reports, PDFs, slides, etc.).

1. Pure visual layout & structure

Goal: require understanding of layout, headings, and non-textual structure, not just OCR.

- "On the page that contains the bar chart comparing quarterly revenues, what is the title of the chart, and how many bars are shown?"
- "Find the page where the main title is centered at the top and a two-column layout is used. Summarize the key claim made in the left column."
- "Locate the page that includes a figure labeled 'Figure 3'. Describe what is shown in that figure in two sentences."
- "Find the first page that contains both a table and an image. Explain how they relate to each other."

Alternative task variants:

- Ask for page number only ("Which page shows …?").
- Ask for a short caption rewrite instead of a summary.

2. Tables, charts, and numerical reasoning

Goal: test retrieval of pages with structured elements + basic reasoning on them.

- "Retrieve the page with the table summarizing benchmark datasets and report the number of queries and documents for each dataset." (mirroring ViDoRe's table-style pages)
- "On the page listing nDCG@5 scores across methods, which model achieves the highest average nDCG@5, and by how many points does it outperform the second-best model?"
- "Find the page that shows a latency versus corpus size plot. At 1,000 pages, what is the approximate query latency for ColPali compared to BGE-M3?"
- "Locate the table that reports performance with token pooling. What pool factor retains at least 97% of the original performance?"

Alternative task variants:

- Ask for ranking ("Order the models by performance on the Healthcare task from best to worst.").
- Ask for trend description ("Does latency grow linearly with corpus size in the figure?").

3. Cross-modal “query is text, evidence is visual”

Goal: queries refer to semantics that appear only visually (figure content, axes labels, layouts).

- “Find the page where the query token ‘hour’ is highlighted in an interpretability heatmap. Explain which parts of the figure are most activated.”
- “Locate the infographic-style page with many icons and small text. What is the main message conveyed by this infographic?”
- “Retrieve the page that contains an image of a line plot with an x-axis labeled ‘hours’. What does the plot measure on the y-axis?”
- “Find the page where red heatmap overlays indicate salient regions. Describe in natural language what the model focuses on.”

Alternative task variants:

- Use very underspecified queries (“Page that emphasizes ‘token pooling’ visually; describe the figure.”).
- Use visual-only cues (“Page with three side-by-side bar charts; summarize the differences.”).

4. Multilingual and domain-specific pages

Goal: test cross-lingual retrieval and domain transfer (e.g., French tables, environmental/medical docs as in ViDoRe).

- “Retrieve the French page that contains an environmental report table and state the column headers in English.”
- “Find a page from an administrative (government) document where a key legal term is visually highlighted (bold or larger font). Explain the term in simple English.”
- “Locate a healthcare document page with a dosage table. What dosage is recommended for adults versus children?”
- “Find the French table page that includes CO_2 emissions data. What are the units used, and what is the largest value in the table?”

Alternative task variants:

- Ask for cross-lingual answers (query in English, answer required in French).
- Ask for domain summarization (“Summarize the main risk mentioned in the medical chart.”).

5. System-level, pipeline-aware queries

Goal: stress understanding of retrieval *and* model design choices described in the paper (good for testing text+figures together).

- "Which page explains the three main requirements (R1, R2, R3) for industrial document retrieval systems? Describe each requirement in one short sentence."
- "Find the page that defines the late interaction scoring function. Restate the formula in words and explain its intuition."
- "Retrieve the page comparing Unstructured + Captioning with ColPali. In what situations does ColPali provide an advantage?"
- "Locate the page that describes token pooling. How does increasing the pool factor affect performance and memory usage?"

Alternative task variants:

- Ask for contrastive explanation ("Why do contrastive VLMs lag behind ColPali on visually complex tasks?").
- Ask for design tradeoffs ("Which factors affect the tradeoff between model size, number of patches, and latency?").

6. RAG-style "answer grounded in one or few pages"

Goal: emulate realistic multimodal RAG: system must retrieve and then generate grounded answers.

Use prompts like:

- "Using at most two pages, explain how ColPali differs from traditional text-only retrieval pipelines that rely on Unstructured parsing."
- "Answer: 'Does ColPali require OCR or PDF parsing at indexing time? Justify your answer based on the paper.'"
- "Given the paper, explain why multi-vector embeddings are beneficial for visually rich documents, and mention at least one tradeoff they introduce."
- "Describe how ViDoRe is constructed (types of tasks and modalities) and why it is more realistic for RAG-like scenarios than prior benchmarks."

Alternative task variants:

- Require citation style in the answer ("Include the page number where evidence comes from.").
- Force the model to say "I don't know" on out-of-scope questions to test hallucination control.

7. Evaluation-oriented meta-questions

Goal: help you debug/score your system.

You can use these as *evaluation tasks* rather than user-facing prompts:

- “Is the answer explicitly supported by the retrieved page(s)? If yes, quote the supporting snippet; if no, mark as hallucinated.”
- “Did the system retrieve a page that actually contains the described figure/table/layout? If not, mark as retrieval failure.”
- “Does the answer mention quantitative values (scores, latencies, counts)? Cross-check them against the retrieved tables/figures and flag discrepancies.”

Alternative task variants:

- Turn these into automatic checks on structured data (e.g., compare extracted numbers to table values).
- Ask a separate LLM/VLM to judge ‘retrieval faithfulness’ given query, pages, and answer.