VinQA: Visual Elements Interleaved Answer Generation for Question Answering on Complex Real-World Multimodal Documents

Anonymous ACL submission

Abstract

The recent advancement of Multimodal Large Language Models (MLLMs) has enabled the extension of Retrieval-Augmented Generation (RAG) to handle multimodal inputs. Prior work has explored retrievers designed to retrieve multimodal contexts relevant to a given query. These contexts typically consist of multiple document pages with various modalities, including text and diverse visual elements such as charts, tables, diagrams, and photos. However, relatively little attention has been paid to generating visual element interleaved answers from such complex multimodal contexts. In this paper, we introduce Visual Elements Interleaved Answer Generation in Question Answering (VinQA) dataset. VinQA is constructed by simulating a Multimodal RAG pipeline over real-world documents, yielding complex multimodal contexts. The answers interleave visual elements at appropriate positions, along with their textual descriptions. We evaluate various proprietary and open-source models on VinQA test set using two encoding methods: *Page Encoding*, which encodes document pages as images to capture full visual appearance, and *Modality Encoding*, which encodes each modality separately for fine-grained understanding. The evaluation assesses grounded answer quality and the effective integration of visual elements. Our results show that *Modality Encoding* generally outperforms *Page Encoding* in the zero-shot setting. However, after training on VinQA training set, both methods exhibit substantial improvements with the performance gap becoming marginal.

1 Introduction

Retrieval-Augmented Generation (RAG) (Guu et al., 2020; Lewis et al., 2021; Yu et al., 2023) has emerged as an effective framework for supplementing large language models (LLMs) with external knowledge. It follows a two-stage paradigm, where a retriever fetches context relevant to the input query, and a generator produces answers based on the retrieved context. By retrieving up-to-date information from external sources, the generator can produce grounded answers even for queries requiring knowledge not seen during pretraining.

Recently, with the emergence of Multimodal Large Language Models (MLLMs) (Yin et al., 2024; Bai et al., 2025; Zhu et al., 2025), the RAG framework has been extended to Multimodal RAG (Cho et al., 2024; Faysse et al., 2025; Yu et al., 2025b; Suri et al., 2025), which processes not only textual content but also visual elements (e.g., charts, tables, diagrams, and photos). However, while prior work has primarily focused on enhancing the ability of retrievers to process multimodal inputs, relatively little attention has been paid to how generators handle retrieved multimodal context.

In this paper, we focus on the generator to produce grounded answers from retrieved multimodal context in real-world scenarios. Figure 1 illustrates how the generator processes the retrieved multimodal context. Designing this process requires consideration of the following three key factors:

Complex Multimodal Contexts: The retrieved context can potentially span multiple document pages. Furthermore, each page typically contains multiple modalities (e.g., text, charts, tables, diagrams, and photos) arranged in diverse layouts. The generator must selectively identify and integrate only the relevant information from these complex multimodal contexts.

Encoding methods: There are two methods for encoding the retrieved multimodal context as input to the MLLM. The first method is *Page Encoding*, which encodes document pages as images to capture the full visual appearance, including layout and structure. The second method is *Modality Encoding*, which encodes each modality separately for a more fine-grained understanding of each modality.

Visual elements interleaved answer: The generator must identify relevant visual elements from

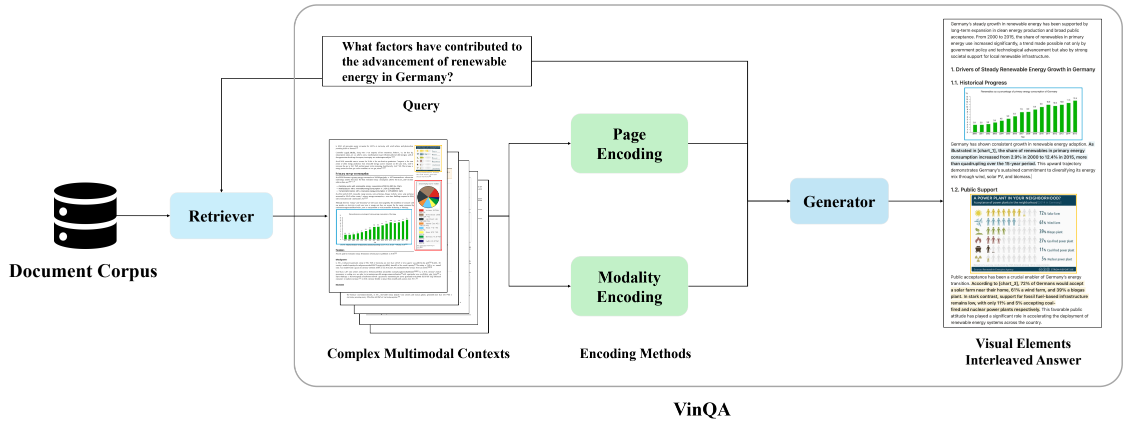


Figure 1: Overview of the Multimodal RAG and VinQA pipeline. The VinQA pipeline involves constructing the input from the retrieved complex multimodal contexts using either *Page Encoding* or *Modality Encoding*, and generating a visual elements interleaved answer.

the retrieved multimodal context and interleave them into the answer at contextually appropriate positions, accompanied by textual descriptions.

Based on these factors, we introduce the Visual Elements Interleaved Answer Generation in Question Answering (VinQA) dataset. We first construct a document corpus by collecting real-world documents from diverse domains (e.g., financial report, presentation, textbook). Next, we curate a set of questions and simulate the retriever step of the Multimodal RAG pipeline to construct complex multimodal contexts.

Since LLMs tend to generate more reliable grounded answers from textual inputs compared to MLLMs dealing with complex multimodal contexts, we leverage this capability to generate answers grounded in the multimodal context. To do so, we first textualize the retrieved multimodal context into a fully text-based format. Specifically, each visual element is converted into textual form via captioning and visual description using an MLLM and is tagged with a unique identifier to enable consistent referencing during answer generation. We then prompt the LLM to generate an answer that cites these identifiers at contextually appropriate positions, along with faithful descriptions. During post-processing, the referenced visual elements are inserted above the paragraph where each citation occurs, resulting in a visual elements interleaved answer. We perform this process separately for constructing both the training and test datasets.

As a result, our experiments demonstrate that training with the VinQA dataset improves performance across multiple evaluation metrics: GroUSE (Muller et al., 2025) for measuring answer quality in RAG generation, Unanswerable F1 for assessing answerability, and Visual Source F1 (Hu et al., 2025) for evaluating visual element citation accuracy. Our analysis reveals that *Modality Encoding* generally outperforms *Page Encoding*, particularly in more complex contexts and in citing table modalities. In addition, evaluation with G-Eval (Liu et al., 2023b) shows that the model also improves in integrating visual elements into answers appropriately, placing them in contextually suitable positions and providing faithful textual descriptions.

2 Related Work

2.1 Multimodal Document QA

Recent works have explored the complete Multimodal RAG pipeline for question answering, in which the generator uses the retrieved multimodal document pages to generate answers (Cho et al., 2024; Suri et al., 2025). VisRAG (Yu et al., 2025b) investigates various methods for generating answers from retrieved document pages and demonstrates that encoding each page as an image input to an MLLM, similar to our *Page Encoding* method, leads to improved performance. However, unlike our work, these studies do not compare alternative encoding methods such as *Modality Encoding*, and their generated answers are composed solely of text without incorporating visual elements.

In addition, some studies perform question answering over all pages of a document, instead of relying solely on retrieved document pages (Ma et al., 2025; Deng et al., 2024; Ding et al., 2024; Zhu et al., 2024). M-LongDoc (Chia et al., 2024) constructured multimodal QA datasets for lengthy documents across various domains and trained models by encoding text and tables separately, similar to our *Modality Encoding* method. However, its answers are also limited to text, without incorporating any visual elements.

2.2 Multimodal Citation Text Generation

Citation text generation refers to the task of generating an answer that includes explicit references to the given context. While various studies have focused on citation text generation from text-only contexts (Gao et al., 2023; Liu et al., 2023a; Zhang et al., 2024), recent works have started to extend this task to multimodal contexts (Yu et al., 2025a).

MCiteBench (Hu et al., 2025) extends citation text generation to multimodal contexts, requiring an MLLM to generate answers by citing relevant information from both text and images (e.g., figure, table) provided in the context. However, the context in MCiteBench is restricted to only five text snippets or images, resulting in a simplified setup that does not fully reflect the complexity of real-world multimodal contexts.

M-DocSum (Yan et al., 2025) constructed a multimodal benchmark focused on summarization tasks, where models generate summaries by explicitly interleaving text and visual elements from documents. M-DocSum was conducted exclusively on a limited domain of arXiv papers, thus having limitations in generalizing to real-world documents.

3 Visual Elements Interleaved Answer Generation in Question Answering (VinQA) Dataset

In this section, we introduce VinQA dataset constructed by simulating a Multimodal RAG pipeline over a corpus of real-world document. Figure 2 provides an overview of the full data construction process, including document collection, question generation, answer generation, and verification. The following sections describe each step in detail.

3.1 Document Collection

To construct VinQA, we first collect a set of multimodal documents from real-world sources (e.g., financial report, presentation, textbook). Accordingly, we systematically harvest the primary sources referenced by six established document QA datasets—MMLongBench-Doc (Ma et al., 2025), TAT-DQA (Zhu et al., 2022), VisDoM (Suri et al., 2025), SlideVQA (Tanaka et al., 2023), MMDocIR (Dong et al., 2025), and VisRAG (Yu et al., 2025b)—and remove duplicates.

After acquiring documents, we classify them into seven categories by topic and format, as shown in Table 1. The corpus is split into training and test corpus. The test corpus is derived via domain-balanced stratified sampling, with strict document-level separation from the training corpus.[1]

Category	Train	Test
Corpus		
Total	**131,906**	**9,373**
Guidebook	19,616	1,278
Financial report	1,554	1,681
Paper	35,331	2,081
Presentation	48,703	1,953
Research report	398	565
Wikipedia	16,508	1,815
Textbook	9,796	–
QA		
Total	**42,700**	**1,712**
Answerable	39,700	1,312
Unanswerable	3,000	400
Page level (Answerable only)		
Single-page	4,030	189
Multi-page	35,670	1,123
Modality level (Answerable only)		
Text-only	12,949	400
Multi-modal	26,751	912
Single-modal	21,164	784
- Chart	4,170	221
- Figure	9,984	245
- Table	7,010	318
Cross-modal	5,587	128

Table 1: Statistics of VinQA. Corpus statistics are shown by domain based on document page images. Page-level categories indicate the number of pages required to answer the question, while modality-level categories indicate the number of different modalities needed.

3.2 Data Preprocessing

In the data preprocessing stage, we convert document page images into text suitable for input to an LLM. First, we perform text OCR using Qwen2.5-VL (Bai et al., 2025) to extract text from each document page, and apply visual element detection using DocLayout (Zhao et al., 2024) to detect visual elements. Additionally, to transform the visual elements into textual form, we employ GPT-4o (Ope-

[1]Textbook data are provided as separate pages without clear document boundaries, so we cannot split them properly. Therefore, we exclude them from the test set.

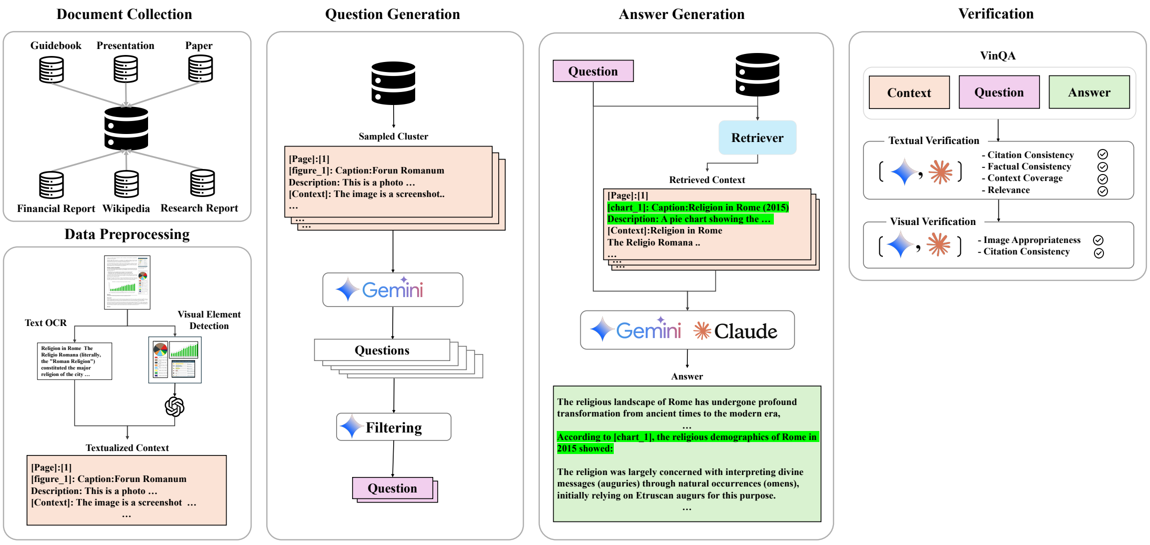


Figure 2: Overview of VinQA construction process. We design the data construction process with the consideration of simulating a Multimodal RAG pipeline over real-world documents.

nAI et al., 2024) to generate class labels[2], captions, and descriptive summaries of the visual content.[3] Note that each textualized visual element is tagged with a unique identifier so that the identifier can later be used for citation during answer generation.

3.3 Question and Answer Generation

We aim to construct the dataset by simulating a realistic Multimodal RAG pipeline. First, we generate questions that could naturally be asked based on our collected document corpus. Specifically, we randomly sample a document page from the corpus and cluster it with similar pages by comparing their image embeddings. We then use Gemini 2.0 Flash Thinking (Google DeepMind, 2024) to generate questions relevant to each cluster. Following the method from VisRAG (Yu et al., 2025b), we filter out context-dependent questions that do not clearly reference specific entities, as these would be difficult to retrieve relevant context.

Using the generated questions, we retrieve the top-K most relevant document pages with a multimodal retriever, Colpali (Faysse et al., 2025), and generate grounded answers with Gemini 2.0 Flash Thinking and Claude 3.7 (Anthropic, 2024b). During answer generation, we prompt the models to cite visual elements using their identifiers, along with faithful descriptions placed at contextually appropriate positions. In the post-processing step, the corresponding visual elements are inserted above the paragraph where each citation appears. Additionally, we construct data for unanswerable QA.[4]

3.4 Data Verification

To ensure the quality of our generated QA data, we perform a multi-step verification procedure. Textual verification is conducted using Gemini 2.0 Flash Thinking and Claude 3.7, following common criteria for citation accuracy, factuality, consistency, and relevancy. Additionally, for the test set, we perform visual verification, evaluating whether relevant images were correctly used and whether cited statements accurately matched the referenced visual elements.[5]

4 Methods

Given a user query q and a retrieved context composed of n document page images $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, the generator is tasked with producing an answer $A = \{x_1, x_2, \ldots, x_k\}$, where each x_i is either a text span or a visual element. This section introduces two methods for encoding $\mathcal{P}$ as input to the generator for producing the answer A: *Page Encoding* and *Modality Encoding*. Figure 3 illustrates the overall workflow of the generator based on these two encoding methods.

[2]We use only the visual elements predicted as `table` or `figure` by DocLayout. To enable fine-grained analysis, we reclassify them into `chart`, `table`, and `figure` (e.g., photos, diagrams, and other non-chart/table elements), as the original labels are not sufficiently precise.

[3]The prompt is shown in Figure 9 of Appendix C.

[4]Detailed procedures for data generation are provided in Appendix B and the prompts used in the data generation process can be found in Figure 10–13 of Appendix C.1.

[5]Detailed verification procedures and the prompts used are provided in Appendix B.4 and C.2, respectively.

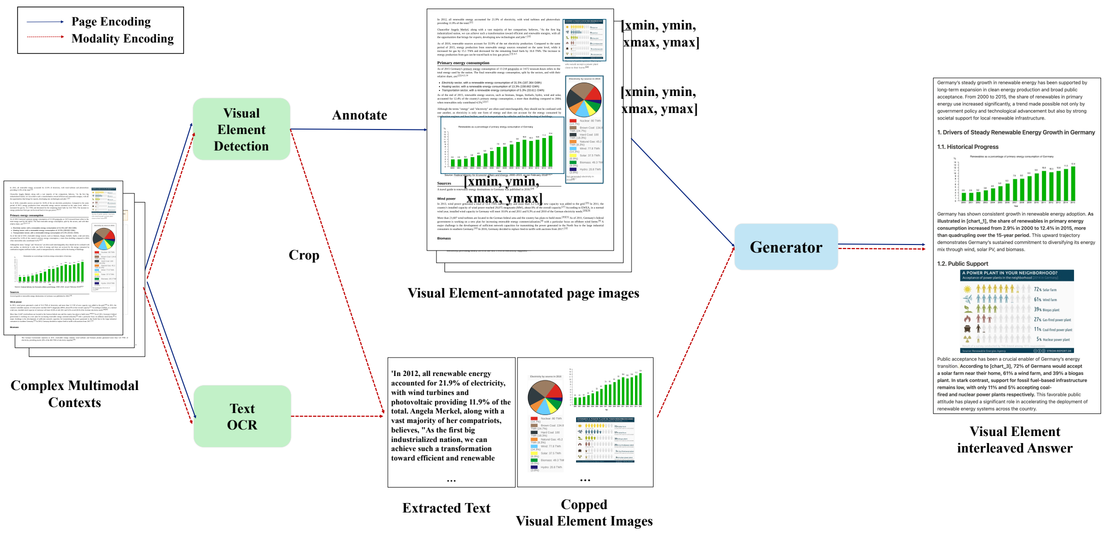


Figure 3: Visual Elements Interleaved Answer Generation based on encoding methods. The blue solid arrows represent the *Page Encoding* method, and the red dotted arrows represent the *Modality Encoding* method.

4.1 Page Encoding

Inspired by methods such as VisRAG (Yu et al., 2025b), we directly utilize the document page images $\mathcal{P}$ as input to the model. While this approach is very simple, it has the advantage of preserving all visual information present in the page image, including diverse layouts and spatial arrangements. In addition, we employ a visual element detection module, DocLayout (Zhao et al., 2024), to identify visual element regions within each page. The detected visual elements are annotated with bounding box coordinates, which are passed to the generator along with the corresponding page image as auxiliary information.

The processed input is represented as:

$$\{(p_i, \mathrm{BBoxList}_i)\}_{i=1}^{n}$$

where p_i denotes the page image of the i-th document page, which is encoded into visual tokens, and $\mathrm{BBoxList}_i = \{b_i^{(1)}, b_i^{(2)}, \dots\}$ represents the set of bounding box coordinates corresponding to visual elements within p_i, each of which is encoded into text tokens. Note that each bounding box is assigned a unique visual element identifier, which allows the generator to later cite the corresponding visual element.[6]

During answer generation, the model cites a visual element identifier whenever these elements are relevant to the question and essential to the answer. When a visual element is referenced, the corresponding image, cropped using its bounding box coordinates, is inserted directly before the paragraph in which the citation appears.

4.2 Modality Encoding

Unlike *Page Encoding*, we first extract text via OCR and detect visual element regions, which are then cropped based on their bounding boxes. The extracted text is then encoded into text tokens, and the cropped visual element images are encoded into visual tokens. Note that a unique identifier is assigned to each cropped visual element image. While this approach may result in some loss of layout information or the spatial arrangement present within the page image, it enables fine-grained understanding of each modality by processing text and visual elements independently.

The processed input is represented as:

$$\{(t_i, \mathrm{V}_i,)\}_{i=1}^{n}$$

where t_i denotes the extracted text from the i-th page, and $\mathrm{V}_i = \{v_i^{(1)}, v_i^{(2)}, \dots\}$ is the set of visual element images cropped from p_i. Similar to the *Page Encoding* approach, the model utilizes visual element identifiers to cite relevant visual elements during answer generation.[7]

5 Experiments

5.1 Main Results

Implementations We adopt Qwen2.5-VL-7B (Bai et al., 2025) as the base model and train it for

[6]The processed input can be found in Figure 17 of Appendix C.

[7]The processed input can be found in Figure 18 of Appendix C.

both encoding methods on the VinQA dataset for 3 epochs using 16 A100 GPUs. Detailed training hyperparameters are provided in Appendix A.

Evaluation Metric First, we adopt the evaluation framework proposed in GroUSE (Muller et al., 2025) to assess the answer quality of the generator in multimodal RAG pipelines. Specifically, we measure Relevancy, Completeness, and Faithfulness of the generated answer along with its multimodal context.[8] These metrics evaluate whether the answer correctly addresses the question (Relevancy), includes all necessary information from the retrieved context (Completeness), and remains grounded in the source content without hallucinations (Faithfulness). Since GroUSE employs an LLM-based evaluation method, we use a textualized context with all visual elements converted into textual descriptions so that the three criteria inherently consider both textual and visual information.

Second, we compute the Unanswerable F1 by checking whether the predicted answer correctly reflects that the question is unanswerable, based on a comparison with the gold answer.

Third, as in MCiteBench (Hu et al., 2025), we compute the Visual Source F1 by comparing the predicted and gold visual element references, aiming to measure the model's ability to cite appropriate visual elements in its answers.

$$\text{Source Precision} = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{gt}}|}{|\mathcal{C}_{\text{pred}}|},$$

$$\text{Source Recall} = \frac{|\mathcal{C}_{\text{pred}} \cap \mathcal{C}_{\text{gt}}|}{|\mathcal{C}_{\text{gt}}|},$$

where $\mathcal{C}$pred and $\mathcal{C}$gt refer to the predicted and ground truth visual element citation sets, respectively.

Overall Performance Table 2 shows the overall performance on the VinQA test set. We evaluate state-of-the-art proprietary models—GPT-4.1, GPT-4.1-mini (OpenAI, 2024), Gemini 2.0 Flash (Google DeepMind, 2024), and Claude 3.5 Sonnet (Anthropic, 2024a)[9]—as well as recent open-source models including InternVL3 (Zhu et al., 2025) and Qwen2.5-VL (Bai et al., 2025).

When comparing the two encoding methods, *Modality Encoding* generally outperforms *Page Encoding* across most metrics and for the majority of models. However, this trend does not hold for models trained on VinQA, which is further analyzed in Section 5.2.

In both the *Page Encoding* and *Modality Encoding* methods, GPT-4.1 shows the best performance in GroUSE average score, Gemini 2.0 Flash in Unanswerable F1, and Claude 3.5 Sonnet in Visual Source F1, indicating that proprietary models generally outperform open-source models in the zero-shot setting without any fine-tuning on the VinQA dataset. When trained on VinQA, the model shows substantial performance improvements across almost all evaluation metrics for both encoding methods. While its performance on GroUSE Avg and Visual Source F1 remains lower than that of proprietary models, it achieves state-of-the-art results in Unanswerable F1.

In the case of GroUSE, both Relevancy and Completeness exhibit substantial performance gains, whereas Faithfulness shows minimal change. This suggests that VinQA enhances the model's ability to generate grounded answers by retrieving question-relevant information from multimodal contexts, but does not significantly improve its ability to generate answers consistent with the given context. This limited improvement may be due to the inherent difficulty the model faces in accurately grounding individual sentences and visual elements within lengthy and complex multimodal contexts. Furthermore, VinQA significantly improves Unanswerable F1, achieving state-of-the-art performance, which demonstrates its effectiveness in helping the model assess question answerability based on the context. It also leads to notable gains in Visual Source F1, indicating that the model becomes better at retrieving relevant visual elements.

5.2 Analysis

Does the Model Perform Well in Complex Multimodal Contexts?

First, we analyze how robust the model is to complex contexts that contain large amounts of text and numerous high-resolution images corresponding to visual elements. Figure 4 (a) and (b) illustrate the average GroUSE performance of *Page Encoding* and *Modality Encoding*, respectively, with respect to the context token length. The context token length refers to the input token length under the *Modality Encoding* method, computed as the sum of text and visual tokens. It reflects the extent of textual length, image resolution, and the

[8] We exclude GroUSE's Negative Rejection and Positive Acceptance criteria, as our evaluation already includes gold-labeled unanswerable data.

[9] We restrict our evaluation to non-thinking models, excluding models such as o3 and Claude 3.7, which may benefit from additional reasoning steps.

Model	GroUSE				Unanswerable	Visual Source
	Relevancy	Completeness	Faithfulness	Avg	F1	F1
Page Encoding						
Proprietary Models						
GPT-4.1	**4.91**	**4.32**	**0.80**	**4.48**	77.89	62.33
GPT-4.1-mini	4.79	4.30	0.76	4.37	46.64	34.89
Gemini 2.0 Flash	4.31	3.10	0.64	3.66	85.57	38.05
Claude 3.5 Sonnet	4.61	4.15	0.73	4.23	69.86	**63.31**
Open-source Models						
InternVL3-8B	4.25	3.07	0.37	3.26	76.54	17.02
Qwen2.5-VL-7B	4.18	3.13	0.59	3.55	70.46	30.24
Qwen2.5-VL-7B (VinQA)	4.68	4.14	0.61	4.09	**90.52**	54.76
Modality Encoding						
Proprietary Models						
GPT-4.1	**4.93**	**4.46**	**0.89**	**4.65** (+0.17)	70.13 (-7.76)	**74.94** (+12.61)
GPT-4.1-mini	4.88	4.45	0.85	4.58 (+0.21)	76.13 (+29.49)	60.64 (+25.75)
Gemini 2.0 Flash	4.54	3.81	0.88	4.28 (+0.62)	85.64 (+0.07)	60.41 (+22.46)
Claude 3.5 Sonnet	4.63	4.17	0.85	4.40 (+0.17)	69.32 (-0.54)	69.94 (+6.63)
Open-source Models						
InternVL3-8B	4.36	3.38	0.53	3.62 (+0.36)	75.85 (-0.69)	32.23 (+15.21)
Qwen2.5-VL-7B	4.31	3.11	0.60	3.61 (+0.06)	73.88 (+3.42)	40.18 (+9.94)
Qwen2.5-VL-7B (VinQA)	4.69	4.10	0.58	4.04 (-0.05)	**90.40** (-0.12)	57.72 (+2.96)

Table 2: Overall performance on the VinQA test set. The values in parentheses show the difference between *Modality Encoding* and *Page Encoding*.

number of images. Although the model's overall performance tends to decline as the token length increases, we observe consistent performance improvements across all input length ranges after training with VinQA. This indicates that VinQA effectively enhances the model's ability to handle complex contexts, regardless of their length.

Second, we analyze the model's robustness to complex visual modalities within the context. Figure 5 illustrates the Visual Source F1 across different types of visual element citations. For this analysis, we leverage the class labels assigned to visual elements during dataset construction to evaluate performance across three modality categories: Chart, Table, and Figure (e.g., photo, diagram, and other elements that do not fall under chart or table). The Mixed modality refers to the case where the answer correctly cites two or more categories. Overall, VinQA leads to consistent performance gains across all modalities. Notably, the model initially struggled with Figure modality, but training with VinQA significantly improved its performance, thereby narrowing the gap compared to other modalities. Furthermore, the performance gains in the Mixed category indicate that the model can effectively handle complex contexts composed of multiple categories of visual elements.

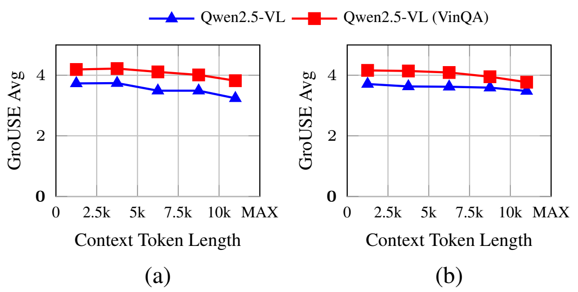


Figure 4: Average GroUSE performance by context token length for (a) *Page Encoding* and (b) *Modality Encoding*.

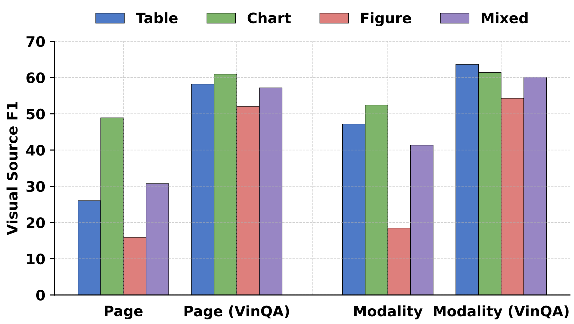


Figure 5: Visual Source F1 across modality types.

Which is Better: Page Encoding or Modality Encoding?

Figure 6 (a) and (b) show the performance differences between *Modality Encoding* and *Page En-*

coding in terms of GroUSE Avg across different context token lengths and Visual Source F1 across different visual modalities, respectively. To mitigate performance variance across models, we also report the model average score computed across GPT-4.1, GPT-4.1-mini, Gemini 2.0 Flash, Claude 3.5 Sonnet, InternVL-3, and Qwen2.5-VL.

In the case of the GroUSE Avg, *Modality Encoding* consistently outperforms *Page Encoding* in terms of model average score across all context token lengths, and the performance gap increases as the context length grows. This indicates that when the context contains a large amount of complex textual content and visual elements, it becomes more effective to encode each modality separately rather than encoding the entire page as an image, which may become too visually dense. Interestingly, after training with VinQA, the *Page Encoding* performs slightly better or comparably to *Modality Encoding* across all token lengths. This suggests that VinQA helps the model become more robust to complex contexts even when using *Page Encoding*.

In Visual Source F1, *Modality Encoding* achieves significantly higher model average scores than *Page Encoding*, particularly for the Table and Mixed modalities. This suggests that in cases such as text-rich tables or when multiple modalities must be cited together, encoding each modality separately can lead to better performance. However, raining with VinQA significantly narrows the gap, indicating improved robustness of *Page Encoding* in citing diverse modalities.

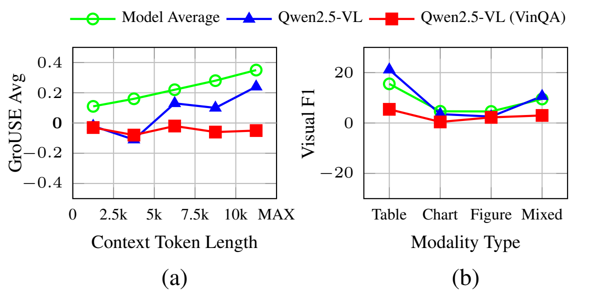


Figure 6: The performance gap between *Modality Encoding* and *Page Encoding* method across (a) Context token length, and (b) Modality type.

Are visual elements appropriately interleaved in the Answer?

The performance of Visual Source F1 is primarily evaluated by comparing the predicted image citations with the gold references. However, this metric does not fully capture whether the cited visual elements are appropriately integrated into the answer. To address this, we additionally evaluate the generated answers using G-Eval (Liu et al., 2023b), focusing on three criteria: Effectiveness (how well the cited image and its accompanying description contribute to the answer), Position (how appropriately the image is placed within the context of the answer), and Expression (how faithfully the accompanying textual description reflects the visual content). Note that, to assess visual aspects, we also include images corresponding to visual elements as part of the input for G-Eval.[10]

Table 3 shows the G-Eval results across the three evaluation criteria. We observe that training with VinQA leads to significant improvements in all criteria, indicating that the model not only becomes better at selecting relevant images but also improves in placing them at appropriate positions and generating faithful textual descriptions, ultimately enhancing the overall answer quality.

Model	**Visual G-Eval**		
	Effectiveness	**Position**	**Expression**
	Page Encoding		
Qwen2.5-VL-7B	1.94	0.43	0.33
Qwen2.5-VL-7B (VinQA)	**3.06**	**0.67**	**0.63**
	Modality Encoding		
Qwen2.5-VL-7B	2.38	0.51	0.46
Qwen2.5-VL-7B (VinQA)	**3.17**	**0.69**	**0.66**

Table 3: Visual G-Eval performance on evaluation for Visual element interleaved answer.

6 Conclusion

We propose VinQA, a dataset for visual elements interleaved answer generation in question answering. Through experiments, we demonstrate that: (1) models trained on VinQA can effectively handle complex multimodal contexts; (2) *Modality Encoding* outperforms *Page Encoding* overall, though the gap narrows after training; and (3) the model generates answers with appropriately placed visual elements and faithful textual descriptions. However, GroUSE Avg and Visual Source F1 remain generally lower than those of proprietary models, even after training on VinQA. To reduce this performance gap, we plan to scale up the model size and expand the training data in future work.

[10] The prompts for G-Eval are provided in Appendix C.4.

Limitations

Limited improvement in faithfulness performance. When evaluating GroUSE performance, the model trained on our dataset demonstrated significant improvements in both Relevancy and Completeness metrics. However, the improvement in Faithfulness was limited. This limitation arises primarily because the model's inherent difficulty in precisely grounding individual sentences and visual elements within lengthy and complex multimodal contexts. Additionally, our data generation process explicitly prioritizes constructing QA data with high question relevancy and appropriate citation of visual elements within multimodal contexts, rather than explicitly addressing faithfulness issues. Furthermore, despite multiple rounds of machine-based verification, the training set may still contain hallucinated data, as it did not undergo extensive noise filtering. Addressing this limitation will likely require more rigorous filtering of erroneous instances and careful refinement of the dataset to enhance faithfulness.

Performance bottleneck due to scaling constraints. Our model, based on Qwen2.5-VL-7B, shows lower baseline performance than proprietary models such as GPT-4.1 and Claude 3.7 across most metrics, except for Unanswerable F1. This is due to the inherently stronger capabilities of proprietary models relative to open-source models like Qwen, and also because we adopt a relatively smaller model size. Results presented in Table 2 show that GPT-4.1-mini—a smaller-scale model—performs noticeably worse than GPT-4.1, suggesting that employing larger-scale models (e.g., 32B or 72B parameters) could similarly yield significant performance improvements for our approach. Furthermore, our training dataset currently consists of approximately 42K QA pairs. Thus, expanding the dataset size could also yield further improvements. Another key limitation is that we restrict the maximum input resolution during training to match our available computational resources. With more resources, increasing this resolution could enable the model to process higher-quality images. This may improve performance, particularly for Page Encoding, which must handle high-resolution document page images, and also for Modality Encoding.

Absence of human verification. Due to resource constraints, our dataset construction process did not include human verification, representing a notable limitation. Instead, we relied on an extensive multi-step machine verification process to ensure data quality. However, incorporating comprehensive human verification, particularly for the test set, would enable a more accurate and reliable assessment of model performance. Future work could address this limitation by conducting thorough human verification on the entire test set.

Ethic statements

We collected multimodal document corpus from various sources as explained in Section 3, providing a permissive licenses for using data. We also utilized multiple APIs for QA generation and evaluation process as mentioned in Section 3. All these APIs are publicly available.

References

Anthropic. 2024a. Claude 3 model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf. Accessed: 2024-04-05.

Anthropic. 2024b. Claude 3.7 sonnet. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2024-05-05.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2024. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *Preprint*, arXiv:2411.06176.

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

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and 1 others. 2024. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Google DeepMind. 2024. Gemini: Our december 2024 update. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 2024-04-05.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of ICML*, pages 3929–3938.

Caiyu Hu, Yikai Zhang, Tinghui Zhu, Yiwei Ye, and Yanghua Xiao. 2025. Mcitebench: A benchmark for multimodal citation text generation in mllms. *Preprint*, arXiv:2503.02589.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2025. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Sacha Muller, Antonio Loison, Bilel Omrani, and Gautier Viaud. 2025. GroUSE: A benchmark to evaluate evaluators in grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4510–4534, Abu Dhabi, UAE. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Gpt-4.1 overview. https://openai.com/index/gpt-4-1/. Accessed: 2024-04-05.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. *Preprint*, arXiv:2412.10704.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Haolong Yan, Kaijun Tan, Yeqing Shen, Xin Huang, Zheng Ge, Xiangyu Zhang, Si Li, and Daxin Jiang. 2025. M-docsum: Do lvlms genuinely comprehend interleaved image-text in document summarization? *arXiv preprint arXiv:2503.21839*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025a. Mramg-bench: A beyondtext benchmark for multimodal retrieval-augmented multimodal generation. *arXiv preprint arXiv:2502.04176*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025b. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of ACL*, pages 2421–2436.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *Preprint*, arXiv:2409.02897.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *Preprint*, arXiv:2410.12628.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. 2024. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Hyperparameters

Configuration	*Page Encoding*	*Modality Encoding*
Epoch	3	3
Optimizer	AdamW	AdamW
Learning Rate	1e-05	1e-05
Learning Rate Scheduler	cosine	cosine
Warm-up Ratio	0.1	0.1
Global Batch Size	16	16
Grad Acc Steps	16	16
Numerical Precision	bfloat16	bfloat16
Image Resolution	2508800	1003520

Table 4: Hyperparameters for training Qwen2.5-VL on VinQA.

B Details of VinQA dataset construction

B.1 Visual-element to textual form transformation

In the data preprocessing process, we generate class label, caption, and description for the visual elements in the document page using GPT-4o. Each visual element was classified into categories such as chart, table, photo, diagram, icon, etc. In our dataset construction, we grouped photo and diagram under the class label "figure" and excluded elements classified as "icon" and "etc". GPT-4o receives two images as input: one containing the full-page image marked with a red bounding box around the target visual element, and the other containing only the cropped image of the visual element. The corresponding prompt is presented in Figure 9.

B.2 Question Generation

To generate diverse and domain-balanced queries, we first sample page images uniformly across all domains in the corpus and assign each as a reference page. For every reference page, we utilize a ColQwen[11], multimodal retriever, to gather the ten most visually and semantically similar pages within the corpus. From these ten, four pages were randomly selected and combined with the reference page, yielding a five-page cluster. Each cluster thus encompasses a coherent but non-redundant context for question generation. From the 131,906 page train corpus, 30,000 reference pages were selected, and an equal number of clusters were consequently constructed; in parallel, 2,000 clusters were constructed from the 9,373 page test corpus.

We prompt Gemini 2.0 Flash Thinking with instructions focused on the following points for generating questions: 1) questions that target the core content of the given context; 2) questions whose answers can be derived from information distributed across multi pages; 3) when the context contains charts, tables or, generate questions that integrate multiple modalities and contexts. We also include eight questions as few-shot examples in the prompt and direct the model to generate five questions for each context. Figure 7 shows an example of input context, and Figure 10 shows the prompt designed for the generation of questions. Subsequently, we sampled three of the five questions generated for each cluster and used Gemini to verify and filter them, removing any questions that were ambiguous, not self-contained, or that referenced unseen context (e.g., “based on the document” or “according to the table”), and retained only those that passed this filtering step. The prompt for filtering questions is present in Figure 11. Out of the 90,000 generated train questions, 66,988 remained after verification, and out of the 6,000 test questions, 4,632 were filtered.

[11] https://huggingface.co/vidore/colqwen2-v1.0

[Page]:[1]
[figure_1]:Caption:None
Description:The image shows a tall, narrow tower with several levels. The structure is made of bricks, and there is a small spire at the top. The tower stands against a clear blue sky.
[chart_1]:Caption:Religion in Rome (2015)
Description:- A pie chart showing the distribution of religions in Rome as of 2015. - Red: Catholicism (82.0%) - Black: Other or non-religious (8.7%) - Blue: Eastern Orthodoxy (4%) - Pink: Protestant (0.8%) - Purple: Judaism (0.7%) - Green: Islam (3.8%)
[Context]:Religion in Rome
The Religio Romana (literally, the "Roman Religion") constituted the major religion of the city in antiquity. The first gods held sacred by the Romans were Jupiter, the highest, and Mars, the god of war, and father of Rome's twin founders, Romulus and Remus, according to tradition. …

[Page]:[2]
[figure_2]:Caption:Forun Romanum
Description:This is a photo with the caption "Forun Romanum" indicating the location as Rome, Holy See and Italy. It is part of the UNESCO World Heritage Site, listed under various cultural criteria. The inscription year is 1980 (4th Session) with extensions noted in 1990 and 2015. The area is 1,430.8 ha (3,536 acres), and coordinates are 41°53′24.8″N 12°29′32.3″E.
[figure_3]:Caption:None
Description:The image shows a map of Rome with a red marker indicating a specific location within the city. There are various lines and markings typically representing roads and geographical features, along with a mini-map of Italy showing the location of Rome within the country.
[Context]:The image is a screenshot of a Wikipedia page titled "Culture of Rome." Here is the extracted text: …

[Page]:[3]
[Context]:The Western religions are the religions that originated within Western culture, which are thus historically, culturally, and theologically distinct from Eastern, African and Iranian religions. The term Abrahamic religions (Christianity, Judaism and Islam) is often used instead of using the East and West terminology, as these originated in the Middle East. …

[Page]:[4]
[figure_4]:Caption:Marcus Aurelius (head covered) sacrificing at the Temple of Jupiter
Description:The image shows a carved relief depicting a group of Roman figures in classical attire. The central figure, identified as Marcus Aurelius with his head covered, appears to be performing a sacrificial ritual at the Temple of Jupiter. The background includes architectural elements such as columns and a pediment structure typical of Roman temples. Several other figures surround Aurelius, engaged in the ceremonial act.
[table_1]:Caption:Religion in ancient Rome
Description:A header section with the title "Religion in ancient Rome" in bold white text on a dark red background. Below the title is a photo with the caption "Marcus Aurelius (head covered) sacrificing at the Temple of Jupiter" in blue italics and black text. Below the photo is a table divided into several categories with pink headers: 1. Practices and beliefs (bold): - libation - votum - temples - festivals - ludi - funerary practices - imperial cult - mystery religions 2. Priesthoods (bold): - Pontifices - Augures - Vestales - Flamines - Fetiales - Epulones - Fratres Arvales 3. Deities (bold): - Twelve major gods (bold) - Capitoline Triad - Aventine Triad - Underworld - indigitamenta (italic) - Agriculture - Birth Two subcategories under Deities: - Deified leaders (bold): - Julius Caesar - Augustus - Other deified persons (bold): - Antinous 4. Related topics (bold): - Glossary of ancient Roman religion (partially visible at the bottom)
[Context]:The text extracted from the image is as follows: …

[Page]:[5]
[table_2]:Caption:Freedom of religion
Description:The table is titled "Freedom of religion" and contains clickable or expandable sections: "Concepts," "Status by country," and "Religious persecution" (with an option to hide). Below these sections, a list of related topics is provided, including Traditional African religions, Atheism, Bahá'í Faith, Buddhism, Christianity (Christophobia), post–Cold War era, Catholicism (Catholic Church), and Mormonism. The table has a light purple background with bold section headers and blue text for clickable links or items.
[Context]:Anti-Judaism
From Wikipedia, the free encyclopedia
Anti-Judaism describes a range of historic and current ideologies which are totally or partially based on opposition to Judaism, on the denial or the abrogation of the Mosaic covenant, and the replacement of Jewish …

Figure 7: Input context example for Question and Answer generation.

B.3 Answer Generation

For 80% of the generated queries, we retrieve the top 5 pages using ColQwen to create answerable QA, while for the remaining 20%, we retrieve the pages ranked 15th to 20th to construct challenging unanswerable QA pairs. We specifically select these lower-ranked pages because they contain partially relevant contexts, making the resulting unanswerable QA more difficult and realistic. These 5 retrieved pages form the context.

We perform answer generation using Gemini 2.0 Flash Thinking and Claude 3.7. The model is provided with a context and a single question, along with the following instructions in the prompt: 1) generate an answer by utilizing as much relevant information as possible from the given context in relation to the question; 2) when citing content from a specific page, include the page index (e.g., [1], [2]) within the response sentence; 3) when referencing charts, tables, or figures include the modality-specific tag (e.g., [chart_1], [table_2], [figure_3]) provided in the context within the response sentence; 4) structure the answer with an introduction, body, and conclusion, where the body is further divided into sections to provide a well-structured response format. For questions aimed at constructing unanswerable QA pairs, a different instruction is provided. While the rest of the process remains the same, the model is instructed to determine whether the question could be answered based solely on the given context, and to generate the reasoning behind this judgment to increase reliability. If the model determines the question to be answerable, it is instructed to generate an answer using the same instructions as for answerable QA pairs. We exclude the questions that are deemed answerable from this process. The prompts designed for the generation of answers are shown in Figure 12 and 13.

As a result, the train corpus contains 53,556 answerable and 10,554 unanswerable QA pairs, while the test corpus consists of 3,704 answerable QA and 751 unanswerable QA pairs. Figure 8 shows an example of the final answer generated from this process.

The religious landscape of Rome has undergone profound transformation from ancient times to the modern era, shifting from a polytheistic society to one predominantly Christian. This evolution reflects broader historical changes in the Mediterranean world and Europe, with significant implications for Rome's cultural and social development.

1. Religious Composition of Modern Rome (2015)

According to [chart_1], the religious demographics of Rome in 2015 showed:

•Catholicism dominates the religious landscape at 82.0%
•Other religions or non-religious individuals constitute 8.7%
•Eastern Orthodoxy represents 4.0% of the population
•Islam accounts for 3.8%
•Protestant denominations make up 0.8%
•Judaism represents 0.7% of Rome's population

This data reveals that modern Rome is overwhelmingly Christian, with the Catholic Church being the dominant religious institution. The presence of Eastern Orthodoxy and Protestantism demonstrates the diversification within Christianity itself.

…

2.2. Religious Practices and Institutions

Ancient Roman religion was highly ritualistic and integrated with civic life. [table_1] indicates that religious practices included libations, votum (vows), temple worship, festivals, and funerary practices. The Roman state maintained various priesthoods including the Pontifices, Augures, Vestales (Vestal Virgins), Flamines, and others to oversee religious affairs.

The religion was largely concerned with interpreting divine messages (auguries) through natural occurrences (omens), initially relying on Etruscan augurs for this purpose.[1]

2.3. Imperial Cult and Deification

During the Imperial period, the cult of the emperor became an important element of Roman religion. As seen in [figure_4], emperors like Marcus Aurelius performed sacrifices at temples, and some emperors were themselves deified after death. [table_1] specifically lists Julius Caesar and Augustus among deified leaders, showing how political power became intertwined with religious authority.

…

Figure 8: Answer example of VinQA. The blue part indicates the citations of either the page numbers or the visual element identifiers of the charts, tables, and within the context.

B.4 Data Verification

During the textual verification step, every Question–Context–Answer triple—where the context consisted exclusively of text, as in the data generation process—is checked for 1) citation errors, 2) statements unsupported or contradicted by the context, 3) omissions of contextually relevant information, and 4) extraneous or query-irrelevant content; only data free of issues on all four criteria were preserved. For the training set, verification is performed solely by Gemini, whereas for the test set, verification is additionally conducted by Claude. Only data meeting all four criteria are retained. The prompt is shown in Figure 14.

During the visual verification step for the test set, whenever the supporting context included a chart, table, or figure, the corresponding image is supplied to the model as input. Visual verification is performed by both Gemini and Claude. The acceptance criteria are also adapted to the visual modality: the models must verify that 1) every image relevant to the question is properly used in the answer, and 2) each cited statement accurately reflects the information presented in its referenced image. Only QA pairs satisfying both visual-reasoning criteria are retained. The prompt is shown in Figure 16.

For unanswerable QA data in the test set, we perform a distinct textual verification with Claude 3.7: each Question–Context pair is inspected to determine whether the context provides enough precise information to answer the question definitively. Any pair that met this condition was deemed incorrectly labeled and discarded. The prompt is shown in Figure 15.

After multi-step machine filtering, the training set comprised 39,700 answerable and 10,554 unanswerable QA pairs, while the test set comprised 1,822 answerable and 723 unanswerable pairs. After the verification process, due to the excessive proportion of unanswerable and text-reference-only QA pairs, we reduced their number to balance the dataset, and the detailed statistics of the finalized dataset are presented in Table 1.

C Prompt Template

C.1 Prompts for Data generation

Figure 9–13 shows the prompts used during the VinQA dataset generation process.

Page Image with red box : **{Page Image}**
Cropped red box image : **{Modality Image}**

Given the page image with the red box and the cropped red box image, you are responsible for explaining the red box image. First, classify the red box image into one of the following categories (category): chart, table, diagram, icon, or photo. Second, generate a caption, which is text in the document explaining the red box image, such as a title, caption, or any other relevant explanation. If there is not any relevant text, just return None. Third, generate Detailed_Description that includes all elements within the red box, such as values, text, and any other relevant details. Do not mention "red box image."
Generated answer format should be Category:str\nCaption:str\nDetailed_Description

Figure 9: The prompts for generating visual-element's class label, caption, and description.

C.2 Prompts for Data verification

Figure 14, 15, and 16 show the prompts used for data verification.

C.3 Prompt examples for encoding method

Figure 17 shows an example prompt input for the *Page Encoding* method, while Figure 18 corresponds to the *Modality Encoding* method.

C.4 Prompts for evaluation

Figure 19, 20, and 21 show the prompts used for vision-based G-Eval.

[Document]
<u>{Input context}</u>

[Instruction]
Please create 5 questions based on the document above, following these guidelines:

1. Reflecting Core Document Content
- You must examine the entire document and create questions related to its core content.
- Questions must be answerable by synthesizing information directly found throughout the provided document.
- Do not create questions that require interpretation or inference not explicitly answered within the document.

2. Multi-page Based Questions
- Prefer questions that can be answered by integrating information from multiple pages within the document rather than those limited to content on a single page.
- It is not necessary to utilize the entire content of the document, but questions reflecting information from multiple pages are preferable.
- Avoid generate questions that use 'and' to ask for two independent pieces of information within a single question as much as possible. (e.g., "What is ~, and how is ~?")

3. Questions Involving Various Modalities (Charts, Tables, Figures)
- If the document contains charts, tables, figures, you must create at least one question whose answer incorporates information from each modality type.
- Create questions that integrate multiple modalities and contexts within the document rather than focusing on only one modality.

4. Question Format
- Generate a total of 5 questions.
- Write each question in a numbered list format, with each question as a single sentence without newlines.
- Do not generate any explanations or statements outside of the question list.
- **The questions must not point to specific parts(e.g., [Document],[Page],[table_1],[chart_1],[figure_1],[Context]) of the provided document; do not include phrases such as "According to the document," "According to the table," "Based on the chart,", "Based on figure," "Based on the document," "as mentioned in document," "shown in figure," "discussed in the document" in questions.**
- Please generate questions by referring to the examples below.
- The question should resemble a search query on GPT without any given context. GPT should be able to retrieve the relevant context using only this question.

Example 1: **<u>{Example question}</u>**
Example 2: **<u>{Example question}</u>**
Example 3: **<u>{Example question}</u>**
Example 4: **<u>{Example question}</u>**
Example 5: **<u>{Example question}</u>**
Example 6: **<u>{Example question}</u>**
Example 7: **<u>{Example question}</u>**
Example 8: **<u>{Example question}</u>**

Figure 10: The prompt for Question generation. The text marked with both bold and underline represents the parts provided as prompt inputs.

[Questions]
{Question}

Your role is to determine whether the given machine generated questions are suitable for a Retrieval-Augmented Generation (RAG) setting.

- The questions should not contain unclear and ambiguous information. For example, a question like “What datasets were utilized in the NLP experiments, and what are some examples of the custom prompts designed for tasks within these datasets?” is not clear, because the term "NLP experiment" is too generic and lacks specificity.
- Since this is a RAG environment, the questions must be self-contained. Question should be formulated as if it were a standalone search query submitted to GPT, without any accompanying context.
- Questions must not contain phrases that indicate something based on unknown context, such as "based on the document," "according to the table," or "provided chart."

For each question, generate a reasoning chain to assess whether it meets the above each criteria. Then, determine whether the question is appropriate using the following format:

[Final Response]
[Appropriate] or [Not appropriate]

Figure 11: The prompt for Question filtering.

[Document]
{Input context}

[Question]
{Question}

[Instruction]
Please structure and generate a detailed answer to the given question, referring to the provided document as much as possible.

1. Answer Guidelines
- Generate a detailed and information-rich response by including as much specific information from the provided document as possible.
- **BUT, Do not include content that deviates from the intent of the question.**
- The response must be written in a professional tone and remain consistent throughout.
- In the entire response, "[Page]", "[Context]", "Caption" and "Description" must not be mentioned in any sentence.

2. Referencing Document Content
- For sentences that use context from the provided document, cite the page number (e.g.,[1],[2]..) at the end of the sentence where the information is used:
    * Apply this only to key sentences, and if the same citation is repeated, include it only in the last instance.
    * Do not cite interpretations or insights that do not exist in the document.
    * If you need to cite more than two pages in one sentence, you need to attach them consecutively as separate tags, as in the example below.
    * Example citation format: This sentence is based on information from a specific page in the document.[1][2]
- For sentences that use modality information (e.g., chart, table, figure), directly mention the corresponding tag (e.g., [chart_1], [table_2], [figure_3]..) within the sentence:
    * Modality tags should only be used as subjects or objects within a sentence and should not appear alone after a sentence.
    * Avoid repeating citations for the same tag across consecutive sentences.
    * Example citation format: According to [chart_1], the response is structured in this way.

3. Answer Format
- Structure your response into introduction, main body, and conclusion.
- Write in Markdown format to maximize readability.
- Do not generate an overly long response by elaborating excessively on the answer to the question.
- The introduction should be a short paragraph summarizing the key points of the response to the given question:
  * Start directly with the introduction without creating a separate title for it.
- The main body should construct the overall content of the response:
  * Divide the key points into sections for clarity and readability.
  * Sections should be divided only to an extent that enhances readability.
  * Avoid duplication of content across different sections.
  * Use ## for main sections with numbered formatting (e.g., "## 1.")
  * Use ### for subsections with decimal formatting (e.g., "### 1.1").
  * Section titles must not include tags such as [chart_1] or [table_1] or [figure_1].
  * If lists are required under subsections, use * for bullet points.
- The conclusion should summarize how the key points of the response align with the question in a short paragraph:
  * Do not create a separate title for the conclusion.
  * Do not mention any tags ([chart_1], [table_1], [figure_1], [Page], [Context], Caption, Description) from the provided document in the conclusion.

Figure 12: The prompt for Answer generation (Answerable QA).

[Document]
{Input context}

[Question]
{Question}

[Instruction]

Step 1: etermine whether you can generate a detailed answer to the given question based on the given document. If even a part of the question cannot be answered completely, the response must be "[Unanswerable]."
1. Carefully review the given document to check if it contains information relevant to the question.
2. When generating an answer to the question, determine whether a complete answer can be provided solely based on the content of the given document.
3. If question is unanswerable, first generate a brief paragraph explaining why it is unanswerable. (Starting with : "Unanswerable explanation: ")
4. After the explanation, write "Result: [Unanswerable]". Do not generate any other sentences afterward.
5. If the question is unanswerable, do not proceed to Step 2 and terminate at Step 1.
6. If the question is answerable, write "Result: [Answerable]", and proceed to Step 2.

Step 2: If question is not unanswerable, please structure and generate a detailed answer to the given question, referring to the provided document as much as possible. Generate answer following the guidelines below.

1. Answer Guidelines
- Generate a detailed and information-rich response by including as much specific information from the provided document as possible.
- **BUT, Do not include content that deviates from the intent of the question.**
- The response must be written in a professional tone and remain consistent throughout.
- In the entire response, "[Page]" and "[Context]" must not be mentioned in any sentence.

2. Referencing Document Content
- For sentences that use context from the provided document, cite the page number (e.g.,[1],[2]..) at the end of the sentence where the information is used:
    * Apply this only to key sentences, and if the same citation is repeated, include it only in the last instance.
    * Do not cite interpretations or insights that do not exist in the document.
    * If you need to cite more than two pages in one sentence, you need to attach them consecutively as separate tags, as in the example below.
    * Example citation format: This sentence is based on information from a specific page in the document.[1][2]
- For sentences that use modality information (e.g., chart, table, figure), directly mention the corresponding tag (e.g., [chart_1], [table_2], [figure_3]..) within the sentence:
    * Modality tags should only be used as subjects or objects within a sentence and should not appear alone after a sentence.
    * Avoid repeating citations for the same tag across consecutive sentences.
    * Example citation format: According to [chart_1], the response is structured in this way.

3. Answer Format
- Start the answer content with the phrase "[Generated Answer]:\n" and write the response below it.
- Structure your response into introduction, main body, and conclusion.
- Write in Markdown format to maximize readability.
- Do not generate an overly long response by elaborating excessively on the answer to the question.
- The introduction should be a short paragraph summarizing the key points of the response to the given question:
  * Start directly with the introduction without creating a separate title for it.
- The main body should construct the overall content of the response:
  * Divide the key points into sections for clarity and readability.
  * Sections should be divided only to an extent that enhances readability.
  * Avoid duplication of content across different sections.
  * Use ## for main sections with numbered formatting (e.g., "## 1.")
  * Use ### for subsections with decimal formatting (e.g., "### 1.1").
  * Section titles must not include tags such as [chart_1] or [table_1] or [figure_1].
  * If lists are required under subsections, use * for bullet points.
- The conclusion should summarize how the key points of the response align with the question in a short paragraph:
  * Do not create a separate title for the conclusion.
  * Do not mention any tags ([chart_1], [table_1], [figure_1], [Page], [Context]) from the provided document in the conclusion.

Figure 13: The prompt for Answer generation (Unanswerable QA).

Question: **{Question}**
Context: **{Input context}**
Answer: **{Answer}**

Given a Question, Context, and Answer, evaluate the following criteria:
1. Is there any citation (**{Citation tag list}**) incorrectly reference the corresponding information in the context? Verify each citation in the answer.
2. Does the answer include any statements that are not supported by the context or contradict it? Verify each statement in the answer.
3. Is there any information in the context that is relevant to the question but missing from the answer? Verify each page in the context.
4. Does the answer include content that is not directly related to the question or unnecessarily detailed? Verify each section in the answer.

First generate rationales for each criterion. And then respond to each number using the following format:

[Final Answer]
[1-Yes] or [1-No]
[2-Yes] or [2-No]
[3-Yes] or [3-No]
[4-Yes] or [4-No]

Figure 14: The prompt for Textual verification (Answerable QA). Citation tag list indicates all page number tags and the chart, table, and figure modality tags that appear in the answer.

Question: **{Question}**
Context: **{Input context}**
Answer: **{Answer}**

Given a Question, Context, and Answer, evaluate the following criteria:
1. Does the context contain enough precise information to answer the question definitively?
• If YES, the pair is mislabeled as unanswerable.

First generate rationales for a criterion. And then respond to each number using the following format:

[Final Answer]
[1-Yes] or [1-No]

Figure 15: The prompt for Textual verification (Unanswerable QA).

Question: **{Question}**
Context: **{Input context}**
Answer: **{Answer}**

Given a Question, Context, and Answer, evaluate the following criteria:
1. Did the response fail to make appropriate use of any image in the context ($image_list) that is relevant to the question and therefore should have been included in the answer? For each image in the context, verify whether it is relevant to the question and whether any relevant image was properly utilized in the answer.
2. Do any sentences in the answer that include a citation ($citation_list) fail to match the facts shown in the image corresponding to each citation within the context? Verify each citation in the answer.

First generate rationales for each criterion. And then respond to each number using the following format:

[Final Answer]
[1-Yes] or [1-No]
[2-Yes] or [2-No]

Figure 16: The prompts for Visual verification. For visual verification, the input context differs from textual verification: any chart, figure, or table modality found in the context is replaced with its image, so the resulting context contains visual elements interleaved with the text.

[Page]:[1]
[figure_1]:<bbox>[602, 889, 940, 979]</bbox>
[chart_1]:<bbox>[731, 223, 939, 431]</bbox>
[Context]:<image>

[Page]:[2]
[figure_2]:<bbox>[673, 517, 934, 648]</bbox>
[figure_3]:<bbox>[677, 834, 930, 979]</bbox>
[Context]:<image>

[Page]:[3]
[Context]:<image>

[Page]:[4]
[figure_4]:<bbox>[762, 298, 914, 498]</bbox>
[table_1]:<bbox>[734, 228, 943, 977]</bbox>
[Context]:<image>

[Page]:[5]
[table_2]:<bbox>[672, 804, 939, 979]</bbox>
[Context]:<image>

Question:How does the religious composition of modern Rome in 2015 compare to the primary religions practiced in ancient Rome?
Find and use information related to the question in the given document to write an answer. If a page in the document contains a chart, table, or figure, the element's location(i.e., bounding box) on that page is provided. If you use information from a chart, table or figure in the given document, write an answer by directly mentioning the corresponding tag (e.g. [chart_1], [table_2], [figure_3]). Write an answer that is clear and systematic, and emphasizes key information.

Figure 17: The prompt example for *Page Encoding* method. The <image> part refers to the corresponding page image input that is converted into visual tokens.

[Page]:[1]
[figure_1]:<image>
[chart_1]:<image>
[Context]:Religion in Rome
The Religio Romana (literally, the "Roman Religion") constituted the major religion of the city in antiquity. The first gods held sacred by the Romans were Jupiter, the highest, and Mars, the god of war, and father of Rome's twin founders, Romulus and Remus, according to tradition. …

[Page]:[2]
[figure_2]:<image>
[figure_3]:<image>
[Context]:The image is a screenshot of a Wikipedia page titled "Culture of Rome." Here is the extracted text: …

[Page]:[3]
[Context]:The Western religions are the religions that originated within Western culture, which are thus historically, culturally, and theologically distinct from Eastern, African and Iranian religions. The term Abrahamic religions (Christianity, Judaism and Islam) is often used instead of using the East and West terminology, as these originated in the Middle East. …

[Page]:[4]
[figure_4]:<image>
[table_1]:<image>
[Context]:The text extracted from the image is as follows: …

[Page]:[5]
[table_2]:<image>
[Context]:Anti-Judaism
From Wikipedia, the free encyclopedia
Anti-Judaism describes a range of historic and current ideologies which are totally or partially based on opposition to Judaism, on the denial or the abrogation of the Mosaic covenant, and the replacement of Jewish …

Question:How does the religious composition of modern Rome in 2015 compare to the primary religions practiced in ancient Rome?
Find and use information related to the question in the given document to write an answer. If you use information from a chart, table or figure in the given document, write an answer by directly mentioning the corresponding tag (e.g. [chart_1], [table_2], [figure_3]). Write an answer that is clear and systematic, and emphasizes key information.

Figure 18: The prompt example for *Modality Encoding* method. The <image> part refers to the input corresponding to each visual element identifier, which is converted into visual tokens.

Question : **{Question}**
Context : **{Input Context}**
Answer : **{Answer}**

Task
Imagine you are a multimodal QA evaluation expert. Your task is to evaluate the effectiveness of each cited image within an answer to the given query. To explain with more detail, images are cited in an Answer using special tag formats such as [category_x] with their corresponding description. These tags are listed in Image Context in the format [category_x]: <context>..., where the placeholder <context> represents the ground truth description of the corresponding image tag. Additionally, the word 'category' is expressed as one of the categories: chart, table, or figure, and 'x' is a natural number. Your task is to evaluate whether the description corresponding to the cited tag in the response is relevant to what the question asks and whether it sufficiently helps explain the answer. The evaluation results should be output in the form of reasons and scores.

Answer Input Format : [text_1] [image_1] [text_2] [image_2]...
Explanation: Each [text_x] is a piece of pure text context, and each [image] represents an image.
The images will be provided in the same order as the placeholders [image].

Scoring Criteria of Effectiveness
When scoring, strictly adhere to the following standards, with a range of 0 to 5:
- 1 point, Harmful: The selected image in the answer are harmful to answering the query, such as causing serious misunderstanding for the reader.
- 2 point, Irrelevant: The selected image in the answer are mostly unrelated to the query and the answer, with little to no connection overall.
- 3 point, Partially Effective: The selected image in the answer are somewhat effective in helping the reader understand the answer to the query.
- 4 point, Mostly Effective: The selected image in the answer are largely consistent with the answer to the query and effectively help the reader better understand the answer.
- 5 point, Highly Effective: The selected image in the answer provide crucial details for answering the query. They not only align with the answer but also offer highly effective supplementary information that aids in understanding the query-answer pair from a multimodal perspective. Provide a brief reason for the evaluation along with a score from 1 to 5. Ensure you do not use any evaluation criteria beyond the query and answer.

Output Format
Please output two lines for each result: the first line is your reasoning for the score, and the second line is the score. Strictly follow this format without any additional content. If no image is used in the response, reply with "No Cited Images".

Output Example (Example with two images)
[chart_1] visually represents the continuously increasing sales of AI semiconductors, which is relevant to the query asking about the potential for AI industry growth. Additionally, the answer asserts that the AI industry is continuously advancing and supports this claim by citing the sales of AI semiconductors such as GPUs. Therefore, the content of [chart_1] is closely related to the query and serves as a highly effective citation, as it is essential to the response.
<chart_1_score>5</chart_1_score>

[table_3] is a table summarizing the annual number of car sales. It is not relevant to the query asking about the potential growth of the AI industry, nor does it align with the answer asserting that the AI industry is continuously advancing. Therefore, [table_1] can be considered an irrelevant citation, as it is unrelated to both the query and the answer.
<table_3_score>2</table_3_score>

Figure 19: The prompts for image citation G-Eval (Effectiveness). The blue part indicates the individual scores assigned to each interleaved image referenced in the answer.

Question : **{Question}**
Context : **{Input Context}**
Answer : **{Answer}**

Task
Imagine you are a multimodal QA evaluation expert. Your task is to evaluate whether the position of each selected image within an Answer to the given Query is appropriate. To explain with more detail, images are cited in an Answer using special tag formats such as [category_x] with their corresponding description. These tags are listed in Image Context in the format [category_x]: <context>..., where the placeholder <context> represents the ground truth description of the corresponding image tag. Additionally, the word 'category' is expressed as one of the categories: chart, table, or figure, and 'x' is a natural number. Specifically, the Answer contains both text and images. Your task is to evaluate whether the cited tags in the response are appropriately placed so that their corresponding descriptions align with the surrounding context without contradiction. The evaluation results should be output in the form of reasons and scores of each image.

Answer Input Format : [context_1] [image_1] [context_2] [text_1] [context_3] ...
Explanation: Each [context_x] is a piece of pure answer text context, and each [image] represents an image. The images will be provided in the same order as the placeholders [image]. The cited text will be provided in the same order as the placeholder [text].

Image Context Input Format
[context_above] [image] [context_bottom]
Explanation: This format represents the contextual information surrounding the image within its original document. It provides supplementary information to assist in evaluating the image.

Revised Evaluation Criteria
Strictly follow the criteria below to assign a score of 0 or 1:
- 0 point, Inappropriate Position: The image is irrelevant to both the preceding and following context, or the position of the image does not enhance content understanding or visual appeal. The insertion of the image does not align with the logical progression of the text and fails to improve the reading experience or information transmission.
- 1 point, Appropriate Position: The image is contextually relevant to at least one of the surrounding contexts (preceding or following), and it enhances content understanding or visual effect. The position of the image aligns with the logical flow of the text and is inserted appropriately, improving the overall information delivery. If the description of the image is detailed, it further clarifies the connection between the image and the text, enhancing the overall expressive effect.

Output Format
Provide a brief justification for the evaluation and a score of either 0 or 1. Ensure no evaluation criteria beyond the provided Query and Answer are used. Please output two lines for each cited image: the first line is your reasoning for the score, and the second line is the score. Strictly follow this format without any additional content. If no image is used in the response, reply with "No Cited Images".

Output Example (Example with two images)
[figure_1] displays a distant aerial view of the site, but the surrounding context focuses on intricate design details of the main entrance. The image placement does not align with the described content and does not improve comprehension.
<figure_1_score>0</figure_1_score>

[figure_2] shows a close-up of one of the pillars, which is directly referenced in the following context about the structure's details. The image placement aligns with the description, enhancing understanding.
<figure_2_score>1</figure_2_score>

Figure 20: The prompts for image citation G-Eval (Position Correctness). The blue part indicates the individual scores assigned to each interleaved image referenced in the answer.

Question : **{Question}**
Context : **{Input Context}**
Answer : **{Answer}**

Task
Imagine you are a multimodal QA evaluation expert. Your task is to evaluate whether the description of each image is accurate. To explain with more detail, images are cited in an Answer using special tag formats such as [category_x] with their corresponding description. These tags are listed in Image Context in the format [category_x]: <context>..., where the placeholder <context> represents the ground truth description of the corresponding image tag. Additionally, the word 'category' is expressed as one of the categories: chart, table, or figure, and 'x' is a natural number. Your task is to compare the descriptions of the cited tags in the response with the previously provided ground truth descriptions and evaluate whether there are any inaccuracies. The evaluation results should be output in the form of reasons and scores of each image.

Answer Input Format : [text_1] [image_1] [text_2] [image_2] ...
Explanation: Each [text_x] is a piece of pure text context, and each [image] represents an image. The images will be provided in the same order as the placeholders [image].

Image Context Input Format
[context_above] [image] [context_bottom]
Explanation: This format represents the contextual information surrounding the image within its original document. It provides supplementary information to assist in evaluating the image.

Revised Evaluation Criteria
Strictly follow the criteria below to assign a score of 0 or 1:
- 0 point, Inappropriate Expression: The description of the image in the response does not match the actual content of the image.
- 1 point, Appropriate Expression: The description of the image in the response matches the actual content of the image.

Output Format
Provide a brief justification for the evaluation and a score of either 0 or 1. Ensure no evaluation criteria beyond the provided query and answer are used. Please output two lines for each selected image: the first line is your reasoning for the score, and the second line is the score. Strictly follow this format without any additional content. If no image is used in the response, reply with "No Cited Images".

Output Example (Example with two images)
[chart_1] visually represents the annually increasing population of South Korea. However, the response describes [chart_1] as containing information about the growing car sales in the United States. Therefore, this does not match what is actually shown in [chart_1].
<chart_1_score>0</chart_1_score>

[table_1] presents the annual number of newborn births, and upon reviewing the specific figures, a yearly increasing trend can be observed. The response cites [table_1], asserting that the number of newborns increases each year, which perfectly aligns with the actual data in [table_1].
<table_1_score>1</table_1_score>

Figure 21: The prompts for image citation G-Eval (Expression Correctness). The blue part indicates the individual scores assigned to each interleaved image referenced in the answer.

D Qualitative Analysis

Figure 22–24 provide the output examples of our model, including comparison with baseline and failure cases.

D.1 Evaluation

Figure 22 presents evaluation results from Qwen2.5-VL-7B and our model trained on VinQA. Compared to Qwen2.5-VL-7B, our model more effectively cites visual elements and generates descriptions in appropriate positions, resulting in more accurate and information-rich responses.

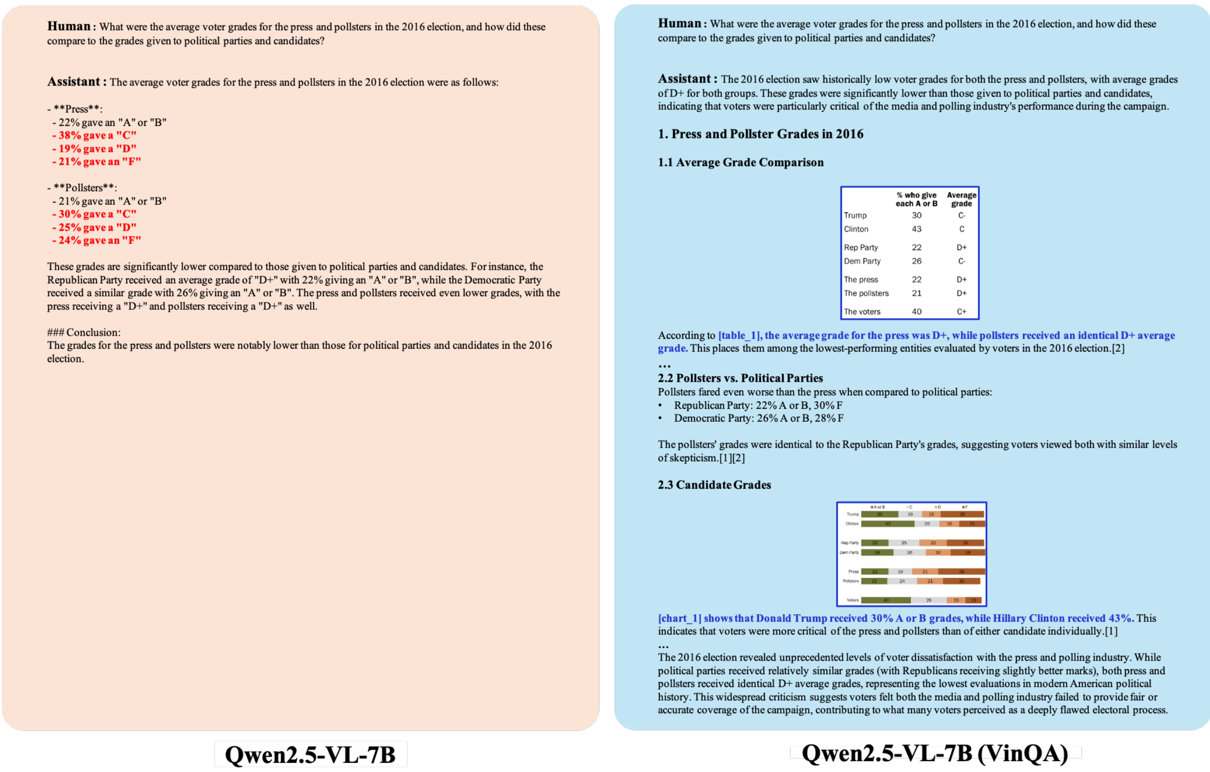


Figure 22: Comparison of predicted answers between Qwen2.5-VL-7B and our model trained on VinQA. The blue and red text respectively indicate the portions correctly and incorrectly predicted by the model.

D.2 Error Analysis

To identify the challenges, we analyze the inference results of our model. Representative failure cases are presented in Figure 23 and 24.

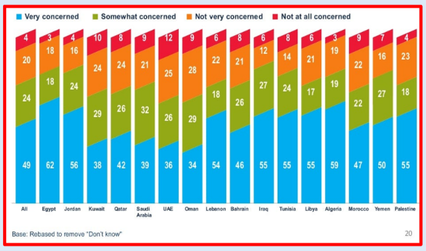
Faithfulness Our model generally shows strong performance in citing relevant visual elements and providing appropriate explanations. However, in some examples, hallucinations related to visual elements are observed. In Figure 23, our model successfully retrieved visual element related to the given question but failed to recognize the detailed values. Such issues primarily occur when the input image exhibits a complex structure, such as charts containing a large amount of information in small text.

Citation While our model generally performs well, it occasionally cites incorrect images, leading to inconsistencies with the retrieved context. In Figure 24, the model correctly retrieved and utilized relevant textual content in the answer, but incorrectly cited the visual element from the table modality. This issue stems from the inherent complexity of integrating visual and textual information, particularly when dealing with ambiguous or highly detailed visuals.

Human : Which country had the highest percentage of respondents expressing very high concern about unemployment, and what was the corresponding perception of ease of starting a business in that same country compared to previous ?

Assistant : Based on the Arab Youth Survey data, Algeria had the highest percentage of respondents expressing very high concern about unemployment at 59%. In contrast, 66% of Algerian respondents believed that people of their generation were more likely to start a business than previous generations.

1. Highest Concern About Unemployment



According to [chart_1], Algeria had the highest percentage of respondents expressing "Very concerned" about unemployment among all surveyed countries. Specifically, 59% of Algerian youth indicated they were very concerned about unemployment, making it significantly higher than other nations in the region.[1]

This high level of concern about unemployment in Algeria aligns with the broader regional trend shown in the survey, where unemployment concerns were relatively high across most Arab countries. The data reveals that while Algeria had the highest percentage of very concerned respondents, many other countries also showed significant levels of concern about unemployment.

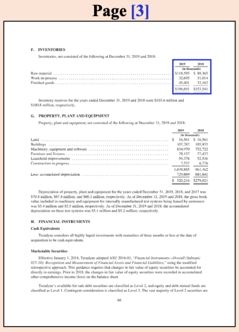
...

The data shows that unemployment concerns increased steadily over time, with 42% of respondents expressing very high concern in 2011, rising to 49% in 2014, demonstrating growing apprehension about employment prospects in the Arab world during this period.[4]

Algeria's high concern about unemployment (59%) was accompanied by a relatively positive perception of entrepreneurship (66%), suggesting that while Algerian youth were worried about finding jobs, they still maintained optimism about starting businesses. This combination of unemployment concerns and entrepreneurial optimism appears to be unique to Algeria among the surveyed countries, though the survey data does not provide information about specific reasons behind these contrasting perspectives.

Figure 23: Faithfulness error in the predicted answer from our model trained on VinQA. The text highlighted in red inaccurately describes the corresponding visual element.

Page [3]



Page [3] Extracted Text

[Page]:[3]
[Context]:F. INVENTORIES
Inventories, net consisted of the following at December 31, 2019 and 2018:
| | 2019 (in thousands) | 2018 (in thousands) |
| --- | --- | --- |
| Raw material | $118,595 | $ 89,365 |
| Work-in-process | 32,695 | 31,014 |
| Finished goods | 45,401 | 33,162 |
| **$196,691 | | **$153,541** |**

Human : What were the reported values for finished goods, work in process, and raw materials inventories as of December 31, 2019 and December 31, 2018?

Assistant : The document provides multiple instances of inventory reporting across different companies and formats. Based on the information provided, I can identify several sets of inventory values for finished goods, work-in-process, and raw materials as of December 31, 2019 and 2018.

1. Inventory Values from Different Tables

...

1.4. Fourth Set of Inventory Values

	Balance at beginning of the period	Additions to expense	(Write offs) / Recoveries	Balance at end of the period
Year ended December 31, 2019	$ 163	30	(84)	$ 109
Year ended December 31, 2018	$ 149	37	(23)	$ 163

[table_4] shows inventory values in thousands of dollars:

- **Raw materials: $118,595 (2019) and $89,365 (2018)**
- **Work-in-process: $32,695 (2019) and $31,014 (2018)**
- **Finished goods: $45,401 (2019) and $33,162 (2018)**[3]

...

The differences between these values likely reflect different accounting periods, reporting standards, or business contexts for each company or entity.

The document provides multiple perspectives on inventory valuation across different companies and reporting formats, demonstrating the importance of context when interpreting financial data. Each set of values represents the inventory position of a specific entity at the end of 2019 and 2018, with varying levels of detail and presentation.

Input Context **Conversation**

Figure 24: Citation error in the predicted answer from our model trained on VinQA. The model correctly retrieves question-relevant content from Page [3], as shown in the blue text, and uses it appropriately in the answer. However, it incorrectly cites an unrelated visual element, highlighted in red.

E GroUSE and Unanswerable F1 performance by Context token length

In this work, we evaluate seven models on our VinQA test set using two encoding strategies—*Page Encoding* and *Modality Encoding*—across five context-token-length intervals (0–2.5k, 2.5–5k, 5–7.5k, 7.5–10k, 10k–). Table 5 presents the overall GroUSE performance and Unanswerable F1 scores across all models by context token length.

F Visual Source performance by Modality type

We evaluate seven models on our VinQA test set using two encoding strategies—*Page Encoding* and *Modality Encoding*—across four modality types (Table, Chart, Figure, Mixed). Table 6 shows the Visual Source performance of all models by modality type.

Model	Context Token Length	GroUSE Relevancy	GroUSE Completeness	GroUSE Faithfulness	GroUSE Avg	Unanswerable F1
Page Encoding						
GPT-4.1	0–2500	4.95	4.30	0.89	4.60	72.13
	2501–5000	4.90	4.37	0.86	4.58	78.95
	5001–7500	4.93	4.37	0.85	4.57	81.20
	7501–10000	4.88	4.35	0.74	4.39	79.71
	10000–	4.89	4.15	0.59	4.14	63.33
GPT-4.1-mini	0–2500	4.79	4.19	0.84	4.45	56.60
	2501–5000	4.75	4.29	0.79	4.39	48.00
	5001–7500	4.84	4.30	0.79	4.44	44.69
	7501–10000	4.77	4.34	0.76	4.38	49.54
	10000–	4.82	4.37	0.57	4.16	31.11
Gemini 2.0 Flash	0–2500	4.36	3.39	0.86	4.07	90.14
	2501–5000	4.34	3.13	0.78	3.87	84.26
	5001–7500	4.26	3.09	0.65	3.65	88.97
	7501–10000	4.28	3.10	0.48	3.43	85.71
	10000–	4.36	2.77	0.35	3.17	72.22
Claude 3.5 Sonnet	0–2500	4.74	4.24	0.78	4.37	65.38
	2501–5000	4.55	4.20	0.78	4.29	70.79
	5001–7500	4.66	4.16	0.77	4.30	68.57
	7501–10000	4.58	4.08	0.69	4.13	76.69
	10000–	4.56	4.08	0.60	4.01	59.26
InternVL3-8B	0–2500	4.36	3.34	0.48	3.54	78.38
	2501–5000	4.26	3.31	0.40	3.39	80.69
	5001–7500	4.21	2.99	0.34	3.19	80.00
	7501–10000	4.20	2.86	0.33	3.13	77.25
	10000–	4.27	2.78	0.26	3.03	51.11
Qwen2.5-VL-7B	0–2500	4.29	3.12	0.70	3.73	72.73
	2501–5000	4.19	3.20	0.71	3.74	66.67
	5001–7500	4.15	3.11	0.56	3.49	70.18
	7501–10000	4.19	3.19	0.53	3.49	75.34
	10000–	4.12	2.94	0.42	3.24	69.23
Qwen2.5-VL-7B (VinQA)	0–2500	4.70	4.16	0.67	4.19	87.18
	2501–5000	4.73	4.25	0.67	4.22	90.43
	5001–7500	4.67	4.17	0.62	4.11	95.31
	7501–10000	4.65	4.06	0.58	4.01	89.66
	10000–	4.65	3.99	0.46	3.82	80.00
Modality Encoding						
GPT-4.1	0–2500	4.92	4.41	0.88	4.62	66.67
	2501–5000	4.92	4.55	0.90	4.69	67.44
	5001–7500	4.95	4.50	0.91	4.70	69.77
	7501–10000	4.93	4.43	0.89	4.64	76.47
	10000–	4.89	4.31	0.87	4.56	67.80
GPT-4.1-mini	0–2500	4.88	4.37	0.89	4.61	75.00
	2501–5000	4.88	4.50	0.87	4.62	76.84
	5001–7500	4.90	4.43	0.88	4.62	76.52
	7501–10000	4.89	4.50	0.84	4.58	76.81
	10000–	4.83	4.42	0.73	4.39	72.13
Gemini 2.0 Flash	0–2500	4.49	3.71	0.94	4.32	88.57
	2501–5000	4.50	3.75	0.89	4.27	88.48
	5001–7500	4.61	3.87	0.89	4.34	84.21
	7501–10000	4.49	3.90	0.85	4.26	89.17
	10000–	4.59	3.77	0.81	4.21	72.00
Claude 3.5 Sonnet	0–2500	4.77	4.28	0.92	4.57	65.38
	2501–5000	4.54	4.16	0.91	4.45	66.67
	5001–7500	4.66	4.14	0.82	4.35	72.81
	7501–10000	4.65	4.21	0.79	4.34	72.18
	10000–	4.61	4.12	0.81	4.32	60.71
InternVL3-8B	0–2500	4.43	3.46	0.58	3.74	69.23
	2501–5000	4.36	3.58	0.60	3.79	78.03
	5001–7500	4.38	3.43	0.49	3.60	83.70
	7501–10000	4.33	3.39	0.50	3.57	78.43
	10000–	4.32	2.84	0.44	3.31	51.92
Qwen2.5-VL-7B	0–2500	4.35	3.07	0.68	3.71	62.92
	2501–5000	4.34	3.15	0.60	3.63	73.73
	5001–7500	4.28	3.11	0.62	3.62	77.29
	7501–10000	4.32	3.16	0.57	3.59	80.00
	10000–	4.26	2.97	0.55	3.48	62.50
Qwen2.5-VL-7B (VinQA)	0–2500	4.78	4.19	0.63	4.16	89.74
	2501–5000	4.69	4.23	0.63	4.14	92.37
	5001–7500	4.71	4.19	0.59	4.09	94.62
	7501–10000	4.68	3.99	0.55	3.95	88.40
	10000–	4.58	3.76	0.49	3.77	76.92

Table 5: Overall performance across context token length.

Model	Modal Type	Precision	Recall	F1
Page Encoding				
GPT-4.1	Table	75.91	62.53	68.57
	Chart	75.67	64.95	69.90
	Figure	74.49	53.36	62.86
	Mixed	75.99	60.35	67.27
GPT-4.1-mini	Table	67.90	42.37	52.17
	Chart	72.98	57.97	64.61
	Figure	76.55	26.77	39.66
	Mixed	71.49	41.89	52.82
Gemini 2.0 Flash	Table	69.39	37.88	49.00
	Chart	72.22	42.22	53.28
	Figure	74.46	33.09	45.81
	Mixed	71.71	27.60	49.33
Claude 3.5 Sonnet	Table	75.06	61.53	67.62
	Chart	73.99	65.83	69.47
	Figure	72.81	64.62	68.47
	Mixed	74.08	63.71	68.50
InternVL3-8B	Table	51.15	10.73	17.73
	Chart	67.82	31.16	42.70
	Figure	63.15	18.09	28.12
	Mixed	62.02	18.79	28.84
Qwen2.5-VL-7B	Table	58.53	16.73	26.02
	Chart	73.32	36.68	48.89
	Figure	69.17	9.00	15.92
	Mixed	66.96	19.94	30.72
Qwen2.5-VL-7B (VinQA)	Table	81.32	45.33	58.21
	Chart	76.25	50.79	60.96
	Figure	76.14	39.56	52.06
	Mixed	78.14	45.07	57.16
Modality Encoding				
GPT-4.1	Table	88.37	73.61	80.31
	Chart	76.87	71.21	73.93
	Figure	79.32	70.57	74.68
	Mixed	82.28	71.98	76.78
GPT-4.1-mini	Table	82.18	63.30	71.51
	Chart	76.44	65.91	70.78
	Figure	68.67	35.26	46.59
	Mixed	77.26	55.27	64.44
Gemini 2.0 Flash	Table	87.54	62.01	72.59
	Chart	74.93	55.71	63.90
	Figure	76.09	42.24	54.32
	Mixed	80.58	54.05	64.70
Claude 3.5 Sonnet	Table	88.06	73.78	80.29
	Chart	76.35	71.80	74.00
	Figure	75.95	71.10	73.44
	Mixed	80.94	72.38	76.42
InternVL3-8B	Table	83.27	20.19	32.49
	Chart	78.89	32.16	45.69
	Figure	67.79	12.25	20.75
	Mixed	78.22	21.08	33.21
Qwen2.5-VL-7B	Table	86.00	32.99	47.18
	Chart	76.32	39.93	52.42
	Figure	67.48	10.70	18.47
	Mixed	79.28	27.98	41.36
Qwen2.5-VL-7B (VinQA)	Table	84.32	51.11	63.64
	Chart	76.63	51.21	61.39
	Figure	76.25	42.09	54.29
	Mixed	79.67	48.32	60.16

Table 6: Overall Visual Source performance across modality types.