ViDoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents

Qiuchen Wang[1], Ruixue Ding, Zehui Chen[1], Weiqi Wu[2], Shihang Wang, Pengjun Xie, Feng Zhao[1*]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2]Shanghai Jiao Tong University

Abstract

Understanding information from visually rich documents remains a significant challenge for traditional Retrieval-Augmented Generation (RAG) methods. Existing benchmarks predominantly focus on image-based question answering (QA), overlooking the fundamental challenges of efficient retrieval, comprehension, and reasoning within dense visual documents. To bridge this gap, we introduce **ViDoSeek**, a novel dataset designed to evaluate RAG performance on visually rich documents requiring complex reasoning. Based on it, we identify key limitations in current RAG approaches: (i) purely visual retrieval methods struggle to effectively integrate both textual and visual features, and (ii) previous approaches often allocate insufficient reasoning tokens, limiting their effectiveness. To address these challenges, we propose **ViDoRAG**, a novel multi-agent RAG framework tailored for complex reasoning across visual documents. ViDoRAG employs a Gaussian Mixture Model (GMM)-based hybrid strategy to effectively handle multi-modal retrieval. To further elicit the model's reasoning capabilities, we introduce an iterative agent workflow incorporating exploration, summarization, and reflection, providing a framework for investigating test-time scaling in RAG domains. Extensive experiments on ViDoSeek validate the effectiveness and generalization of our approach. Notably, ViDoRAG outperforms existing methods by over 10% on the competitive benchmark. The code is available at `https://github.com/Alibaba-NLP/ViDoRAG`.

1 Introduction

Retrieval-Augmented Generation (RAG) enhances Large Models (LMs) by enabling them to use external knowledge to solve problems. As the expression of information becomes increasingly diverse, we often work with visually rich documents that contain diagrams, charts, tables, etc. These visual elements make information easier to understand and are widely used in education, finance, law, and other fields. Therefore, researching RAG within visually rich documents is highly valuable.

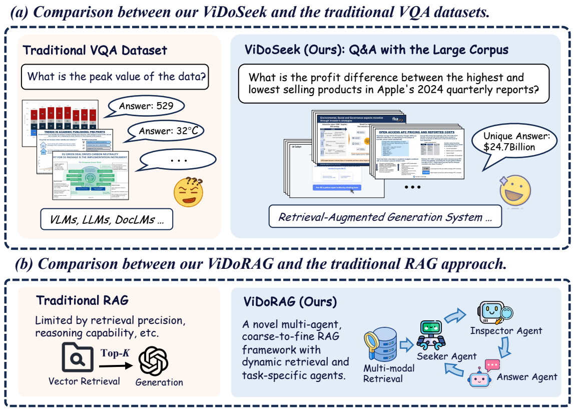

Figure 1: Comparison of our work with the existing datasets and methods. (a) In traditional datasets, each query must be paired with specific images or documents. In our ViDoSeek, each query can obtain a unique answer within the large corpus. (b) Our ViDoRAG is a multi-agent, coarse-to-fine framework specifically optimized for visually rich documents.

In practical applications, RAG systems often need to retrieve information from a large collection consisting of hundreds of documents, amounting to thousands of pages. As shown in Figure 1, existing Visual Question Answering (VQA) benchmarks aren't designed for such large corpus. The queries in these benchmarks are typically paired with one single image (Methani et al., 2020; Masry et al., 2022; Li et al., 2024; Mathew et al., 2022) or document(Ma et al., 2024), which is used for evaluating Q&A tasks but not suitable for evaluating RAG systems. The answers to queries in these datasets may not be unique within the whole corpus.

To address this gap, we introduce ViDoSeek, a novel dataset designed for visually rich document retrieval-reason-answer. In ViDoSeek, each query has a unique answer and specific reference pages.

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9113–9134
November 4-9, 2025

It covers the diverse content types and multi-hop reasoning that most VQA datasets include. This specificity allows us to better evaluate retrieval and generation performance separately.

Moreover, to enable models to effectively reason over a large corpus, we propose ViDoRAG, a multi-agent, coarse-to-fine retrieval-augmented generation framework tailored for visually rich documents. Our approach is based on two critical observations: **(i) Inefficient and Variable Retrieval Performance.** Traditional OCR-based retrieval struggles to capture visual information. With the development of vision-based retrieval, it is easy to capture visual information (Faysse et al., 2024; Yu et al., 2024a; Zhai et al., 2023). However, there lack of an effective method to integrate visual and textual features, resulting in poor retrieval of relevant content. **(ii) Insufficient Activation of Reasoning Capabilities during Generation.** Previous studies on inference scaling for RAG focus on expanding the length of retrieved documents (Jiang et al., 2024; Shao et al., 2025; Xu et al., 2023). However, due to the characteristics of VLMs, only emphasizing on the quantity of knowledge without providing further reasoning guidance presents certain limitations. There is a need for an effective inference scale-up method to efficiently utilize specific action spaces, such as resizing and filtering, to fully activate reasoning capabilities.

Building upon these insights, ViDoRAG introduces improvements in both retrieval and generation. We propose Multi-Modal Hybrid Retrieval, which combines both visual and textual features and dynamically adjusts results distribution based on Gaussian Mixture Models (GMM) prior. This approach achieves the optimal retrieval distribution for each query, enhancing generation efficiency by reducing unnecessary computations. During generation, our framework comprises three agents: the seeker, inspector, and answer agents. The seeker rapidly scans thumbnails and selects relevant images with feedback from the inspector. The inspector reviews, then provides reflection and offers preliminary answers. The answer agent ensures consistency and gives the final answer. This framework reduces exposure to irrelevant information and ensures consistent answers across multiple scales.

Our major contributions are as follows:

- We introduce ViDoSeek, a benchmark specifically designed for visually rich document retrieval-reason-answer, fully suited for evaluation of RAG within large document corpus.
- We propose ViDoRAG, a novel RAG framework that utilizes a multi-agent, actor-critic paradigm for iterative reasoning, enhancing the noise robustness of generation models.
- We introduce a GMM-based multi-modal hybrid retrieval strategy to effectively integrate visual and textual pipelines.
- Extensive experiments demonstrate the effectiveness of our method. ViDoRAG significantly outperforms strong baselines, achieving over 10% improvement, thus establishing a new state-of-the-art on ViDoSeek.

2 Related Work

Visual Document Q&A Benchmarks. Visual Document Question Answering is focused on answering questions based on the visual content of documents (Antol et al., 2015; Ye et al., 2024; Wang et al., 2024). While most existing research (Methani et al., 2020; Masry et al., 2022; Li et al., 2024; Mathew et al., 2022) has primarily concentrated on question answering from single images, recent advancements have begun to explore multi-page document question answering, driven by the increasing context length of modern models (Mathew et al., 2021; Ma et al., 2024; Tanaka et al., 2023; Fang et al., 2025b). However, prior datasets were not well-suited for RAG tasks involving large collections of documents. To fill this gap, we introduce ViDoSeek, the first large-scale document collection QA dataset, where each query corresponds to a unique answer across a collection of $\sim 6k$ images.

Retrieval-augmented Generation. With the advancement of large models, RAG has enhanced the ability of models to incorporate external knowledge (Lewis et al., 2020; Chen et al., 2024b; Wu et al., 2025; Fang et al., 2025a; Zhang et al., 2024; Wang et al., 2025). In prior research, retrieval often followed the process of extracting text via OCR technology (Chen et al., 2024a; Lee et al., 2024; Robertson et al., 2009). Recently, the growing interest in multimodal embeddings has greatly improved image retrieval tasks (Faysse et al., 2024; Yu et al., 2024a). Additionally, there are works that focus on In-Context Learning in RAG (Agarwal et al., 2025; Yue et al., 2024; Team et al., 2024; Weijia et al., 2023). Our work builds upon these

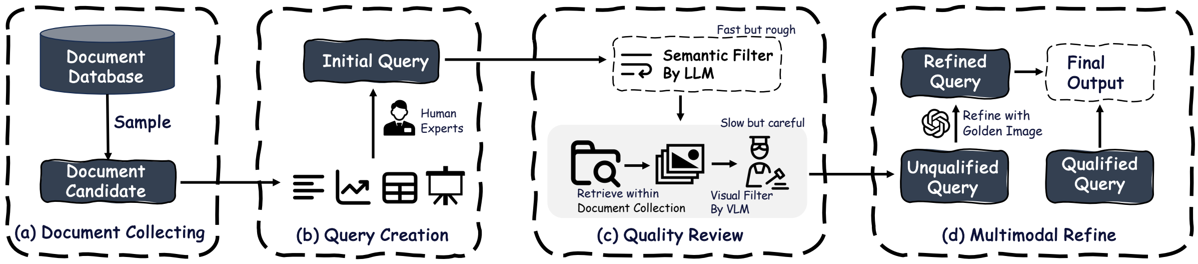

Figure 2: **Data Construction pipeline.** (a) We sample and filter documents according to the requirements to obtain candidates. (b) Then experts construct the initial query from different contents. (c) After that, we prompt GPT-4 to directly determine whether the query is a general query. The remaining queries are carefully reviewed with top-K recall images. (d) Finally, unqualified queries are refined paired with golden image by GPT-4o.

developments by combining multi-modal hybrid retrieval with a coarse-to-fine multi-agent generation framework, seamlessly integrating various embedding and generation models into a scalable framework.

3 Problem Formulation

Given a query as q, and we have a collection of documents $\mathcal{C} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$ which contains M documents. Each document $\mathcal{D}_m$ consists of N pages, each image representing an individual page, defined as $\mathcal{D}_m = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$. The total number of images included in the collection is $\sum_{m=1}^{M} |\mathcal{D}_m|$. We aim to retrieve the most relevant information efficiently and accurately and generate the final answer a to the query q.

4 ViDoSeek Dataset

Existing VQA datasets typically consist of queries paired with a single image or a few images. However, in practical application scenarios, users often pose questions based on a large-scale corpus rather than targeting an individual document or image. To better evaluate RAG systems, we prefer questions that have unique answers when retrieving from a large corpus. To address this need, we introduce a novel **Vi**sually rich **Do**cument dataset specifically designed for RAG systems, called ViDoSeek. We provide the pipeline for constructing the dataset (§4.1) and a detailed analysis of the dataset (§4.2).

4.1 Dataset Construction.

To construct the ViDoSeek dataset, we developed a four-step pipeline to ensure that the queries meet our stringent requirements. As illustrated in Figure 2, our dataset comprises two parts: one annotated from scratch by our AI researchers, and the other derived from refining queries in the existing open-source dataset SlideVQA (Tanaka et al., 2023). For the open-source dataset, we initiate the query refinement starting from the third step of our pipeline. For the dataset we build from scratch, we follow the entire pipeline beginning with document collection. The following outlines our four-step pipeline:

Step 1. Document Collecting. As slides are a widely used medium for information transmission today, we selected them as our document source. We began by collecting English-language slides containing 25 to 50 pages, covering 12 domains such as economics, technology, literature, and geography. And we filtered out 300 slides that simultaneously include text, charts, tables, and two-dimensional layouts which refer to flowcharts, diagrams, or any visual elements composed of various components and are a distinctive feature of slides.

Step 2. Query Creation. To make the queries more suitable for RAG over a large-scale collection, our experts were instructed to construct queries that are specific to the document. Additionally, we encouraged constructing queries in various forms and with different sources and reasoning types to better reflect real-world scenarios.

Step 3. Quality Review. In large-scale retrieval and generation tasks, relying solely on manual annotation is challenging due to human brain limitations. To address this, we propose a review module that automatically identifies problematic queries.

Step 4. Multimodal Refine. In this final step, we refine the queries that did not meet our standards during the quality review. We use carefully designed VLM-based agents to assist us throughout the entire dataset construction pipeline.

Table 1: **Comparison of existing dataset with ViDoSeek.**

DATASET	DOMAIN	CONTENT TYPE	REFERENCE TYPE	LARGE DOCUMENT COLLECTION
PlotQA (Methani et al., 2020)	Academic	Chart	Single-Image	✗
ChartQA (Masry et al., 2022)	Academic	Chart	Single-Image	✗
ArxivQA (Li et al., 2024)	Academic	Chart	Single-Image	✗
InfoVQA (Mathew et al., 2022)	Open-Domain	Text, Chart, Layout	Single-Image	✗
DocVQA (Mathew et al., 2021)	Open-Domain	Text, Chart, Table	Single-Document	✗
MMLongDoc (Ma et al., 2024)	Open-Domain	Text, Chart, Table, Layout	Single-Document	✗
SlideVQA (Tanaka et al., 2023)	Open-Domain	Text, Chart, Table, Layout	Single-Document	✗
ViDoSeek(Ours)	Open-Domain	Text, Chart, Table, Layout	Multi-Documents	✓

4.2 Dataset Analysis

Dataset Statistics. ViDoSeek is the first dataset specifically designed for question-answering over large-scale document collections. It comprises approximately $\sim 1.2k$ questions across a wide array of domains, addressing four key content types: Text, Chart, Table, and Layout. Among these, the Layout type poses the greatest challenge and represents the largest portion of the dataset. Additionally, the queries are categorized into two reasoning types: single-hop and multi-hop. Further details of the dataset can be found in the Appendix D and E.

Comparative Analysis. Table 1 highlights the limitations of existing datasets, which are predominantly tailored for scenarios involving single images or documents, lacking the capacity to handle the intricacies of retrieving relevant information from large collections. ViDoSeek bridges this gap by offering a dataset that more accurately mirrors real-world scenarios. This facilitates a more robust and scalable evaluation of RAG systems.

5 Method

In this section, drawing from insights and foundational ideas, we present a comprehensive description of our **ViDoRAG** framework, which integrates two modules: Multi-Modal Hybrid Retrieval (§5.1) and Multi-Scale View Generation (§5.2).

5.1 Multi-Modal Hybrid Retrieval

For each query, our approach involves retrieving information through both textual and visual pipelines, dynamically determining the optimal value of top-K using a Gaussian Mixture Model (GMM), and merging the retrieval results from both pipelines.

Adaptive Recall with Gaussian Mixture Model. Traditional methods rely on a static hyperparameter, $\mathcal{K}$, to retrieve the top-K images or text chunks from a corpus. A smaller $\mathcal{K}$ might fail to capture sufficient references needed for accurate responses, as the most relevant nodes are not always ranked at the top. Conversely, a larger $\mathcal{K}$ can slow down inference and introduce inaccuracies due to noise. Additionally, manually tuning $\mathcal{K}$ for different scenarios is troublesome.

Our objective is to develop a straightforward yet effective method to automatically determine $\mathcal{K}$ for each modality, without the dependency on a fixed value. We utilize the similarity $\mathcal{S}$ of the embedding E to quantify the relevance between the query and the document collection $\mathcal{C}$:

$$\mathcal{S}(q, \mathcal{C}) = \{s_i | cos(E_q, E_{p_i}), p_i \in \mathcal{C}\} \quad (1)$$

where s_i represents the cosine similarity between the query $\mathcal{Q}$ and page p_i. In the visual pipeline, a page corresponds to an image, whereas in the textual pipeline, it corresponds to chunks of OCR text. We propose that the distribution of $\mathcal{S}$ follows a GMM and we consider they are sampled from a bimodal distribution $\mathcal{P}(s)$ shown in Figure 3:

$$\mathcal{P}(s) = w_F \cdot \mathcal{N}(s \mid \mu_F, \sigma_F^2) + w_T \cdot \mathcal{N}(s \mid \mu_T, \sigma_T^2) \quad (2)$$

where $\mathcal{N}$ represents a Gaussian distribution, with w, μ, σ^2 indicating the weight, mean, and variance, respectively. The subscripts T and F refer to the distributions of pages with high and low similarity. The distribution with higher similarity is deemed valuable for generation. The Expectation-Maximization (EM) algorithm is utilized to estimate the prior probability $\mathcal{P}(T|s, \mu_T, \sigma_T^2)$ for each modality. The dynamic value of $\mathcal{K}$ is defined as:

$$\mathcal{K} = |\{p_i \in \mathcal{C} \mid p_i \sim \mathcal{N}(\mu_T, \sigma_T^2)\}| \quad (3)$$

Considering that the similarity score distribution for different queries within a document collection may not strictly follow a standard distribution, we establish upper and lower bounds to manage outliers. The EM algorithm is employed sparingly, less than $\sim 1\%$ of the time. Dynamically adjusting $\mathcal{K}$ enhances generation efficiency compared to a static setting. Detailed analysis is available in §7.2.

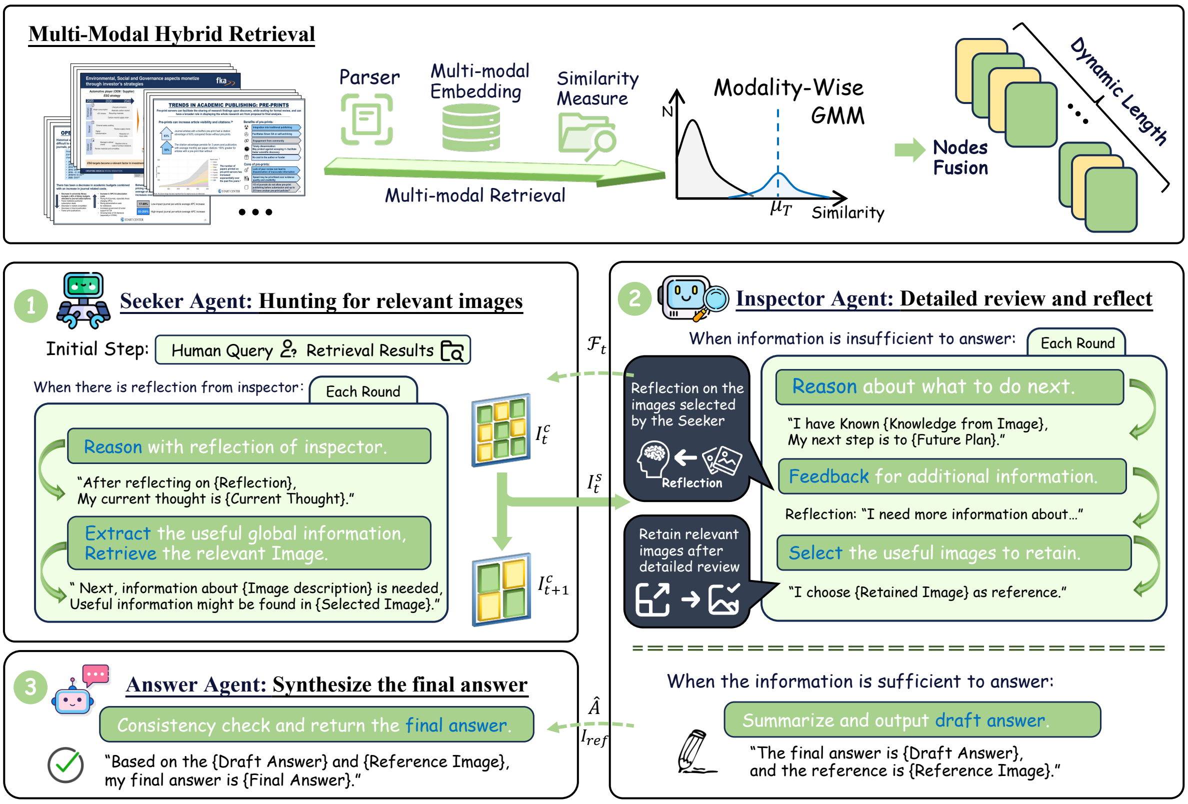

Figure 3: **Overview of our ViDoRAG.** The Multi-Modal Hybrid Retrieval combines visual and textual features and dynamically adjusts the results distribution via GMM. The Multi-Agent Generation involves three agents—Seeker, Inspector, and Answer—which iteratively refine and summarize the answer through coarse-to-fine reasoning.

Textual and Visual Hybrid Retrieval. In the previous step, nodes were retrieved from both pipelines. In this phase, we integrate them:

$$\mathcal{R}_{hybrid} = Sort[\mathcal{F}(\mathcal{R}_{Text}, \mathcal{R}_{Visual})] \quad (4)$$

where $\mathcal{R}_{Text}$ and $\mathcal{R}_{Visual}$ denote the retrieval results from the textual and visual pipelines, respectively. The function $\mathcal{F}(\cdot)$ signifies a union operation, and $Sort(\cdot)$ arranges the nodes in their original sequence, as continuous pages often exhibit correlation (Yu et al., 2024b).

The textual and visual retrieval pipelines demonstrate varying levels of performance for different features. Without adaptive recall, the combined retrieval $\mathcal{R}_{hybrid}$ can become excessive. Adaptive recall ensures that effective retrievals are concise, while traditional pipelines yield longer recall results. This strategy optimizes performance relative to context length, underscoring the value of adaptive recall in hybrid retrieval.

5.2 Multi-Agent Generation with Iterative Reasoning

During the generation, we introduce a multi-agent framework which consists of three types of agents: the Seeker Agent, the Inspector Agent, and the Answer Agent. As illustrated in Figure 3, this framework extracts clues, reflects, and answers in a coarse-to-fine manner from a multi-scale perspective. More details are provided in Appendix H.

Seeker Agent: Hunting for relevant images. The Seeker Agent is responsible for selecting from a coarse view and extracting global cues based on the query and reflection from the Inspector Agent. We have made some improvements to ReAct (Yao et al., 2022) to facilitate better memory management. The action space is defined as the selection of the images. Initially, the agent will reason only based on the query $\mathcal{Q}$ and select the most relevant images $\mathbf{I}_0^{\mathrm{s}}$ from the candidate images $\mathbf{I}_0^{\mathrm{c}}$, while the initial memory $\mathcal{M}_0$ is empty. In step t, the candidate images $\mathbf{I}_{t+1}^{\mathrm{c}}$ are the complement of previously selected images $\mathbf{I}_t^{\mathrm{s}}$, defined as $\mathbf{I}_{t+1}^{\mathrm{c}} = \mathbf{I}_t^{\mathrm{c}} \setminus \mathbf{I}_t^{\mathrm{s}}$. The seeker has received the reflection $\mathcal{F}_{t-1}$ from the inspector, which includes an evaluation of the selected images and a more detailed description of the requirements for the images. The Seeker integrates feedback $\mathcal{F}_{t-1}$ from the Inspector, which includes an evaluation of the selected images and a description of image requirements, to further refine the selection $\mathbf{I}_t^s$ and update the memory $\mathcal{M}_{t+1}$:

$$\mathbf{I}_{t+1}^c,\ \mathcal{M}_{t+1} = \Theta(\mathbf{I}_t^c, \mathcal{Q}, \mathcal{M}_t, \mathcal{F}_{t-1}) \tag{5}$$

where $\mathcal{M}_{t+1}$ represents the model's thought content in step t under the ReAct paradigm, maintaining a constant context length. The process continues until the Inspector determines that sufficient information is available to answer the query, or the Seeker concludes that no further relevant images exist among the candidates.

Inspector Agent: Review in detail and Reflect. In baseline scenarios, increasing the top-K value improves recall@K, but accuracy initially rises and then falls. This is attributed to interference from irrelevant images, referred to as noise, affecting model generation. To address this, we use Inspector to perform a more fine-grained inspection of the images. In each interaction with the Seeker, the Inspector's action space includes providing feedback or drafting a preliminary answer. At step t, the inspector reviews images at high resolution, denoted as $\Theta(\mathbf{I}_t^c \cup \mathbf{I}_{t-1}^r, \mathcal{Q})$ where $\mathbf{I}_{t-1}^r$ are images retained from the previous step and $\mathbf{I}_t^c$ are from the Seeker. If the current information is sufficient to answer the query, a draft answer $\hat{\mathcal{A}}$ is provided, alongside a reference to the relevant image:

$$\hat{\mathcal{A}},\ \mathbf{I}^{ref} = \Theta(\mathbf{I}_t^c \cup \mathbf{I}_{t-1}^r, \mathcal{Q}) \tag{6}$$

Conversely, if more information is needed, the Inspector offers feedback $\mathcal{F}_t$ to guide the Seeker in better image selection and identifies images $\mathbf{I}_t^r$ to retain for further review in the next step $t+1$:

$$\mathcal{F}_t,\ \mathbf{I}_t^r = \Theta(\mathbf{I}_t^c \cup \mathbf{I}_{t-1}^r, \mathcal{Q}) \tag{7}$$

The number of images the Inspector reviews is typically fewer than the Seeker's, ensuring robustness in reasoning, particularly for Visual Language Models with moderate reasoning abilities.

Answer Agent: Synthesize the final answer. In our framework, the Seeker and Inspector engage in a continuous interaction, and the answer agent provides the answer in the final step. To balance accuracy and efficiency, the Answer Agent verifies the consistency of the Inspector's draft answer $\hat{\mathcal{A}}$. If the reference image matches the Inspector's input, the draft answer is accepted as the final answer $\mathcal{A} = \hat{\mathcal{A}}$. If the reference image is a subset of the input image, the answer agent should check for consistency between the draft answer $\hat{\mathcal{A}}$ and the reference image, then give the final answer $\mathcal{A}$: If the reference image is a subset of Inspector's the input, the Answer Agent ensures consistency between the draft answer $\hat{\mathcal{A}}$ and the reference image before finalizing the answer $\mathcal{A}$:

$$\mathcal{A} = \Theta(\mathbf{I}_{ref}, \mathcal{Q}, \hat{\mathcal{A}}) \tag{8}$$

The Answer Agent utilizes the draft answer as prior knowledge to refine the response from coarse to fine. The consistency check between the Answer Agent and Inspector Agent enhances the depth and comprehensiveness of the final answer.

6 Experiments

6.1 Experimental Settings

Evaluation Metric For our end-to-end evaluation, we employed a model-based assessment using GPT-4o, which involved assigning scores from 1 to 5 by comparing the reference answer with the final answer. Answers receiving scores of 4 or above were considered correct, and we subsequently calculate accuracy as the evaluation metric. For retrieval evaluation, we use Recall and MRR (Mean Reciprocal Rank) as the metrics. MRR calculates the average of the reciprocal ranks of the first correct answer for a set of queries.

Baselines and Oracle. We selecte Nv-embed-V2 (Lee et al., 2024) and ColQwen2 (Faysse et al., 2024) as the retrievers for the TextRAG and VisualRAG baselines, respectively. Based on their original settings, we choose the top-5 recall results as the generation input, which equals the average length of dynamic recall results. This ensures a fair comparison and highlights the advantages of our method. The **Oracle** serves as the upper bound performance, where the model responds based on the golden page without retrieval or other operations.

6.2 Main Results

As shown in Table. 2, we conducted experiments on both closed-source and open-source models: GPT-4o, Qwen2.5-7B-Instruct, Qwen2.5-VL-7B-Instruct (Yang et al., 2024), Llama3.2-Vision-90B-Instruct. Closed-source models generally outperform open-source models performance. It is worth mentioning that the Qwen2.5-VL has shown excellent instruction following and reasoning capabilities within our framework. In contrast, we found that the Llama3.2-VL requires 90B parameters to

Table 2: **Overall Generation performance.** The evaluations were conducted on various advanced closed-source and open-source models. Upper Bound represents direct inference with the golden pages.

METHOD	REASONING TYPE		ANSWER TYPE				OVERALL
	Single-hop	Multi-hop	Text	Table	Chart	Layout	
Llama3.2-Vision-90B-Instruct							
Upper Bound	83.1	78.7	88.7	73.1	68.1	85.1	81.1
TextRAG	42.6	45.7	67.6	41.8	25.4	45.9	43.9
VisualRAG	61.8	60.5	82.5	48.5	52.2	63.9	61.2
ViDoRAG (Ours)	73.3	68.5	85.1	65.6	56.1	74.7	**71.2**
Qwen2.5-VL-7B-Instruct							
Upper Bound	77.5	78.2	88.4	77.1	69.4	78.8	77.9
TextRAG	59.6	55.7	78.7	53.8	40.7	60.5	57.6
VisualRAG	66.8	64.3	84.9	61.1	52.8	67.5	65.7
ViDoRAG (Ours)	70.4	67.3	81.9	65.2	57.7	71.3	**69.1**
GPT-4o (Closed-Sourced Models)							
Upper Bound	88.8	86.3	97.5	85.7	77.1	89.4	87.7
TextRAG	64.3	62.6	78.7	61.0	48.4	66.1	63.5
VisualRAG	75.7	66.1	90.1	62.4	58.5	75.4	72.1
ViDoRAG (Ours)	83.5	74.1	88.5	73.6	76.4	80.4	**79.4**

Table 3: Retrieval Performance on ViDoSeek.

Retriever	Recall@1	Recall@3	Recall@5	MRR@5
BM25	55.2	77.4	84.5	66.5
BGE-M3(Chen et al., 2024a)	60.2	79.3	87.6	70.5
NV-Embed-V2(Lee et al., 2024)	64.1	83.5	90.3	74.7
VisRAG-Ret(Yu et al., 2024a)	64.4	84.1	91.2	75.2
ColPali(Faysse et al., 2024)	70.6	87.9	92.8	79.6
ColQwen2(Faysse et al., 2024)	75.4	89.7	95.1	83.3

accomplish the same instructions, which may be related to the model's pre-training domain. The results suggest that, while API-based models offer strong baseline performance, our method is also effective in enhancing the performance of open-source models, offering promising potential for future applications. To further demonstrate the robustness of the framework, we constructed a pipeline using data to rewrite queries from Slide-VQA, making the queries suitable for scenarios involving large corpora. The experimental results are presented in the analysis.

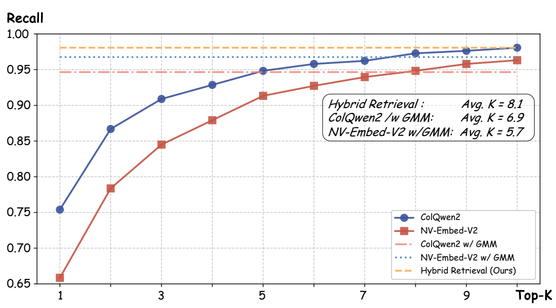

Figure 4: Retrieval performance across different retrievers and hybrid retrieval, along with ablations on GMM.

6.3 Retrieval Evaluation

In Table 3, we report the detailed performance for various retrievers, including OCR-based and visual-based. Due to the uncertainty of dynamical retrieval across queries, we use the average length of results for analysis. Our goal is to incorporate more relevant information within a shorter context while minimizing the impact of noise and reducing computational cost without losing valuable information. As shown in Figure 4, Dynamic Retrieval can achieve better recall performance with a smaller context length, while Hybrid Retrieval combines the results of two pipelines achieving state-of-the-art performance.

7 Analysis

7.1 Ablations

Table 4 presents the impact of different retrievers and generation methods on performance. We have decomposed the retrieval into two components, Dynamic and Hybrid. Naive refers to the method of direct input, which is most commonly used as a baseline. Dynamic indicates using GMM to fit the optimal recall distribution based solely on the visual pipeline. Hybrid refers to merging the visual and the textual retrieval results directly, which leads to suboptimal results due to long contexts. Experiments demonstrate that the effectiveness and scalability of our proposed modules, as well as their combination, can comprehensively enhance end-to-end performance from various perspectives.

Table 4: Ablation study on ViDoSeek benchmark.

Retrieval			Generation		Accuracy
Naive	Dynamic	Hybrid	Naive	Multi-Agent	
✓			✓		72.1
	✓		✓		72.8
		✓	✓		74.1
	✓	✓	✓		74.3
✓				✓	77.3
	✓	✓		✓	**79.4**

7.2 Time Efficiency

How does dynamic retrieval balance latency and accuracy? In traditional RAG systems, using a small Top-K value may result in missing critical information, whereas employing a larger value can introduce noise and increase computational overhead. ViDoRAG dynamically determines the number of documents to retrieve based on the similarity distribution between the query and the corpus. This approach ensures that only the most relevant documents are retrieved, thereby reducing unnecessary computations from overly long contexts and accelerating the generation process. As shown in Table 5, we compare retrieval with and without GMM based on the Naive method. The experiments indicate that GMM may reduce recall due to distribution bias. However, because it significantly shortens the generation context, it effectively improves performance in end-to-end evaluations.

Table 5: Evaluation of Dynamic Retrieval Methods.

Method	Accuracy ↑	Avg. Pages ↓
w/o GMM	72.1	10
w/ GMM	**72.8**	**6.76**

Latency Analysis of the Multi-Agent Generation. There is an increase in delay due to the iterative nature of the multi-agent system, as shown in Figure 5. Each agent performs specific tasks in a sequential manner, which adds a small overhead compared to traditional straightforward RAG. However, despite the increase in latency, the overall performance improves due to the higher quality of generated answers, making the trade-off between latency and accuracy highly beneficial for complex RAG tasks.

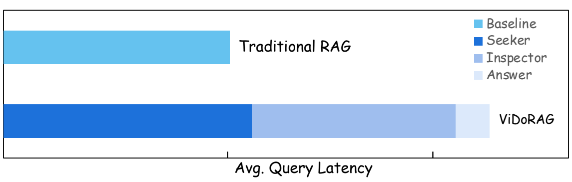

Figure 5: Latency Analysis on Generation.

7.3 Modalities and Strategies of Generation

As shown in Figure 6, the vision-based pipeline outperforms the text-based pipeline across all types, even for queries related to text content. Generally speaking, due to models' inherent characteristics, the reasoning ability of LLMs is stronger than that of VLMs. However, the lack of visual information makes it difficult for models to identify the intrinsic connections between pieces of information. This also poses a challenge for the generation of content based on visually rich documents. While obtaining visual information, VidoRAG further enhances the reasoning capabilities of VLMs, striking a balance between accuracy and computational load.

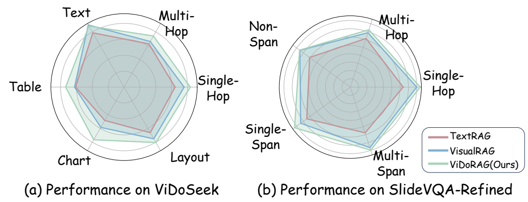

Figure 6: Performance across different types of queries on our ViDoSeek and the refined SlideVQA datasets.

7.4 Performance with Test-time Scaling

Figure 7 illustrates the number of interaction rounds between the seeker and the inspector within ViDoRAG based on different models. Due to the limited instruction capabilities of some models, we sampled 200 queries for the experiment. Models with stronger performance require fewer reasoning iterations, while weaker models often need additional time to process and reach a conclusion. Conditioning the model on a few demonstrations of the task at inference time has been proven to be a computationally efficient approach to enhance model performance (Brown et al., 2020; Min et al., 2021). The results indicate that predefining tasks and breaking down complex tasks into simpler ones is an effective method for scaling inference.

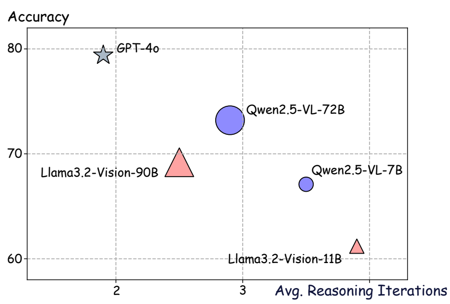

Figure 7: Scaling behavior with ViDoRAG.

8 Conclusion

In this work, we introduced ViDoRAG, a novel multi-agent RAG framework tailored for visually rich documents. By proposing a coarse-to-fine reasoning process and a multi-modal retrieval strategy, ViDoRAG significantly outperforms existing methods, achieving new SOTA on the ViDoSeek benchmark. Future work will focus on further optimizing the framework's efficiency while maintaining high accuracy, and exploring its potential in diverse real-world applications, such as education and finance, where visually rich document RAG is crucial.

Limitations

In addition to the advanced improvements mentioned above, our work has several limitations: **(1) Potential Bias in Query Construction.** The queries in ViDoSeek were constructed by human experts, which may introduce bias in the types of questions and the way they are phrased. This could affect the model's ability to handle more diverse and natural language queries from real-world users. **(2) Computational Overhead of ViDoRAG.** The multi-agent framework, while effective in enhancing reasoning capabilities, introduces additional computational overhead due to the iterative interactions between the seeker, inspector, and answer agents. This may limit the scalability of the framework in scenarios with strict latency requirements. **(3) Model Hallucinations.** Despite the improvements in retrieval and reasoning, the models used in ViDoRAG can still generate hallucinated answers that are not grounded in the retrieved information. This issue can lead to incorrect or misleading responses, especially when the model is overconfident in its generated content.

In summary, while ViDoRAG demonstrates significant improvements in visually rich document retrieval and reasoning, there are still areas for further enhancement, particularly in terms of generalization to diverse document types, reducing potential biases in query construction, optimizing the computational efficiency of the multi-agent framework, and addressing the issue of model hallucinations. Future work will focus on addressing these limitations to further improve the robustness and applicability of the model.

Ethical Considerations

Our data does not contain any private or sensitive information, and all content is derived from publicly available sources. Additionally, the construction and refinement of the dataset were conducted in a manner that respects copyright and intellectual property rights.

Acknowledgments

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. 2025. Many-shot in-context learning. *In Advances in Neural Information Processing Systems*, 37:76930–76966.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *In Advances in Neural Information Processing Systems*, 33:1877–1901.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. 2024b. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*.

Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. 2025a. Attentionrag: Attention-guided context pruning in retrieval-augmented generation. *arXiv preprint arXiv:2503.10720*.

Yixiong Fang, Tianran Sun, Yuling Shi, Min Wang, and Xiaodong Gu. 2025b. Lastingbench: Defend benchmarks against knowledge leakage. *arXiv preprint arXiv:2506.21614*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre

Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *In Advances in Neural Information Processing Systems*, 33:9459–9474.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing*, 1(1):105–115.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2025. Scaling retrieval-based language models with a trillion-token datastore. *In Advances in Neural Information Processing Systems*, 37:91260–91299.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Shi Weijia, Min Sewon, Yasunaga Michihiro, Seo Minjoon, James Rich, Lewis Mike, and Yih Wen-tau. 2023. Replug: Retrieval-augmented black-box language models. *ArXiv: 2301.12652*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024a. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024b. In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2024. Inference scaling for long-context retrieval augmented generation. *arXiv preprint arXiv:2410.04343*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. 2024. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*.

Appendix

A Case Study

As shown in Figure 9, the example demonstrates the use of our ViDoRAG to address questions related to various visually rich content. After two rounds of reasoning, the seeker agent and inspector agent successfully locate the reference image from the candidate images provided by the hybrid retriever. Then, the answer agent reviews and summarizes the inspector's draft answer, providing the final response. This multi-hop query shows the robustness and effectiveness of our method.

B More Analysis on Model-based Evaluation

In order to more accurately evaluate the performance of the framework, we chose the model-based evaluation and carefully designed evaluation criteria and prompts. Here is additional experiment and detailed analysis on model-based evaluation.

Evaluation Based on Different Models. We conducted multiple evaluations using different LLMs on the same set of generated results. The experimental results are shown in Table 6. From the experimental results, it can be seen that model-based evaluation exhibits a slight bias in scoring, but it does not affect the final assessment. The model scores based on its 5-score scale standard, and then we calculate accuracy by setting a threshold 4. The results show that the calculated accuracy is more robust than direct scoring. Using accuracy as the evaluation result is convincing. The table above also shows evaluation results using different models. The results indicate that more advanced models are better aligned with the scoring criteria. Typically, when conducting model-based evaluations, we select models with superior performance.

Table 6: **Results based on different evaluators.**

MODELS	METRIC	SCORE 1	2	3	4	5	ACCURACY ($score \geq 4$)
GPT-4o	Mean	2.4	9.7	8.7	21.7	57.4	79.1
	Std.	0.13	0.10	0.31	0.99	0.66	0.33
GPT-4	Mean	2.2	10.2	10.2	22.5	54.8	77.3
	Std.	0.33	0.35	0.15	0.53	0.44	0.10

Evaluation eesults on different methods. As shown in Table 7, we use different models separately for model-based evaluation to assess whether different models have the ability to distinguish between various methods. The model-based evaluator can effectively distinguish the performance of different RAG pipelines, and its results can serve as a reference. For the models with stronger performance, different evaluators can assess the same RAG method strictly according to the scoring rules, and there is almost no bias in the model.

Table 7: **Consistency assessment among different evaluators.**

METHOD	EVALUATOR	ACCURACY	STD.
TextRAG	GPT-4o	63.4	0.31
	Qwen-Max	63.3	0.22
VisualRAG	GPT-4o	72.1	0.34
	Qwen-Max	71.9	0.28
ViDoRAG	GPT-4o	79.1	0.33
	Qwen-Max	79.2	0.32

Evaluation experiments with various metrics. As shown in Table 8, we use different metrics to evaluate the experimental results, including EM, F1 and ANLS. The results show the performance of different frameworks evaluated using different metrics. Both model-based evaluation and other indicators demonstrate that our framework has achieved state-of-the-art performance. Among these, we consider ANLS to be the best evaluation metric apart from Model-Based Evaluation Accuracy. EM and F1 are more suitable for assessing mathematical answers and short answers, while for long answers, due to the bias in generated answers, using Model-Based Evaluation is a better choice.

Table 8: **Results on Different Metrics.**

METHOD	EM	F1	ANLS	MODEL-BASED
TextRAG	5.1	17.9	20.5	63.5
VisualRAG	10.1	24.5	31.1	72.1
ViDoRAG	32.2	46.6	57.8	79.4

Comparison between automated evaluation and human evaluation. As shown in Figure 9, we sample a batch of queries from different types to conduct repeated experiments in order to compare the differences between human evaluation and automated evaluation. For this evaluation, we used the same criteria to conduct the experiment, and the results are as follows. We have found that human evaluations can be highly unstable, depending on factors such as mood, thoughts, and even levels of fatigue.

The Table 10 summarizes the differences between human evaluation and automated evaluation. Automated evaluation is more convenient than human evaluation when strict rules are established:

Table 9: Evaluation Performance Metrics.

Method	Mean Accuracy	Standard Deviation
Human Evaluation	72.1	4.33
Automated Evaluation	74.1	0.14

Table 10: **Comparison between Human and Automated Evaluation.**

DIMENSION	HUMAN EVALUATION	AUTOMATED EVALUATION
Speed and Cost	Slower and more costly	Faster and more cost-effective.
Consistency	May vary between different evaluators or even the same evaluator at different times due to fatigue or subjective judgment.	Highly consistent across multiple evaluations, when the model and prompts remain unchanged.
Bias	May be prone to human biases.	More objective once the evaluation criteria are defined.

Overall, with strict standards and scoring strategies in place, automated evaluation can completely replace human evaluation and even perform better than human.

C Additional Experiments Details

Backbones. To thoroughly validate the effectiveness of ViDoRAG, we conducted experiments on various models across various baselines, including both closed-source and open-source models: GPT-4o, Qwen2.5-7B, Llama3.2-3B, Qwen2.5-VL-7B-Instruct, Llama3.2-Vision-90B. For OCR-based pipelines, we use PPOCR (Ma et al., 2019) to recognize text within documents. Optionally, VLMs can also be employed for text recognition, as their OCR capabilities are quite strong.

Experimental Environments. We conducted our experiments on a server equipped with 8 A100 GPUs and 96 CPU cores. Open-source models require substantial computational resources.

Retrieval Implementation Details. Due to the context length limitations of the model, we use the Top-$2K$ pages to fit the GMM and we restrict the output chunks of the GMM algorithm to be between $K/2$ and K, we set $K = 10$ in practice.

D More Details on Datasets

D.1 Annotation Case

As shown in Figure 8, this is an example from our dataset. In addition to the query and the golden answer, it also includes the document and page source of the question.

Annotated Data Format

```
## JSON Format
{
    "uid": "04d8bb0db929110f204723c56e5386c1d8d21587_2",
    "query": "What is the temperature of Steam explosion of
      Pretreatment for Switchgrass and Sugarcane bagasse
      preparation?",
    "reference_answer": "195-205 Centigrade",
    "meta_info": {
        "file_name": "04d8bb0db929110f204723c586c1d8d21587.pdf
        ",
        "reference_page": [
            10
        ], # may contain multiple pages
        "source_type": "2d_layout",
        "query_type": "Multi-Hop"
    }
}
```

Figure 8: Annotation case in ViDoSeek.

D.2 Details on ViDoSeek

More Dataset Statistics. The statistical about ViDoSeek is presented in Table 12. We categorize queries from a logical reasoning perspective into single-hop and multi-hop. Text, Table, Chart and Layout represent different sources of reference.

Dataset Difficulty. ViDoSeek sets itself apart with its heightened difficulty level, attributed to the multi-document context and the intricate nature of its content types, particularly the Layout category. The dataset contains both single-hop and multi-hop queries, presenting a diverse set of challenges. Consequently, ViDoSeek serves as a more comprehensive and demanding benchmark for RAG systems compared to previous works.

Table 11: **Statistics of ViDoSeek.**

STATISTIC	NUMBER
Total Questions	1142
Single-Hop	645
Multi-Hop	497
Pure Text	80
Chart	157
Table	175
Layout	730

D.3 Details on SlideVQA-Refined

Dataset Statistics. We supplemented our experiments with the SlideVQA dataset to demonstrate the scalability of our method. SlideVQA categorizes queries from a logical reasoning perspective into single-hop and multi-hop. Non-span, single-span, and multi-span respectively refer to answers derived from a single information-dense sentence, reference information that is sparse but located on the same page, and reference information distributed across different pages. The statistical information about dataset is presented in Table 12.

Table 12: **Statistics of SlideVQA-Refined.**

STATISTIC	NUMBER
Total Questions	2020
Single-Hop	1486
Multi-Hop	534
Non-Span	358
Single-Spin	1347
Multi-Span	315

Dataset Difficulty. The SlideVQA dataset focuses on evaluating the RAG's ability to understand both visually sparse and visually dense information. When multi-hop questions involve reference information spread across different pages, it presents a significant challenge to the RAG system, further demonstrating the effectiveness of our approach.

E Data Construction Details

To construct the ViDoSeek dataset, we developed a four-step pipeline to ensure that the queries meet our requirements.

Step 1. Document Collecting. We collected English-language slides containing 25 to 50 pages, covering 12 domains such as economics, technology, literature, geography, etc.

Step 2. Query Creation. To make the queries more suitable for RAG over a large-scale collection, our experts constructed queries based on the following requirements: (i) Each query must have a unique answer when paired with the document. (ii) The query must include unique keywords that point to the specific document and pages. (iii) The query should require external knowledge. Additionally, we encouraged constructing queries in various forms and with different sources and reasoning types to better reflect real-world scenarios. Our queries not only focus on types of references, including text, tables, charts, and layouts, but also provide a classification of reasoning types, including single-hop and multi-hop.

Step 3. Quality Review. To effectively evaluate the generation and retrieval quality of our RAG system, we require queries that yield unique answers, preferably located on a specific page or within a few pages. However, in large-scale retrieval and generation tasks, relying solely on manual annotation is challenging due to human cognitive limitations. To address this, we propose a review module that automatically identifies problematic queries. This module consists of two steps: (i) We prompt LLMs to filter out queries that may have multiple answers across the document collection; for example, the question *What is the profit for this company in 2024?* might have a unique answer within a single document but could yield multiple answers in a multi-document setting. (ii) For the remaining queries, we retrieve the top-k slides for each query and use a VLM to determine whether each slide can answer the query. If only the golden page can answer the question, we consider it to meet the requirements. If pages other than the golden page can answer the query, we have experts manually evaluate and refine them. Please see Figure 10 and Figure 11 for detailed prompts.

Step 4. Multimodal Refine. In this final step, we refine the queries that did not meet our standards during the quality review. The goal is to adjust these queries so they satisfy the following requirements: (i) The refined query should point to specific pages within the large collection with minimal additional information; (ii) The refined query must retain its original meaning. We use carefully designed VLM-based agents to assist us throughout the entire dataset construction pipeline. The prompt is presented in Figure 10 and Figure 11, respectively. We will first perform filtering based on semantics, and then conduct a fine-grained review using a multimodal reviewer. Please see Figure 12 for detailed prompts.

F Retrieval Performance Across Various Data Types

Apart from purely visual elements and text, tables are elements that lie between text and two-dimensional distributions. In the retrieval stage, from the text retrieval perspective, the structured nature of tables allows the retrieval system to quickly locate keywords and match queries with table content, enhancing precision.

From the visual retrieval perspective, the 2D layout of tables enables vision models to identify their

Table 13: **Comparison between Existing Works and Our ViDoRAG.**

DIMENSION	EXISTING WORKS	OUR VIDORAG
Retrieval Modality	Single Modality (Text or Visual)	Multi-Modality (both text and visual)
Context Length	Static Top-K requiring manual adjustment	Dynamic top-k based on feature relevance
Generation Paradigm	Limited action space, overly reliant on textual reasoning capabilities, lacking visual perception.	Multi-modal generation framework with visual feature-based action space, supporting visual scaling and coarse-to-fine reasoning.
Reasoning Approach	Text-based reasoning only, struggling with visual information	Emphasizes visual coarse-to-fine reasoning, fully leveraging the reasoning capabilities of VLM models with limited context length.

Table 14: Retrieval Performance on Table Type.

Retriever	Recall@1	Recall@3	Recall@5	MRR@5
BM25	56.5	77.1	86.3	68.1
BGE-M3 (Chen et al., 2024a)	64.5	82.3	92.1	74.5
NV-Embed-V2 (Lee et al., 2024)	69.7	88.5	92.6	79.1
VisRAG-Ret (Yu et al., 2024a)	75.4	90.3	95.4	83.5
ColPali (Faysse et al., 2024)	79.4	94.3	97.7	86.9
ColQwen2 (Faysse et al., 2024)	85.7	96.6	98.9	91.4

structure and spatial relationships, facilitating rapid screening of relevant table images. The experimental results in Figure 14 show that for table-type queries, the NV-Embed-V2 retriever achieved a Recall@5 of 92.6% and an MRR@5 of 79.1%, while the ColQwen2 retriever achieved a Recall@5 of 98.9% and an MRR@5 of 91.4%. Their retrieval results still have a mutually exclusive set, demonstrating the complementary relationship in the final retrieval performance of the two modalities. In the ViDoRAG framework, integrating text and visual retrieval capabilities substantially enhances the retrieval performance of tabular data with shorter context lengths as shown in Figure 4 of our manuscript.

As shown in Figure 15, in the generation stage, our framework demonstrates a general improvement across all types of queries, including those involving tabular data. Understanding tables requires both spatial positional information and specific information extraction. Our ViDoRAG treats tables as two-dimensional visual elements, enabling it to effectively integrate spatial and textual information during the reasoning process. Compared to TextRAG and VisualRAG, our framework achieves a significant improvement in accuracy for table-type queries, reaching 73.6% with GPT-4o.

G The Difference Between Our ViDoRAG and Existing Works

As shown in Table 13, our method introduces **four innovative aspects** aimed at addressing key challenges in visual document retrieval and reasoning.

Table 15: Comparison of Different Methods.

Method	Llama3.2-Vision-90B-Instruct	Qwen2.5-VL-7B-Instruct	GPT-4o
TextRAG	41.8	53.8	61.0
VisualRAG	48.5	61.1	62.4
ViDoRAG	65.6	65.2	73.6

Multi-Modal Hybrid Retrieval. Our method is specifically designed for multi-modal retrieval. It takes into account the issue of insufficient granularity in visual retrieval and the inability of text retrieval to capture visual information. To date, current work in this field has not provided corresponding solutions to these problems.

The existing work typically relies solely on either text or visual features, and is unable to capture features from both modalities. Additionally, the length of the context needs to be manually adjusted and cannot be automatically determined according to the query.

Our Multi-Modal Hybrid Retrieval incorporates both textual and visual features, dynamically adjusting retrieval results based on the similarity distribution between the query and the document collection. This mechanism ensures that only the most relevant documents are retrieved, reducing noise and improving generation efficiency. This is a significant improvement compared to static top-K retrieval methods that utilized a single modality.

Multi-Agent Generation with Iterative Reasoning. Our method offers an effective solution for the model's visual perception, defining the agent's action space based on visual features. This includes visual scaling up and down, as well as Coarse-to-Fine reasoning, which is the most significant difference compared to existing works.

The existing multi-agent methods are limited to text modality, and those actor-critic-based multi-agent frameworks mainly focus on exploring the boundaries of knowledge of models and reducing noise interference.

Simply placing images into the context like text does not fully exploit the reasoning capabilities of VLMs. The multi-agent approach for text cannot truly address the key challenges at the multimodal QA task. Our multi-agent framework is a novel multimodal generation framework that defines agents based on a visually specified action space, including visual scaling up and down. Our framework emphasizes visual Coarse-to-Fine reasoning, fully leveraging the reasoning capabilities of current VLM models with limited context length.

H More Details about Multi-Agent Generation with Iterative Reasoning

We designed prompts to drive VLMs-based agents, and through our experiments, we found that some open-source models require the design of few-shot examples to learn specific thought patterns. See detailed prompts in Figure 13, Figure 14 and Figure 15.

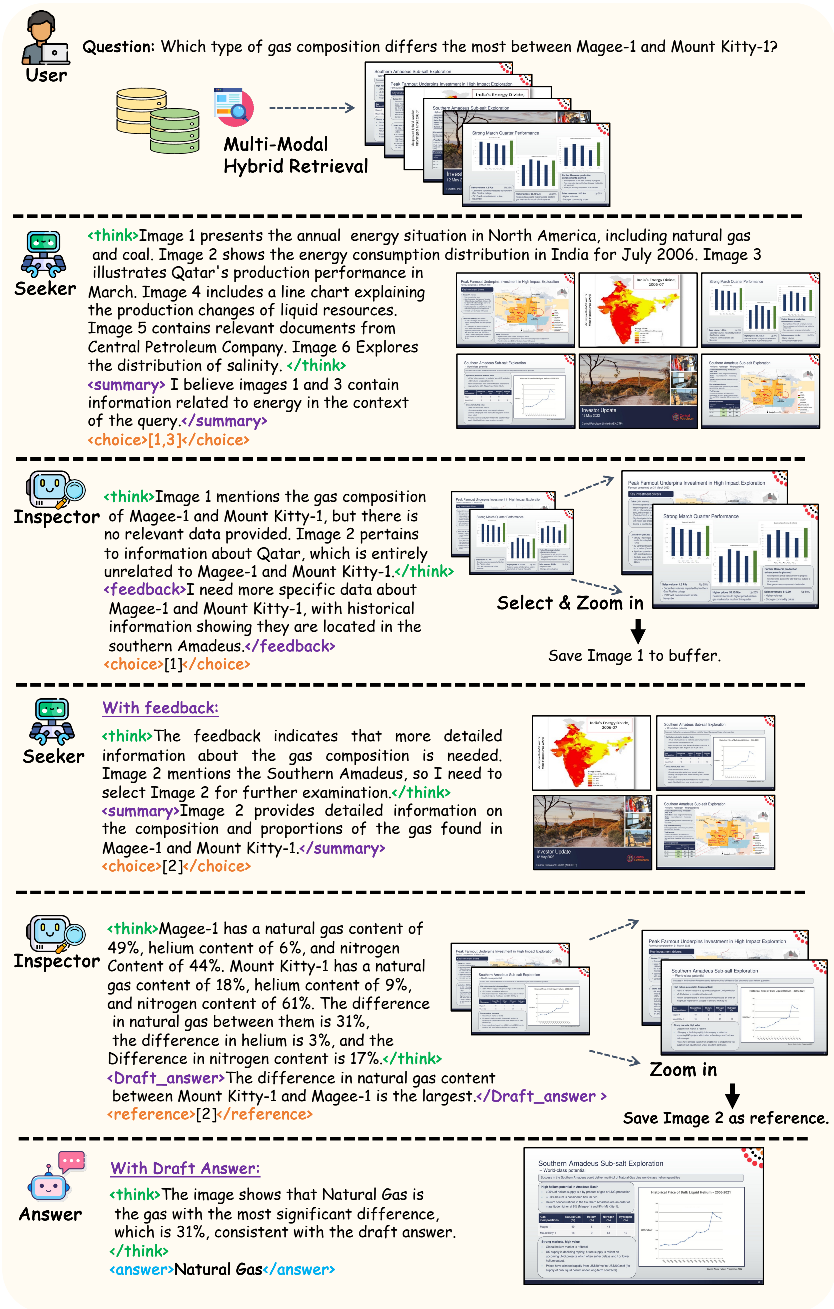

Figure 9: **Case of ViDoRAG.**

Query Reviewer Prompt.

System Prompt:

Task

I have some QA data here, and you can observe that the questions can be divided into two categories:
The category #A: When you see this question alone without a given document, you are sure to find a unique document in a corpus to provide a unique answer. The question having some key words to help you locate the document from corpus.
The category #B: When you see this question alone without a given document, you will find hard to locate a document to give a deterministic answer for this question, because you will find multiple candidate documents in a corpus, which may lead to different answers for this question. The question do not have any special key words to help you locate the document from corpus.

Examples

The number mentioned on the right of the leftside margin? #B
What is the date mentioned in the second table? #B
What is the full form of PUF? #A
What is the number at the bottom of the page, in bold? #B
Who presented the results on cabin air quality study in commercial aircraft? #A
What is the name of the corporation? #B
Which part of Virginia is this letter sent from? #B
who were bothered by cigarette odors? #A
which cigarette would be better if offered on a thicker cigarette? #A
Cigarettes will be produced and submitted to O/C Panel for what purpose? #A
What is the heading of first table? #B
What is RIP-6 value for KOOL KS? #A
Which test is used to evaluate ART menthol levels that has been shipped? #A
How much percent had not noticed any difference in the odor of VSSS? #A
what mm Marlboro Menthol were subjectively smoked by the Richmond Panel? #A
What are the steps of Weft Preparation between Spinning bobbin and Weaving? #A
What level comes between Middle Managers and Non-managerial Employees? #A
What are the six parts of COLLABORATION MODEL of the organization where James has a role of leading the UK digital strategy? #A

User Prompt:

Query: **{Query Description}**

Figure 10: Prompt of Query Reviewer.

Multi-Modal Reviewer Prompt.

System Prompt:

Please check the image, tell me whether the image can answer my question.

User Prompt:

Query: **{Query Description}**
Image: **{Relevant Image}**

Figure 11: Prompt of Multi-Modal Reviewer.

Multi-Modal Query Refiner Prompt.

System Prompt:

Task

Rewrite the following question so that it contains specific keywords that clearly point to the provided document, ensuring that it would likely match this document alone within a larger corpus.

Instruction

- Do not add any additional information or context to the question.
- You should not change the meaning of the question.
- If the question is already specific and unique, you may leave it unchanged.
- Please make the sentences you have rewritten more diverse and fluent.

Examples

- Original question: GIS data integration is part of which process?
- Rewritten question: Citizen Science shows which process the GIS data integration is part of?

- Original question: What percentage of apps ranked in the top five for including what resulted in a 10,3% Ranking Increase?
- Rewritten question: According to the App Store Optimization what percentage of apps ranked in the top five for including what resulted in a 10,3% Ranking Increase?

- Original question: Who is the author of the book, the title of which is the same as the section title of the presentation?
- Rewritten question: Who is the author of the book, the title of which is the same as the section title of the presentation by Michael Sahota and Olaf Lewitz?

- Original question: Which region of the world accounts for the highest percentage of revenues in the year 12% GROWTH is achieved?
- Rewritten question: Which region of the world accounts for the highest percentage of revenues in the year 12% GROWTH is achieved?

- Original question: What directly follows "conduct market research to refine" in the figure?
- Rewritten question: What directly follows "conduct market research to refine" in the figure within the Social Velocity Strategic Plan Process?

- Original question: How can the company which details 24 countries in the report be contacted?
- Rewritten question: How can the company which details 24 countries in the Global Digital Statistics 2014 report, be contacted?

- Original question: What substances are involved in the feeding of substrates?
- Rewritten question: What substances are involved in the feeding of substrates during the production of penicillin?

User Prompt:

Query: **{Query Description}**
Document: **{Document Description}**
Image: **{Image File}**

Figure 12: Prompt of Multi-Modal Refiner.

Seeker Agent Prompt.

System Prompt:

Character Introduction

You are an artificial intelligence assistant with strong ability to find references to problems through images. The images are numbered in order, starting from zero and numbered as 0, 1, 2 ... Now please tell me what information you can get from all the images first, then help me choose the number of the best picture that can answer the question.

Response Format

The number of the image is starting from zero, and counting from left to right and top to bottom, and you should response with the image number in the following format:

```
{
    "reason": Evaluate the relevance of the image to the question step by step,
    "summary": Extract the information related to the problem,
    "choice": List[int]
}
```

Response Example # open-source models sometimes need few-shot instructions.

```
Example 1: Question: Who is the person playing a musical instrument in restaurant?
Response to Example 1:
{
    "reason": "Image 0 shows that KFC on Renmin Road has a birthday party on February 3rd. I can
    know that there are musical instruments playing in Shanghai hotels during meals from Image 1.
    Image 2 shows that this is an invitation letter for the music performance of the New Year's
    Concert at Qintai Art Museum on December 31st. The question is related to the restaurant, and
    Image 2 is not relevant to the question.",
    "summary": "KFC on Renmin Road has a birthday party on February 3rd;Shanghai hotels have
    musical instruments playing during meals;The Qintai Art Museum will hold a New Year's concert
    on December 31st.",
    "choice": [0, 1]
}

Example 2: Question: What time is the train departing from hangzhou to beijing?
Response to Example 2:
{
    "reason": "Image 0 shows that Beijing has a temperature of 18 degrees Celsius. Image 0 is a
    train ticket from hangzhou to beijing showing a departure time of 14:30. Image 1 is a photo of
    a train station clock, but it's blurry and hard to read the exact time. Image 2 shows a train
    schedule with multiple departure times listed. Image 3 is the timetable of Hangzhou Xiaoshan
    International Airport, and this image is not related to the issue. I think Image 0 is the most
    relevant to the question.",
    "summary": "The train ticket shows a departure time of 14:30;The train station clock is
    blurry;Train schedule shows time.",
    "choice": [0]
}

Example 3: Question: Where can I find a bookstore that sells rare books?
Response to Example 3:
{
    "reason": "Image 0 is a street view of a shopping mall with various stores, but no bookstores
    are visible. Image 1 shows a sign for a bookstore called "Rare Finds Bookstore" specializing
    in rare books. Image 2 is a map with multiple bookstores marked, but it doesn't specify if
    they sell rare books. Image 3 is a photo of a library, which is not a place to buy books.
    Image 5 is a rare books list, which includes the names and prices of various books. ",
    "summary": "The shopping mall has no visible bookstores;Rare Finds Bookstore specializes in
    rare books;Map shows multiple bookstores but doesn't specify rarity;Library is not for buying
    books;The price list includes the prices and names of rare books.",
    "choice": [1, 5]
}
```

User Prompt:

Query: **{Query Description}**
Images: **{Candidate Images}**
Reflection: **{Feedback From Inspector}**

Figure 13: Prompt of Seeker Agent.

Inspector Agent Prompt.

System Prompt:

Character Introduction

You are an artificial intelligence assistant with strong ability to answer questions through images. Please provide the answer to the question based on the information provided.

Task Description

- If the images can answer the question, please answer the question directly.
- If the images are not enough to answer the question, please tell me which pictures are related to the question.

Response Format

- If the images can answer the question, please answer the question directly:

```
{
    "reason": Solve the question step by step,
    "answer": Answer the question briefly with several words,
    "reference": List[int]
}
```

- If the images are not enough to answer the question, please tell me what additional information you need, and tell me which pictures are related to the question:

```
{
    "reason": Evaluate the relevance of the image to the question one by one, and solve the
    question step by step,
    "information": Carefully clarify the information required,
    "choice": List[int]
}
```

Response Example # open-source models sometimes need few-shot instructions.

```
- Example 1:
{
    "reason": "The image only provides information about the Bohr Model and does not include
    details about subshells in the Modern Quantum Cloud Model.",
    "information": "More information about the Bohr Model.",
    "choice": []
}

- Example 2:
{
    "reason": "The images provide information about the #swallowaware campaign, including its aims
    and how they were measured. However, specific details on the success metrics are not clearly
    visible in the provided images.",
    "information": "More information about the success metrics of the #swallowaware campaign.",
    "choice": [0, 1]
}

- Example 3:
{
    "reason": "We first found the restaurant name on the menu, and then we located the restaurant
    in the city center on the map.",
    "answer": "city center",
    "reference": [2, 3]
}

- Example 4:
{
    "reason": "The entire process, from input, processing to output, ultimately produces a product
    with a purity of 42%.",
    "answer": "42%",
    "reference": [0]
}
```

User Prompt:

Query: **{Query Description}**

Plan: **{Thought From Last Step.}**

Images: **{Images Pending Review.}**

Figure 14: Prompt of Inspector Agent.

Answer Agent Prompt.

System Prompt:

Character Introduction

You are an artificial intelligence assistant with strong ability to answer questions through images. Please provide the answer to the question based on the information provided and tell me which pictures are your references.

Response Format

Please provide the answer in JSON format:

```
{
    "reason": Solve the question step by step,
    "answer": Answer the question briefly with several words,
    "reference": List[int]
}
```

User Prompt:

Query: **{Query Description}**

Draft Answer: **{Draft Answer From Inspector}**

Images: **{Reference Images}**

Figure 15: Prompt of Answer Agent.

Model-based Evaluation Prompt.

System Prompt:

Task

You are an expert evaluation system for a question answering chatbot, and you are given the following information:

- a user query,
- a generated answer,
- and a reference answer to use for reference in your evaluation.

Your job is to judge the relevance and correctness of the generated answer.

Output a single score that represents a holistic evaluation.

You must return your response in a line with only the score.

Do not return answers in any other format.

On a separate line provide your reasoning for the score as well.

Instruction

Follow these guidelines for scoring: - Your score has to be between 1 and 5, where 1 is the worst and 5 is the best.

- If generated answer is not relevant to the user query, you should give a score of 1.
- If generated answer is relevant but contains mistakes, you should give a score between 2 and 3.
- If generated answer is relevant and fully correct, you should give a score between 4 and 5.

Response Example

4.0

The generated answer has the exact same metrics as the reference answer, but it is not as concise.

User Prompt:

Query: **{Query Description}**

Reference Answer: **{Reference Answer}**

Generated Answer: **{Model's Final Answer}**

Figure 16: Prompt of Model-based Evaluation.