mRAG: Elucidating the Design Space of Multi-modal Retrieval-Augmented Generation

Anonymous EMNLP submission

Abstract

Large Vision-Language Models (LVLMs) have made remarkable strides in multimodal tasks such as visual question answering, visual grounding, and complex reasoning. However, they remain limited by static training data, susceptibility to hallucinations, and inability to verify claims against up-to-date, external evidence, compromising their performance in dynamic real-world applications. Retrieval-Augmented Generation (RAG) offers a practical solution to mitigate these challenges by allowing the LVLMs to access large-scale knowledge databases via retrieval mechanisms, thereby grounding model outputs in factual, contextually relevant information. Here in this paper, we conduct the first systematic dissection of the multimodal RAG pipeline for LVLMs, explicitly investigating (1) the retrieval phase: on the modality configurations and retrieval strategies, (2) the re-ranking stage: on strategies to mitigate positional biases and improve the relevance of retrieved evidence, and (3) the generation phase: we further investigate how to best integrate retrieved candidates into the final generation process. Finally, we extend to explore a unified agentic framework that integrates re-ranking and generation through self-reflection, enabling LVLMs to select relevant evidence and suppress irrelevant context dynamically. Our full-stack exploration of RAG for LVLMs yields substantial insights, resulting in an average performance boost of 5% without any fine-tuning.

1 Introduction

Recent advancements in Large Vision-Language Models (LVLMs) have significantly enhanced their capabilities in processing and generating multimodal content, substantially benefiting real-world applications such as visual question answering (VQA) (Zhang et al., 2024a; Sinha et al., 2024; Bai et al., 2023; Chen et al., 2024c), visual grounding (Wang et al., 2025; Xu et al., 2024; Yang et al., 2023), complex task planning (Yang et al., 2024; Zhaxizhuoma et al., 2024; Li et al., 2023), and physical reasoning (Chen et al., 2024a; Zhou et al., 2025; Chow et al., 2025; Gao et al., 2024). Despite these remarkable strides, however, LVLMs inherently suffer from several fundamental limitations, primarily stemming from their reliance on static, frozen training data (Abootorabi et al., 2025; Mei et al., 2025), insufficient semantic grounding capabilities (Liao et al., 2024; Wan et al., 2024), and inadequate alignment across modalities (Alonso et al., 2025; Zhu et al., 2024). Specifically, these limitations lead to practical challenges, such as prone to producing factual hallucinations—outputs that appear plausible but are factually incorrect (Favero et al., 2024; Rawte et al., 2025; Sahoo et al., 2024; Bai et al., 2024), struggling with outdated knowledge for time-sensitive question answering (Siyue et al., 2024; Wu et al., 2024; Uddin et al., 2024), and lacking robust verification mechanisms to validate claims using external evidence (Prabhu et al., 2024; Sahu et al., 2024; Cekinel et al., 2025).

Retrieval-Augmented Generation (RAG) offers a promising and practical solution to mitigate these limitations by equipping LVLMs to access, retrieve, and integrate external, up-to-date knowledge sources(Lewis et al., 2020; Gao et al., 2023; Chen et al., 2024b). Specifically, RAG incorporates retrieval mechanisms to fetch contextually relevant information from large-scale knowledge bases, significantly reducing the likelihood of factual hallucinations and enhancing the accuracy of generated outputs. Recently, multi-modal extensions of RAG (referred to as **mRAG**) have emerged, integrating textual, visual, and other modalities into the retrieval-generation pipeline, substantially expanding the versatility of LVLMs across many domains. For instance, multimodal RAG has successfully enabled evidence-based medical diagnostics (Xia et al., 2024a,b), decision-making in autonomous driving (Yuan et al., 2024), and industry applications (Riedler and Langer, 2024).

Prior Work. Despite this rapid progress, existing research in multimodal RAG remains fragmented and lacks a comprehensive exploration of its full design space. Firstly, there exists limited

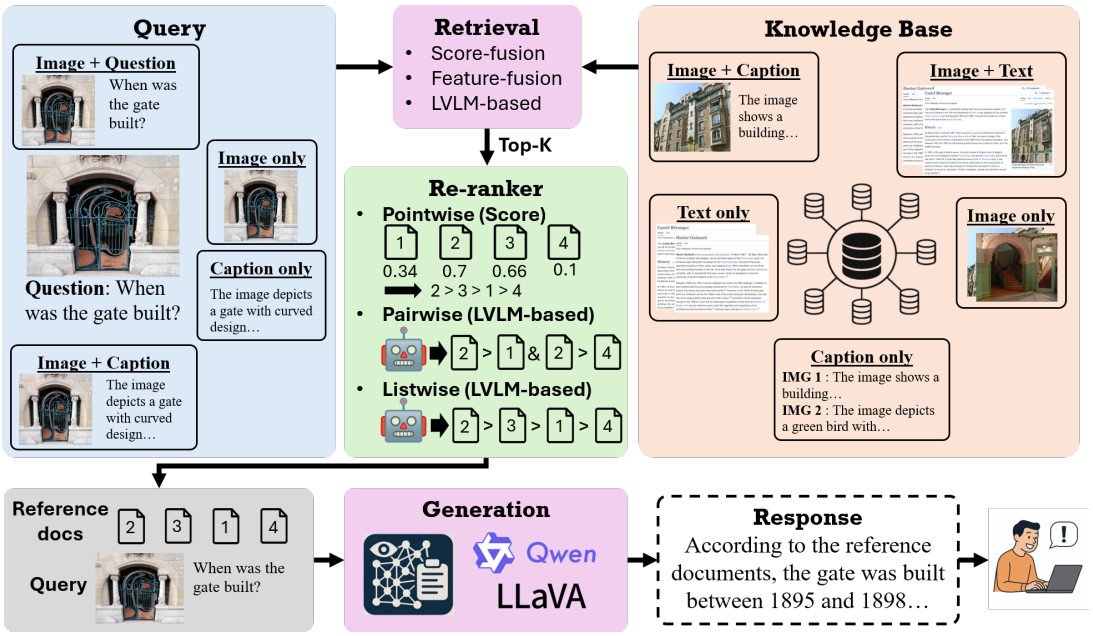

Figure 1: The multi-modal RAG (mRAG) pipeline utilized in our journey to exploit the design space of each component thereof: ❶ **Retrieval** (§3), ❷ **Re-ranking** (§4), and ❸ **Generation** (§5).

empirical validation of how multi-modal alignment strategies impact retrieval effectiveness in mRAG workflows. While some studies propose fusion techniques combining visual and textual embeddings (Wei et al., 2024; Lu et al., 2024), they do not analyze how these methods perform across different combinations of modalities. Second, re-ranking approaches in existing mRAG frameworks predominantly rely on straightforward relevance scoring mechanisms, assigning absolute scores based on query-candidate similarity (Mortaheb et al., 2025). Alternative ranking strategies, such as pairwise and listwise methods (Gangi Reddy et al., 2024; Qin et al., 2023; Ren et al., 2025; Zhuang et al., 2024), have remained underexplored in multimodal contexts. Lastly, current mRAG frameworks typically isolate the retrieval, re-ranking, and generation phases, resulting in suboptimal coordination between evidence selection and answer generation.

Our Work. To this end, we present a systematic study that elucidates the comprehensive design space of mRAG for LVLMs, methodically dissecting each critical phase of the mRAG pipeline. We start with a baseline design as shown in Fig. 1, and perform detailed investigations into: ❶ the **retrieval** phase, analyzing multiple modality configurations and retrieval strategies; ❷ the **re-ranking** phase, evaluating different approaches aimed at mitigating positional biases and enhancing evidence relevant; and ❸ the **generation** phase, exploring optimal methods for integrating retrieved candidates into the final model outputs. Through this structured exploration, we identify crucial insights and best practices across each phase, ultimately converging on an **optimized mRAG pipeline** that involves the following recipe:

> **Recipe:** integration of **(1)** EVA-CLIP as retriever, **(2)** listwise LVLM-based re-ranking, and **(3)** only providing most relevant document for generation yields a +2.32%/+0.65% response accuracy increase on benchmark datasets including E-VQA and InfoSeek.

Finally, building on our best mRAG pipeline, we present an initial exploration into a unified **Agentic mRAG** framework by incorporating a self-reflection mechanism. We systematically compare multiple strategies, utilizing powerful LVLM-based re-rankers to enhance candidate ordering. Additionally, we examine how retrieval quality impacts answer accuracy, specifically highlighting how irrelevant candidates degrade performance even when correct information is present. Motivated by these insights, we propose a unified agentic framework that integrates re-ranking and generation via iterative self-reflection. This unified approach enables LVLMs to dynamically assess candidate relevance, selectively leveraging beneficial context while suppressing irrelevant information.

2 Preliminaries

Before exploring best practices for mRAG, we provide an overview of the general dataset setup and the evaluation metrics.

2.1 Dataset Constructions

Original Dataset. Following prior studies (Caffagni et al., 2024; Yan and Xie, 2024), we adopt VQA as the task of our study. We chose two knowledge-based VQA datasets.

- **Encyclopedic-VQA (E-VQA)** (Mensink et al., 2023) comprises 221k unique visual question-answer pairs. These images are sourced from iNaturalist 2021 (Van Horn et al., 2021) and Google Landmarks Dataset V2 (Weyand et al., 2020). The visual questions emphasize fine-grained category distinctions and instance-level recognition, requiring alignment between visual content and structured knowledge. A knowledge base of 2M Wikipedia articles is provided.
- **InfoSeek** (Chen et al., 2023) is designed to evaluate models on knowledge-intensive, information-seeking questions that cannot be answered using only visual content or common sense. It contains 1.3M curated image-question-answers corresponding to 100K Wikipedia articles.

Table 1: Statistics of the distilled dataset.

	InfoSeek		E-VQA	
	Original	Distilled	Original	Distilled
#articles	100K	50K	2M	50K
#images	371K	184K	6.6M	171K

Distilled Dataset. The scale of knowledge bases in E-VQA and InfoSeek introduces significant computational demands when employing LVLMs as retrievers, particularly for vector search operations, requiring full knowledge base encoding before retrieval phase. As a result, we distill a 50k-article subset through sampling. This process ensures that all evaluation queries remain answerable within the reduced knowledge base. Additionally, we sample articles in a manner that preserves the original category distribution. The statistics of the distilled datasets are presented in Table 1. For evaluation, in line with previous research (Yan and Xie, 2024), we use 4,750 test cases for E-VQA, and 5,000 cases for InfoSeek. More details are in Appendix A.1.

2.2 Evaluation Metrics

Retrieval. We assess retrieval performance using Recall@K and Mean Reciprocal Rank (MRR). Recall@K measures the percentage of the correct article found in the top-K retrieved candidates across the evaluation queries. MRR calculates the reciprocal of the rank at which the first correct article is retrieved. It provides a clear measure of how quickly a relevant article is found to ensure LVLMs receive critical contextual information early after the re-ranking phase.

Visual Question Answering. We evaluate VQA performance through complementary metrics addressing both lexical and semantic accuracy. ROUGE-L is used to compare the model response with reference answers. However, this metric may not fully capture answers that are phrased differently but convey the same meaning. To capture semantic correctness, we employ InternVL3 (Chen et al., 2024c; Zhu et al., 2025) and GPT-4.1 (Achiam et al., 2023) as automated judges to assess if LVLM's answer is semantically correct, providing a more comprehensive evaluation of VQA performance. All evaluations are performed using a temperature setting of 0.

3 Retrieval Configurations and Strategies

The retrieval phase in mRAG requires careful consideration of input modalities and their alignment with candidates in the knowledge base because the complexity of processing heterogeneous data types, such as images and text, exhibits distinct semantic structures and embedding space distributions. This section systematically examines how different modality pairings between user queries and candidates in the knowledge base impact retrieval performance in mRAG. Unlike traditional approaches that rely on fine-tuned models adapted to specific knowledge bases, our investigation focuses on zero-shot retrieval capabilities using frozen pre-trained models. Furthermore, we evaluate multi-modal configurations where queries and candidates in the knowledge base may independently combine text, image, or hybrid modalities, testing models' inherent abilities to establish semantic alignment across modalities without parameter updates. Once the query and candidates' embeddings are generated, we employ the FAISS library (Douze et al., 2024) with dot product similarity for retrieving top-K candidates and their corresponding wiki articles.

3.1 Modality Configurations

Retrieval effectiveness depends on how information is encoded in queries and candidates in the knowledge base. In this study, we investigate five modality configurations to systematically evaluate multimodal retrieval. The **image-only** (I) configuration relies solely on images and is applicable to both query and candidate sides. The **image + question** (IQ) setting, which combines image with the user's question to enable joint vision-language reasoning, is used solely on the query side, as questions are inherently tied to the query image and not present in the knowledge base. The **image + text** (IT) configuration fuses images with associated textual information, such as article passages, and is only applied to knowledge base candidates. The **image + caption** (IC) setup augments images with

Table 2: Results on foundational stage. Following (Wei et al., 2024; Lin et al., 2024), we report Recall@5 for both datasets. The $I \leftrightarrow IT$ modality configuration achieves peak Recall@5 scores (underlined) for both datasets using EVA-CLIP$_{SF}$, outperforming all other strategies including MLLM-based methods.

Dataset	Task (Query ↔ KB)	Retrieval Strategy					
		CLIP$_{SF}$	EVA-CLIP$_{SF}$	BGE-CLIP$_{SF}$	BLIP$_{FF}$	BGE-MLLM	GME
InfoSeek	$I \leftrightarrow I$	67.5	77.84	49.92	57.2	39	56.52
	$I \leftrightarrow IT$	73.6	<u>81.58</u>	41.02	64.22	11.36	62.32
	$IQ \leftrightarrow I$	67.5	77.8	13.36	42.1	18.78	74.94
	$IQ \leftrightarrow IT$	27.2	76.94	0.7	33.92	10.72	81.48
E-VQA	$I \leftrightarrow I$	62.8	75.9	46.46	54.3	33.8	53.6
	$I \leftrightarrow IT$	72.29	<u>80.69</u>	35.28	59.81	13.37	50.84
	$IQ \leftrightarrow I$	63.3	76.75	11.34	40.44	15.11	61.93
	$IQ \leftrightarrow IT$	31.2	77.2	6.8	38.32	21.35	77.03

Table 3: Results on expansion stage using EVA-CLIP. Underlined scores are the best Recall@5 based on raw data, shown in Table 2.

Dataset	Task (Query ↔ KB)	Recall@K		
		K=1	K=5	K=10
InfoSeek	$I \leftrightarrow I$	57.7	77.84	82.08
	$I \leftrightarrow IT$	<u>63.42</u>	<u>81.58</u>	<u>85.22</u>
	$I \leftrightarrow IC$	56.32	77.3	81.39
	$I \leftrightarrow C$	31.1	52.16	58.9
	$IC \leftrightarrow I$	56.12	76.63	81.21
	$IC \leftrightarrow IC$	53.02	74.1	79.23
	$IC \leftrightarrow C$	23.9	42.23	49.74
	$IC \leftrightarrow IT$	**64.44**	**82.43**	**85.32**
	$C \leftrightarrow I$	21.69	38.97	46.3
	$C \leftrightarrow IC$	21.75	38.91	46.56
	$C \leftrightarrow C$	17.77	32.27	39.01
	$C \leftrightarrow IT$	26.71	42.4	48.78
E-VQA	$I \leftrightarrow I$	54.5	75.9	81
	$I \leftrightarrow IT$	<u>61.85</u>	<u>80.69</u>	<u>85.51</u>
	$I \leftrightarrow IC$	54.32	75.12	80.63
	$I \leftrightarrow C$	23.77	41.18	48.88
	$IC \leftrightarrow I$	53.26	75.71	80.36
	$IC \leftrightarrow IC$	49.26	70.99	78.04
	$IC \leftrightarrow C$	20.42	34.82	42.59
	$IC \leftrightarrow IT$	**62.61**	**80.8**	**86.47**
	$C \leftrightarrow I$	17.81	30.86	38.46
	$C \leftrightarrow IC$	18.84	34.8	41.62
	$C \leftrightarrow C$	12.61	25.89	33.05
	$C \leftrightarrow IT$	19.2	33.81	42

generated captions from a caption model, providing explicit semantic cues to complement the image; this configuration is applicable to both queries and candidates. Finally, the **caption-only** (C) configuration uses generated captions alone, and can also be applied to either side. We do not consider a question-only configuration, as questions are always grounded in the corresponding query image.

3.2 Retrieval Strategies

Score Fusion. This approach involves combining scores derived from different modalities. For instance, $CLIP_{SF}$ in (Wei et al., 2024) employs a dual encoder where visual and textual modalities are processed independently through separate unimodal encoders, producing two distinct embedding vectors of the same dimensionality. The fusion mechanism operates by computing a weighted linear combination of these unimodal embeddings to produce a unified representation vector. Formally, given a visual encoder ϕ_{vis} and a text encoder ϕ_{txt}, the fusion score S is shown as $S = \phi_{vis}(I) + \phi_{txt}(\tau)$, where $\tau \in \{C, T, Q\}$ depending on the configuration of the modality.

Feature Fusion. Unlike score fusion, which combines modality-specific similarities post-hoc, feature fusion integrates multimodal features during the encoding phase. This fusion approach generates a single feature representation for multi-modal queries or candidates by applying mixed-modality layers. However, this approach typically requires fine-tuning the fusion layers on the target knowledge base, while our work prioritizes evaluating frozen models' inherent ability to bridge modality gaps using their pretrained representations. Therefore, we utilize a pretrained $BLIP_{FF}$ model from (Wei et al., 2024). $BLIP_{FF}$ is trained on diverse datasets and is capable of retrieving heterogeneous outputs in both text and image modalities.

LVLM-based Retriever. Modern LVLMs integrate visual encoders (typically vision transformers) with pretrained language models, enabling them to natively process image-text token sequences. This ability makes them particularly suited for zero-shot retrieval scenarios where frozen pretrained parameters must bridge modality gaps and map diverse modalities into a unified token space. Our evaluation focuses on benchmarking these retrieval models' (Lin et al., 2024; Jiang et al., 2024; Zhou et al., 2024; Zhang et al., 2024b) zero-shot retrieval performance across modality configurations, testing their ability to align queries and candidates without task-specific fine-tuning.

In this study, we evaluated the retrieval performance of six distinct approaches across three types above, including three score fusion methods: CLIP$_{SF}$ (Wei et al., 2024), EVA-CLIP$_{SF}$ (Sun et al., 2023, 2024), and BGE-CLIP$_{SF}$ (Zhou et al., 2024) (For score fusion, image and text embeddings are assigned equal weight), one feature

fusion BLIP$_{FF}$ (Wei et al., 2024), and two LVLM-based retrievers, BGE-MLLM (Zhou et al., 2024) and GME (Zhang et al., 2024b).

3.3 Results

Our evaluation follows a two-stage design to isolate and quantify the impact of different modality combinations on retrieval performance. The **foundational stage** establishes baseline performance by evaluating three core configurations: I, IT, and IQ, all without caption augmentation. This stage identifies which retrieval strategies perform the best on raw inputs. The **expansion stage** then introduces caption-augmented configurations: IC and C, to test whether automatically generated captions enhance retrieval robustness. This sequential experiment ensures any observed improvements that is directly attributed to image caption rather than variance in the foundational stage. In this work, Qwen2-VL-2B-Instruct (Bai et al., 2023; Wang et al., 2024) is employed to generate image captions. The prompt is shown in Appendix B.1.

From Table 2, the $I \leftrightarrow IT$ configuration demonstrates superior performance across both datasets when paired with EVA-CLIP$_{SF}$ (Sun et al., 2023, 2024). This suggests EVA-CLIP's pretrained vision-language alignment excels at bridging pure image queries with image + text candidates in the knowledge base, and thus motivates our selection of EVA-CLIP as the default retriever in this study.

In the expansion stage from Table 3, we observe that augmenting image queries with generated caption (IC) yields modest improvements over raw image queries (I). This shows that image captions provide complementary semantic signals that enhance the query. However, applying captions to both query and candidates, $IC \leftrightarrow IC$, degrades performance drastically even compared to $I \leftrightarrow I$. With the low retrieval accuracy of $C \leftrightarrow C$, this is likely due to caption discrepancies between query and candidates that amplify visual differences.

> **Takeaway:** It is clear that a large-scale CLIP model (in this study, EVA-CLIP) is a robust zero-shot retriever. Furthermore, augmenting image **on the query side** with generated captions improves Recall@1 accuracy by 1% over image-only.

4 Re-ranking

While relevant candidates may appear in the top-K retrieval results, modern LVLMs still exhibit a positional attention bias, called the "lost-in-the-middle" effect (Liu et al., 2024). This effect persists even when correct articles are retrieved but present in the middle, as LVLMs disproportionately focus on candidates in the beginning during answer generation. To mitigate this issue, re-ranking aims at pushing the most relevant candidate to the beginning, aligning with LVLMs' inherent attention patterns. This step is critical because LVLMs' performance on knowledge-intensive tasks degrades sharply when key information appears later in the input sequence.

4.1 Experimental Setup

Following prior work (Liu et al., 2025), we evaluate three re-ranking approaches. **Pointwise Ranking** computes absolute relevance scores for individual query-candidate pairs and sorts the scores. **Pairwise Ranking** compares candidate pairs through relative preference judgments, asking models to select the more relevant option for each query. **Listwise Ranking** operates on full candidate lists, requiring models to holistically assess and reorder all retrieved items simultaneously.

In this section, the query modality is fixed to image + question (IQ) and the candidate modality is image + text (IT). We utilize MM-Embed (Lin et al., 2024), Q-Former from EchoSight (Yan and Xie, 2024), and Qwen2-VL-7B-Instruct (Bai et al., 2023; Wang et al., 2024) as the re-ranker, and take the top-5 retrieval candidates from Section 3 to measure Recall@1 and MRR improvements after re-ranking. Note that the re-ranker from EchoSight is specifically fine-tuned on InfoSeek and E-VQA, while MM-Embed and Qwen2-VL-7B-Instruct are zero-shot re-rankers. This controlled setup isolates the effectiveness of re-ranking from initial retrieval quality. For pointwise ranking, we compute absolute relevance scores by extracting the last-layer embeddings of both queries and candidates from MM-Embed and EchoSight, and then calculating the dot product similarity score. For pairwise and listwise ranking, we prompt Qwen2-VL-7B-Instruct to re-rank. The prompt template is shown in Appendix B.2.

4.2 Results

Table 4 presents a comparison of re-ranking strategies, focusing on their performance on Recall@1 and MRR for both the InfoSeek and E-VQA datasets. The baseline (w/o re-ranking) establishes a starting point, with Recall@1 of 64.44 and 62.61 and MRR of 0.71 and 0.694 for InfoSeek and E-VQA, respectively.

For InfoSeek, the zero-shot MM-Embed re-ranker and pairwise ranking degrade performance substantially compared to the baseline. In contrast, EchoSight's fine-tuned re-ranker achieves near-baseline Recall@1 and improves MRR by 0.011 becasue of the fine-tuning on both knowledge

Table 4: Re-ranking results across different strategies and datasets. The baseline, w/o Re-rank, is the original retrieval result with $IC \leftrightarrow IT$ from Section 3.

Dataset	Strategy	Metric	
		Recall@1 ↑	MRR ↑
InfoSeek	w/o Re-rank	64.44_{+0}	0.71_{+0}
	MM-Embed	$33.18_{-31.26}$	$0.47_{-0.24}$
	EchoSight	$64.40_{-0.04}$	$0.72_{+0.01}$
	Pairwise	$51.84_{-12.6}$	$0.63_{-0.08}$
	Listwise	$65.88_{+1.44}$	$0.73_{+0.02}$
E-VQA	w/o Re-rank	62.61_{+0}	0.69_{+0}
	MM-Embed	$43.85_{-18.76}$	$0.56_{-0.13}$
	EchoSight	$69.81_{+7.2}$	$0.74_{+0.05}$
	Pairwise	$56.34_{-6.27}$	$0.67_{-0.02}$
	Listwise	$66.42_{+3.81}$	$0.72_{+0.03}$

bases. Surprisingly, listwise ranking with QwenVL surpasses EchoSight's performance, demonstrating that LVLMs are inherently a good re-ranker.

Similarly, for E-VQA, EchoSight's fine-tuned re-ranker outperforms listwise ranking but requires training on both knowledge bases, while listwise offers competitive zero-shot performance without fine-tuning. Our experimental findings are consistent with previous work (Ma et al., 2023), indicating that while LVLMs are not effective for initial retrieval, they are good when used for re-ranking retrieved candidates in a zero-shot manner.

Takeaway: Listwise ranking with LVLMs is an effective zero-shot re-ranking strategy, may even surpass the performance of fine-tuned re-rankers and yielding an average of 2.6% improvement in Recall@1 accuracy over two datasets.

5 Generation

The generation phase synthesizes retrieved knowledge into accurate, contextually grounded answers, where retrieval quality may directly impact answer correctness. This section evaluates the generation capabilities and explores how retrieval influences the quality of the generated responses. We evaluate four conditions. **Generation without retrieval** answers the question solely without any knowledge provided, serving as the lower bound. **Generation with initial retrieval** synthesizes responses using top-K retrieved documents before re-ranking. **Generation after re-ranking** uses optimized candidate ordering to enhance answer accuracy by aligning with VLMs' positional attention biases. **Generation with gold document** takes the document containing the answer to the query as a reference to measure the upper bound accuracy.

5.1 Experimental Setup

We evaluate two state-of-the-art LVLMs: Qwen2-VL-7B-Instruct (Wang et al., 2024) and LLaVA-OneVision (Li et al., 2024) to assess how retrieval information improves answer quality. Both models operate in zero-shot mode, leveraging their pretrained multimodal understanding without task-specific fine-tuning. To assess answer quality, we compute ROUGE-L (Lin, 2004) score against reference answers. However, this traditional metric may not sufficiently capture semantic meanings. As a result, we also employ InternVL3-14B (Chen et al., 2024c; Zhu et al., 2025) and GPT-4.1 (Yu et al., 2023; Duan et al., 2024; Achiam et al., 2023) as automated judges. The judge receives the generated answer and the reference answer, then checks if the generated answer is correct (answer aligns with reference answer) or incorrect (factually wrong or irrelevant). The prompt template for the judge is shown in Appendix B.3.

5.2 Results

Figure 2 shows the response accuracy with Qwen2-VL-7B-Instruct. The results demonstrate a critical divergence between retrieval accuracy and response accuracy. While retrieval accuracy (Ret. Acc.) monotonically increases with larger K, response accuracy (Res. Acc.) does not improve and even declines. For instance, in E-VQA after re-ranking, Ret. Acc. at top-1 achieves 66.42% and increases to 80.8% at top-5, with 14.38% improvement. However, ROUGE-L score drops from 0.416 to 0.392 and Res. Acc. decreases 0.17% and 2.11% with InternVL3 and GPT 4.1 evaluation, correspondingly.

A similar trend is also observed at the bottom of Figure 2. Our experiments suggest that LVLMs indeed exhibit a strong positional bias, prioritizing information from the initial positions of the input context, so adding more documents may make LVLMs overlook key details or be confused by the irrelevant documents. Thus, re-ranking is necessary to push the most relevant to the beginning.

Takeaway: While re-ranking retrieved results boosts generation accuracy by at least 1%, adding more documents does not improve generation accuracy, even if the correct answer is present among them. Thus, providing only the most relevant document as a reference is optimal.

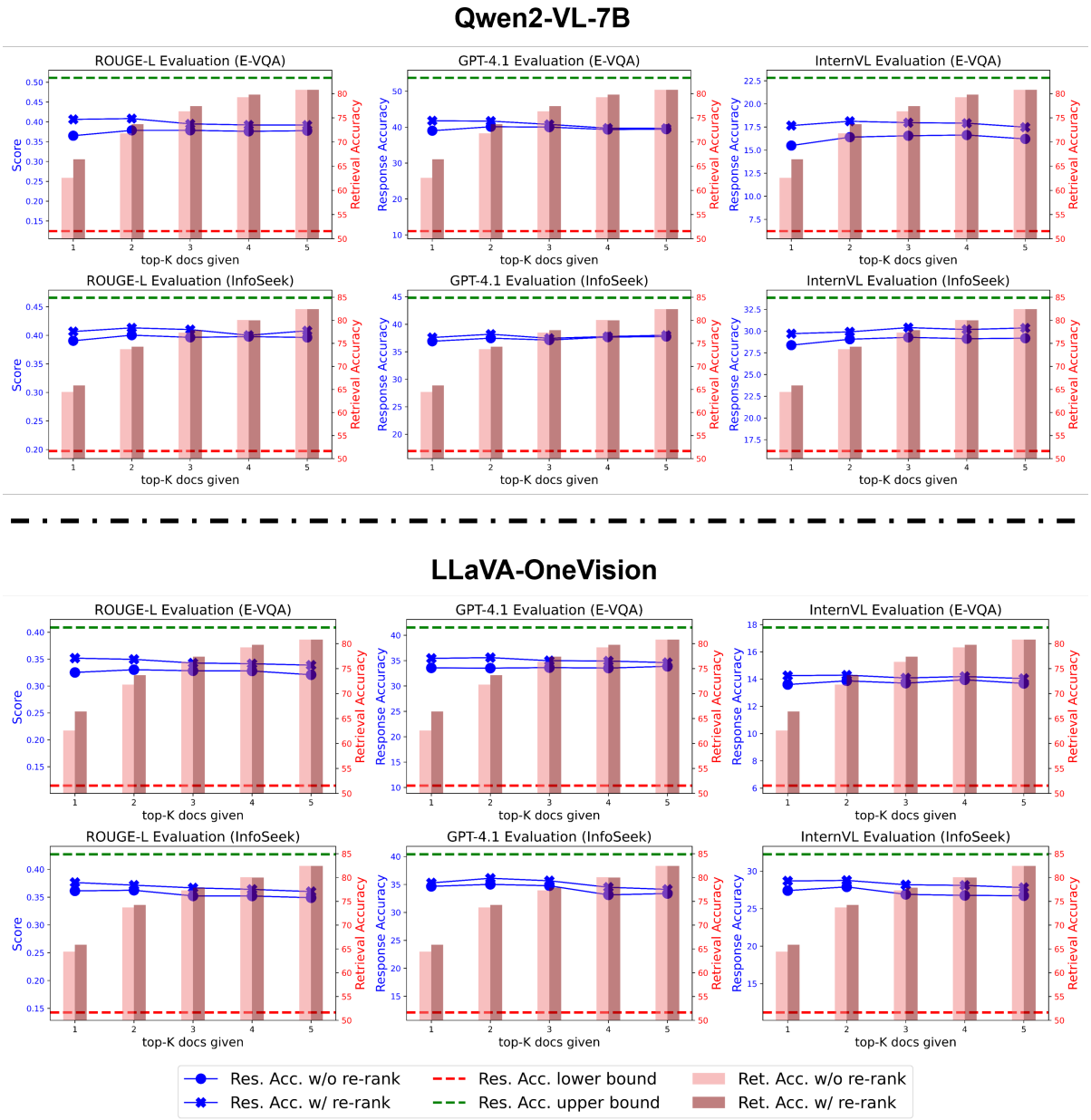

Figure 2: Generation performance of Qwen2-VL-7B-Instruct (top) and LLaVA-OneVision (bottom) across different evaluation metrics on E-VQA and InfoSeek. The line plots correspond to the left y-axis, while the bar plots correspond to the right y-axis. The green dashed line marks the performance of **Generation with gold document** and the red dashed line marks the **Generation without retrieval** of achievable response accuracy. The left y-axis, shown in blue, is the ROUGE-L score and Response Accuracy (Res. Acc.) for **Generation with initial retrieval** and **Generation after re-ranking**. The right y-axis, in red, represents the Retrieval Accuracy (Ret. Acc.) as the number of top-K retrieved documents varies.

6 Unifying Re-ranking and Generation

Our experiments in Section 5 demonstrate that adding less-relevant documents may not benefit response accuracy, motivating our attempt to explore the potential of unifying re-ranking and generation into a single agentic framework. This approach incorporates a self-reflection loop where the model evaluates both the query and retrieved documents through multiple iterations, and decides the most relevant document. Unlike prior methods like (Yu et al., 2024), which rely on instruction-tuned and text-only LLMs, we explore the possibility of LVLMs to dynamically assess query-document relevance and prioritize critical evidence without specific training.

6.1 Experimental Setup

Similar to Section 5, we employ Qwen2-VL-7B (Wang et al., 2024) and LLaVA-OneVision (Li

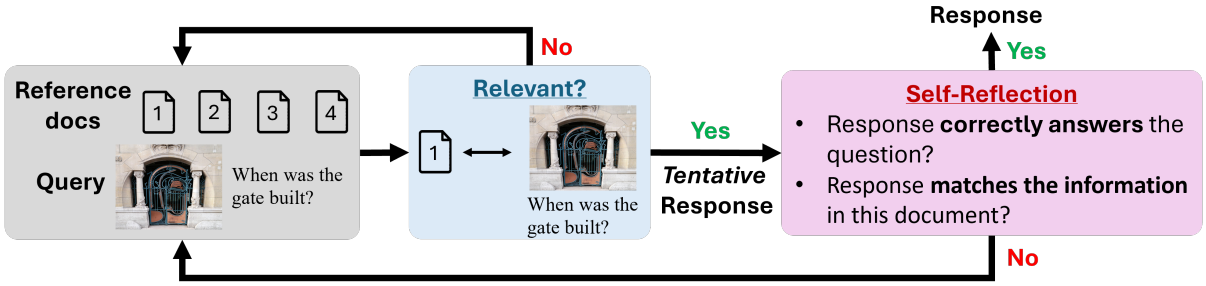

Figure 3: The self-reflection process of unifying re-ranking and generation in a single agentic framework.

Table 5: Response accuracy on the **E-VQA** dataset when unifying re-ranking and generation, compared to separate approaches, across two LVLMs.

Model	Strategy	Evaluation Method		
		ROUGE-L	GPT	InternVL
Qwen	non-Unified	0.41	41.77	17.65
	Unified	**0.43**	**45.66**	**19.49**
LLaVA	non-Unified	0.35	35.44	14.25
	Unified	**0.37**	**40.69**	**16.57**

Table 6: Response accuracy on the **InfoSeek** dataset.

Model	Strategy	Evaluation Method		
		ROUGE-L	GPT	InternVL
Qwen	non-Unified	0.41	37.6	29.7
	Unified	**0.42**	**39.5**	**31**
LLaVA	non-Unified	0.38	35.28	28.68
	Unified	**0.39**	**37.86**	**29.72**

et al., 2024) as the base LVLMs. At each iteration, the model assesses whether the current document contains evidence directly addressing the query. If a document is relevant, the model generates a tentative answer and checks its validity against the document's content via a self-reflection prompt. A valid response is returned, while an invalid response prompts the model to consider the next document in the retrieved set. If none of the documents provide relevant information, the model returns "Model fails to answer the question" and ends the process. Figure 3 depicts the pipeline of the unification of re-ranking and generation. The prompts are shown in Appendix B.4. To evaluate the performance, we take the ROUGE-L score and response accuracy with top-1 document given after re-ranking from Section 5 as the baseline, named **non-Unified**, and also employ InternVL3 and GPT 4.1 as automated judges to assess whether the response is semantically *Correct* or *Incorrect*.

6.2 Results

From Table 5 and 6, we observe that the unified agent consistently outperforms decoupled re-ranking and generation pipelines across both datasets and LVLM architectures. This demonstrates that integrating self-reflection capabilities directly into the generation process enables LVLMs to validate response relevance against retrieved documents and the query. By iteratively filtering irrelevant evidence while prioritizing critical information through document-level attention, positional bias is avoided in the decoupled approach.

Takeaway: The unified agentic framework outperforms decoupled pipelines, boosting the response accuracy by 5%/2% for E-VQA/InfoSeek through LVLM's iterative self-reflection.

7 Conclusion

In this paper, we systematically revisited the mRAG pipeline, focusing on zero-shot settings for LVLMs. Our study dissected the retrieval phase, revealing that large-scale CLIP models are highly effective as zero-shot retrievers, and that augmenting image queries with generated captions can provide modest gains. We further analyzed re-ranking strategies and found that listwise re-ranking with LVLMs offers strong zero-shot performance. Our generation experiments demonstrated that candidate ordering has a direct impact on answer accuracy, with re-ranking being essential to ensure relevant evidence is prioritized. However, we observed that adding less-relevant documents is not beneficial. As a result, we introduced an unified agentic framework that integrates re-ranking and generation via self-reflection, enabling LVLMs to dynamically filter irrelevant context and enhance answer accuracy without task-specific fine-tuning.

Limitations

Although our systematic study provides several suggestions on each phase in mRAG pipeline, several limitations should be mentioned. First, our evaluation is conducted in a zero-shot setting using frozen pre-trained models, which may not fully capture

the performance upper bound achievable with task-specific fine-tuning and the model may provide hallucainated responses. Second, the reliance on distilled datasets, while necessary for computational feasibility in this work, could introduce distributional biases that do not entirely reflect real-world scenarios with larger and more diverse knowledge bases. Third, while LVLM-based judges provide scalable evaluation, they may not perfectly align with human judgment, especially for nuanced or open-ended questions. Future work focusing on improving multi-modal alignment and developing human-centered evaluation frameworks for mRAG remains to be explored.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iñigo Alonso, Ander Salaberria, Gorka Azkune, Jeremy Barnes, and Oier Lopez de Lacalle. 2025. Vision-language models struggle to align entities across modalities. *arXiv preprint arXiv:2503.03854*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633, Abu Dhabi, UAE. Association for Computational Linguistics.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, Miami, Florida, USA. Association for Computational Linguistics.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

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. *Advances in neural information processing systems*, 33:9459–9474.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. 2023. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Can feedback enhance semantic grounding in large vision-language models? *arXiv preprint arXiv:2404.06510*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Qi Liu, Haozhe Duan, Yiqun Chen, Quanfeng Lu, Weiwei Sun, and Jiaxin Mao. 2025. Llm4ranking: An easy-to-use framework of utilizing large language models for document reranking. *arXiv preprint arXiv:2504.07439*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. 2025. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*.

Viraj Prabhu, Senthil Purushwalkam, An Yan, Caiming Xiong, and Ran Xu. 2024. Trust but verify: Programmatic vlm evaluation in the wild. *arXiv preprint arXiv:2410.13121*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Vipula Rawte, Aryan Mishra, Amit Sheth, and Amitava Das. 2025. Defining and quantifying visual hallucinations in vision-language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 501–510.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2025. Self-calibrated listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 3692–3701.

Monica Riedler and Stefan Langer. 2024. Beyond text: Optimizing rag with multimodal inputs for industrial applications. *arXiv preprint arXiv:2410.21943*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:2405.09589*.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8228–8248, Miami, Florida, USA. Association for Computational Linguistics.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. Guiding vision-language model selection for visual question-answering across tasks, domains, and knowledge types. *arXiv preprint arXiv:2409.09269*.

Zhang Siyue, Xue Yuxiang, Zhang Yiming, Wu Xiaobao, Luu Anh Tuan, and Zhao Chen. 2024. Mrag: A modular retrieval framework for time-sensitive question answering. *arXiv preprint arXiv:2412.15540*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven R Corman, and Chitta Baral. 2024. Unseentimeqa: Time-sensitive question-answering beyond llms' memorization. *arXiv preprint arXiv:2407.03525*.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. 2024. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. 2025. Learning visual grounding from generative vision and language model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8057–8067. IEEE.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. 2024. Time-sensitve retrieval-augmented generation for question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2544–2553.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. In *CoRL*.

Yibin Yan and Weidi Xie. 2024. EchoSight: Advancing visual-language models with Wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, Miami, Florida, USA. Association for Computational Linguistics.

Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2024. Guiding long-horizon task and motion planning with vision language models. *arXiv preprint arXiv:2410.02193*.

Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. 2023. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19165–19174.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in

multi-modal large language model. *arXiv preprint arXiv:2402.10828*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Zhaxizhuoma Zhaxizhuoma, Pengan Chen, Ziniu Wu, Jiawei Sun, Dong Wang, Peng Zhou, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. 2024. Alignbot: Aligning vlm-powered customized task planning with user reminders through fine-tuning for household robots. *arXiv preprint arXiv:2409.11905*.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, and Jinqiao Wang. 2025. Physvlm: Enabling visual language models to understand robotic physical reachability. *arXiv preprint arXiv:2503.08481*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

A Dataset and Knowledge Base Construction

In this section, we provide details on the dataset and knowledge base construction.

A.1 Dataset

The articles in both E-VQA and InfoSeek are in English, and are about encyclopedic knowledge derived from Wikipedia and Google Landmark datasets. In line with previous research (Yan and Xie, 2024), we use a total of 4,750 test cases for our evaluation. Our primary focus is on single-hop questions from the original E-VQA dataset (Mensink et al., 2023), ensuring that the model can directly identify the answer by referencing the retrieved documents. For the InfoSeek dataset (Chen et al., 2023), since the original release does not include a test set, we sample 5,000 test cases from its validation set to facilitate a fair and consistent evaluation.

A.2 Knowledge Base

Given the test cases, we first select articles containing the target answers from the original dataset, then sample additional articles to construct a knowledge base of 50,000 entries. As shown in Table 7, we report the approximate construction time and retriever model size for our distilled dataset. For LVLM-based retrievers (BGE-MLLM and GME), we use two NVIDIA RTX A6000 GPUs in parallel with a batch size of 3. For other retrievers, we use a single NVIDIA RTX A6000 GPU with a batch size of 4. The result demonstrates that LVLM-based retrievers require significantly longer processing time to construct a 50,000-entry knowledge base, which directly motivates our decision to distill the dataset for efficient experimentation.

B Prompt Templates

B.1 Image captioning

Figure 4 presents the prompt to caption the image **on the query side**. For the knowledge base side, as no question is provided, we caption the image with essential clues only. Figure 5 provides the prompt for captioning images in the knowledge base.

B.2 Re-ranking

This section lists the prompt templates used in the re-ranking phase. Figure 6 and 7 show the prompt for pairwise and listwise re-ranking, respectively. N is the number of documents to provide.

B.3 Generation

This section shows the prompts used during the generation phase. Figure 8 and 9 show the prompt for model generation and automated judge, respectively.

Given the image and question:
<image>
<question>
To answer the question, generate an image caption that provides the essential clues within three sentences.

Figure 4: Image captioning prompt on the query side.

Given the image:
<image>
Generate an image caption that provides the essential clues within three sentences.

Figure 5: Image captioning prompt on the knowledge base side.

Given an image and a question to the image as following:
<image>
<question>
Now given two documents below, [Document A] and [Document B]. Which of the following documents is more relevant or the answer is in the document?

[Document A]: ***<Doc A>***
[Document B]: ***<Doc B>***

The output should be exactly [Document A] or [Document B], do not include any other text.
Output [Document A] or [Document B]:

Figure 6: Pairwise re-ranking prompt.

B.4 Unifying re-ranking and generation

Figure 10 outlines the prompt for assessing document relevance to the input query. If the model decides the document is relevant and generates a response, a self-reflection prompt, shown in Figure 11, evaluates the validity of the tentative response. Valid response is kept, and an invalid one prompts the model to shift to the following document. If no document is found, the model outputs "Model fails to answer the question", terminating the process.

Table 7: Details of constructing the knowledge base.

	CLIP_{SF}	EVA-CLIP_{SF}	BGE-CLIP_{SF}	BLIP_{FF}	BGE-MLLM	GME
Model Size	0.4B	7B	0.4B	2.7B	7.57B	8.2B
Time - E-VQA (GPU hours)	2	10.1	2	4.4	39.9	40.1
Time - InfoSeek (GPU hours)	2.1	10.4	2.1	4.7	41.4	41.7

You are provided with *N* documents, each indicated by a numerical identifier [Document x], where x represents a number and should be at least 1 and at most *N*.

Rank the documents based on their relevance to the visual question answering task:

<image>
<question>

Related Documents:

[Document 1]: ***<Doc 1>***

...

[Document *N*]: ***<Doc N>***

Please carefully read all the documents and rank the *N* documents above based on their relevance to the visual question answering task above.
All the passages should be included and listed using identifiers, in descending order of relevance. The output format should be [Document x] > [Document y] > ..., e.g., [Document 4] > [Document 2] > ..., x and y should be at least 1 and DO NOT exceed *N*. Only respond with the ranking results, do not include anything else or explain.

Your ranking:

Figure 7: Listwise re-ranking prompt.

C Experimental Costs

The cost for the response evaluation using GPT 4.1 cost approximately $20 in total.

D Licenses

The datasets we used, InfoSeek and E-VQA, are licensed under Apache License 2.0 and CC BY 4.0, respectively. The retrieval models, CLIP, EVA-CLIP, BGE-CLIP, BLIP, BGE-MLLM, and GME, are under MIT License, MIT License, MIT License, MIT License, MIT License, Apache license 2.0, correspondingly. The re-ranker models, MM-Embed (Lin et al., 2024) and Qwen2-VL-7B-Instruct, are licensed under CC-BY-NC-4.0 and Apache License 2.0. Q-former from

Read the provided documents under RELATED DOCUMENTS section and answer the question according to what is provided. Do not use your own knowledge!

RELATED DOCUMENTS:
<Provided Docs>

Now given the following image and question, answer the question with the image by referring to the RELATED DOCUMENTS in a few words.

<image>

<question>

Figure 8: Generation prompt.

You are now a good judge on comparing the response from a student and several reference answers. Compare the generated RESPONSE to the REFERENCE answers. Evaluate if the generated response correctly conveys the same meaning with one of the listed REFRENCE answers, even if the wording is different.
Return one of these labels: 'Correct' or 'Incorrect'. Please directly return your label, do not include the reasoning process!!

Generated RESPONSE : ***<model responses>***

Candidate REFERENCE answers (Correct answers): ***<references>***

The evaluation is:

Figure 9: Judge prompt used for the InternVL3 and GPT 4.1.

EchoSight (Yan and Xie, 2024) was released without an accompanying license. The response generation models, Qwen2-VL-7B-Instruct and LLaVA-OneVision, are both licensed under Apache License 2.0. The InternVL3-14B model is released under MIT License.

Our use of the released datasets and models are all consistent with their intended use.

Table 8: ROUGE-L score and response accuracy with **Qwen2-VL-7B-Instruct** as the generation model in Figure 2.

Dataset	**Evaluation Methods**								
	ROUGE-L			GPT 4.1			InternVL3		
	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank
E-VQA	1	0.3651	0.4063	1	39.01	41.77	1	15.49	17.65
	2	0.3786	0.4083	2	40.11	41.68	2	16.4	18.12
	3	0.3787	0.3949	3	40	40.76	3	16.55	17.97
	4	0.3762	0.3922	4	39.34	39.71	4	16.63	17.91
	5	0.3778	0.3919	5	39.43	39.66	5	16.21	17.48
	lower bound	0.1245		lower bound	11.05		lower bound	6.23	
	upper bound	0.5111		upper bound	53.73		upper bound	22.82	
	unified	0.4305		unified	45.66		unified	19.49	
	ROUGE-L			GPT 4.1			InternVL3		
	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank
InfoSeek	1	0.3905	0.4067	1	36.92	37.6	1	28.38	29.7
	2	0.4003	0.4131	2	37.48	38.2	2	29.06	29.91
	3	0.3965	0.4101	3	37.14	37.46	3	29.28	30.4
	4	0.3978	0.4001	4	37.68	37.78	4	29.12	30.18
	5	0.3963	0.408	5	37.78	38.02	5	29.18	30.36
	lower bound	0.1975		lower bound	16.96		lower bound	16.2	
	upper bound	0.4659		upper bound	44.84		upper bound	33.84	
	unified	0.4151		unified	39.5		unified	30.96	

Table 9: ROUGE-L score and response accuracy with **LLaVA-OneVision** as the generation model in Figure 2.

Dataset	**Evaluation Methods**								
	ROUGE-L			GPT 4.1			InternVL3		
	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank
E-VQA	1	0.325	0.352	1	33.56	35.44	1	13.6	14.25
	2	0.33	0.35	2	33.52	35.59	2	13.87	14.29
	3	0.328	0.343	3	33.62	34.98	3	13.7	14.08
	4	0.328	0.342	4	33.53	34.92	4	13.95	14.19
	5	0.321	0.339	5	33.87	34.6	5	13.68	14.04
	lower bound	0.115		lower bound	10.38		lower bound	6.19	
	upper bound	0.409		upper bound	41.54		upper bound	17.79	
	unified	0.372		unified	40.69		unified	16.57	
	ROUGE-L			GPT 4.1			InternVL3		
	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank	top-k given	w/o re-rank	w/ re-rank
InfoSeek	1	0.361	0.376	1	34.66	35.28	1	27.4	28.68
	2	0.362	0.372	2	35.02	36.11	2	27.9	28.76
	3	0.352	0.367	3	34.77	35.66	3	26.9	28.18
	4	0.352	0.364	4	33.14	34.5	4	26.76	28.1
	5	0.349	0.36	5	33.38	34.12	5	26.72	27.8
	lower bound	0.143		lower bound	12.12		lower bound	11.18	
	upper bound	0.427		upper bound	40.4		upper bound	32.24	
	unified	0.385		unified	37.86		unified	29.72	

Given an image and question, read the provided document under RELATED DOCUMENT section and decide whether the answer can be derived from the document. Do not use your own knowledge.

RELATED DOCUMENT:
<Provided Docs>

Now given the following image and question, access whether the RELATED DOCUMENT is describing the image and provides answer to the question.

<image>

<question>

Return one of these labels: 'Yes' or 'No'. Please directly return your label, do not include the reasoning process!!

Your evaluation is:

Figure 10: The evaluation prompt to assess the relation between query and provided document **before the response generation**.

Given the following reference document, image, question, and response, please evaluate whether the response correctly answers the question by referring to the document. Do not use your own knowledge.

RELATED DOCUMENT:
<Provided Docs>

Now given the following image, question, and response again, assess whether the response correctly answers the question by referring to the above Reference document.

<image>

Question: ***<question>***

Response: ***<model answer>***

Return one of these labels: 'Yes' or 'No'. Please directly return your label, do not include the reasoning process!!

Your evaluation is:

Figure 11: The self-reflection prompt to verify the model response answers to the query and is derived from the provided document.