Progressive Multimodal Search and Reasoning for Knowledge-Intensive Visual Question Answering

Anonymous ACL submission

Abstract

Knowledge-intensive visual question answering (VQA) requires external knowledge beyond image content, demanding precise visual grounding and coherent integration of visual and textual information. Although multimodal retrieval-augmented generation has achieved notable advances by incorporating external knowledge bases, existing approaches largely adopt single-pass frameworks that often fail to acquire sufficient knowledge and lack mechanisms to revise misdirected reasoning. We propose PMSR (Progressive Multimodal Search and Reasoning), a framework that progressively constructs a structured reasoning trajectory to enhance both knowledge acquisition and synthesis. PMSR uses dual-scope queries conditioned on both the latest record and the trajectory to retrieve diverse knowledge from heterogeneous knowledge bases. The retrieved evidence is then synthesized into compact records via compositional reasoning. This design facilitates controlled iterative refinement, which supports more stable reasoning trajectories with reduced error propagation. Extensive experiments across six diverse benchmarks (Encyclopedic-VQA, InfoSeek, MMSearch, LiveVQA, FVQA, and OK-VQA) demonstrate that PMSR consistently improves both retrieval recall and end-to-end answer accuracy.

1 Introduction

The emergence of multimodal large language models (MLLMs) has driven significant progress in multimodal understanding and reasoning. Nonetheless, recent models continue to struggle with knowledge-intensive visual question answering (VQA) tasks, which require external knowledge beyond the visual content in the image. These questions require a tightly coupled process of (1) grounding visual entities, (2) retrieving relevant external knowledge, and (3) synthesizing visual and textual evidence to produce an answer.

Multimodal Retrieval-Augmented Generation (RAG) has become a natural solution to this challenge. In the standard RAG process, the model retrieves image–text pairs from an external knowledge base given the input image and question, and then conditions the MLLM on the retrieved context to generate an answer. Recent work has strengthened RAG via improved multimodal retrievers, hierarchical filtering, and reranking (Cocchi et al., 2025; Zhang et al., 2024; Ling et al., 2025; Chen et al., 2024; Liu et al., 2024b; Yan and Xie, 2024; Yang et al., 2025).

However, this *retrieve-then-read* process is problematic for knowledge-intensive VQA, where initial retrieval is often insufficient, as imperfect retrievers frequently fail to gather all necessary knowledge or introduce distracting passages (Zhang et al., 2023; Shi et al., 2023; Cuconasu et al., 2024; Yoran et al., 2024). These limitations are further amplified in multimodal RAG, where distractors in both modalities can mislead reasoning and degrade performance. Textual distractors dominate the model's attention and bias it toward irrelevant passages, whereas visual distractors can corrupt visual grounding and misdirect reasoning (Deng et al., 2025; Bae et al., 2025).

Motivated by these limitations, an emerging line of work has explored agentic approaches that leverage reasoning for iterative, tool-augmented retrieval (Li et al., 2024; Geng et al., 2025; Wu et al., 2025; Hong et al., 2025a). In these frameworks, agents reason and act iteratively, conditioning each action on the accumulated interaction history, including prior reasoning traces and tool outputs. However, errors in query generation, filtering information, and evidence summarization frequently accumulate in these multi-round interactions (Jiang et al., 2024a). Since these frameworks condition each step on the full interaction history, they primarily rely on context accumulation, retaining intermediate reasoning and tool outputs in an ever-growing context. As a result, early errors can propagate through the unstructured history and gradually drift subsequent retrieval and reasoning.

We propose PMSR (Progressive Multimodal Search and Reasoning), which progressively constructs a structured reasoning trajectory to enhance knowledge acquisition and synthesis. Unlike prior approaches that condition each step on the full interaction history, PMSR maintains the reasoning state as a trajectory of compact records synthesized from retrieved evidence, and leverages this trajectory to guide subsequent retrieval and reasoning. To acquire diverse knowledge, PMSR formulates dual-scope queries conditioned on reasoning states at both the record and the trajectory levels. These queries then retrieve knowledge from heterogeneous knowledge bases (KBs), which PMSR synthesizes via compositional reasoning into a compact reasoning record. The updated record is appended to the reasoning trajectory to guide subsequent iterations. The process terminates adaptively when additional iterations become redundant.

We conduct extensive experiments on six knowledge-intensive VQA benchmarks, including Encyclopedic-VQA (E-VQA), InfoSeek, MM-Search, LiveVQA, FVQA, and OK-VQA. Experimental results demonstrate that PMSR consistently improves retrieval recall and end-to-end answer accuracy over multimodal baselines across various benchmarks, achieving outstanding performance on five benchmarks. Our ablations confirm that the components work synergistically, and trajectory analysis shows that PMSR more often corrects early failures and reduces drift across iterations.

2 Related work

2.1 Multimodal RAG

Multimodal retrieval-augmented generation (RAG) for knowledge-intensive VQA has largely followed a *retrieve-then-read* paradigm: external knowledge is retrieved in single-step and then read by the model to answer the question. Early work primarily focused on improving single-step retrieval by learning more effective multimodal embeddings (Wei et al., 2024; Lin et al., 2024, 2025; Liu et al., 2024b; Jiang et al., 2025, 2024c). Subsequent approaches extended this paradigm by incorporating coarse-to-fine retrieval strategies. Hierarchical systems such as Wiki-LLaVA (Caffagni et al., 2024) and EchoSight (Yan and Xie, 2024) adopt coarse-to-fine retrieval pipelines with image-based retrieval followed by multimodal or text-based reranking, while OMGM (Yang et al., 2025) further develops this paradigm through a multi-step pipeline that explicitly models multiple knowledge granularities via successive multimodal and textual reranking.

More recent studies have focused on enhancing the *read* phase by leveraging the reasoning capabilities of MLLMs. For instance, ReflectiVA (Cocchi et al., 2025) and mR^{2}AG† (Zhang et al., 2024) use self-reflection to evaluate retrieval adequacy and evidence relevance. MMKB-RAG (Ling et al., 2025) generates semantic tags to filter irrelevant evidence. Wiki-PRF (Hong et al., 2025b) adopts reinforcement learning to retain only relevant information. Despite these advances, the majority of multimodal RAG approaches remain a static *retrieve-then-read* paradigm, constraining the model's ability to refine retrieval as reasoning evolves.

2.2 Multimodal Agents

The emergence of agentic paradigms has shifted research from *retrieve-then-read* to agent-based frameworks, where an agent iteratively combines step-by-step reasoning with actions, enabling it to solve complex problems by interacting with external tools (Yao et al., 2022). Early explorations of this direction appeared in *iterative RAG* for text-only question answering (Wang et al., 2024; Xiong et al., 2024; Yue et al., 2025; Trivedi et al., 2022; Yu et al., 2024; Jiang et al., 2024d; Zhang et al., 2025b; Liu et al., 2024a), where models decompose complex queries into sub-queries and iteratively perform retrieval within predefined workflows.

Building on these ideas, multimodal agents bring retrieval and tool use to vision–language settings by coupling retrieval with tool-augmented interaction, enabling models to iteratively reformulate queries and select external tools during multi-step reasoning. OmniSearch (Li et al., 2024) introduces adaptive planning that routes multimodal queries across multiple search tools. More recent work learns search and tool-use policies via reinforcement learning: WebWatcher (Geng et al., 2025) and MMSearch-R1 (Wu et al., 2025) optimize retrieval trajectories over tool interactions, internalizing decision-making across steps. DeepEyesV2 (Hong et al., 2025a) further integrates perception, search, and code execution within a unified agentic reasoning loop.

However, these agentic approaches condition each step on a long interaction history, such that earlier intermediate outputs remain in the conditioning

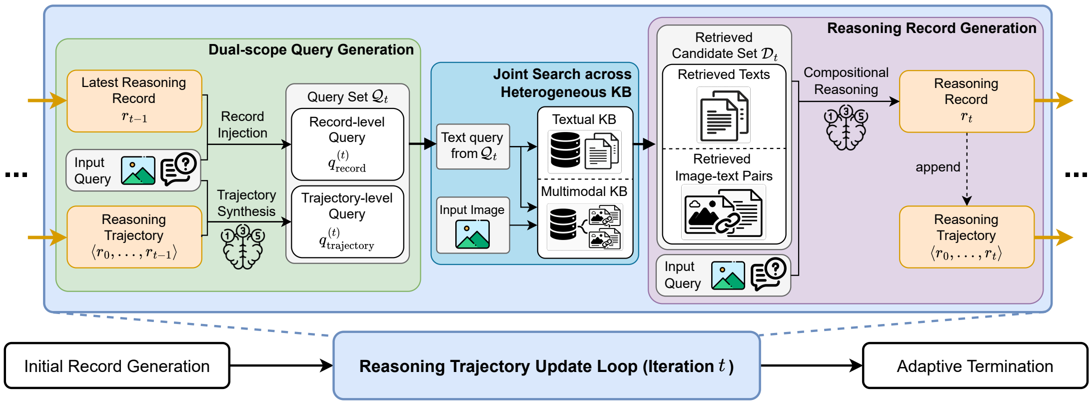

Figure 1: Overview of PMSR with the reasoning trajectory update loop at iteration t. PMSR consists of three stages: initial record generation, iterative reasoning trajectory updates, and adaptive termination. At each iteration, the reasoning trajectory update loop generates dual-scope queries conditioned on the latest reasoning record and the trajectory, retrieves knowledge from heterogeneous textual and multimodal KBs, and synthesizes the retrieved candidates into a new reasoning record. The newly generated record is appended to the trajectory to guide subsequent iterations. The process terminates adaptively when further iterations provide limited additional evidence.

and continue to influence subsequent retrieval and reasoning. In contrast, our method progressively retrieves evidence using dual-scope queries over heterogeneous KBs and condenses retrieved knowledge as a record to update the reasoning trajectory, mitigating drift from earlier intermediate outputs.

3 Method

PMSR (Progressive Multimodal Search and Reasoning) is a framework for knowledge-intensive VQA where it progressively constructs a reasoning trajectory, as illustrated in Figure 1.

3.1 Initial Reasoning Record Generation

To bootstrap the reasoning trajectory, PMSR first constructs an initial reasoning record using an MLLM. Unlike later iterations that build upon the trajectory, this step combines the model's parametric knowledge with externally retrieved knowledge.

Given the input query composed of an image I and a question Q, the MLLM generates a visually grounded description relevant to the question:

$$d_0 = \mathcal{G}_{\text{desc}}(Q, I). \tag{1}$$

We then expand the query by concatenating Q with this description, forming an enriched query $q_{\text{init}} = [Q; d_0]$. Using q_{init}, PMSR retrieves an initial candidate set $\mathcal{D}_0$ from the heterogeneous KBs.

Finally, the retrieved candidate set is synthesized into the first reasoning record using a dedicated reasoning operator:

$$r_0 = \mathcal{G}_{\text{reason}}(Q, I, \mathcal{D}_0). \tag{2}$$

This produces a coherent summary of the relevant facts, initializing the reasoning trajectory $\langle r_0 \rangle$.

3.2 Dual-scope Query Formulation

After initialization, PMSR progressively guides knowledge search by generating new queries conditioned on the evolving reasoning trajectory. PMSR decomposes query generation into two complementary scopes: a record-level query grounded in the latest reasoning record and a trajectory-level query derived from the accumulated reasoning trajectory. The record-level query supports local refinement based on newly acquired knowledge, while the trajectory-level query preserves broader intent accumulated across iterations. The set of queries generated at iteration t is given by:

$$\mathcal{Q}_t = \left\{ q^{(t)}_{\text{record}},\ q^{(t)}_{\text{trajectory}} \right\}. \tag{3}$$

Record-level query. The record-level query conditions on the latest reasoning record r_{t-1}, using its most recent deductions to expand the query to retrieve additional knowledge relevant to the current reasoning state. In the standard PMSR setting for knowledge-intensive VQA, this is implemented by concatenating the input question with r_{t-1}:

$$q^{(t)}_{\text{record}} = [Q;\ r_{t-1}]. \tag{4}$$

For the web-equipped variant of PMSR, this operator is adapted to produce a compact reformulation suitable for search engine constraints.

Trajectory-level query. The trajectory-level query leverages the reasoning trajectory to retrieve knowledge guided by the evolving reasoning records. Formally, a dedicated operator synthesizes information from this trajectory to generate a context-specific query:

$$q^{(t)}_{\text{trajectory}} = \mathcal{G}_{\text{trajectory}}\left(Q, I, \langle r_0, \ldots, r_{t-1} \rangle\right). \quad (5)$$

In contrast to the record-level query, it incorporates broader context accumulated over prior reasoning steps.

3.3 Joint Search across Heterogeneous KBs

To support compositional reasoning with diverse external knowledge, PMSR performs a joint search over heterogeneous KBs using the dual-scope query set generated at each iteration. Given the query set $\mathcal{Q}_t$, PMSR retrieves candidates from a textual KB and a multimodal KB.

Retrieval from textual KB. For each query $q_t \in \mathcal{Q}_t$, we retrieve passages p from the textual KB using text-text semantic similarity:

$$S_{\text{txt}} = \text{sim}_{\text{text}}(q_t, p), \quad (6)$$

where sim_{text} denotes cosine similarity in a text embedding space.

Retrieval from multimodal KB. The multimodal KB consists of image-text pairs (I_c, t_c). For each query $q_t \in \mathcal{Q}_t$ and the input image I, we compute a decoupled similarity score:

$$S_{\text{mm}} = \lambda \; \text{sim}_{\text{text}}(q_t, t_c) + (1 - \lambda) \; \text{sim}_{\text{img}}(I, I_c), \quad (7)$$

where sim_{img} denotes cosine similarity in an image embedding space. We use a fixed weighting of $\lambda = 0.5$. The text term adapts retrieval to the dual-scope query, while the image term preserves visual relevance to the input image.

Combined retrieval. We retrieve up to N_{txt}=20 text passages and N_{mm}=10 image-text pairs per iteration and aggregate them into the candidate set $\mathcal{D}_t$. We evenly split the retrieval budget between record- and trajectory-level queries. Additional implementation details are provided in Appendix D.

3.4 Reasoning Record Generation

After retrieving candidates from heterogeneous KBs, PMSR constructs a reasoning record from the newly retrieved knowledge.

At iteration t, the retrieved candidate set $\mathcal{D}_t$ is synthesized using a dedicated reasoning operator:

$$r_t = \mathcal{G}_{\text{reason}}(Q, I, \mathcal{D}_t). \quad (8)$$

The operator $\mathcal{G}_{\text{reason}}$ integrates retrieved visual and textual knowledge conditioned on the input query, producing a reasoning record. PMSR supports compositional reasoning by aggregating diverse knowledge from heterogeneous KBs into records to guide subsequent iterations.

Importantly, each reasoning record r_t is generated solely from the newly retrieved candidate set $\mathcal{D}_t$, without directly conditioning on previous reasoning records. The resulting record is appended to the reasoning trajectory, yielding $\langle r_0, \ldots, r_{t-1}, r_t \rangle$ for subsequent iterations.

3.5 Adaptive Termination via Information Saturation

To improve inference efficiency, we introduce an adaptive termination criterion based on information saturation, where increasing similarity between newly generated and earlier queries indicates redundant retrieval. This similarity is quantified by the saturation score, defined as

$$\delta^{(t)}_{\text{query}} = \max_{q \in \mathcal{Q}_t,\, q' \in \mathcal{Q}_j,\, j<t} \text{sim}_{\text{text}}(q, q'). \quad (9)$$

The iterative process terminates when

$$\delta^{(t)}_{\text{query}} \geq \tau. \quad (10)$$

Unless otherwise stated, we set $\tau = 0.9$ in all experiments. Upon termination at iteration T, the MLLM generates the final answer conditioned on Q, I, and the reasoning trajectory $\langle r_0, \ldots, r_T \rangle$.

The prompt templates used to instantiate the MLLM operators ($\mathcal{G}_{\text{desc}}$, $\mathcal{G}_{\text{trajectory}}$, and $\mathcal{G}_{\text{reason}}$) are provided in Appendix A. For the web-equipped variant of PMSR, the prompt and details of implementations are provided in Appendix B and C.

4 Experiments

To evaluate the performance of our proposed PMSR framework, we conduct experiments on several challenging benchmark datasets using a diverse set of evaluation metrics.

4.1 Experiment Setup

Datasets. We evaluate PMSR on an extensive suite of knowledge-intensive VQA benchmarks covering encyclopedic, factual, and real-world information-seeking scenarios. Our experiments use the InfoSeek validation split of M2KR (Lin et al., 2024), the OK-VQA validation split, and the single-hop questions of the E-VQA test

Method	InfoSeek			E-VQA		
	R@5	R@10	R@20	R@5	R@10	R@20
Wiki-LLaVA (Caffagni et al., 2024)	-	66.1	71.9	-	9.9	13.2
LLM-RA (Jian et al., 2024)	53.8	-	-	-	-	-
mR²AG (Zhang et al., 2024)	-	65.0	71.0	-	-	-
ReflectiVA (Cocchi et al., 2025)	<u>77.6</u>	-	**86.4**	36.1	-	<u>49.8</u>
EchoSight† (Yan and Xie, 2024)	74.0	77.4	77.9	<u>47.9</u>	48.8	48.8
OMGM (Yang et al., 2025)	73.9	<u>80.0</u>	84.8	41.2	<u>49.8</u>	**58.7**
OMGM† (Yang et al., 2025)	**80.8**	**83.6**	<u>84.8</u>	**55.7**	**58.1**	**58.7**
ReAuSE (Long et al., 2025)	59.5	-	-	-	-	-
			Cumulative Recall			
Ours (Qwen3-VL-4B)*		<u>93.9</u>			<u>64.3</u>	
Ours (Qwen3-VL-8B)*		**94.6**			**67.3**	

Table 1: Recall comparison on the InfoSeek validation and E-VQA test sets. † indicates methods that utilize reranking. For PMSR (*), we report cumulative recall at adaptive termination. Best and second-best results are highlighted in bold and underlined, respectively.

split (Mensink et al., 2023; Chen et al., 2023; Marino et al., 2019), following standard practice. Moreover, we extend our evaluation of PMSR to four search-oriented benchmarks: FVQA-test, the InfoSeek Human subset, LiveVQA, and MMSearch (Jiang et al., 2024a; Wu et al., 2025; Fu et al., 2025). These benchmarks target real-world questions requiring factual grounding, time-sensitive news, and long-tail knowledge. The details of each benchmark are provided in Appendix E.

Knowledge bases and retrievers. For a fair comparison, we use fixed heterogeneous KBs across all experiments. The multimodal KB consists of 2M Wikipedia image-text pairs provided in InfoSeek, while the textual KB comprises approximately 21M Wikipedia passages from FlashRAG (Jin et al., 2024).

For retrieval, we adopt dense similarity search. Multimodal retrieval uses SigLIP2 (Tschannen et al., 2025) for image embeddings and Qwen3-Embedding (Zhang et al., 2025a) for text embeddings, while textual retrieval uses E5-base-v2 (Wang et al., 2022). Comparisons with retriever baselines are provided in Appendix D.

Multimodal large language models. To assess how performance scales with reasoning capacity while ensuring fair comparison, we evaluate two tiers of MLLM backbones: open-source models from the Qwen-VL series (Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a)) and the proprietary Gemini-2.5-Flash (Comanici et al., 2025).

Evaluation metrics. We evaluate PMSR using standard accuracy and retrieval metrics across

Method	OK-VQA	
	PRR@5	PRR@10
DPR (Karpukhin et al., 2020)	66.9	76.4
ReViz-ICT (Luo et al., 2023)	61.9	72.6
GeMKR (Long et al., 2024)	70.8	79.1
FLMR (Lin et al., 2023)	68.1	78.0
Pre-FLMR (Lin et al., 2024)	68.6	-
ReAuSE (Long et al., 2025)	**88.0**	**91.3**
OMGM† (Yang et al., 2025)	73.4	-
	Cumulative Recall	
Ours (Qwen3-VL-4B)*	92.1	
Ours (Qwen3-VL-8B)*	**97.1**	

Table 2: Recall comparison on the OK-VQA benchmark using Wikipedia as the knowledge source. † indicates methods that utilize reranking. For PMSR (*), we report the cumulative recall at adaptive termination.

benchmarks. For accuracy, we report the official BERT matching score (BEM) (Bulian et al., 2022) on E-VQA and exact match (EM) on InfoSeek. We additionally report cover exact match (CEM) (Jiang et al., 2024b; Yue et al., 2025) as a complementary metric that checks whether the ground-truth answer appears in the model output. For OK-VQA, FVQA-test, InfoSeek Human, MMSearch, and LiveVQA, we adopt an LLM-as-Judge protocol following MMSearch-R1 (Wu et al., 2025) using GPT-4o; the evaluation prompts are provided in Appendix F.

For retrieval performance, we measure recall based on the presence of ground-truth evidence in the retrieved context. We report entity recall for InfoSeek and E-VQA, and Pseudo-Relevance Recall (PRR) (Luo et al., 2021) for OK-VQA following PreFLMR (Lin et al., 2024). To assess progressive knowledge acquisition, we further report cumulative recall under adaptive termination.

Method	Retriever	Model	InfoSeek		E-VQA
			Val	**M2KR**	**Single-hop**
Wiki-LLaVA (Caffagni et al., 2024)	CLIP-ViT-L	LLaVA-1.5-7B	28.9	-	21.8
EchoSight† (Yan and Xie, 2024)	EVA-CLIP-8B	Mistral-7B	31.3	-	35.5
LLM-RA (Jian et al., 2024)	EVA-CLIP-8B	BLIP2-Flan-T5XL	23.1	-	-
mR²AG† (Zhang et al., 2024)	CLIP-ViT-L	LLaVA-1.5-7B	40.2	-	-
ReflectiVA (Cocchi et al., 2025)	EVA-CLIP-8B	LLaVA-MORE-8B	40.1	-	35.5
MMKB-RAG† (Ling et al., 2025)	PreFLMR ViT-G	Qwen2-VL-7B	36.7	34.7	39.7
RET-2 (Caffagni et al., 2025)	RET-2	LLaVA-MORE-8B	22.8	-	28.5
Wiki-PRF(w/ RL) (Hong et al., 2025b)	EVA-CLIP-8B	VLM-PRF-7B	42.5	-	40.1
OMGM† (Yang et al., 2025)	EVA-CLIP-8B	LLaVA-1.5-7B	**43.5**	-	50.2
OMGM†	EVA-CLIP-8B	GPT-4o	42.1	-	**51.2**
Ours	SigLIP2-g	Qwen3-VL-4B	-	38.3*	40.9
		Qwen3-VL-8B	-	41.5*	46.4
		Gemini-2.5-Flash	-	**50.5***	**59.9**

Table 3: Overall accuracy on InfoSeek and E-VQA. **Val** denotes the full InfoSeek validation set (137K), and **M2KR** the 5K subset. E-VQA results are reported on the single-hop subset. † indicates methods that utilize reranking; * indicates that LLaVA-MORE-8B (ReflectiVA) is used as the final answer generator for EM evaluation.

4.2 Retrieval Performance on VQA Benchmarks

Table 1 reports the retrieval performance of PMSR on the InfoSeek and E-VQA benchmarks. Across both datasets, PMSR achieves consistent and substantial improvements over prior methods.

On InfoSeek, the 4B model reaches a cumulative recall of 93.9%, outperforming the previous best result reported by ReflectiVA (86.4% at R@20). On E-VQA, the 4B model achieves 64.3% cumulative recall, exceeding OMGM† by 5.6 percentage points. Scaling the backbone from 4B to 8B further yields consistent gains, improving cumulative recall to 94.6% on InfoSeek and 67.3% on E-VQA.

We report cumulative recall for PMSR to reflect progressive evidence accumulation under adaptive stopping. For completeness, we also report per-iteration recall and contributions of each KB in Appendix G. To quantify the efficiency gains enabled by adaptive stopping, Section 5.3 presents an ablation study.

As shown in Table 2, PMSR demonstrates strong and consistent retrieval performance on the OK-VQA benchmark, despite large differences from other benchmarks in knowledge type, question formulation, and grounding requirements. On OK-VQA, PMSR achieves 92.1% and 97.1% cumulative recall with the 4B and 8B models, closely matching its performance on InfoSeek and E-VQA. This cross-domain stability contrasts sharply with prior retrieval-augmented models, which often perform well only within their target domain. The results indicate that PMSR's progressive retrieval strategy generalizes effectively across knowledge types without requiring dataset-specific tuning.

Additionally, Appendix H reports end-to-end answer accuracy under an LLM-as-Judge protocol. Appendix I further provides an analysis of the contribution of the latest reasoning record via a sensitivity study on the interpolation weight λ.

4.3 Accuracy on Knowledge-Intensive VQA

We report end-to-end answer accuracy of PMSR on the InfoSeek and E-VQA benchmarks in Table 3. Across both datasets, PMSR achieves strong performance compared to prior retrieval-augmented approaches, highlighting the effectiveness of progressive, reasoning-guided retrieval.

On the E-VQA benchmark, PMSR with Qwen3-VL-8B achieves 46.4% accuracy, which is comparable to strong prior baselines. When the trajectory is generated using a more capable model (Gemini-2.5-Flash), accuracy increases to 59.9%, surpassing the previous best by 8.7%.

On the InfoSeek benchmark, PMSR also demonstrates substantial improvements. Performance scales with the capacity of the reasoning backbone, with the Qwen3-VL-8B configuration achieving 41.5% accuracy and the Gemini-2.5-Flash configuration reaching 50.5% accuracy.

Importantly, for InfoSeek, we evaluate accuracy using LLaVA-MORE-8B as a final answerer, regardless of which MLLM is used to generate the reasoning records. This controlled setup isolates the contribution of PMSR: improvements on InfoSeek reflect reasoning trajectory produced by PMSR, rather than differences in the answer generation model. Accordingly, stronger trajectory-

Method	FVQA test	InfoSeek Human	MM Search	Live VQA
OmniSearch (GPT-4o) (Li et al., 2024)	-	-	49.7	40.9
MMSearch-R1 (Wu et al., 2025)	58.4	55.1	53.8	48.4
WebWatcher (Geng et al., 2025)	-	-	49.1	51.2
DeepEyesV2 (Hong et al., 2025a)	60.6	51.1	**63.7**	-
Ours	**61.2**	**58.2**	54.3	**54.2**

Table 4: Performance on search-oriented multimodal benchmarks. Results for OmniSearch are taken from WebWatcher (Geng et al., 2025). Unless otherwise noted, all methods use Qwen2.5-VL-7B as the backbone.

generation configurations (e.g., Gemini-2.5-Flash) yield higher accuracy because they construct more informative and better-grounded reasoning trajectories, which the same answerer can exploit more effectively. Additional qualitative examples illustrating these trajectories are provided in Appendix O.

4.4 Results on Search-Oriented Benchmarks

We further evaluate PMSR on search-oriented benchmarks that require multimodal grounding and open-domain knowledge acquisition. As shown in Table 4, using the same Qwen2.5-VL-7B backbone, PMSR achieves 61.2% and 58.2% accuracy on FVQA and InfoSeek Human, respectively, surpassing recent agent-based baselines. On MMSearch, PMSR attains 54.3% accuracy, remaining competitive with specialized multimodal search agents. On LiveVQA, which targets real-world, time-sensitive information seeking over diverse news sources, PMSR reaches 54.2% accuracy, the highest among the methods reported in Table 4.

5 Ablations

For ablation studies, we conduct experiments using Qwen3-VL-8B to validate the robustness of individual components. Unless otherwise specified, the retrieval budget is fixed across single-query and dual-scope query settings, and adaptive termination is used with $\tau = 0.9$ (up to a maximum of 5 iterations).

5.1 Impact of Iterative Performance

To quantify the effect of progressive search and reasoning, Table 5 reports performance as the number of iterations increases. Across both InfoSeek and E-VQA, PMSR exhibits monotonic improvements in both accuracy (CEM/BEM) and retrieval recall. The largest improvements occur in the first one to two iterations, after which improvements become smaller but remain consistent. These results indicate that progressively accumulating reasoning records and using them to guide subsequent retrieval provides measurable benefits. To further examine how intermediate reasoning evolves across iterations, we analyze reasoning trajectory dynamics in Section 6.1.

Iter.	InfoSeek		E-VQA	
	CEM	Recall	BEM	Recall
0	48.3	91.7	37.1	59.2
1	53.6 (+5.3)	93.2 (+1.5)	42.2 (+5.1)	63.4 (+4.2)
2	54.8 (+1.2)	94.1 (+0.9)	45.4 (+3.2)	65.3 (+1.9)
3	55.4 (+0.6)	94.5 (+0.4)	46.2 (+0.8)	66.4 (+1.1)
4	56.3 (+0.9)	95.0 (+0.5)	47.1 (+0.9)	67.4 (+1.0)

Table 5: Performance of PMSR across iterations on InfoSeek and E-VQA. Results are reported for a fixed sequence of 5 iterations; Iteration 0 corresponds to the initial reasoning record.

5.2 Ablation of components

Dual-scope Query	Hetero. KB	Iter.	BEM	Recall
×	×	×	30.7	48.7
✓	×	×	32.4	48.8
×	✓	×	34.4	58.4
✓	✓	×	37.3	58.8
✓	×	✓	34.6	56.3
×	✓	✓	43.1	64.8
✓	✓	✓	**46.4**	**67.4**

Table 6: Component ablation of PMSR on the E-VQA test set. We ablate dual-scope query formulation, retrieval over heterogeneous KBs, and progressive search and reasoning over iterations. Removing dual-scope query uses only the trajectory-level query; removing heterogeneous KBs restricts retrieval to the multimodal KB; removing iterations corresponds to single-pass RAG.

Table 6 examines the contribution of key components of PMSR on E-VQA. Adding a textual KB alongside the multimodal KB substantially improves retrieval recall (48.7→58.4) and increases BEM (30.7→34.4), highlighting the importance of heterogeneous KBs. Enabling dual-scope query formulation further boosts performance, with larger gains observed under heterogeneous KB retrieval (BEM 34.4→37.3; recall 58.4→58.8). Finally, introducing progressive search and reasoning over iterations yields additional improvements, and the full model achieves the best overall performance. These results indicate that repeated reasoning-guided retrieval and accumulation of reasoning records provide complementary benefits beyond any single component.

5.3 Ablation of Adaptive Termination

Method	InfoSeek			E-VQA		
	Avg. Iter.	CEM	Recall	Avg. Iter.	BEM	Recall
Fixed	5.0	56.3	95.0	5.0	47.1	67.5
Adaptive	3.3	55.1	94.6	3.5	46.4	67.3

Table 7: Impact of adaptive termination compared with a fixed-iteration strategy (fixed: 5 iterations vs. adaptive: $\tau = 0.9$).

We evaluate the efficiency of adaptive termination by comparing it against a fixed-iteration strategy with the same backbone (Qwen3-VL-8B). As shown in Table 7, adaptive termination with the default threshold $\tau = 0.9$ reduces the average number of iterations from 5.0 to 3.3 on InfoSeek and 3.5 on E-VQA, while maintaining comparable accuracy and retrieval recall.

6 Analysis

6.1 Analysis of Reasoning Trajectory

Trajectory Type	InfoSeek	E-VQA
Stable-Correct	43.95%	47.73%
Persistent-Fail	36.79%	4.91%
Correction	11.64%	30.06%
Conflict	7.63%	17.31%

Table 8: Distribution of reasoning trajectory types based on per-iteration correctness patterns (CEM on InfoSeek; BEM on E-VQA): *Stable-Correct* (correct at all iterations), *Persistent-Fail* (incorrect at all iterations), *Correction* (recovers from incorrect reasoning and is correct at the final iteration), and *Conflict* (correct in some iterations but incorrect at the final iteration).

We analyze reasoning trajectories by evaluating the correctness of each intermediate reasoning record across iterations, rather than only the final prediction, to characterize how reasoning evolves. As shown in Table 8, *Correction* occurs more frequently than *Conflict*, suggesting that the proposed framework more often recovers from early incorrect reasoning than propagates it to later iterations. Moreover, a substantial portion of trajectories is categorized as *Stable-Correct*, indicating that once correct grounding and relevant knowledge are established, the reasoning process tends to preserve correctness across subsequent iterations.

Notably, on E-VQA, the combined proportion of *Stable-Correct* and *Correction* trajectories exceeds 70%, indicating that reasoning records progressively gather sufficient knowledge to address the question. However, this proportion is considerably higher than the final answer accuracy, revealing a gap between the quality of intermediate reasoning records and the model's ability to fully utilize them for answer prediction. To further examine this gap, we present a model sensitivity analysis of the contextual noise of distractors in Appendix M. Furthermore, we provide a trajectory type comparison with the web search agent in Appendix N under the same KB and retriever.

6.2 Analysis of Adaptive Termination

Trajectory Category	InfoSeek	E-VQA
Stable-Correct	1.96	2.26
Correction	1.60	1.87

Table 9: Average additional iterations after convergence to a correct reasoning record under adaptive termination.

To assess whether adaptive termination halts once sufficient knowledge has been acquired, we measure the number of iterations executed after the reasoning process has already converged to a correct state. We focus on *Stable-Correct* and *Correction* trajectories, in which a correct reasoning record is reached at some iterations and maintained thereafter.

As shown in Table 9, adaptive termination typically occurs shortly after convergence on both InfoSeek and E-VQA. Because adaptive termination requires at least one subsequent iteration to assess saturation by comparing newly generated outputs with prior states, the procedure executes, on average, one to two additional iterations before termination.

7 Conclusion

In this paper, we introduced PMSR, a progressive multimodal search and reasoning framework for knowledge-intensive VQA. PMSR constructs a structured reasoning trajectory composed of compact reasoning records synthesized from diverse evidence to enhance both knowledge acquisition and synthesis. This design enables controlled, iterative refinement of retrieval and reasoning, promoting more stable trajectories that can correct early mistakes and reduce drift over successive iterations. Extensive experiments on six knowledge-intensive VQA benchmarks demonstrate consistent improvements in retrieval recall and end-to-end answer accuracy over strong baselines, highlighting the effectiveness of PMSR.

8 Limitations

While the PMSR framework demonstrates significant improvements in retrieval recall and answer accuracy across several knowledge-intensive VQA benchmarks, several limitations warrant consideration for future research. First, the proposed framework relies on iterative retrieval and reasoning, which introduces additional inference overhead compared to single-pass RAG methods. Although the adaptive termination mechanism mitigates redundant iterations, the overall computational cost remains higher in cases where convergence is slow.

Second, PMSR retrieves from heterogeneous sources, but its overall performance remains sensitive to retrieval quality and query formulation. In this work, we rely on standard retrieval components and do not integrate recent MLLM-based multimodal retrievers that learn fused multimodal embeddings for joint retrieval. Incorporating such retrievers is a promising direction, particularly for reasoning-guided queries, where transformed queries may benefit from joint image–text retrieval.

Finally, PMSR remains limited by the reasoning and grounding capabilities of the underlying MLLM. While PMSR provides progressive reasoning records and a structured, iterative knowledge acquisition process, smaller or less capable backbones may still struggle with accurate visual grounding and compositional reasoning over multiple visual-text associations, limiting how effectively retrieved relevant knowledge can be leveraged for correct predictions.

References

Jiyun Bae, Hyunjong Ok, Sangwoo Mo, and Jaeho Lee. 2025. Do reasoning vision-language models inversely scale in test-time compute? a distractor-centric empirical analysis. *arXiv preprint arXiv:2511.21397*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Recurrence-enhanced vision-and-language transformers for robust multimodal document retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9286–9295.

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.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Florin Cuconasu, Giovanni Trappolini, F. Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.

Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S. Yu, and Ranjay Krishna. 2025. Seeking and updating with live visual knowledge. *Preprint*, arXiv:2504.05288.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025a. Deepeyesv2: Toward agentic multimodal model. *Preprint*, arXiv:2511.05271.

Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. 2025b. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, and 1 others. 2024a. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024b. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *arXiv preprint arXiv:2412.12881*.

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

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. 2024d. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. *arXiv preprint arXiv:2407.13101*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2vec: Training vision-language models for massive multi-modal embedding tasks. In *The Thirteenth International Conference on Learning Representations*.

Jiajie Jin, Yutao Zhu, Guanting Dong, Yuyao Zhang, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, Zhicheng Dou, and Ji-Rong Wen. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *arXiv preprint arXiv:2405.13576*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S. Yu, Fei Huang, and Jingren Zhou. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5294–5316, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024a. RA-ISF: Learning to answer and understand from retrieval augmentation via iterative self-feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4730–4749, Bangkok, Thailand. Association for Computational Linguistics.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2024b. Lamra: Large multimodal model as your advanced retrieval assistant. *arXiv preprint arXiv:2412.01720*.

Xinwei Long, Zhiyuan Ma, Ermo Hua, Kaiyan Zhang, Biqing Qi, and Bowen Zhou. 2025. Retrieval-augmented visual question answering via built-in autoregressive search engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24723–24731.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 38, pages 18733–18741.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. *arXiv preprint arXiv:2306.00424*.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

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.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

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.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. 2025. Omgm: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. *arXiv preprint arXiv:2505.07879*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Auto-rag: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2025. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024. mr^2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuocheng Zhang, Yang Feng, and Min Zhang. 2025b. Levelrag: Enhancing retrieval-augmented generation with multi-hop logic planning over rewriting augmented searchers. *arXiv preprint arXiv:2502.18139*.

A Prompts for PMSR Framework

We present the prompt templates used in the PMSR framework. All prompts are designed to be model-agnostic. In the templates below, terms enclosed in `{braces}` denote dynamic content populated at runtime.

A.1 Initial Reasoning Record Generation

Initial Description Prompt

```
Question: {question}
```

```
Instruction:
Concisely describe the image which is
relevant to the question.
```

Figure A1: The prompt used to generate the initial visual description (d_0) for the first query expansion.

To bootstrap the iterative process, we first instruct the model to generate a query-focused description of the image (Figure A1). This serves as the initial grounding for the first retrieval step.

A.2 Dual-scope Query Formulation

As described in Section 3.2, PMSR constructs the query set $\mathcal{Q}_t$ using two complementary operators: a *record-level* query and a *trajectory-level* query.

Record-level query. This operator conditions on the most recent reasoning record to use its most recent deductions to guide successive retrieval. It is implemented by concatenating the original question Q with the latest reasoning record r_{t-1}, i.e., $q_{\text{record}}^{(t)} = [Q; r_{t-1}]$. No additional instruction prompt is required.

Trajectory-level query. To implement the trajectory-level query operator $\mathcal{G}_{\text{trajectory}}$, we use the structured prompt shown in Figure A2. The prompt instructs the MLLM to analyze the accumulated reasoning trajectory (provided in the `{knowledge}` field) together with the original question. By separating an explicit `Analysis` section from the `Output`, the model is encouraged to identify missing or underspecified information in $\langle r_0, \ldots, r_{t-1} \rangle$ before generating a context-specific query for subsequent knowledge search.

A.3 Reasoning Record Generation

At each iteration, we synthesize the retrieved evidence into a concise "Reasoning Record." This prompt is designed to retain only correct and relevant information for the next step (Figure A3).

Trajectory-level Query Operator Prompt ($\mathcal{G}_{\text{trajectory}}$)

```
Input Context:
**Query**: {question}
**Knowledge**: {knowledge}
```

```
Instruction:
Please first analyze all the information
in a section named Analysis (## Analysis).
Generate more accurate question based on
the Knowledge to search more information
helpful to addressing Query.

Your response should be in the following
format:

## Analysis
Analysis query and correct knowledge to
search more accurately.

## Output
Question: context-specific new question
```

Figure A2: The prompt used for the trajectory-level query operator $\mathcal{G}_{\text{trajectory}}$. The `{knowledge}` field is populated with the accumulated reasoning trajectory $\langle r_0, \ldots, r_{t-1} \rangle$ to allow the model to identify gaps before generating a new search query.

Reasoning Record Synthesis ($\mathcal{G}_{\text{reason}}$)

```
Input Context:
Question: {question}
Knowledge: {knowledge}
```

```
Instruction:
Based on the image, description, and
knowledge, summarize the correct and
relevant information with respect to the
image and question.
```

Figure A3: The prompt used for the Reasoning Record Generation operator $\mathcal{G}_{\text{reason}}$, which synthesizes a reasoning record from newly retrieved multimodal evidence.

A.4 Final Answer Generation

Once the iterations are complete (or adaptive termination is triggered), we use the reasoning trajectory to derive the final answer (Figure A4).

B Prompts for Web Search

We employ two specialized prompts to optimize the web search process. The first ensures that search queries are concise and keyword-optimized, while the second condenses retrieved long content into text passages.

Final Answer Generation

```
Input Context:
Question: {question}
Relevant Knowledge: {reasoning_records}

Instruction:
Please answer the following question using
the provided information and image.
Based on the information, provide a
detailed answer to the question.
```

Figure A4: The final prompt used to generate the answer by synthesizing the original question, image, and the full chain of reasoning records.

B.1 Search Query Condensation

When a generated query exceeds a predefined length threshold (e.g., 400 characters), we use a condensation prompt to rewrite it into a form suitable for search engines. This helps extract the main entities and essential intent from a verbose reasoning step (Figure A5).

Search Query Condensation

```
Instruction:
Rewrite this query to be a concise question
for search engine including main entity
and essential point.

Input Context:
Query: {prompt}
```

Figure A5: The prompt used to condense verbose or complex reasoning outputs into an effective keyword-based search query.

B.2 Web Content Summarization

To efficiently handle the noise and length of raw web pages, we use a summarization prompt. This prompt instructs the MLLM to generate a summary relevant to the original query (Figure A6).

C Web-equipped Implementation

To evaluate PMSR on search-oriented benchmarks that require access to time-varying or out-of-KB information, we implement a web-equipped variant that augments PMSR with web search. This variant preserves the core PMSR pipeline (query formulation $\rightarrow$ retrieval $\rightarrow$ reasoning record generation), while adapting query construction and evidence processing to practical constraints of web search (e.g., query length limits and noisy webpage content).

Web Content Summarization

```
Input Context:
Query: {prompt}

Instruction:
Summarize the following web content,
focusing on information relevant to the
query. Provide a concise summary in a
single paragraph:

Target Content:
Title: {title}
Content: {content}
```

Figure A6: The prompt used to summarize retrieved web pages. By explicitly conditioning the summary on the input query, the model filters irrelevant candidates and focuses on the evidence needed for the answer.

Image search tool. We use the ScrapingDog API to interface with Google Lens at the initial iteration (t=0) only, in order to obtain an initial set of visually related webpages. Given the input image, the tool returns visually similar pages with metadata such as thumbnails and titles. We then fetch the corresponding page contents and generate question-conditioned summaries, which are used as the initial visual evidence to bootstrap the first reasoning record. In subsequent iterations ($t > 0$), PMSR performs multimodal retrieval over the Wikipedia-based multimodal KB.

Text search tool. For textual retrieval, we employ Ollama Web Search following a *search-parse-summarize* pipeline. Given a query, the system retrieves relevant URLs, parses their contents, and summarizes each page using GPT-OSS 120B (5.1B active parameters). The same model is also used to rewrite long queries into concise forms to meet search-engine constraints.

C.1 Initial Reasoning Record Generation.

At the initial iteration ($t = 0$), we follow the protocol of MMSearch-R1 by anchoring retrieval with image-based search. Specifically, we submit the input image I to Google Lens to obtain visually similar webpages, then fetch and summarize their contents to form the initial candidate set $\mathcal{D}_0$. We generate the initial reasoning record r_0 by applying the standard reasoning operator $\mathcal{G}_{\text{reason}}$ on $(Q, I, \mathcal{D}_0)$, thereby bootstrapping the reasoning trajectory with visually grounded evidence.

C.2 Dual-scope Query Formulation

In the standard PMSR framework, the record-level query is constructed by concatenating the original question with the latest reasoning record. However, web search engines impose constraints on query length and formatting, making direct concatenation impractical. To address this, we introduce a rewriting operator $\mathcal{G}_{\text{record}}$ that compresses (Q, r_{t-1}) into a concise search string:

$$q_{\text{record}}^{(t)} = \mathcal{G}_{\text{record}}(Q, r_{t-1}). \tag{C.11}$$

The model is prompted to produce a keyword-focused query suitable for web search. For the trajectory-level query, we use the same operator $\mathcal{G}_{\text{trajectory}}$ as in Section 3.2, which synthesizes the accumulated reasoning trajectory into a context-specific query.

C.3 Web Search and Evidence Summarization

Each web query is executed to retrieve a relevant list of URLs. To summarize relevant information, we apply a summarization pipeline:

1. **Content extraction:** We scrape the raw HTML content of the top-k retrieved webpages.
2. **Query-conditioned summarization:** Each page is summarized to produce a relevant summary of the input query.

The resulting text summaries and the text retrieved from image search are treated as the candidate set $\mathcal{D}_t$. We then generate the next reasoning record r_t using $\mathcal{G}_{\text{reason}}(Q, I, \mathcal{D}_t)$, following the same procedure as in Section 3.4.

D Multimodal Retrieval Implementation

To construct the multimodal knowledge base, we process the Wikipedia corpus used in InfoSeek, which is derived from the 2022-10-01 Wikipedia dump. For each image–text pair, we generate normalized image and text embeddings to compute a decoupled similarity score. Specifically, image embeddings are extracted using a pretrained SigLIP2 model, while text embeddings are obtained from Wikipedia section summaries using a Qwen3-Embedding encoder.

To efficiently implement decoupled similarity retrieval, we concatenate the normalized image and text embeddings into a joint multimodal representation and index them using FAISS with an IndexFlatIP (inner product) index. At query time, the input image and refined text query are encoded separately using the same encoders and concatenated to form a single multimodal query vector. A maximum inner product search (MIPS) is then performed to retrieve the top-k candidates. With the default weight $\lambda = 0.5$, this retrieval procedure is equivalent to the decoupled similarity score, enabling efficient and scalable retrieval while preserving consistency with our scoring function.

To prevent image duplication, we additionally deduplicate the knowledge base against the evaluation splits using perceptual hashing. We compute a perceptual hash for every image in the KB and for all query images in the InfoSeek validation split and the E-VQA test split, and treat images with matching hashes as duplicates. This reveals a small number of overlapping images: 17 in the InfoSeek validation split and 35 in the Encyclopedic-VQA (E-VQA) test split. We remove the duplicate images from the KB before building the FAISS index.

Retriever	Dataset	Query Modality	R@5	R@10	R@20
OMGM	InfoSeek	image-to-text	73.9	80.0	84.8
	E-VQA	image-to-text	41.2	49.8	58.7
EVA-CLIP-8B	InfoSeek	image-to-image	67.1	73.0	77.9
	E-VQA	image-to-image	31.3	41.0	48.8
SigLIP2-g Qwen3-Embedding-0.6B	InfoSeek	image-to-image	66.7	72.7	77.5
		image+text	69.4	76.2	81.1
	E-VQA	image-to-image	36.2	41.9	46.4
		image+text	43.1	48.7	54.5

Table A1: Performance of multimodal similarity on the InfoSeek validation split and E-VQA test split.

Table A1 reports retriever performance on the InfoSeek validation split and the E-VQA test split. The results show that multimodal queries that jointly incorporate image and text similarity consistently achieve higher recall than unimodal (image-only) queries.

E Details of Datasets

InfoSeek. InfoSeek is a large-scale benchmark designed to evaluate visual information-seeking capabilities. It consists of automatically generated and human-annotated questions grounded in Wikidata entities, paired with corresponding images and factual answers. Question templates cover hundreds of relational types, ensuring broad coverage of entity attributes, locations, and fine-grained factual properties. Each question–image pair is retained only when supporting evidence exists in Wikipedia, resulting in a dataset well aligned with real-world encyclopedic knowledge. Following the evaluation protocol of PreFLMR, we evaluate on 5K questions from the M2KR subset of the InfoSeek validation split.

Encyclopedic VQA. Encyclopedic VQA focuses on fine-grained entity understanding across natural and landmark categories. Each entity is associated with multiple images and supported by textual evidence drawn from a large, controlled Wikipedia-derived knowledge base. The dataset includes both single-hop and multi-hop questions, enabling evaluation of visual grounding combined with factual reasoning. Consistent with prior work, we evaluate on the official E-VQA test split using only single-hop questions, which provides a clean setting for assessing retrieval quality and answer accuracy via the BEM metric.

FVQA-test. FVQA-test is a curated evaluation set of 2K questions constructed to emphasize factual reasoning grounded in visual evidence. It combines three sources: human-verified samples selected from an automatically generated FVQA pool, re-annotated examples drawn from the InfoSeek Human Split, and newly collected instances by human annotators. Together, these subsets span diverse categories of factual knowledge, requiring the model to jointly interpret the image content and retrieve the appropriate supporting fact. This controlled, carefully validated setup allows precise assessment of factual multimodal reasoning.

OK-VQA. OK-VQA evaluates knowledge-intensive question answering, where answers cannot be derived from the image alone. Questions cover common-sense, cultural, geographic, and scientific knowledge, requiring external information sources to supplement visual understanding. The dataset is widely used to benchmark retrieval-augmented visual reasoning, as models must identify the relevant factual concept and connect it to the visual context in order to generate correct answers. We report results on 5K questions from the validation split, which is commonly used for benchmarking retrieval-augmented VQA systems.

InfoSeek Human. The InfoSeek Human subset, composed of 2K questions used in MMSearch-R1, was drawn from the InfoSeek Human split that demands open-domain retrieval. These samples include entity-level and relational queries where relevant evidence must be located across large textual corpora. The subset captures the retrieval-intensive aspects of InfoSeek while removing questions whose answers can be inferred solely from the image, providing a targeted test bed for evaluating search and reasoning under multimodal constraints.

LiveVQA. LiveVQA evaluates real-world information-seeking under time-sensitive news contents. The benchmark is built from contemporary articles across major global news outlets, each paired with images and automatically generated questions that range from basic visual recognition to multi-hop reasoning over the article's text. Its emphasis on up-to-date events, diverse categories, and mixed reasoning styles makes LiveVQA an effective test of a model's ability to retrieve current information and integrate it with visual cues. We evaluate performance on the 3,602 questions from the preview split, covering all news categories and reasoning types.

MMSearch. MMSearch contains manually curated examples spanning a wide range of real-world domains, divided into knowledge-oriented and news-oriented queries. The benchmark includes a subset of visual questions that require models to perform multimodal retrieval over both general knowledge and rare, specialized facts. Many questions are chosen specifically because leading LLMs struggle to answer them without external search. This makes MMSearch particularly suitable for evaluating agentic or iterative retrieval systems designed for complex information-seeking tasks. For evaluation, we use the visual subset of 171 questions, which isolates multimodal information-seeking scenarios requiring retrieval beyond image content.

F Prompts for LLM-as-Judge

For OK-VQA, FVQA-test, InfoSeek Human, MMSearch, and LiveVQA, we follow the LLM-as-Judge evaluation framework introduced in MMSearch-R1 (Wu et al., 2025). Given an input question, a ground-truth answer, and a model prediction, a judging LLM evaluates whether the prediction is correct, producing both a binary decision and a brief justification. The evaluation focuses on semantic equivalence rather than exact string matching, while enforcing strict correctness for core factual content, names, and numerical values.

Prompt Template for LLM-as-Judge

Input Format:
Question: {question}
Ground Truth Answers: {gold_answer}
Model Response: {model_response}

Evaluation Instructions:
You are an AI assistant tasked with evaluating the correctness of model responses given the Question and Ground Truth answer. Your judgment should follow these principles:

1. Consider the question, and ground truth answer holistically before evaluating the model's response.
2. Your decision should be strictly **Yes or No**, based on whether the model's response is factually accurate and aligns with the ground truth answer.
3. If the model response is a more specific form that includes the ground truth answer, it is correct.
4. If the model response includes all key information but adds minor details, it is correct as long as the extra details are factually correct.
5. If the model response contradicts, modifies, or omits critical parts of the answer, it is incorrect.
6. For numerical values, ensure correctness even when presented in different units.
7. For names, check for first and last name correctness. If the middle name is extra but correct, consider it correct.
8. For yes/no questions, the response must exactly match "Yes" or "No" to be correct.

Evaluate whether the Model Response is correct based on the Question and Ground Truth Answer. Follow the predefined judgment rules and provide a clear Yes/No answer along with a justification.

Output Format:
<reason>Detailed reasoning following the evaluation principles.</reason>
<judge>Yes/No</judge>

Figure A7: **Prompt template for LLM-as-Judge evaluation.** We employ this structured prompt to enforce strict factual consistency while allowing minor semantic variations. The placeholders `{question}`, `{gold_answer}`, and `{model_response}` are populated dynamically for each sample.

Concretely, we use the prompt template shown in Figure A7, where `{question}`, `{gold_answer}`, and `{model_response}` are filled with the corresponding values for each sample.

G Per-iteration Recall Analysis

Table A2 reports a per-iteration recall for the textual KB, multimodal KB, and their heterogeneous combination, offering a fair comparison of how each source contributes during iterative retrieval. Across both InfoSeek and E-VQA, each knowledge source shows incremental gains over iterations, while the heterogeneous setting consistently achieves the highest recall at every step. These results indicate that PMSR's iterative refinement leverages diverse associations from both sources, and that its improvements arise from their complementary signals.

Dataset	Knowledge Source	Recall			
		Iter. 1	Iter. 2	Iter. 3	Iter. 4
InfoSeek	Textual KB	77.9	79.9	80.5	80.8
	Multimodal KB	88.3	88.5	88.2	88.7
	Heterogeneous KB	90.4	90.5	90.4	90.7
E-VQA	Textual KB	35.7	39.2	40.4	41.4
	Multimodal KB	51.4	51.3	51.9	52.4
	Heterogeneous KB	58.5	58.8	59.5	60.6

Table A2: Per-iteration recall of different knowledge sources in PMSR using Qwen3-VL-8B. Heterogeneous KB combines textual KB (R@20) and multimodal KB (R@10) at each iteration.

Method	Knowledge Source	OK-VQA
MMSearch-R1-7B (Wu et al., 2025)	Google Search Google Lens	59.9
DeepMMSearch-R1-7B (Narayan et al., 2025)	Google Search Google Lens	67.8
PMSR (Qwen3-VL-8B)	Wikipedia	66.0

Table A3: Accuracy comparison on the OK-VQA benchmark, comparing our Wikipedia-based approach with the web-equipped baselines from (Narayan et al., 2025).

H Accuracy on OK-VQA

To assess synthesis quality beyond retrieval recall, we further evaluate end-to-end accuracy on OK-VQA using *LLM-as-Judge* protocol. Given the open-ended nature of this benchmark, we compare PMSR against agentic systems that utilize live web search tools. As shown in Table A3, PMSR achieves competitive performance using only the Wikipedia knowledge source, recording 66.0% accuracy with the Qwen3-VL-8B backbone.

I Record-level Query Analysis

Table A4 compares retrieval over the multimodal KB using the question alone versus the record-level query that appends the latest reasoning record. Incorporating the reasoning record consistently improves Recall@5/10/20 (+3.5/+3.4/+1.7 points). These results indicate that the reasoning record provides an additional reasoning-guided signal that helps retrieve more relevant knowledge.

Text Signal	R@5	R@10	R@20
Question only	43.1	48.7	54.5
Question + reason. record	46.6	52.1	56.2

Table A4: Effect of incorporating the latest reasoning record into the record-level query of PMSR for multimodal retrieval on E-VQA, using only the multimodal KB with the default weight λ=0.5.

Model	Retriever Size	BEM	Recall
Qwen3-VL-4B	Small	39.8	61.7
	Large	40.9	68.0
Qwen3-VL-8B	Small	45.5	63.3
	Large	47.1	67.4

Table A5: Impact of retriever scaling on E-VQA, comparing a small retriever (SigLIP2-SO400m + ModernBert-GTE) and a large retriever (SigLIP2-g + Qwen3-Emb).

J Scaling Multimodal Retrievers

We investigate the impact of retriever capacity on overall performance. To this end, we compare a small retriever (SigLIP2-So400m + ModernBERT-GTE) with a large retriever (SigLIP2-giant + Qwen3-Embedding-0.6B). As summarized in Table A5, this scaling yields consistent gains in recall on the E-VQA benchmark. Crucially, these improvements in retrieval show increases in BEM accuracy across both model sizes.

K Top-k Sensitivity Analysis

Model	Metric	$k = 10$	$k = 20$
Qwen3-VL-4B	Acc	40.9	42.4
	Recall	64.3	68.7
Qwen3-VL-8B	Acc	46.4	46.3
	Recall	67.3	72.8

Table A6: Top-k sensitivity analysis of PMSR on the E-VQA benchmark, showing the impact of retrieval budget on different Qwen3-VL models. k denotes the number of retrieved image-text pairs.

Table A6 analyzes the effect of increasing the retrieval budget on performance on E-VQA. Expanding the number of retrieved image–text pairs from $k = 10$ to $k = 20$ consistently improves retrieval recall for both Qwen3-VL-4B and Qwen3-VL-8B, indicating increased evidence coverage. However, answer accuracy shows marginal gains for the 4B model and remains largely unchanged for the 8B model.

L Sensitivity of the Multimodal Similarity Weight

λ	R@5	R@10	R@20
0.3	0.482	0.528	0.570
0.4	0.479	0.527	0.570
0.5	0.466	0.521	0.562
0.6	0.446	0.497	0.546
0.7	0.425	0.474	0.524

Table A7: Ablation of the weight λ for combining the retrieval score with the last reasoning record on E-VQA.

Table A7 reports Recall@5/10/20 for $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Performance peaks at smaller λ (0.3-0.4) and gradually decreases as λ increases, suggesting that maintaining visual relevance is important. At the same time, recall remains competitive across a broad range of λ, indicating that the record-level query provides a useful text signal that complements visual matching.

M Sensitivity to Contextual Noise

To further examine the gap observed in Section 6.1, we analyze the model's sensitivity to contextual noise introduced by retrieved *distractors*. In particular, we test whether adding extra retrieved context can degrade answer prediction even when the oracle evidence is already present in text passages.

We conduct a controlled sensitivity analysis on the E-VQA subset, restricting the evaluation to samples for which oracle textual evidence is exactly available. This setup allows us to isolate the impact of distracting context while holding the presence of correct supporting evidence constant. As summarized in Table A8, conditioning the model solely on the oracle text yields an accuracy of 86.6%. However, augmenting this context with retrieved image–text pairs reduces accuracy to 78.7%, suggesting that visually similar but semantically irrelevant images can interfere with correct entity grounding. When additional textual context retrieved from Google Search is further incorporated, accuracy decreases to 72.5%. Overall, the results highlight that contextual distractors can substantially impair evidence utilization even when correct supporting text is available.

N Trajectory-Type Comparison on E-VQA

To better understand the behavioral differences between progressive record-based updating and

Context Configuration	Accuracy
Oracle Section Text	86.6%
Oracle + 10 Retrieved Pairs	78.7%
Oracle + 10 Retrieved Pairs (w/ Web Search)	72.5%
Oracle + 20 Retrieved Pairs (w/ Web Search)	64.3%

Table A8: Sensitivity analysis on E-VQA using Qwen3-VL-8B. All configurations include the ground-truth (Oracle) evidence. Retrieved pairs use 10 image-text pairs from multimodal KB. The results show that increasing the context of relevant pairs introduces distraction, progressively degrading accuracy.

global-trajectory-only updating under the same KB and retriever, we compare the distributions of reasoning trajectory types on E-VQA between PMSR and WebWatcher.

WebWatcher exhibits a substantial *Correction* rate (31.2%), indicating that it can revise its trajectory and recover from some initially incorrect states. However, it also shows a high *Persistent-Fail* rate (51.8%), meaning that many examples remain incorrect across all iterations. This observation indicates that early failure steps can persist across iterations and continue to influence subsequent actions and reasoning, which may make some initial failures difficult to overcome. Supporting this, *Persistent-Fail* trajectories in WebWatcher take additional iterations on average (2.85 extra iterations, ranging from 1 to 9) without improving final correctness, suggesting that more steps do not necessarily enable recovery in these cases. In contrast, PMSR exhibits a much lower *Persistent-Fail* rate (4.9%) and higher *Stable-Correct* rate (47.7%), while maintaining a comparable *Correction* rate (30.1%).

O Qualitative Examples

This section presents qualitative examples illustrating how PMSR progressively formulates dual-scope queries and constructs structured reasoning records over iterations. Examples are drawn from InfoSeek, FVQA, and E-VQA, using PMSR instantiated with Qwen3-VL-8B. Each figure corresponds to one case and visualizes the iterative trajectory (reasoning records, dual-scope queries, and prediction).

To improve readability, we condense each reasoning trajectory in the figures by retaining at most two representative updates (i.e., up to $t \leq 2$) and omitting minor intermediate details. Specifically, we preserve the key transitions that drive progressive search and reasoning: (i) the initial record that bootstraps the trajectory, (ii) an intermediate update where dual-scope queries retrieve new evidence that revises or sharpens reasoning, and (iii) the final update that resolves the question. For each retained step, we report essential points of the reasoning record (grounded entities, newly retrieved facts, and the resulting inference), while omitting auxiliary text such as partial evidence lists, redundant descriptions, and formatting artifacts. This condensed presentation highlights how PMSR progressively refines its retrieval and reasoning across iterations.

Case 1 (E-VQA: diet of a sea star). As shown in Figure A8, the initial reasoning record r_0 relies on generic sea-star knowledge and contains only loosely related evidence, which is insufficient to answer the question. In subsequent iterations, PMSR decomposes retrieval into dual scopes: the record-level query targets the most recent uncertainty by refining species-level grounding, while the trajectory-level query preserves the overall intent of retrieving diet knowledge for the grounded entity. This reasoning-guided retrieval surfaces species-specific passages for *Pacific blood star* (*Henricia leviuscula*), enabling PMSR to update the record with precise dietary information and converge on the correct answer, *sponges and small bacteria*.

Case 2 (E-VQA: geographic region of a shrike). This example is challenging due to visually similar shrike species, which can induce errors in early grounding and retrieval (Figure A9). Across iterations, PMSR retrieves knowledge of specific species that better match the visual cues, enabling later records to recover from early confusion and finalize the correct region (*North America*).

Case 3 (E-VQA: native range of a plant). The initial record exhibits an early grounding failure, identifying an incorrect visual entity (e.g., *box-elder*), which leads to only coarse and partially mismatched regional knowledge (Figure A10). Subsequent dual-scope queries improve retrieval toward discriminative visual attributes (e.g., the distinctive large, spherical fruit) while maintaining the trajectory's objective of resolving the plant's native range. Across iterations, PMSR identifies the plant as *Maclura pomifera* (Osage orange) and describes its native distribution, enabling later records to correct the initial grounding and retrieve corresponding knowledge. The final trajectory converges

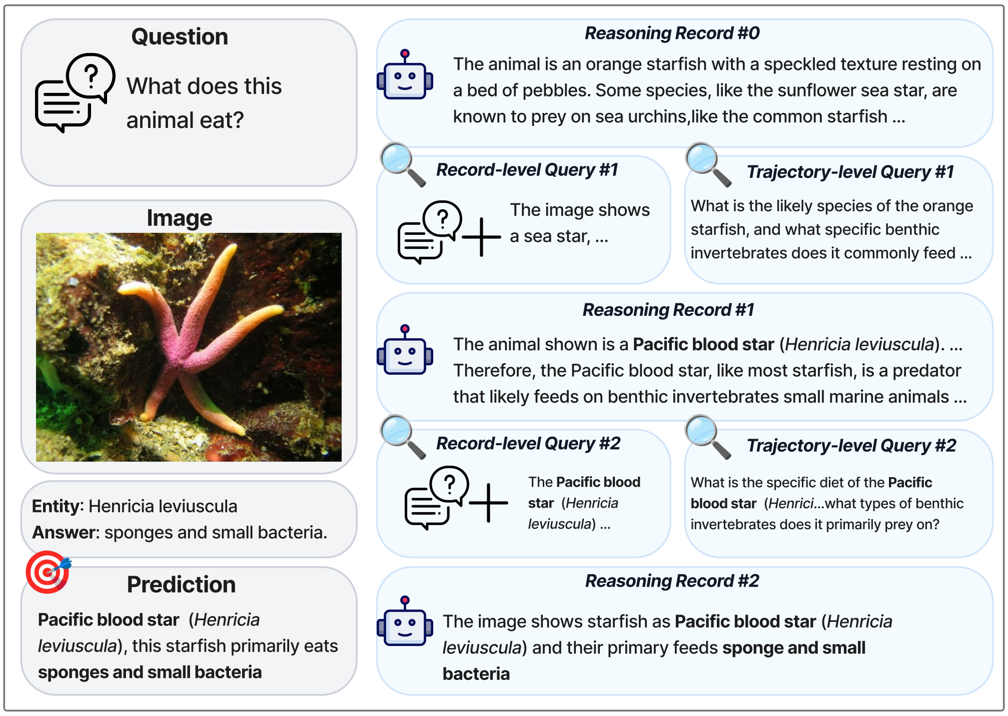

Figure A8: **E-VQA case: diet of a sea star.** PMSR progressively refines visual grounding and retrieves entity-specific evidence via dual-scope queries, enabling the reasoning records to converge to the correct diet.

to the correct answer (*south-central United States*), illustrating how PMSR recovers from early grounding errors through progressive retrieval and record updates.

Case 4 (F-VQA: chemical class of molecules). The initial reasoning record provides a broad classification that is correct but underspecified for the question (Figure A11). PMSR then uses dual-scope querying: the record-level query seeks discriminative evidence (i.e., the shared structural signature), while the trajectory-level query focuses on the depicted molecules that share a common motif. Across iterations, PMSR retrieves evidence highlighting that the depicted molecules contain sulfur atoms within an organic framework, enabling the synthesized record to resolve the intended class (*organosulfur compounds*). This example illustrates how PMSR refines from generic to specific knowledge through iterative retrieval and reasoning-record updates.

Case 5 (InfoSeek: downstream water body of an urban river). In Figure A12, the initial record grounds the scene as the Rotterdam cityscape and identifies the river as the Nieuwe Maas, but the question requires the *immediate* water body it drains into rather than the eventual outlet. In subsequent iterations, PMSR's record-level query focuses on confirming the river identity and its downstream connection, while the trajectory-level query targets the broader river-network relation. Across iterations, PMSR retrieves knowledge corresponding to the Nieuwe Maas river system, indicating that it joins the Oude Maas near Vlaardingen and drains into the Het Scheur, which then continues as the Nieuwe Waterweg toward the North Sea. Synthesizing this knowledge, the latest reasoning record resolves the intended answer as *Het Scheur*.

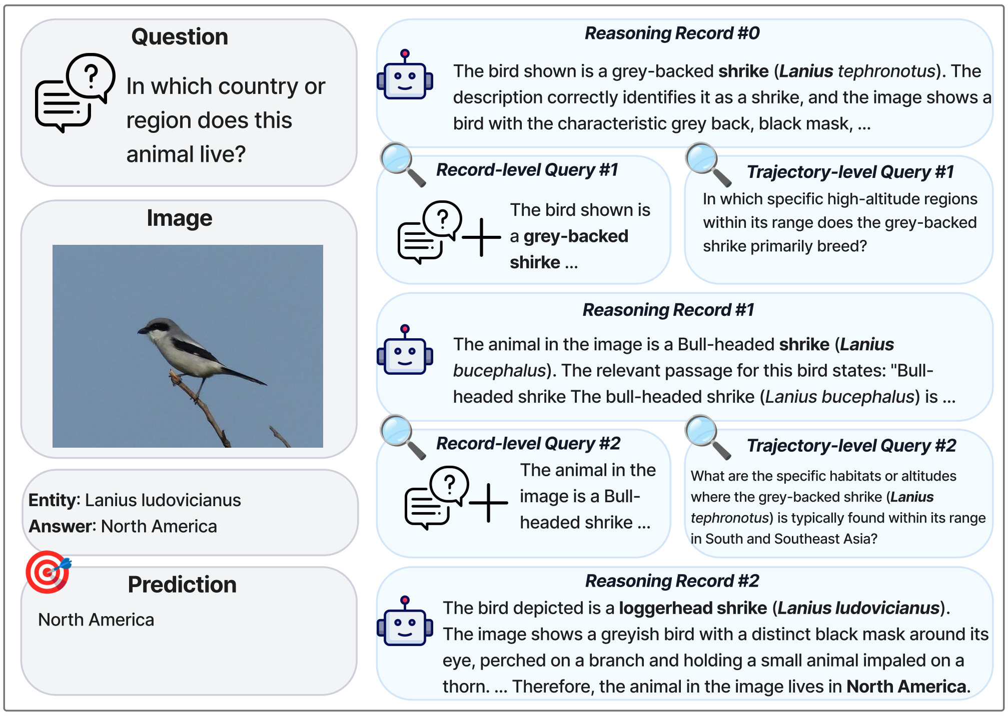

Figure A9: **E-VQA case: geographic region of a shrike.** Dual-scope queries mitigate early mis-grounding among visually similar species and retrieve the knowledge of the ground entity.

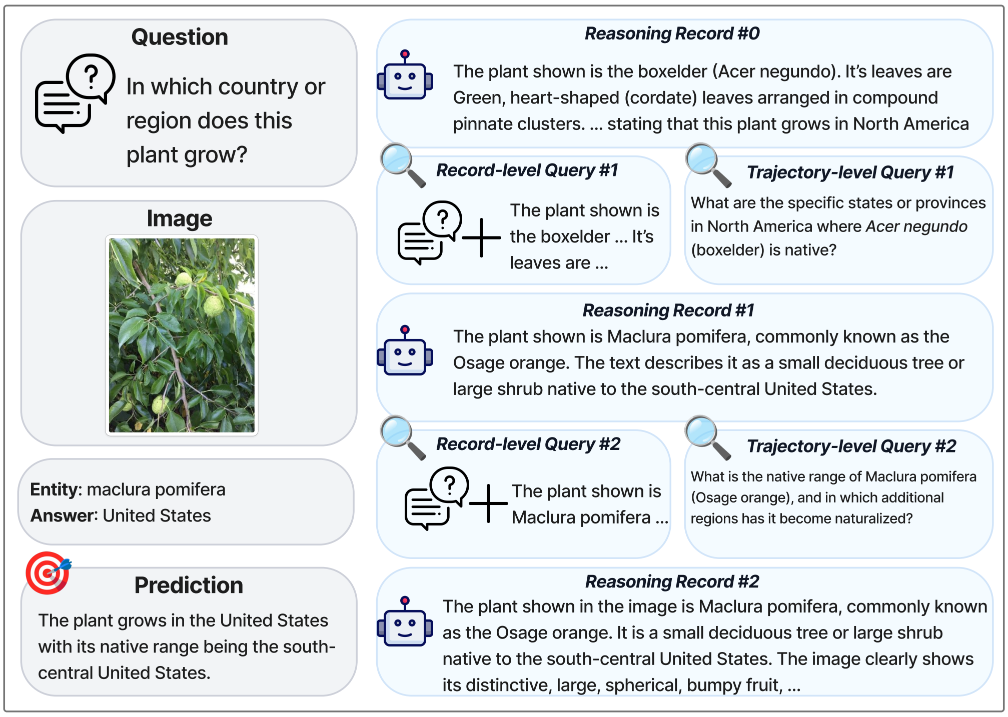

Figure A10: **E-VQA case: native range of a plant.** PMSR progressively aligns retrieved encyclopedic evidence with discriminative visual attributes to resolve the plant identity and its native distribution.

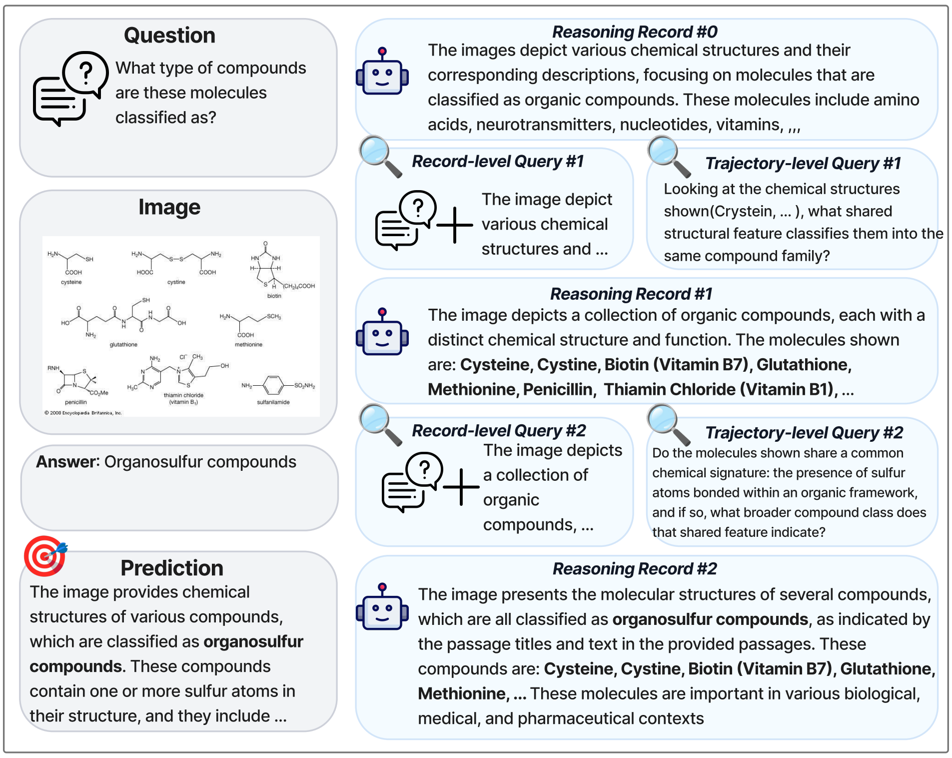

Figure A11: **F-VQA case: chemical class of molecules.** PMSR narrows the classification from an overly broad concept to the targeted class by retrieving knowledge about shared structural motifs and updating the reasoning record accordingly.

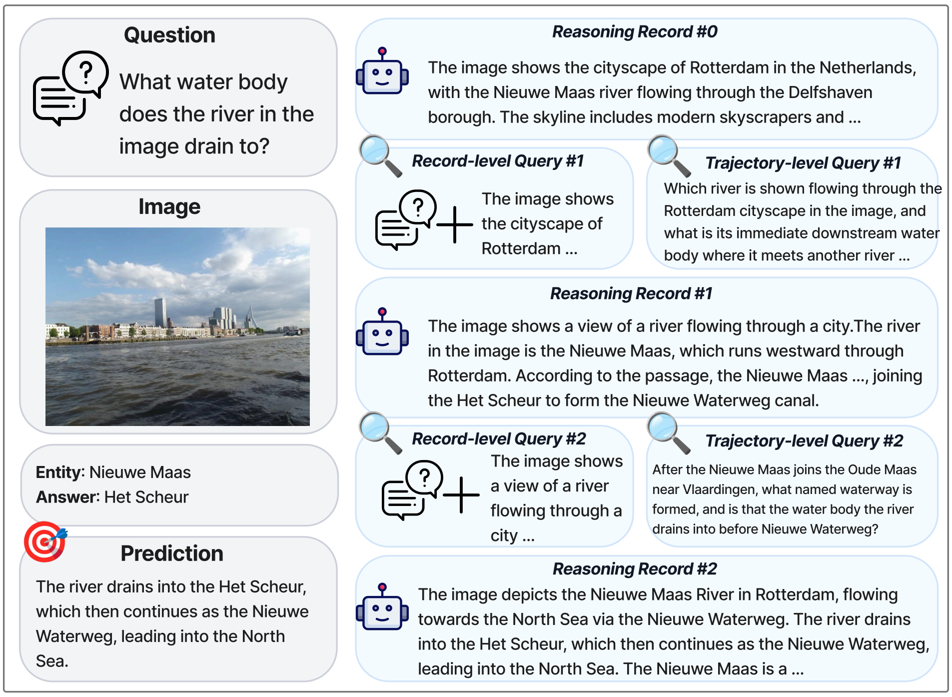

Figure A12: **InfoSeek case: downstream water body of an urban river.** PMSR refines visual grounding from coarse geographic grounding (Rotterdam cityscape) to a fine-grained prediction of the river's immediate downstream water body.

P Computational Cost Analysis

We analyze the per-sample computational cost of PMSR with three iterations on a commodity server (NVIDIA RTX 3090, Intel Xeon Gold 6248R) using local retrieval. Under this setup, PMSR requires 17.2 seconds per iteration to process one sample end-to-end. Importantly, this latency is not inherent to the framework but mainly reflects an unoptimized implementation.

The overhead stems primarily from (i) *serial* query execution across multiple knowledge bases with vanilla index configurations, and (ii) model inference without an optimized serving stack (e.g., vLLM or SGLang). Both factors are amenable to system-level optimization, such as parallelizing retrieval across sources, tuning/accelerating indexing and search, and using optimized inference runtimes.