CLIMB: Confidence-Guided Complementary Evidence for Multimodal Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have shown strong visual reasoning abilities, but knowledge-intensive visual question answering often requires external textual evidence beyond the image and the model's parametric knowledge. Existing multimodal RAG systems commonly rely on Top-K retrieval or reranking, which may return redundant passages and provide limited control over whether an answer update is sufficiently supported by the retrieved evidence. We propose *CLIMB*, a training-free inference-time framework for multimodal RAG. CLIMB first constructs a compact complementary evidence pool using an MMR-style objective that balances query relevance and passage-level redundancy. It then performs confidence-controlled refinement within this fixed pool: an R/E/C critic scores passages by relevance, evidence specificity, and cross-modal alignment, while an evidence-grounded confidence estimator accepts an updated answer only when the estimated confidence increases. This design provides a simple stopping criterion and reduces unnecessary refinement without modifying the underlying retriever or MLLM. Experiments on Encyclopedic-VQA and InfoSeek show that CLIMB consistently improves over retrieval-augmented multimodal baselines. Ablations further indicate that complementary pooling, critic-based scoring, and iterative confidence-controlled refinement each contribute to the final performance.

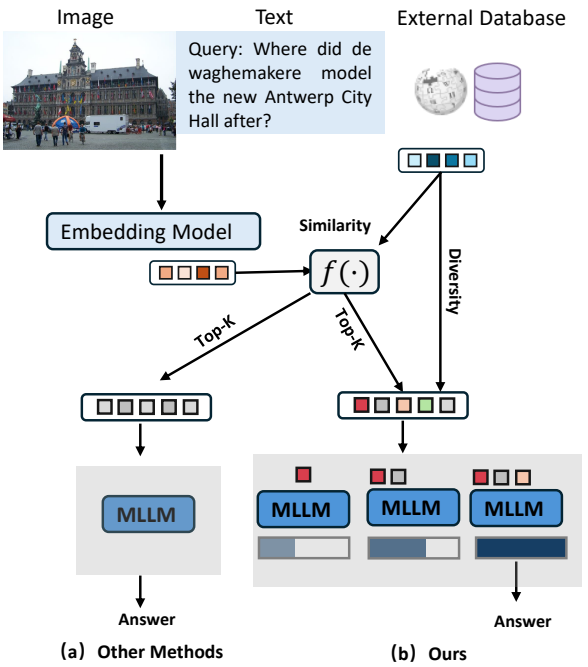


Figure 1: Comparison of multimodal RAG approaches. (a) Standard methods (Cocchi et al., 2025; Caffagni et al., 2024b; Yan and Xie, 2024) rely on Top-K similarity selection, which often produces redundant evidence with limited coverage and lacks explicit quality control for answer generation. (b) Our CLIMB framework addresses these limitations through two key innovations: diversity-aware evidence construction that balances relevance and complementarity to build a non-redundant passage pool, and confidence-controlled decoding that uses multi-faceted scoring and monotonic confidence constraints to ensure reliable answer updates.

1 Introduction

Multimodal large language models (MLLMs) (Liu et al., 2024a; Li et al., 2024; Abdin et al., 2024) have made rapid progress in visual reasoning and grounded dialogue, yet knowledge-intensive visual question answering remains challenging (Caffagni et al., 2024a; Wu et al., 2023; Yin et al., 2024). In these tasks, the image provides important grounding cues, but the answer often depends on external facts such as entity attributes, historical details, geographic information, or fine-grained taxonomy. Such information is difficult to cover fully through parametric knowledge alone, especially for long-tail entities and time-sensitive facts (Ye et al., 2024; Liu et al., 2024b; Li et al., 2023; Chen et al., 2023; Mensink et al., 2023).

Retrieval-augmented generation (RAG) addresses this limitation by providing external textual evidence at inference time (Lewis et al., 2020;

Zhao et al., 2023; Liu et al., 2025). In multimodal RAG, however, performance depends not only on retrieving relevant passages, but also on assembling a compact evidence set that is complementary, non-redundant, and sufficiently reliable for answer generation. A retrieval set that contains several near-duplicate passages may have high similarity scores but still fail to cover the specific facets needed to answer the question.

Despite steady advances, common multimodal RAG pipelines exhibit limitations rooted in design choices. The prevailing practice ranks passages by image-text similarity and conditions the MLLM on the top-ranked set. Such Top-K assembly concentrates scores on near-duplicates (Yan and Xie, 2024), producing redundant contexts with limited semantic coverage, reflecting the core challenge of *evidence complementarity*. Enhancements in retrievers and hierarchical search reduce the candidate space (Caffagni et al., 2024b), yet the final assembly step rarely suppresses pairwise redundancy in a principled way. Iterative RAG variants (Arefeen et al., 2024) attempt to address insufficient evidence through multiple retrieval-and-rewrite cycles, but face the challenges of *quality control* and *termination criteria*: these systems increase latency and variance while offering limited transparency about when updates should be trusted or when to stop refinement. Most architectures also lack explicit mechanisms for reliability assessment, leaving systems unable to determine when evidence is sufficient. These issues motivate an inference-time mechanism that jointly considers evidence complementarity and answer-level verifiability.

To tackle these challenges, we propose *CLIMB*, a framework that treats multimodal RAG as two coupled decisions: which evidence should be assembled, and when an answer update should be accepted. CLIMB first builds a fixed complementary micro-pool using an MMR-style objective that balances query relevance with redundancy suppression. This pool provides a compact set of candidate passages that are individually relevant while covering different aspects of the question. Within this fixed pool, CLIMB performs confidence-controlled refinement. A passage-level R/E/C critic scores each candidate by relevance, evidence specificity, and cross-modal alignment. The model then generates an answer from the selected subset, estimates an evidence-grounded confidence score, and accepts a revised answer only if the estimated confidence improves. Because all refinement steps operate inside the fixed pool, CLIMB avoids repeated external retrieval and provides a bounded, interpretable stopping rule. Empirically, CLIMB delivers consistent gains on knowledge-intensive VQA benchmarks. Across datasets such as Encyclopedic-VQA(Mensink et al., 2023) and InfoSeek(Chen et al., 2023), the framework improves over Top-K retrieval and iterative RAG variants, and it preserves performance on instances where additional evidence is unnecessary. Ablations indicate that critic is a primary driver of coverage and answer quality, while confidence-controlled acceptance contributes to stability, calibration, and early stopping. Our contributions are summarized as follows:

1. We present CLIMB, which constructs a complementary evidence pool via an MMR-style objective that promotes relevance while suppressing redundancy.

2. Building on this pool, we introduce confidence-controlled decoding with an R/E/C scoring interface and an acceptance rule that updates only when a scalar confidence increases, enabling monotone improvement and natural early stopping.

3. We systematize multimodal RAG as a complementarity-aware problem, providing a modular pipeline that integrates with standard encoders and retrievers and aligns with hierarchical retrieval and reflective-control designs.

2 Related Work

Multimodal large language models (MLLMs) extend text-only LLMs by coupling a vision encoder with a language backbone, enabling grounded question answering, visual dialogue, and open-ended multimodal generation (Radford et al., 2021; Liu et al., 2023; McKinzie et al., 2024; Sun et al., 2024). Architecturally, systems differ in how they bridge modalities, from query-former style adaptors and projection layers to end-to-end finetuning, and in how they leverage instruction tuning and conversational data for alignment (Radford et al., 2021; Liu et al., 2024b; Lerner et al., 2024; Qi et al., 2024). Subsequent work improves robustness and utility through better pretraining corpora, mixed-format curricula, and task-oriented instruction sets, yielding models that can act as general-purpose perceptual reasoners under open-world conditions

(Alayrac et al., 2022; Changpinyo et al., 2021; Gadre et al., 2023; Deitke et al., 2024).

Another line of research showed great potential of Retrieval-augmented generation (RAG), which injects non-parametric evidence into the context to expand coverage and mitigate hallucination, first established in text settings by combining sparse or dense retrievers with sequence generators (Lewis et al., 2020; Guu et al., 2020; Izacard et al., 2023; Borgeaud et al., 2022). Early systems such as Fusion-in-Decoder and subsequent large-scale variants demonstrated that learned retrieval over web-scale corpora, together with retrieval-conditioned decoding, substantially improves knowledge-intensive QA (Gao et al., 2023; Arslan et al., 2024; Fan et al., 2024). As instruction-tuned LLMs became the default interface, RAG shifted toward inference-time augmentation: passages are formatted into input sequences, sometimes routed or filtered by reflection and verifier modules that decide whether to retrieve, how much to use, and when to revise, with attention to calibration and reproducibility of the retrieval–decoding loop (Asai et al., 2024; Yang et al., 2025). Extending RAG to multimodal LLMs introduces image-conditioned retrieval and multimodal context assembly. Recent pipelines structure candidates hierarchically, coarse selection guided by image or title cues followed by question- and text-level refinement, while keeping the MLLM backbone intact (Caffagni et al., 2024b; Yan and Xie, 2024). Other approaches integrate reflective signals into decoding to regulate access to external knowledge for knowledge-based VQA (Cocchi et al., 2025). These developments collectively move from simply "finding relevant text" toward assembling an image- and question-aware context that a multimodal decoder can reliably condition on. However, assembling a compact, complementary context and governing acceptance by answer reliability remain under-specified at inference time. Our CLIMB framework targets this gap by pairing an explicitly complementary, fixed evidence pool with confidence-governed decoding, providing a protocol for both diversity-aware assembly and monotone, reliability-driven refinement on top of standard multimodal RAG pipelines.

3 Method

CLIMB is an inference-time framework for multimodal RAG with two stages: complementarity-aware evidence assembly and confidence-controlled decoding. Given an image, a question, and an external corpus, CLIMB first builds a compact fixed-size evidence pool. It then iteratively selects evidence subsets and accepts an updated answer only when the estimated evidence-grounded confidence improves. The base MLLM is kept frozen and treated as a black-box generator. Figure 2 provides an overview.

3.1 Problem Formulation

We consider knowledge-intensive visual question answering. Each example consists of an image I, a question q, and an external text corpus $\mathcal{D}$. The goal is to generate an answer a supported by a subset of passages $\mathcal{S} \subset \mathcal{D}$, and a scalar confidence $U \in [0, 1]$ indicating how well the answer can be verified. CLIMB treats the MLLM as a frozen generator and focuses on inference-time evidence selection. Instead of repeatedly querying the full corpus during refinement, CLIMB first constructs a fixed complementary micro-pool $\mathcal{P} \subset \mathcal{D}$ of size K. All subsequent scoring, facet diagnosis, and answer refinement operate within $\mathcal{P}$.

3.2 Complementary Evidence Pool

Vision–text retrieval typically ranks passages by similarity to the image or the query, often surfacing a cluster of near-duplicates at the top. Such redundancy limits semantic coverage and leaves later reasoning steps without the diverse evidence needed for knowledge-intensive VQA. *CLIMB* therefore constructs a compact *micro-pool* of size K whose members are individually relevant yet mutually non-overlapping.

Given an image I and question q, a standard CLIP-based search retrieves a candidate set $\mathcal{C}_N \subset \mathcal{D}$. We reuse the retriever's similarity as a query–passage relevance score $\mathrm{rel}(I, q, t)$ for each $t \in \mathcal{C}_N$. From the screened candidates, we assemble a micro-pool $\mathcal{P}$ using a Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) objective. Let S denote the set of already admitted items, and let $\mathrm{sim}(t, u)$ be cosine similarity between the text embeddings of t and u. Starting from the most relevant passage, each subsequent passage is chosen by

$$t^\star = \arg\max_{t \in \mathcal{C}_N \setminus S} \{\lambda\, \mathrm{rel}(I, q, t) - (1-\lambda) \max_{u \in S} \mathrm{sim}(t, u)\}. \quad (1)$$

where $\lambda \in [0, 1]$ balances query fit and redundancy

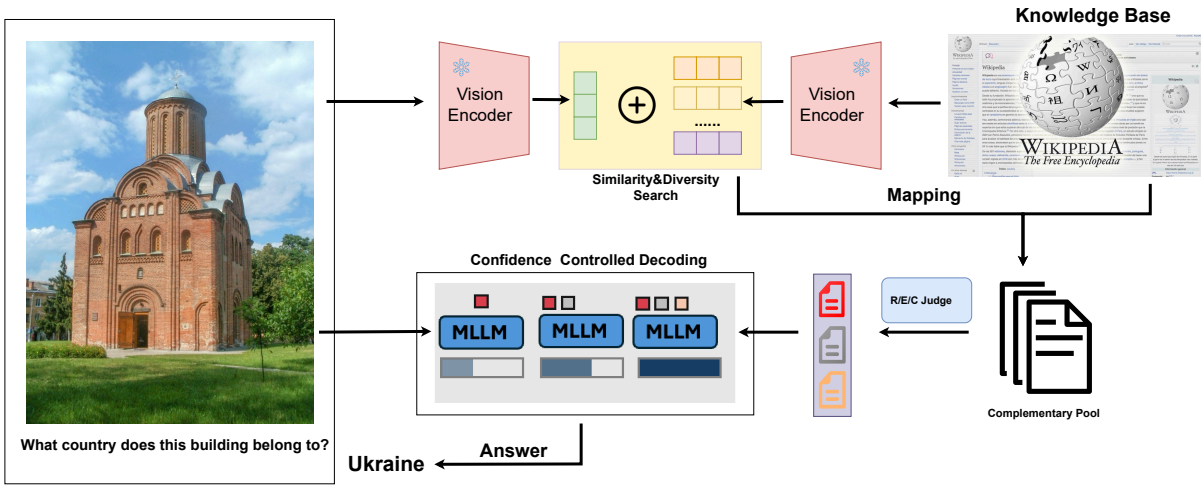


Figure 2: **Overview of *CLIMB*.** Given an image and a question, CLIMB retrieve a small candidate set from the knowledge base and build a complementary evidence pool via similarity–diversity selection. An R/E/C critic scores passages, and a frozen MLLM generates answers from the top-M evidence while a confidence estimator and gap diagnosis drive an iterative refinement loop, accepting updates only when confidence increases.

(default $\lambda = 0.6$). Penalizing the maximum similarity discourages the strongest duplication, so that each admitted passage is encouraged to contribute a distinct facet of the question.

With cached embeddings, this one-pass selection costs $O(NK)$ for a candidate set of size N and a pool of size K. The resulting micro-pool is fixed for the remaining refinement steps. This design keeps the evidence search space bounded and reduces variation caused by repeated external retrieval. We set $K = 5$ by default and analyze its effect in the sensitivity study.

3.3 Evidence Judge and Confidence Control

Once the micro-pool $\mathcal{P}$ is fixed, CLIMB evaluates evidence at the passage level and answer level.

R/E/C evidence judge. For each passage $t \in \mathcal{P}$, CLIMB assigns three interpretable scores. The relevance score $R \in \{1, \ldots, 5\}$ measures how directly t addresses the question. The evidence specificity score $E \in \{1, \ldots, 5\}$ measures whether t contains concrete and verifiable details, such as named entities, dates, numbers, or definitions. The cross-modal alignment score $C \in \{1, \ldots, 5\}$ measures whether t is consistent with the visual content of I. These scores are produced by a constrained judgment prompt to the frozen MLLM. After normalization to $[0, 1]$, they are combined as

$$\phi(t) = \alpha_R \tilde{R}(t) + \alpha_E \tilde{E}(t) + \alpha_C \tilde{C}(t), \quad (2)$$

where α_R, α_E, and α_C are non-negative weights that sum to one. We use equal weights by default. The initial evidence subset is selected as

$$\mathcal{S}_0 = \arg\operatorname*{topM}_{t \in \mathcal{P}} \phi(t), \qquad M = \lceil \sqrt{K} \rceil. \quad (3)$$

Verifiability-oriented confidence. Given an answer a generated from an evidence subset $\mathcal{S}$, CLIMB estimates a scalar confidence score $U \in [0, 1]$ based on three bounded signals: coverage, consistency, and specificity. Coverage measures whether the main factual claims in a are supported by passages in $\mathcal{S}$. Consistency penalizes conflicts among the selected passages or between the passages and the image. Specificity rewards answers that contain concrete and verifiable information. The three scores are averaged to obtain U. During refinement, U is used as a relative acceptance signal: a new answer is accepted only if its confidence strictly increases, and refinement stops once U exceeds a target threshold U_{STAR}. In the iterative refinement procedure that follows, this confidence is used relatively: a new answer is accepted only if its confidence strictly increases, and refinement stops early once U exceeds a target threshold U_{STAR} or a maximum number of rounds is reached.

3.4 Facet Diagnosis and Pool-Internal Refresh

Some errors arise when the current subset $\mathcal{S}_t$ misses a key facet of the question. Rather than performing another external retrieval, *CLIMB* diagnoses under-covered facets and refreshes evidence *within* the fixed micro-pool $\mathcal{P}$.

Given the current state $(I, q, \mathcal{S}_t, a_t)$, we elicit at most two short facet phrases

$$\mathcal{G}_t = \{g_1, g_2\},$$

each describing a missing aspect that prevents full support (we constrain each g_i to be a short phrase). If no meaningful facet can be identified, refinement terminates. For every passage $u \in \mathcal{P}$, we then compute a lightweight facet–text match score $\psi(u \mid \mathcal{G}_t)$ using a combination of lexical overlap (e.g., BM25) and sentence-level embedding similarity. The original evidence scores $\phi(u)$ are updated as

$$\phi'(u) = \phi(u) + \gamma\, \psi(u \mid \mathcal{G}_t), \qquad \gamma \in [0, 1], \quad (4)$$

and the next working subset $\mathcal{S}_{t+1}$ is chosen as the top-M passages under ϕ'.

CLIMB then queries the frozen MLLM once with $(I, q, \mathcal{S}_{t+1})$ to obtain a new answer a_{t+1} and compute its confidence U_{t+1}. The update is accepted only if $U_{t+1} > U_t$; otherwise the procedure stops and returns $(a_t, U_t, \mathcal{S}_t)$. In practice we also cap the number of refinement rounds by a small constant $T_{\max}$ (we use $T_{\max} = 3$) and stop early once $U_t \geq U_{\text{STAR}}$. Because all operations remain inside the fixed micro-pool, *CLIMB* avoids re-retrieval, keeps latency predictable, and enforces a monotone improvement in verifiability across accepted updates.

3.5 Algorithm

The end-to-end procedure of *CLIMB* is summarized in Algorithm 1. The algorithm take $(I, q, \mathcal{D})$ as the input and output the answer, confindence and supporting set. Lines 2-3 perform a small vision–text retrieval followed by lightweight relevance filtering to obtain a candidate set $\mathcal{C}_N$. Lines 4-5 construct a complementary micro-pool $\mathcal{P}$ of size K via BUILDMICROPOOL, which implements the MMR objective in Eq. 1 to balance query relevance against pairwise redundancy. In lines 7-9, passages in $\mathcal{P}$ are scored by the R/E/C critic to produce ϕ, the top-M passages form the initial subset $\mathcal{S}_0$, and a single call to the frozen MLLM yields the initial answer and confidence (a_0, U_0). Lines 11-16 implement the confidence-guided refinement. At each iteration, *CLIMB* first checks whether the current confidence has already exceeded the target threshold U_{STAR}; if so, the procedure terminates. It then diagnoses up to two under-supported facets $\mathcal{G}_t$ from the current state $(I, q, \mathcal{S}_t, a_t)$; if no facet is found, refinement stops. Otherwise, lines 17-19 refresh the evidence scores inside the fixed pool using REFRESHSCORES, select a new subset $\mathcal{S}_{t+1}$, and obtain an updated answer and confidence (a_{t+1}, U_{t+1}). The update is accepted only when $U_{t+1} > U_t$; if confidence fails to increase, the loop exits and returns the last accepted $(a_t, U_t, \mathcal{S}_t)$. This fixed-pool, monotone confidence climb avoids retrieval drift, stabilizes behavior across runs.

Algorithm 1: CLIMB

Input: Image I, Question q, Corpus $\mathcal{D}$
Output: Answer a, Confidence U, Supporting $\mathcal{S}$

1 **Candidate Retrieval**;
2 $\mathcal{C}_N \leftarrow$ RETRIEVE$(I, q, \mathcal{D})$;
3 $\mathcal{C}_N \leftarrow$ FILTERBYREL$(\mathcal{C}_N, \tau)$;
4 **Complementary Pool**;
5 $\mathcal{P} \leftarrow$ BUILDMICROPOOL$(I, q, \mathcal{C}_N, K)$;
6 **Initial Scoring and Answer**;
7 $\phi \leftarrow$ SCOREREC$(I, q, \mathcal{P})$;
8 $\mathcal{S}_0 \leftarrow$ TOPM$(\mathcal{P}, \phi, M)$;
9 $(a_0, U_0) \leftarrow$ GENERATEANDSCORE$(I, q, \mathcal{S}_0)$;
10 **Iterative Confidence Climb**;
11 **for** $t = 0$ **to** $T_{\max} - 1$ **do**
12 **if** $U_t \geq U_{\text{STAR}}$ **then**
13 **break**;
14 $\mathcal{G}_t \leftarrow$ DIAGNOSEFACETS$(I, q, \mathcal{S}_t, a_t)$;
15 **if** $\mathcal{G}_t = \emptyset$ **then**
16 **break**;
17 $\phi' \leftarrow$ REFRESHSCORES$(\mathcal{P}, \phi, \mathcal{G}_t, \gamma)$;
18 $\mathcal{S}_{t+1} \leftarrow$ TOPM$(\mathcal{P}, \phi', M)$;
19 $(a_{t+1}, U_{t+1}) \leftarrow$ GENERATEANDSCORE$(I, q, \mathcal{S}_{t+1})$;
20 **if** $U_{t+1} \leq U_t$ **then**
21 **break**;
22 **return** a_t, U_t, $\mathcal{S}_t$;

4 Experiments

4.1 Datasets

We evaluate CLIMB on two knowledge-intensive VQA benchmarks that require combining image understanding with external Wikipedia knowledge.

Encyclopedic-VQA. Encyclopedic-VQA contains image–question pairs grounded in Wikipedia entities. Questions target fine-grained factual properties such as species attributes, geographic information, and historical facts. Following prior work, we use the standard single-hop setting and report the BERT matching score (BEM).

Model	LLM	Retrieval Mode	E-VQA		InfoSeek		
			Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot LLMs							
Vanilla	Vicuna-7B	-	2.1	2.0	0.3	0.0	0.0
Vanilla	LLaMA-3.1-8B	-	16.5	16.6	2.1	0.0	0.0
Zero-shot Multimodal LLMs							
LLaVA-v1.5	Vicuna-7B	-	16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5	LLaMA-3.1-8B	-	16.0	16.9	8.3	8.9	7.8
Retrieval-Augmented Models							
$\text{DPR}_{\text{V+T}}$ [ECIR24(Lerner et al., 2024)]	Multi-passage BERT	CLIP ViT-B/32	29.1	-	-	-	12.4
RORA-VLM [arXiv24(Qi et al., 2024)]	Vicuna-7B	CLIP+Google Search	-	20.3	25.1	27.3	-
Wiki-LLaVA [CVPRW24(Caffagni et al., 2024b)]	LLaMA-3.1-8B	CLIP ViT-L/14	18.3	19.6	28.6	25.7	27.1
EchoSight [EMNLP24(Yan and Xie, 2024)]	LLaMA-3.1-8B	EVA-CLIP-8B	22.4	21.7	30.0	30.7	30.4
ReflectiVA [CVPR25(Cocchi et al., 2025)]	LLaMA-3.1-8B	CLIP ViT-L/14	24.9	26.7	34.5	32.9	33.7
▶ Climb	LLaMA-3.1-8B	CLIP ViT-L/14	**31.3**	**34.8**	**39.7**	**38.5**	**39.1**

Table 1: Accuracy on the Encyclopedic-VQA test set and the InfoSeek validation set.

InfoSeek. InfoSeek contains fact-seeking visual questions over Wikipedia entities, including both automatically generated and human-authored examples. Compared with Encyclopedic-VQA, it covers broader entity types and question templates. We follow the official split and evaluation protocol, restrict retrieval to the provided knowledge base, and report VQA accuracy and relaxed accuracy when applicable.

4.2 Baselines

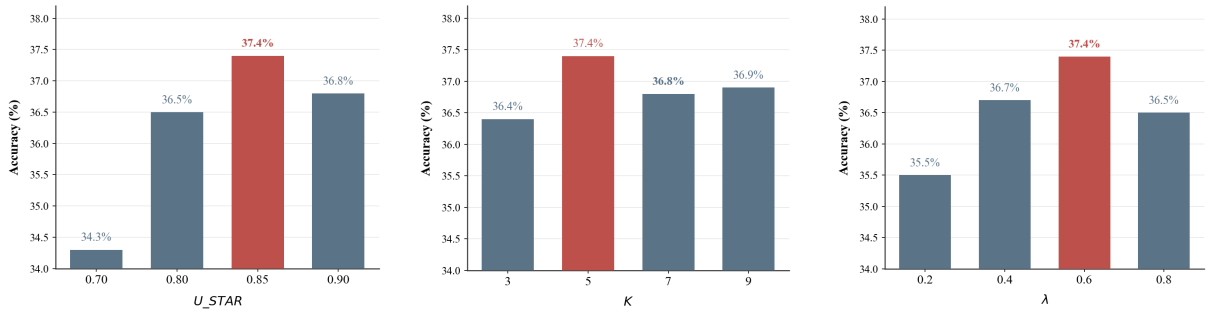
We compare CLIMB with three groups of baselines. First, we include text-only LLMs, Vicuna-7B (Zheng et al., 2023) and LLaMA-3.1-8B (Grattafiori et al., 2024), which receive only the question and have no access to the image or external corpus. Second, we evaluate zero-shot multimodal LLMs, including LLaVA-v1.5 (Liu et al., 2024a) with Vicuna-7B and LLaMA-3.1-8B backbones, which use the image but do not retrieve external knowledge. Third, we compare with retrieval-augmented multimodal systems, including DPRV+T (Lerner et al., 2024), RORA-VLM (Qi et al., 2024), Wiki-LLaVA (Caffagni et al., 2024b), EchoSight (Yan and Xie, 2024), and ReflectiVA (Cocchi et al., 2025). These baselines use external knowledge through retrieval or filtering and therefore provide the most relevant comparison for evaluating CLIMB's complementary evidence assembly and confidence-controlled refinement.

4.3 Implementation Details

We use CLIP ViT-L/14@336px as the visual encoder to obtain image embeddings and image–text similarity scores for retrieval, and LLaMA-3.1-8B model as the answer generator. For each query, we first retrieve N=64 candidate passages from the dataset-specific Wikipedia collection using cosine similarity in the CLIP embedding space, and then construct a complementary evidence pool of size K=5 with the similarity&diversity objective. On top of this pool, we compute relevance, evidence specificity, and cross-modal alignment scores for all passages and calibrate them on a held-out development split to obtain a unified ranking signal. During decoding, the generator is conditioned on the top passages via a fixed structured instruction template, and our confidence module evaluates each candidate answer using coverage completeness, inter-evidence and image consistency, and answer specificity. Facet diagnosis and pool-internal refresh are applied for at most $T_{\max}$=3 rounds, and an updated answer is accepted only when it yields a strictly higher calibrated confidence score. All results are obtained under the knowledge-base-restricted protocol of each benchmark.

4.4 Overall Effectiveness

Main results. Table 1 reports accuracy on Encyclopedic-VQA and InfoSeek. Text-only LLMs provide a very weak baseline, and zero-shot multimodal LLMs improve mainly by grounding the question in the image, but still stay around 17% on Encyclopedic-VQA and below 10% on InfoSeek. This shows that visual grounding and parametric knowledge alone are insufficient for encyclopedic VQA over large knowledge bases. Retrieval-augmented systems narrow this gap: Wiki-LLaVA, EchoSight, and ReflectiVA all benefit from access to Wikipedia, lifting accuracy into the 20–30% range and confirming the importance of explicit evidence retrieval. EchoSight uses a stronger retriever (EVA-CLIP-8B) than Wiki-LLaVA but does not consistently outperform ReflectiVA, suggesting that retrieval strength alone is not enough when the assembled contexts remain redundant. *CLIMB* achieves the best performance across all evaluation splits and consistently improves over the strongest

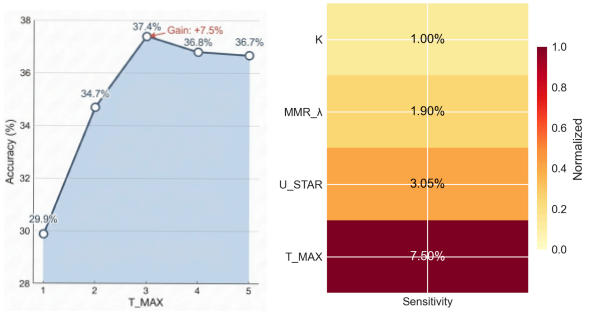


(a) Answer accuracy across different confidence thresholds U_STAR ∈ {0.70, 0.80, 0.85, 0.90}.

(b) Answer accuracy across different micro-pool sizes K ∈ {3, 5, 7, 9}

(c) Answer accuracy across different MMR lambda values λ ∈ {0.2, 0.4, 0.6, 0.8}.

Figure 3: Sensitivity scores for hyperparameters U_{STAR}, K, λ, measured as the accuracy range across different parameter settings.



(a) (b)

Figure 4: Hyperparameter sensitivity of *CLIMB*. (a) Accuracy across different maximum iteration $T_{\max}$. (b) Aggregated sensitivity scores across key hyperparameters, indicating that the framework is stable within a broad range of settings.

retrieval baseline. On Encyclopedic-VQA, it yields absolute gains of 6.4 on the Single-Hop subset and 8.1 on the All setting. The gains on Unseen-Q and Unseen-E are 5.2 and 5.6, respectively, indicating that *CLIMB* not only retrieves useful passages but also regulates how they are selected. On InfoSeek, *CLIMB* further improves accuracy by 5.4 over the best baseline, despite the different knowledge base and question style. These results suggest that complementary evidence construction and confidence-controlled decoding provide benefits beyond those obtained by stronger retrievers alone.

4.5 Ablation Study

Method	E-VQA	InfoSeek
Full	**34.8**	**39.1**
w/o Facet Diagnosis	33.7	38.7
w/o Complementary Evidence Pool	32.4	38.1
w/o Critic	26.9	33.4
w/o Iteration	25.8	33.0

Table 2: Ablation study results on InfoSeek and E-VQA.

Table 2 reports the contribution of each component in *CLIMB*.Removing the facet diagnosis step produces a small but consistent drop on both datasets, indicating that explicitly identifying under-supported facets helps the model redirect attention within the fixed pool. Eliminating the complementary pool (MMR) leads to a larger degradation, showing that suppressing redundancy is important for assembling evidence that covers multiple attributes of the queried entity. The critic, responsible for R/E/C scoring and confidence estimation, is the most influential component: without it, accuracy decreases sharply on both benchmarks, confirming that a notion of verifiability is important for filtering misleading passages and stabilizing decoding. Finally, disabling iterative refinement yields the lowest performance among all ablations, demonstrating that many queries require progressively incorporating additional facets that are not captured in a single pass. Together, these results highlight that complementary evidence construction and confidence-guided refinement jointly enable the robustness of *CLIMB*.

4.6 Sensitivity Analysis

We further investigate how *CLIMB* behaves under variations of its main hyperparameters by using the dataset InfoSeek: the micro-pool size K, the MMR trade-off λ, the confidence threshold U_{STAR}, and the maximum number of refinement rounds $T_{\max}$. Figure 3 and Figure 4(a) summarizes results on InfoSeek. Varying the pool size K reveals the expected coverage–noise trade-off. Very small pools lead to noticeable degradation, as the evidence set cannot cover all facets of the query. As K increases, accuracy improves and then remains stable once a small complementary set is formed. Pushing K further brings little or no gain and can slightly hurt performance, reflecting the fact that additional passages are often redundant or weakly

Figure 5: Question–Image pairs from the InfoSeek, shown with answers produced by GPT-5.2 and *CLIMB*.

relevant. The trade-off parameter λ shows a similar pattern: accuracy is highest in a broad middle range where relevance and redundancy suppression are balanced, and deteriorates when λ approaches extremes that either ignore diversity or over-penalize similar but useful passages. These trends are consistent across dataset, and the default configuration (K=5, λ=0.6) lies well inside the stable region. The confidence threshold U_{STAR} primarily controls when the refinement loop accepts an updated answer. Lower thresholds allow less updates, while higher thresholds enforce more conservative behavior; in practice, accuracy remains stable across a wide interval around the default choice, indicating that the verifiability signal is reasonably calibrated and does not require precise tuning. The maximum number of refinement rounds $T_{\max}$ has a limited effect beyond moving from a single pass to a small number of iterations: performance improves when allowing more than one round, then saturates, and additional iterations bring negligible changes, which matches the observation that most queries require at most two or three refinement steps.

To summarize the overall stability, we also report a sensitivity score that measures the accuracy range induced by sweeping each hyperparameter within the tested bounds in Figure 4(b). The variations are modest on InfoSeek, with slightly larger effects for K and λ and smaller effects for U_{STAR} and $T_{\max}$. This pattern supports the view that *CLIMB*'s gains are driven by its structural design, complementary pooling and confidence-guided refinement, rather than by delicate hyperparameter tuning.

4.7 Qualitative Results

In addition to the quantitative results, we also include qualitative comparison using GPT-5.2 as a reference model in Figure 5. Showing side-by-side outputs from GPT-5.2 and *CLIMB* helps illustrate how complementary evidence and confidence-guided refinement affect the final answers beyond what is visible from aggregate metrics alone.

5 Conclusion

We introduced *CLIMB*, a training-free framework for multimodal retrieval-augmented generation that strengthens knowledge-intensive visual question answering by building a complementary evidence pool and refining answers under a confidence acceptance rule. Rather than scaling model size or retrieval depth, *CLIMB* focuses on which passages to trust and when to update. On both datasets, it steadily surpasses the baselines, with ablations and sensitivity studies indicating that the gains come from this structural design. We hope this view encourages further work on verifiability-aware, complementarity-driven multimodal RAG.

6 Limitations

This work focuses on knowledge-intensive visual question answering with external Wikipedia-style textual evidence. While this setting is widely used for evaluating multimodal RAG, it does not exhaust all multimodal retrieval scenarios. Extending CLIMB to video, multi-image reasoning, or domain-specific corpora may require adapting the retrieval interface and evidence scoring prompts. The confidence score in CLIMB is used as a relative signal for accepting or rejecting answer updates. It should therefore be interpreted as an evidence-grounded refinement criterion rather than as a calibrated probability of correctness. Future work could investigate more specialized confidence estimators while keeping the framework training-free.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. irag: Advancing rag for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4341–4348.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024a. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024b. 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.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

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 *EMNLP*.

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 Computer Vision and Pattern Recognition Conference*, pages 9199–9209.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112.

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).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval*, pages 421–438. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 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, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2781–2790.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, and 1 others. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

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.

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Emu: Generative pretraining in multimodality. In *ICLR*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

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.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. 2025. Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1305–1315.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

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

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and 1 others. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

APPENDIX

A Implementation and Metric Details

This section provides additional details on how we instantiate *Climb* in practice. The system is decomposed into five deterministic modules: (i) corpus indexing and micro-pool construction, (ii) R/E/C evidence scoring, (iii) answer generation with explicit evidence conditioning, (iv) a verifiability-oriented confidence function, and (v) facet diagnosis with an iterative confidence climb. All components operate purely at inference time; no finetuning or gradient-based updates are used.

System Overview. Given an input image I, question q, and a text-only corpus $\mathcal{D}$, the pipeline proceeds as follows:

1. Encode all corpus passages with CLIP ViT-L/14 and build a FAISS index using inner-product similarity.

2. For a query image I, retrieve an initial candidate set and construct a complementary micro-pool of size K=5 using a Maximal Marginal Relevance (MMR) objective.

3. For each passage in the pool, compute three interpretable scores: relevance R, evidence specificity E, and cross-modal consistency C, and aggregate them into a single ranking signal ϕ.

4. Select the top-M passages under ϕ, generate a short textual answer a, and compute a scalar confidence $U \in [0, 1]$ based on verifiability criteria.

5. If U is below a target threshold $U_\star$, diagnose under-supported facets, refresh scores inside the fixed micro-pool, and repeat the generation–confidence step. An update is accepted only if it strictly increases U, yielding a monotone confidence climb with at most $T_{\max}$=3 rounds.

Complementary Micro-pool Construction. We first embed all corpus passages using CLIP ViT-L/14 and store the vectors in a FAISS index. At query time, the image I is encoded into a normalized embedding, and we compute cosine similarity against all passage embeddings. Let $\mathrm{rel}(t)$ denote the similarity between the image embedding and passage t, and let $\mathrm{sim}(t, u)$ denote the cosine similarity between two passages t and u in the text-embedding space.

To avoid redundancy in the evidence context, we construct a micro-pool $\mathcal{P}$ with an MMR-style objective. Starting from the most relevant passage, we iteratively add one passage at a time by maximizing

$$\mathrm{MMR}(t) \;=\; \lambda\,\mathrm{rel}(t) - (1-\lambda)\,\max_{u\in S}\mathrm{sim}(t,u), \tag{5}$$

where S is the set of already selected passages, and $\lambda \in [0, 1]$ balances query fit and redundancy. In our implementation we use K=5 and set λ in the range $[0.5, 0.6]$. With cached embeddings, the overall complexity is $O(NK)$ for N initial candidates.

R/E/C Evidence Scoring. Once the micro-pool $\mathcal{P}$ is fixed, we assign each passage $t \in \mathcal{P}$ three scores in $[1, 5]$:

- Relevance $R(t)$: how directly and comprehensively the passage addresses the question q.

- Evidence Specificity $E(t)$: how rich the passage is in concrete, verifiable details (named entities, dates, numeric values, etc.).

- Cross-modal Consistency $C(t)$: how well the passage aligns with visual cues in I (category, attributes, location, or other scene constraints).

These scores are produced by a structured evaluator that returns a JSON object with fields `R`, `E`, `C`, and a brief textual rationale. All calls use temperature 0, which makes the scoring highly stable across runs. We then normalize the scores to $[0, 1]$ and compute a unified ranking signal

$$\phi(t) \;=\; \alpha_R\,R(t) + \alpha_E\,E(t) + \alpha_C\,C(t), \tag{6}$$

with fixed weights $\alpha_R, \alpha_E, \alpha_C$ chosen on a small development split. The top-M passages under ϕ form the initial working subset $\mathcal{S}_0$.

Answer Generation and Confidence Estimation. For a given working subset $\mathcal{S}$, we generate an answer in two explicit steps:

1. Key information extraction: the system first identifies key facts in the concatenated evidence that directly address the question (e.g., entity names, dates, numeric values). The output is a short, structured summary of only the relevant facts.

2. Short answer decoding: conditioned on the extracted key information (and optionally the image), the system produces the shortest possible answer (typically a single word or short phrase) without explanations or hedging phrases.

We then compute a scalar confidence $U \in [0, 1]$ for the tuple $(I, q, \mathcal{S}, a)$. The confidence function is implemented as a verifier that returns a JSON object with fields such as:

- `confidence`: overall confidence score in $[0, 1]$,
- `completeness`: coverage of the question by the evidence,
- `reliability`: quality and specificity of the supporting evidence,
- `reasoning`: a short explanation of the assessment.

If JSON parsing fails, we fall back to a conservative default. In all reported experiments, we use the `confidence` field as the scalar U that drives the confidence-climb decision.

Facet Diagnosis and Iterative Confidence Climb Many errors arise when the current subset $\mathcal{S}_t$ misses a key facet of the question. Instead of issuing new retrieval calls, we perform facet diagnosis within the fixed pool $\mathcal{P}$. Given $(I, q, \mathcal{S}_t, a_t)$, we extract at most two short facet phrases

$$\mathcal{G}_t = \{g_1, g_2\}, \tag{7}$$

each constrained to be no longer than a few tokens, describing missing aspects that prevent full support. A relaxed mode is used by default: if the answer already roughly addresses the question, the diagnoser returns "no significant gaps" and the loop terminates.

When non-empty $\mathcal{G}_t$ is returned, we re-score passages inside $\mathcal{P}$ by combining the original score $\phi(t)$ with a facet–text match score $\psi(t \mid \mathcal{G}_t)$, which mixes lexical overlap (BM25-style) and embedding similarity:

$$\phi'(t) = \phi(t) + \gamma\, \psi(t \mid \mathcal{G}_t), \tag{8}$$

where γ is a small positive constant. The next subset $\mathcal{S}_{t+1}$ is obtained by taking the top-M passages under ϕ', and we regenerate an answer a_{t+1} with an updated confidence U_{t+1}. We enforce a monotone acceptance rule:

$$U_{t+1} > U_t \;\Rightarrow\; \text{accept and continue}, U_{t+1} \leq U_t \;\Rightarrow\; \text{stop}. \tag{9}$$

In practice, we set $U_\star$=0.85 as a target confidence and cap the number of refinement rounds by $T_{\max}$=3.

B Datasets

This section provides additional details about the datasets used in our experiments. We describe the construction process, data sources, and annotation characteristics of Encyclopedic-VQA and InfoSeek, with an emphasis on the types of visual grounding and factual reasoning they require. These details are included to clarify the benchmark-specific assumptions and to ensure the experimental setup can be fully reproduced.

Encyclopedic-VQA. Encyclopedic-VQA is a large-scale, entity-centric benchmark designed to test fine-grained factual reasoning in multimodal settings. Each instance consists of (i) an image depicting a real-world entity, (ii) a question targeting a specific factual attribute, and (iii) a short answer extracted verbatim from a Wikipedia article. Images are sourced from several curated recognition datasets, including iNaturalist 2021, Google Landmarks, and other domain-specific corpora, ensuring substantial variation in species, landmarks, geological formations, artworks, and historical objects.

Each image is linked to a corresponding Wikipedia entity page, and the dataset constructs its retrieval corpus by segmenting roughly two million pages into paragraph-level units. These segments constitute the complete evidence pool for all experiments, and retrieval is strictly restricted to this controlled corpus.

The question design emphasizes factual specificity. Many questions require identifying fine-grained attributes such as biological taxonomy (e.g., "Order?", "Genus?"), geographical information ("Which river flows through this site?"), historical metadata ("Who inaugurated this building?"),

physical measurements, or conservation status. Importantly, most questions require combining visual grounding (e.g., identifying the depicted species) with factual lookup (e.g., locating the required line in the Wikipedia entry). The official evaluation focuses on the single-hop setting, where answers are supported by a single paragraph from the article, allowing controlled analysis of evidence retrieval.

InfoSeek. InfoSeek serves as a complementary benchmark emphasizing broader question styles and larger-scale variability. It contains approximately 1.3M automatically generated image–question–answer triplets, augmented with a curated set of human-authored questions for quality assurance. Entities span more than ten thousand Wikipedia pages, including species, buildings, artworks, geographic regions, public figures, and artifacts.

Each example is paired with a textual corpus consisting of around one hundred thousand aligned Wikipedia pages. Compared to Encyclopedic-VQA, the InfoSeek knowledge base is smaller but covers a wider breadth of topics. The dataset intentionally introduces greater lexical and structural diversity in question templates—for example, describing the same fact using different surface forms, or mixing attribute, relational, temporal, and descriptive queries.

InfoSeek is especially challenging because the model must (i) correctly recognize the visual entity from the image (fine-grained categories such as subspecies, architectural variants, instrument models), (ii) identify the relevant page among many visually similar candidates, and (iii) pinpoint the correct fact within the retrieved document. The evaluation protocol follows the official validation split and restricts retrieval to the provided corpus to ensure comparability.

Comparison and Usage. Together, the two datasets provide complementary perspectives on multimodal knowledge-intensive reasoning. Encyclopedic-VQA emphasizes controlled factual grounding with a very large retrieval corpus and fine-grained entity alignment, while InfoSeek emphasizes large-scale variability, surface diversity in questions, and robustness to paraphrases and distributional shifts.