Ground Then Rank: Revisiting Knowledge-Based VQA with Training-Free Entity Identification

Anonymous ACL submission

Abstract

Knowledge-Based Visual Question Answering (KB-VQA) requires grounding visual queries to external knowledge beyond directly observable content in images. While recent multi modal large language models (MLLMs) show strong perceptual abilities, they struggle on KB-VQA tasks requiring groundings from both fine-grained entity and evidence levels. Most existing multi-modal retrieval augmented generation(MM-RAG) methods tightly couple entity discrimination and section-level evidence ranking into a single re-ranking stage, leading to high cost and limited generalization. In this work, we revisit existing MM-RAG solutions from a workflow perspective and argue both entity-level and fact-level groundings are both key bottlenecks. We observe that although MLLMs often fail under open-ended entity naming, they can better identify the correct entity when selecting from a small set of candidate names. Based on this insight, we propose a simple and training-free *identify-before-answer IBA* framework that decouples entity identification from section-level re-ranking. Our approach prompts an MLLM to select high-confidence entities using only candidate names, followed by an off-the-shelf textual re-ranker for evidence selection. Experiments on Encyclopedic-VQA and InfoSeek show that our method consistently outperforms fine-tuned multi-modal re-ranking baselines while reducing training and inference complexity. Additional analyses reveal that the improvements arise not only from better entity identification, but also from selecting more informative evidence once correct entity is fixed. Our implementation is made public to ease reproducibility [1].

1 Introduction

Knowledge-Based Visual Question Answering (KB-VQA) extends standard VQA by requiring external world knowledge beyond what is directly observable in the image (Deng et al., 2025; Kim et al., 2025). While recent multi-modal large language models (MLLMs) achieve strong performance on perception-driven VQA, KB-VQA queries often hinge on identifying the correct real-world entity and grounding fine-grained factual information that cannot be inferred from pixels alone (Mensink et al., 2023; Chen et al., 2023). This reliance on entity-level knowledge makes KB-VQA a challenging benchmark for multi-modal intelligence.

Modern MLLMs (Liu et al., 2023, 2024; Bai et al., 2025) have demonstrated remarkable progress on general VQA tasks, yet they remain unreliable on KB-VQA where relevant knowledge is sparse, long-tailed, and difficult to encode in model parameters (Kuang et al., 2025; Deng et al., 2025). As a result, recent systems predominantly adopt a multi-modal retrieval-augmented generation (MM-RAG) paradigm (Chen et al., 2022; Yu et al., 2025), which first retrieves a set of potentially relevant knowledge entries (e.g., Wikipedia articles) and then re-ranks textual sections to support answer generation, as illustrated in Figure 1.

Despite their success, we argue that existing MM-RAG methods suffer from a fundamental limitation in their workflow. Producing a correct and verifiable answer requires grounding at two distinct levels: (i) *entity-level grounding*, ensuring that the retrieved context refers to the correct entity depicted in the image, and (ii) *section-level grounding*, locating the passage within that entity's article that is relevant to the question. However, most existing approaches (Yan and Xie, 2024; Cocchi et al., 2025; Tian et al., 2025) implicitly couple these two challenging tasks into a single re-ranking step over all candidate sections. This entangled formulation forces a single scoring function to simultaneously discriminate between entities and rank textual evidence, often leading to textually relevant but entity-incorrect contexts, or correct entities paired with

[1] https://anonymous.4open.science/r/ACL26-IBA-3857

irrelevant sections.

In contrast, humans naturally decouple these two steps when solving KB-VQA problems. After an initial retrieval or recall of plausible candidates, people typically first identify or narrow down the entity depicted in the image, and only then examine a small number of relevant articles to locate supporting evidence. This decomposition reduces distractors and simplifies subsequent reasoning, suggesting a more principled paradigm.

Motivated by this observation, we revisit the role of MLLMs in entity identification. While directly naming an entity from an image remains challenging for current models (Caron et al., 2024), we make a surprising empirical observation: simply providing candidate entity names enables MLLMs to identify the correct entity with much higher accuracy. This suggests that MLLMs often possess incomplete yet usable entity knowledge that is difficult to exploit under open-ended generation but can be effectively activated when the task is reframed as a constrained discrimination problem. This behavior bears resemblance to a tip-of-the-tongue-like effect (Brown and McNeill, 1966), which we use only as an intuitive analogy. ToT describes a situation that human experts may also encounter that they have the expertise to the entity but can't directly recall the name from scratch. But once several plausible names (*e.g.,* "Lapsana communis" and "Crepis tectorum" in Figure 1) are presented, they can reason from visual cues with their expertise to select the correct one.

Based on this insight, we propose a simple yet effective *IBA* framework for KB-VQA. Our approach explicitly inserts a lightweight entity identification step into the MM-RAG workflow. After initial retrieval, the MLLM scores candidate entities using only their names, retains a small subset of high-confidence entities, and then applies a pre-trained standard textual re-ranker to select supporting sections within this narrowed scope. This training-free design decouples entity recognition from evidence selection, improving both accuracy and efficiency.

Our contributions are summarized as follows:

- To best of our knowledge, we are the first to report the 'tip-of-the-tongue' phenomenon in MLLMs for KB-VQA, where providing candidate entity names significantly amplifies the model's reasoning capability to better identify the entity in query image.
- Based on this finding, we propose a simple yet effective framework that integrates an explicit identification step into existing MM-RAG workflows, enhancing answer accuracy and computational efficiency without additional fine-tuning or specifically training.
- We validate our approach on two mainstream KB-VQA benchmarks, achieving new state-of-the-art results while improving efficiency compared to existing MM-RAG systems.

2 Related Works

2.1 MLLMs for KB-VQA

Multimodal large language models (MLLMs) extend text-only LLMs with visual encoders and cross-modal alignment mechanisms, enabling joint reasoning over images and text. Recent models such as LLaVA (Liu et al., 2023, 2024) and Qwen-VL (Bai et al., 2025) achieve strong performance on perception-driven VQA benchmarks, where answers can be inferred directly from visual content or broadly learned parametric knowledge.

However, emerging evaluations (Li et al., 2024; Tan et al., 2025) consistently show that even state-of-the-art MLLMs underperform on knowledge-based VQA (KB-VQA) tasks that require fine-grained, entity-centric, or long-tail encyclopedic knowledge. This limitation motivates augmenting MLLMs with external knowledge sources to support explicit grounding and reasoning beyond their parametric capacity

2.2 KB-VQA and MM-RAG based solutions

KB-VQA benchmarks extend conventional VQA by requiring external knowledge not contained in the image alone, such as entity attributes or encyclopedic facts. Early datasets such as OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) emphasize commonsense or general knowledge, which increasingly falls within the training scope of large-scale MLLMs.

More recent benchmarks, including Encyclopedic-VQA (E-VQA)(Mensink et al., 2023) and InfoSeek(Chen et al., 2023), raise the difficulty by requiring explicit grounding to fine-grained Wikipedia entities and supporting sections. To address these challenges, many methods adopt multimodal retrieval-augmented generation (MM-RAG), typically consisting of a retriever, a re-ranking stage, and an answer generator. Representative approaches such as

EchoSight (Yan and Xie, 2024), ReflectiVA (Cocchi et al., 2025), and CoRe-MMRAG (Tian et al., 2025) differ in how relevance is assessed, ranging from explicitly trained multimodal re-rankers to relevance implicitly learned through fine-tuning. Despite their success, these methods generally couple entity discrimination and section selection into a single re-ranking process, which can be costly to train and sensitive to data availability.

2.3 Positioning of our work

In contrast to prior MM-RAG methods, our work revisits the KB-VQA pipeline from a workflow perspective. Rather than entangling entity identification and section-level evidence selection within a single re-ranking module, we explicitly decouple these two stages. By introducing a lightweight identification step before section re-ranking, our approach isolates the entity-level grounding problem and leverages off-the-shelf components without requiring task-specific re-ranker training or MLLM fine-tuning. This design differs fundamentally from prior approaches that rely on learned multimodal relevance functions, and enables more interpretable and transferable KB-VQA pipelines across datasets with varying knowledge distributions.

3 Methodology

In this section, we present our training-free IBA (Identify Before Answer) framework, which explicitly decouples entity-level identification from section-level evidence selection. We first introduce the overall workflow and then describe each component in detail, including problem formulation, initial retrieval, identify-before-re-rank, and answer generation.

3.1 Problem Formulation

Given a query image I and question Q, a KB-VQA system aims to generate an answer y by grounding external knowledge. In retrieval-augmented generation, this is typically achieved by selecting a supporting text snippet from an external knowledge base.

We model the knowledge base as $KB = \{(P_i, I_i)\}_{i=1}^{N}$, where each page P_i consists of multiple textual sections $P_i = \{S_{i,j}\}_{j=1}^{n_i}$ and is associated with a representative image I_i. Due to the large scale of the knowledge base (N can be millions), practical systems first retrieve a small candidate set and then perform fine-grained re-ranking. The objective of KB-VQA is to select the most relevant section $S_{i,j}$ that provides sufficient evidence to support generating the correct answer y.

3.2 Initial Retrieval

The goal of initial retrieval is to obtain a small set of candidate knowledge entries from a massive external knowledge base. This coarse-grained step ensures tractability by narrowing the search space from millions of entries to a manageable top-K set.

Following prior work (Yan and Xie, 2024; Cocchi et al., 2025; Tian et al., 2025), we adopt an image-to-image retrieval strategy. Each knowledge base page is indexed using a frozen EVA-CLIP-8B vision encoder (Sun et al., 2024). Image embeddings are pooled from the final layer and indexed using FAISS (Douze et al., 2025). Given a query image, cosine similarity is used to retrieve the top-K visually similar candidate pages, which are then passed to subsequent identify-before-re-rank stage.

3.3 Identify-Before-Re-Rank

After initial retrieval, existing MM-RAG methods directly perform section-level re-ranking over all K candidate entries. In contrast, we explicitly introduce an entity-level identification step to further reduce the re-ranking scope.

As we have discussed in Section 1, one of the key challenges in KB-VQA is entity identification to secure grounding at the entity level. Even for human experts, directly naming the exact real-world entity depicted in an image can be non-trivial, especially when visual cues are subtle or the entity belongs to a fine-grained category. This difficulty is further amplified for MLLMs under open-ended generation settings, where the model must produce the correct entity name from a vast output space without explicit constraints.

At the same time, modern MLLMs have been trained on large-scale, high-quality corpora that include extensive encyclopedic knowledge, much of which originates from Wikipedia-style resources. Such training endows models with latent expertise that can support entity recognition, but this expertise is often difficult to reliably elicit through unconstrained generation. We hypothesize that the challenge lies not in the absence of knowledge, but in the form of the task: open-ended entity naming imposes a high uncertainty burden, whereas selecting the correct entity from a small candidate set is a more tractable discriminative problem. As empirically validated in Section 4.4, prompting

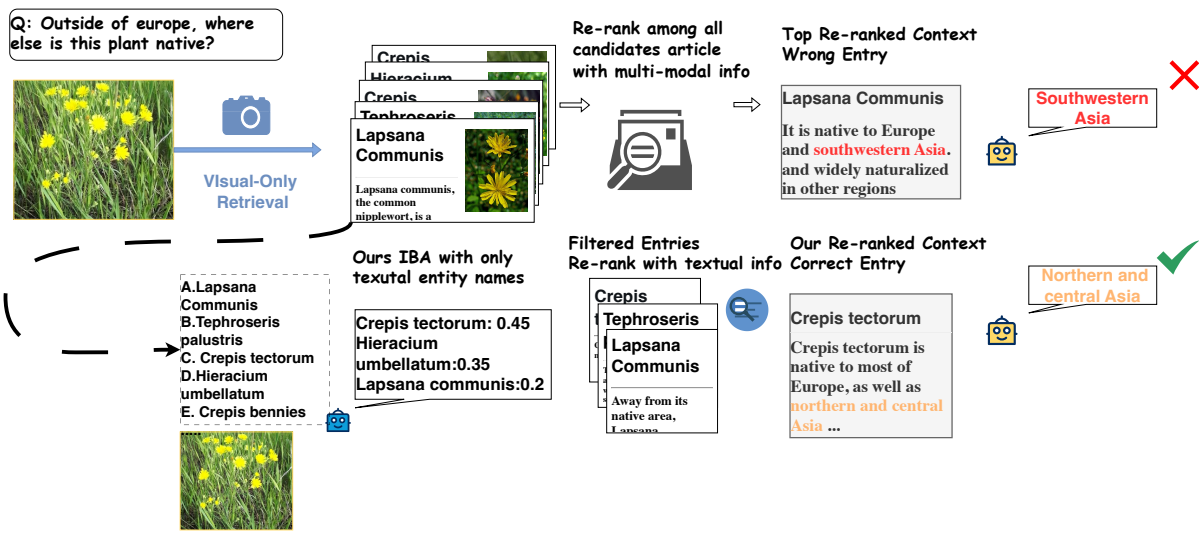

Figure 1: Overall workflow comparison between existing MM-RAG methods and our proposed IBA. **Upper**: Existing MM-RAG pipelines retrieve top-K candidate articles from a large knowledge base and directly perform section-level re-ranking using trained re-rankers or fine-tuned VLMs, before generating the answer from the top-ranked context. **Lower**: Our training-free IBA inserts an explicit entity identification step before re-ranking. Given the query image and retrieved candidate names, the VLM selects a small set of high-confidence entities, which are then used to narrow the scope of section-level re-ranking with an off-the-shelf textual re-ranker.

MLLMs to select from candidate entities yields substantially higher identification accuracy than open-ended naming. This behavior is analogous to the human tip-of-the-tongue phenomenon (Brown and McNeill, 1966), where experts may struggle to recall an exact name spontaneously but can readily identify the correct option when presented with a shortlist of plausible candidates.

Motivated by this observation, we design the identification step by prompting the MLLM with a list of candidate entity names, rather than asking it to generate the entity name freely. This formulation allows the model to focus on assessing relative relevance among plausible candidates, effectively activating its latent encyclopedic knowledge while avoiding the brittleness of open-ended generation.

Accordingly, given the retrieved candidate set $\{(P_i, I_i)\}_{i=1}^{K}$, we prompt the MLLM with: (i) query image I, (ii) textual names of all K candidate entries, and (iii) their initial visual retrieval similarity scores.

The MLLM is asked to assess entity relevance and select the top-j candidates ($j < K$). This produces an identification confidence score $\mathrm{ID}(P_i)$ for each candidate entry, reflecting model's belief of the entity depicted in image corresponds to P_i.

For each identified entry, we compute textual relevance between the question Q and each section $S_{i,j}$ using a pre-trained textual re-ranker such as BGE (Chen et al., 2024). The re-ranker outputs a normalized textual relevance score $\mathrm{T}(S_{i,j}) \in [0, 1]$. Unlike multimodal re-rankers used in prior work (Yan and Xie, 2024), this component is used off-the-shelf without task-specific training.

We combine identification confidence, visual similarity from initial retrieval, and textual relevance to compute a final score:

$$\mathrm{score}(S_{i,j}) = \alpha \cdot \mathrm{ID}(P_i) + \beta \cdot \mathrm{V}(I, I_i) + \gamma \cdot \mathrm{T}(S_{i,j}),$$

where α, β, and γ control the contribution of each signal. For E-VQA, we set $(\alpha, \beta, \gamma) = (0.5, 0.5, 1)$. For InfoSeek, we increase α to emphasize entity identification due to weaker visual alignment between query images and knowledge base images. The top-ranked section is selected as supporting context for answer generation.

Overall, the identify-before-re-rank design provides a simple yet effective alternative to existing MM-RAG pipelines. By explicitly decoupling entity-level identification from section-level evidence selection, our framework avoids the need for training specialized multimodal re-rankers or fine-tuning large vision–language models. This decoupling also improves interpretability, as the contributions of visual similarity, entity identification, and textual relevance can be examined independently. Moreover, restricting section-level scoring to a small set of identified entities substantially reduces computational cost and context length, leading to more efficient inference. Finally, because our

method relies only on off-the-shelf components, it generalizes naturally across different knowledge bases and KB-VQA benchmarks without dataset-specific adaptation.

3.4 Answer Generation

Once the top-ranked supporting section is obtained, we use off-the-shelf large language models to generate the final answer. Our framework does not require fine-tuning the generation model, making it flexible across different backbones. Compared with prior approaches that rely on fine-tuned multimodal generators (Cocchi et al., 2025; Tian et al., 2025), our method improves answer quality by providing more precise and entity-grounded context, rather than modifying the generation model itself.

4 Experiments

4.1 Datasets and External Knowledge Base

We evaluate on two KB-VQA benchmarks: Encyclopedic VQA (E-VQA) (Mensink et al., 2023) and InfoSeek (Chen et al., 2023), where answering requires external knowledge beyond the query image. E-VQA contains 221K image-question pairs (up to five images per question) and covers both single-hop and two-hop questions. Following prior work (Yan and Xie, 2024), we focus on the single-hop setting. Importantly, E-VQA provides a controlled knowledge base of 2M Wikipedia articles with associated images, ensuring that each QA pair is answerable when the correct article is retrieved.

InfoSeek contains 1.3M QA pairs over 11K visual entities from OVEN (Hu et al., 2023). Following existing MM-RAG baselines (Yan and Xie, 2024; Tian et al., 2025; Cocchi et al., 2025), we adopt the 100K-article knowledge base released by Yan and Xie (2024) and report results on the validation split following the same settings.

4.2 Evaluation Metrics

Retrieval. We report Recall@K, which measures whether the ground-truth article appears in the top-K retrieved candidates. Following prior work (Yan and Xie, 2024; Cocchi et al., 2025), a retrieved article is counted as correct only if its URL exactly matches the target page URL. We report Recall@1 as a proxy for *top-1 entity selection accuracy* after re-ranking, reflecting how well a method prioritizes correct entity among retrieved candidates.

Answer generation. For E-VQA, we evaluate open-ended answers using the BEM score (Zhang et al., 2019). For InfoSeek, we follow prior practice (Yan and Xie, 2024; Cocchi et al., 2025) and use VQA accuracy (Goyal et al., 2017; Marino et al., 2019) for time and numerical questions, and BEM score for string questions.

4.3 Implementation Details

Retriever and candidate set. Following prior work (Yan and Xie, 2024; Cocchi et al., 2025), we use EVA-CLIP-8B (Sun et al., 2024) for image-to-image retrieval and retrieve the top-K=20 candidate articles from the knowledge base released by Yan and Xie (2024).

Answer generators. We instantiate our pipeline with two off-the-shelf backbones: Llama-3.1-8B-Instruct and Qwen-2.5-VL-7B-Instruct(same for identification).

Baselines. For EchoSight (Yan and Xie, 2024), we follow the original pipeline and apply its released multimodal re-ranker to re-rank sections from the same top-K retrieved candidates, using the default weighting between initial retrieval similarity and re-ranking scores. For ReflectiVA (Cocchi et al., 2025), we run the officially released model to produce REL tokens and generate answers by conditioning on sections assigned REL tokens within the top-5 retrieved articles.

Zero-shot variants. Following the two-stage prompting design in Core-MMRAG (Tian et al., 2025), we implement several zero-shot variants to probe the role of workflow design. Given the top-5 retrieved articles, *1Stage* prompts the MLLM to directly answer with all articles as context, while *2Stage* first selects the most relevant article and then generates the answer conditioned on that article only. We also include *Para* (no external evidence) and *Article* (directly use the top-5 retrieved articles without explicit re-ranking) variants.

4.4 Retrieval Results

The retrieval results on InfoSeek (Chen et al., 2023) and E-VQA (Mensink et al., 2023) are reported in Tables 1 and 2. EVA-CLIP retrieval yields moderate Recall@20 but much lower Recall@1, showing that while the correct entity is usually present among candidates, it is rarely ranked first by visual similarity alone (e.g., E-VQA Recall@1: 13.4%, Recall@20: 48.8%). This highlights the need for a re-ranking stage to better prioritize the correct entity.

After applying re-ranking, both EchoSight and our proposed IBA substantially improve Recall@1,

confirming the importance of re-ranking for entity prioritization. On InfoSeek, our method outperforms EchoSight by a clear margin, achieving Recall@1 of 58.4% compared to 53.1% (Table 1). This improvement indicates that the explicit identify-before-re-rank design is more effective at prioritizing the correct entity from visually similar candidates.

Table 1: InfoSeek retrieval results. EVA-CLIP denotes the initial image-to-image retrieval using EVA-CLIP-8B. EchoSight applies its trained multimodal re-ranker on top of the same retrieved candidates. Our method prompts the MLLM to select the top-3 entities from the 20 retrieved candidate names.

Method	InfoSeek Recall@k				
	k=1	k=3	k=5	k=10	k=20
EVA-CLIP	45.6	63.1	68.6	74.6	77.9
EchoSight	53.1	69.4	73.9	77.4	77.9
Our IBA	**58.4**	**72.4**	-	-	-

On E-VQA, EchoSight achieves slightly higher Recall@1 (36.5%) than our method (35.5%), as shown in Table 2. This outcome is expected, as EchoSight is specifically trained on E-VQA using curated positive supervision. In contrast, our approach is entirely training-free and directly transferable across datasets. Despite this small gap in Recall@1, the downstream generation results (Section 4.5) show that higher identification accuracy alone does not guarantee better answer quality.

Table 2: E-VQA retrieval results. EVA-CLIP denotes the initial image-to-image retrieval using EVA-CLIP-8B. EchoSight applies its trained multimodal re-ranker on top of the retrieved candidates. Our method prompts the MLLM to select the top-3 entities from the 20 retrieved candidate names.

Method	E-VQA Recall@k				
	k=1	k=3	k=5	k=10	k=20
EVA-CLIP	13.4	26.1	31.9	41.8	48.8
EchoSight	**36.5**	**45.3**	47.9	48.8	48.8
Our IBA	35.5	43.3	-	-	-

Grounded-subset identification analysis. To directly probe entity identification (independent of answer generation), we evaluate on grounded subsets where each question is guaranteed to be answerable from its ground-truth entity page. On E-VQA, our method correctly identifies the ground-truth entity for 934/2,322 grounded questions (40.2%), compared to 593/2,322 (25.5%) under open-ended entity naming. On InfoSeek, we randomly sample 1,000 validation questions, among which 790 are grounded; our method identifies correctly for 578/790 (73.2%) versus 362/790 (45.8%) under open-ended naming. These results further support providing candidate entity names substantially amplifies MLLM-based identification.

4.5 Generation Results

We evaluate answer generation quality on InfoSeek and E-VQA and compare our training-free pipeline with finetuned MM-RAG baselines and zero-shot variants (Table 3).

RAG vs. purely parametric MLLMs. Across both backbones, retrieve-augmented methods substantially outperform purely parametric variants (*Para-**), highlighting that external evidence is essential for KB-VQA. On InfoSeek, *Para-LLaVA* and *Para-Qwen* achieve 9.0 and 25.5 overall, while retrieve-augmented variants reach 30–38. On E-VQA, *Para-LLaVA* and *Para-Qwen* obtain 13.3 and 21.2, and our proposed IBA exceed 43.

Training-free identify-before-answer vs. finetuned MM-RAG. Our training-free pipeline outperforms finetuned baselines on both datasets. On InfoSeek, our *IBA-LLaVA* attains the best overall score (37.8), followed by *IBA-Qwen* (37.2), both surpassing ReflectiVA (36.4) and EchoSight (33.8). On E-VQA, *IBA-Qwen* and *IBA-LLaVA* achieve 43.6 and 43.2, outperforming EchoSight (41.9) and ReflectiVA (38.6). Notably, our method requires no task-specific training or additional parameters, while EchoSight and ReflectiVA rely on finetuned components. A closer look at InfoSeek question types shows that our improvements are consistent on time questions, where correct entity grounding and evidence selection are critical. For example, on unseen-question time queries, *IBA-LLaVA* achieves 38.7 versus 29.7 (ReflectiVA) and 24.9 (EchoSight), and on unseen-entity time queries it achieves 44.5 versus 36.5 and 37.5, respectively. More qualitative results are shown in Appendix B

4.6 Ablation Study

Zero-shot MLLM. To test whether an off-the-shelf MLLM can solve KB-VQA by prompting alone, we follow Tian et al. (2025) and compare four zero-shot variants that differ in how retrieved evidence is used. *Para-** answers using only parametric knowledge (no external evidence). *Article-** answers in one step given the full text of the top-5 retrieved ar-

Table 3: Answer generation results on E-VQA and InfoSeek.

Methods	Backbone	InfoSeek							E-VQA
		Overall	Unseen Question			Unseen Entity			
			time	num	string	time	num	string	
Our proposed IBA									
IBA-Qwen	Qwen-2.5-VL	37.2	34.5	7.8	47.5	40.1	6.6	45.2	**43.6**
IBA-LLaVA	Llama-3.1-8B	**37.8**	38.7	13.5	45.9	44.5	12.8	44.4	43.2
Retrieve Augmented Models requiring Fine-tuning									
ReflectiVA	Llama-3.1-8B	36.4	29.7	10.4	45.6	36.5	12.1	43.2	38.6
EchoSight	Llama-3.1-8B	33.8	24.9	12.3	41.7	37.5	11.3	39.8	41.9
Zero-shot base models									
Para-Llava	Llama-3.1-8B	9.0	0.9	0.5	12.4	2.4	0.5	10.2	13.3
Para-Qwen	Qwen-2.5-VL	25.5	10.9	0.0	35.5	12.1	0.0	32.9	21.2
Zero-shot with Retrieval									
Article-llava	Llama-3.1-8B	18.4	0.0	0.0	26.4	0.0	0.0	24.1	23.0
Article-qwen	Qwen-2.5-VL	34.5	14.3	0.2	46.8	12.1	0.5	45.9	35.6
Zero-shot with Re-rank									
1Stage-llava	Llama-3.1-8B	10.5	0.0	0.0	15.0	0.0	0.0	13.9	4.0
1Stage-Qwen	Qwen-2.5-VL	34.6	32.7	1.5	44.8	40.1	2.1	44.4	34.3
2Stage-llava	Llama-3.1-8B	27.2	6.9	0.9	38.5	5.0	0.2	36.4	23.1
2Stage-Qwen	Qwen-2.5-VL	35.9	8.4	0.0	49.0	4.5	0.3	48.5	34.1

ticles. *1Stage-** further adds the retrieved entry images and explicitly asks the model to use the most relevant reference, effectively requiring implicit multimodal re-ranking inside a single long prompt. *2Stage-** decomposes this into two prompts: the model first selects the most relevant article and then generates answer conditioned on that article only.

External evidence helps, but prompting is not a reliable re-ranker. Across both datasets, moving from *Para-** to *Article-** yields large gains, confirming that KB-VQA cannot be solved reliably from parametric knowledge alone. For LLaVA, InfoSeek improves from 9.0 (*Para-LLaVA*) to 18.4 (*Article-LLaVA*), and E-VQA improves from 13.3 to 23.0. For Qwen, InfoSeek improves from 25.5 (*Para-Qwen*) to 34.5 (*Article-Qwen*), and E-VQA improves from 21.2 to 35.6. However, asking the model to implicitly re-rank evidence within a single prompt is unstable: *1Stage-LLaVA* collapses to 10.5 on InfoSeek and 4.0 on E-VQA, despite using more information than *Article-LLaVA*.

Two-stage prompting helps, but still trails a structured pipeline. Decomposing the interaction partially mitigates the above issue: *2Stage-LLaVA* recovers to 27.2 on InfoSeek and 23.1 on E-VQA, yet remains far behind our full pipeline (37.8 and 43.2, respectively). For Qwen, explicit re-ranking provides limited or inconsistent gains: *1Stage-Qwen* and *2Stage-Qwen* reach 34.6/35.9 on InfoSeek (vs. 34.5 for *Article-Qwen*) and slightly underperform on E-VQA (34.3/34.1 vs. 35.6). Overall, these results suggest that while strong backbones can exploit long retrieved text, prompting alone does not reliably perform evidence selection, motivating an explicit and lightweight identification-and-reranking workflow.

EchoSight's multimodal re-ranker vs. our proposed re-rank after identification. To disentangle the effect of entity identification from section selection, we evaluate several re-ranking variants on the E-VQA grounded subset, where the correct entity page is guaranteed to contain the answer. We compare (i) EchoSight's original pipeline, (ii) our full identify-before-answer model, and two hybrids that keep our identification module but replace our textual re-ranker with EchoSight's released multimodal re-ranker, either without (*IBA- MIS*) or with (*IBA + MIS*) incorporating the MLLM identification score (MIS) into the final ranking. Table 4 summarizes the results.

Table 4: Ablation on E-VQA grounded questions comparing re-ranking choices after entity identification. IBA-/+ MIS replace our textual re-ranker with EchoSight's released multimodal re-ranker, without /with using the MLLM identification score (MIS) in final ranking.

Methods	Identified Ratio	Score
EchoSight	77.8	66.4
IBA- MIS w EchoSight reranker	73.6	63.2
IBA + MIS w EchoSight reranker	75.1	62.2
IBA full	75.7	70.5

We draw two observations from Table 4. First, replacing our textual re-ranker with EchoSight's multi-modal re-ranker consistently reduces answer

Q: What does this fish eat?

EchoSight (Yan and Xie, 2024):
Walleye is a nocturnal feeder, which means it feeds at night. It is most easily caught using live minnows or lures that mimic small fish. This implies that the walleye's diet consists of small fish, likely minnows. ✗
SELECTED Section Title: As food
Walleye is considered to be a palatable freshwater fish, consequently, is fished recreationally and commercially for food. Because of its nocturnal feeding habits, it is most easily caught at night using live minnows or lures that mimic small fish ...
IBA (Ours):
Walleyes eat a variety of food sources depending on their age and development. Here are specific food sources mentioned:* Invertebrates, such as fly larvæ and zooplankton, when the young walleyes are absorbing their yolk sac.* Fish, such as yellow perch or ciscoes, when they become piscivorous after 40-60 days.* Crayfish, minnows, and leeches, in addition to fish, when they become adults. ✓
SELECTED Section Title: Reproduction
... Once the yolk has been fully absorbed, the young walleyes begin to feed on invertebrates, such as fly larvæ and zooplankton. After 40 to 60 days, juvenile walleyes become piscivorous. Thenceforth, both juvenile and adult walleyes eat fish almost exclusively, frequently yellow perch or ciscoes, moving onto bars and shoals at night to feed. ...

Figure 2: Qualitative result E-VQA, where we compare the answers provided by IBA with EchoSight (Yan and Xie, 2024) that more informative context is selected by our IBA.

quality, even when the identified ratio is comparable (around 73–78%). For example, EchoSight and the two hybrid variants obtain scores of 66.4/63.2/62.2, all below our method (70.5). Second, incorporating MIS does not improve the hybrid: *IBA + MIS* slightly increases the identified ratio over *IBA– MIS* (75.1 vs. 73.6) but further lowers the final score (62.2 vs. 63.2). These results suggest that, once the entity is (mostly) correct, effective KB-VQA hinges on selecting *informative* sections rather than visually plausible but weak evidence. To further analyze this effect, Table 5 breaks down performance on grounded questions by whether EchoSight and our method identify the ground-truth entity.

Table 5: Breakdown results on E-VQA grounded questions. Our IBA achieves better overall results mainly from selecting more informative sections than EchoSight. Some samples are shown in Figure 2 and 5.

	Both ✓ (59.3%)	Our ✓ (13.4%)	Echo ✓ (15.5%)	Both ✗ (11.8%)	Overall
Echo	79.6	22.8	80.6	30.9	66.4
IBA	88.2	83.6	23.9	27.6	70.5

Although EchoSight attains a slightly higher identified ratio (77.8% vs. 75.7%), our method achieves a higher overall answer score (70.5 vs. 66.4). When both methods identify the correct entity (59.3% of questions), our score is substantially higher (88.2 vs. 79.6), indicating more informative section selection given the same entity. Our advantage is even more pronounced when only our method identifies the correct entity (13.4%): we maintain a strong score (83.6) while EchoSight often fails to provide useful evidence (22.8). Conversely, EchoSight performs better only when it identifies the correct entity and we do not (15.5% of questions; 80.6 vs. 23.9). When both miss the entity (11.8%), both methods perform poorly.

Overall, these ablations indicate that our gains are not solely due to entity identification. Explicitly decoupling identification from purely textual re-ranking leads to more reliable selection of supportive evidence. Even when the same entity is retrieved, our approach tends to choose sections that directly contain the facts required by the question, whereas a trained multimodal re-ranker can favor visually plausible but less informative sections. Qualitative examples in Figures 2 and 5 further illustrate this difference.

5 Conclusion

In this work, we revisit KB-VQA from a workflow perspective and identify entity and evidence level groundings as critical bottlenecks. While recent MLLMs possess substantial encyclopedic knowledge, we show that this knowledge is difficult to reliably elicit under open-ended entity naming. Instead, MLLMs exhibit significantly stronger identification ability when selecting from a small set of candidate entities. Motivated by this observation, we propose a simple and training-free identify-before-answer framework that explicitly decouples entity identification from section-level evidence selection. By prompting MLLMs with candidate entity names and leveraging an off-the-shelf textual re-ranker for evidence selection, our approach avoids the need for specialized multi-modal re-ranker training while remaining interpretable and robust. Extensive experiments on E-VQA and InfoSeek demonstrate that this decomposition consistently outperforms finetuned MM-RAG baselines. Our analyses also reveal that effective KB-VQA depends not only on identifying the correct entity, but also on selecting informative supporting evidence once the entity is fixed. We hope this work encourages future research to rethink retrieval-augmented reasoning pipelines by explicitly separating distinct grounding and selection stages, and to explore more lightweight designs for knowledge-intensive multi-modal reasoning.

Limitations

Although our proposed IBA surpass existing fine-tuning baselines and demonstrates impressive performance on Knowledge-based VQA like Encyclopedic-VQA and InfoSeek, we also commit the following limitation that there is a dependence on the external knowledge base. In real world application, it could be possible that the knowledge base is not perfect to include all supporting evidence for answering questions. However, there are emerging works focused on integrating agentic workflow into the VQA tasks (Jiang et al., 2024; Wu et al., 2025; Fu et al., 2025) on retrieval stage. Specifically, they will invoke external search tools *i.e.,* the whole Internet will be considered as the external knowledge base, which could be a promising direction for the future works.

References

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.

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

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

Roger Brown and David McNeill. 1966. The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337.

Mathilde Caron, Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. 2024. A generative approach for wikipedia-scale visual entity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17313–17322.

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

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

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? *arXiv preprint arXiv:2302.11713*.

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.

Jiaqi Deng, Zonghan Wu, Huan Huo, and Guandong Xu. 2025. A comprehensive survey of knowledge-based vision question answering systems: The lifecycle of knowledge in visual reasoning task. *arXiv preprint arXiv:2504.17547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Mingyang Fu, Yuyang Peng, Benlin Liu, Yao Wan, and Dongping Chen. 2025. Livevqa: Live visual knowledge seeking. *arXiv preprint arXiv:2504.05288*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen,

Chaoyou Fu, Guanglu Song, and 1 others. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10):1–35.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.

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

Yunxin Li, Xinyu Chen, Baotian Hu, Haoyuan Shi, and Min Zhang. 2024. Cognitive visual-language mapper: Advancing multimodal comprehension with enhanced visual knowledge alignment. *arXiv preprint arXiv:2402.13561*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. 2023. Token pooling in vision transformers for image classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 12–21.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

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

Yuwen Tan, Yuan Qing, and Boqing Gong. 2025. Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck. *arXiv preprint arXiv:2505.24840*.

Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, and 1 others. 2025. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. *arXiv preprint arXiv:2506.02544*.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. 2022. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*, pages 497–515. Springer.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

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*.

Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*.

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

Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025. Mramg-bench: A comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3616–3626.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yu Zhang, Yepeng Liu, Duoqian Miao, Qi Zhang, Yiwei Shi, and Liang Hu. 2023. Mg-vit: a multi-granularity method for compact and efficient vision transformers. *Advances in Neural Information Processing Systems*, 36:69328–69347.

A Additional Related Work

A.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) extend text-only LLMs by integrating modality encoders and alignment mechanisms, enabling joint reasoning over images and text. Early models such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) demonstrated that coupling a pretrained vision encoder with a frozen language model can already yield strong few-shot multimodal capabilities. Subsequent models, including the LLaVA family (Liu et al., 2023, 2024) and Qwen-VL (Bai et al., 2025), further advanced this paradigm through large-scale instruction tuning and improved cross-modal fusion architectures, achieving strong performance across a wide range of multimodal benchmarks.

Despite these advances, recent empirical studies (Li et al., 2024; Tan et al., 2025) consistently show that even state-of-the-art MLLMs underperform on knowledge-based VQA (KB-VQA) benchmarks that require fine-grained, entity-centric, or long-tail factual knowledge. This limitation motivates augmenting MLLMs with external knowledge sources to support explicit grounding and reasoning beyond parametric knowledge alone.

A.2 Knowledge-Based Visual Question Answering

Conventional visual question answering (VQA) benchmarks (Antol et al., 2015; Goyal et al., 2017; Wang et al., 2017) focus on questions that can be answered using visual content and common-sense knowledge. With large-scale pretraining and instruction tuning, modern MLLMs perform competitively on these tasks, as much of the required information is either directly observable or implicitly encoded during training (Yin et al., 2024; Bai et al., 2025).

Knowledge-based VQA (KB-VQA) fundamentally extends this setting by requiring external knowledge not contained in the image alone, such as entity attributes or encyclopedic facts (Deng et al., 2025). Early benchmarks, including OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), introduced questions that depend on commonsense or general knowledge, which increasingly falls within the training scope of large-scale MLLMs.

More recent benchmarks, such as Encyclopedic-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023), further raise the difficulty by requiring fine-grained, entity-level grounding in Wikipedia. These datasets demand explicit identification of the correct entity and selection of supporting sections from retrieved articles. Despite progress in MLLMs, models continue to struggle in this setting due to challenges in discriminating visually similar entities and selecting the most informative evidence, motivating retrieval-augmented approaches.

A.3 MM-RAG-Based Solutions

To address the limitations of parametric knowledge in KB-VQA, many recent methods adopt multimodal retrieval-augmented generation (MM-RAG). As illustrated in Figure 1, MM-RAG pipelines typically consist of three stages: (i) a retriever that performs coarse retrieval from a large-scale knowledge base, (ii) a re-ranking stage that selects the most relevant context among retrieved candidates, and (iii) an answer generator that produces the final response conditioned on the selected evidence.

Representative methods differ primarily in how relevance is modeled. EchoSight (Yan and Xie, 2024) trains a dedicated multimodal re-ranker to select the most relevant sections given an image–question pair, leveraging contrastive learning with curated supervision. However, obtaining high-quality positive supervision for section-level relevance is challenging, and among major KB-VQA benchmarks, only E-VQA provides explicit supporting section annotations.

Other approaches, such as ReflectiVA (Cocchi et al., 2025) and CoRe-MMRAG (Tian et al., 2025), rely on fine-tuning large MLLMs to implicitly assess relevance through internal representations. While effective, these methods require constructing task-specific training data and incur substantial computational cost when fine-tuning large-scale models such as LLaVA (Liu et al., 2024) or Qwen-VL (Bai et al., 2025). These limitations motivate alternative designs that reduce training overhead while improving robustness and transferability.

B Qualitative results

In this section we will provide more qualitative results on both InfoSeek and E-VQA datasets as

infoseek-00016131 **Q**: In which year was this equipment retired from operational service?
Query Entity: Bren light machine gun

ReflectiVA (Cocchi et al., 2025):
1963 ✗
EchoSight (Yan and Xie, 2024):
Not mentioned in the text. ✗
IBA (Ours):
2006 ✓

00073156 **Q**:What is the sea level in metre of this mountain?
Query Entity:Corcovado

ReflectiVA (Cocchi et al., 2025):
396 ✗
EchoSight (Yan and Xie, 2024):
396 ✗
IBA (Ours):
710 ✓

00073009**Q**: what was the date this aircraft enter into service?
Query Entity:Tupolev Tu-154

ReflectiVA (Cocchi et al., 2025):
1970 2 21 ✗
EchoSight (Yan and Xie, 2024):
Not mentioned. ✗
IBA (Ours):
9 February 1972 ✓

00072497 **Q**: In which year did this building come into service?
Query Entity:Niechorze Lighthouse 1866

ReflectiVA (Cocchi et al., 2025):
1814 ✗
EchoSight (Yan and Xie, 2024):
1814 ✗
IBA (Ours):
1866 ✓

00073353 **Q**: What is the length of this lake in kilometre?
Query Entity:Windermere

ReflectiVA (Cocchi et al., 2025):
11.23 ✗
EchoSight (Yan and Xie, 2024):
11 ✗
IBA (Ours):
18 ✓

00072995 **Q:** What is the weight of a male of this bird in gram?
Query Entity:Least grebe 129

ReflectiVA (Cocchi et al., 2025):
692 ✗
EchoSight (Yan and Xie, 2024):
692-925 ✗
IBA (Ours):
129 ✓

00018660 **Q**: What is the closest upper taxonomy of this bird?
Query Entity:Eurasian collared dove

ReflectiVA (Cocchi et al., 2025):
Zenaida ✗
EchoSight (Yan and Xie, 2024):
Genus ✗
IBA (Ours):
Streptopelia ✓

00018687 **Q**: What is this plant named after?
Query Entity:Allamanda

ReflectiVA (Cocchi et al., 2025):
Charles Plumier ✗
EchoSight (Yan and Xie, 2024):
Thevetia ✗
IBA (Ours):
Frédéric-Louis Allamand ✓

00000129 **Q:** What country does this building belong to?
Query Entity:Polish Baltic Philharmonic

ReflectiVA (Cocchi et al., 2025):
Netherlands ✗
EchoSight (Yan and Xie, 2024):
Denmark ✗
IBA (Ours):
Poland ✓

Figure 3: Sample qualitative results on image-question pairs from InfoSeek, where we compare the answers provided by IBA with those from ReflectiVA (Cocchi et al., 2025) and EchoSight (Yan and Xie, 2024).

shown in Figure 3 and 4.

C More examples for informative context

We have presented more examples on Figure 5 where our proposed IBA selects more informative context compared with EchoSight (Yan and Xie, 2024).

D Error Case Study

In this section, we provide more examples of error cases. As shown in Figure 6, for most cases, the error is caused by comprised entity grounding as shown in first two cases. However, it could be possible that the data annotation is not concise enough or evaluation is not comprehensive enough. As shown in the example, both our proposed IBA and EchoSight have secured the correct entity grounding and provided reasonable answer *i.e.,* the exact name of the law signed ‘ Oklahoma City National Memorial Act ’. However, it is judged as wrong via existing evaluation protocols, which could be caused due to the too coarse-grained or vague ground truth answer **law.**

E Prompt Template

E.1 Identification

We explicitly prompt the MLLM to perform entity identification as a constrained selection task, rather than open-ended entity naming. Given a query image and a small set of retrieved candidate entity names, the model is asked to select the most likely entity (or top-k entities) depicted in the image. We also utilize the initial visual retrieval similarity score in the prompts. The identification prompt is formatted as follows:

```
SYSTEM: You are an expert visual entity
recognizer. Look at the image and here
are some potentially relevant options.

Options:
A. <ENTITY_NAME_1> (image similarity: <
SIM_1>)
B. <ENTITY_NAME_2> (image similarity: <
SIM_2>)
C. <ENTITY_NAME_3> (image similarity: <
SIM_3>)
...

Reply with 'Answer: <label1>, <label2>,
...' listing the top <K> option letters
from most to least likely based on the
image.
```

Here, each option corresponds to a candidate entity retrieved from the external knowledge base, optionally augmented with its initial image-to-image

Q: How long is this bird in centimeters?
Query Entity: Sanderling

ReflectiVA (Cocchi et al., 2025):
30–40 ✗
EchoSight (Yan and Xie, 2024):
The breeding adult bird is 14-16 cm in length. ✗
IBA (Ours):
The sanderling is 18–20 cm (7.1–7.9 in) in length. ✓

Q: Who designed this hotel?
Query Entity:The Fullerton Hotel Singapore

ReflectiVA (Cocchi et al., 2025):
The hotel rooms were designed by Hirsch Bedner Associates. ✗
EchoSight (Yan and Xie, 2024):
Hirsch Bedner Associates designed the hotel rooms. ✗
IBA (Ours):
The Fullerton Building was designed by Major P.H. Keys of Keys & Dowdeswell, a Shanghai firm of architects. ✓

Q: What settlements comprised the this abbey estate?
Query Entity: Bolton Abbey

ReflectiVA (Cocchi et al., 2025):
villages ✗
EchoSight (Yan and Xie, 2024):
The Fountains Abbey estate comprised the following settlements:... ✗
IBA (Ours):
The Bolton Abbey estate comprised the following settlements: Addingham,Holme,Beamsley and Gargrave ✓

Q: What tool is associated with the traditional seaside atmosphere of this pier?
Query Entity:Brighton Palace Pier

ReflectiVA (Cocchi et al., 2025):
whelk ✗
EchoSight (Yan and Xie, 2024):
The answer is: Carousel. ✗
IBA (Ours):
The traditional seaside atmosphere of the Brighton Palace Pier is associated with ""bucket and spade"". ✓

Q: What is the architectural style of this mosque?
Query Entity: Great Mosque of Banten

ReflectiVA (Cocchi et al., 2025):
Islamic architecture ✗
EchoSight (Yan and Xie, 2024):
The architectural style of the Great Mosque of Banten is a mix of Indian Mughal pattern and ancient candi decoration. ✗
IBA (Ours):
The architectural style of the Great Mosque of Banten is described as follows joglo style, a traditional Javanese style ✓

Q: When was this park created?
Query Entity: Princes Street Gardens

ReflectiVA (Cocchi et al., 2025):
1843 ✗
EchoSight (Yan and Xie, 2024):
The park was created in 1876, when the town council reacquired the ground and laid it out as a public park. ✗
IBA (Ours):
Princes Street Gardens were created in the 1820s. ✓

Figure 4: Sample qualitative results on image-question pairs from E-VQA, where we compare the answers provided by IBA with those from ReflectiVA (Cocchi et al., 2025) and EchoSight (Yan and Xie, 2024).

retrieval similarity score. The model is required to respond strictly in the prescribed format, enabling us to directly interpret the output as entity-level confidence for subsequent re-ranking.

E.2 Answer Generation

For the answer generation stage, we follow existing works (Yan and Xie, 2024) to apply answer generation templates depending on the dataset.

E-VQA. The prompt we use for LLMs when testing Encyclopedic-VQA (E-VQA) (Mensink et al., 2023) is shown as follows:

```
USER: Context: <CONTEXT>
Question: <QUESTION>
The answer is:
```

InfoSeek. Due to the strict exact-match evaluation used by InfoSeek (Chen et al., 2023), following existing works (Yan and Xie, 2024), we adopt a one-shot prompting strategy and add instructions to ensure the generated answer strictly matches the required format. The prompt used for InfoSeek is:

```
SYSTEM: You always answer the question
the user asks. Do not answer anything
else.

USER:Context: The sounthern side of the
Alps is next to Lake Como.
Question: Which body of water is this
mountain located in or next to?
Just answer the questions, no
explanations needed.
Short answer is: Lake Como

Context: <CONTEXT>
Question: <QUESTION>
Just answer the questions, no
explanations needed.
Short answer is:
```

F Dataset Details

Following existing works(Yan and Xie, 2024; Tian et al., 2025; Cocchi et al., 2025), we use the same test set of InfoSeek (Chen et al., 2023) and E-VQA (Mensink et al., 2023), consists of 71,335 and 4,750 question pairs correspondingly,

G Token Budget and Re-ranking Efficiency

In this section we will introduce more details to compare the cost of our proposed method with existing methods in terms of token and FLOPs.

G.1 Our proposed IBA with BGE Reranker

The Qwen top-k pipeline operates as follows:

1. Retrieve top 20 entities per query; Qwen identification keeps the top 3 entities.

Q: How many tiny reddish yellow warts does each segment of this moth have?
Query Entity: Atolmis rubricollis

EchoSight (Yan and Xie, 2024):
According to the text, there is no mention of tiny reddish-yellow warts on the segments of the red-necked footman moth. The description only mentions the coloration of the thorax, head, abdomen, antennae, legs, and wings, but does not mention any warts. Therefore, the answer is None. ✗
Section Title: Description
The red-necked footman is a small moth that is mostly charcoal grey or deep dark brown (fresh specimens almost black), but has a conspicuous orange thorax, part of which is visible behind the black head as an orange-red collar. The hindwings are brownish grey. The antennae and legs are black and the end of the abdomen is yellowish orange or golden yellow. The wings are tightly folded together around the body and have pleated, squared-off ends. The wingspan is 25 to 35 mm (1.0 to 1.4 in) and the length of the forewings is 15 to 18 mm (0.6 to 0.7 in). Wilhelm Neuburger named Romanian specimens as the synonym Atolmis flavicollis; these were less dark brown and with a light yellow instead of a red collar.
IBA (Ours):
According to the text, each segment of the caterpillar bears six tiny reddish yellow warts. ✓
Section Title: Life cycle
The white eggs of the red-necked footman are laid in small groups in crevices in the branches of trees, especially those of old firs. They grow to a length of about 27 mm (1.1 in). Their head is black with a bold diagonal white stripe on either side. Their main colour is dark greenish grey marbled with cream. Each segment bears six tiny reddish yellow warts which bear black hairs. The caterpillars feed on lichens growing on the trunks and branches of trees, and can be found between August and October...

Q: In which year(s) was this castle rebuilt?
Query Entity: Dudley Castle

EchoSight (Yan and Xie, 2024):
The answer is: None. According to the text, the bulk of the remaining habitable parts of the castle was destroyed by fire in 1750, and there is no mention of the castle being rebuilt. ✗
SELECTED Section Title: Final years and ruin
stable block was constructed on the site at some point before 1700. This was the final building to be constructed in the castle. The bulk of the remaining habitable parts of the castle was destroyed by fire in 1750. However, in the nineteenth century, the site found a new use as a 'Romantic Ruin' and a certain amount of tidying up of the site was carried out by the Earls of Dudley. Battlements on one of the remaining towers were reconstructed and two cannon captured during the Crimean Wars were installed. In the nineteenth and early twentieth century the site was used for fêtes and pageants. In 1937, when the Dudley Zoo was established, the castle grounds were incorporated into the zoo. ...
IBA (Ours):
According to the evidence, the castle keep was rebuilt starting from 1262. ✓
SELECTED Section Title: The keep
The castle keep dates from the rebuilding that started in 1262. It rests on the motte, constructed in the Norman period but somewhat reduced in height afterwards. The original building was slightly rectangular in plan with approximate dimensions 15 metres north to south and 22 metres east to west. The four drum towers on each corner are 9.8 metres in diameter. After the slighting at the end of the civil war, only the north side of the castle and parts of two of the drum towers remain.

Q:What kind of scabious is this plant commonly known as?
Query Entity: Knautia Arvensis

EchoSight (Yan and Xie, 2024):
According to the article, the plant is commonly known as Gipsy Rose. ✗
SELECTED Section Title: Name
Species of scabious were used to treat scabies, and many other afflictions of the skin including sores caused by the bubonic plague. The word scabies comes from the Latin word scabere 'scratch'. Another name for this plant is gipsy rose. The genus Knautia is named after a 17th-century German botanist, Christian Knaut.
IBA (Ours):
Based on the evidence text, the answer is: field scabious. ✓
SELECTED Section Title: Knautia arvensis
Knautia arvensis, commonly known as field scabious, is a herbaceous perennial species of flowering plant in the honeysuckle family Caprifoliaceae.

Figure 5: Sample qualitative results on image-question pairs from E-VQA, where we compare the answers provided by IBA with EchoSight (Yan and Xie, 2024) where we selected more informative context.

Q: In which part of the world does this animal live?
Query Entity: Echinometra Mathaei rubricollis

ReflectiVA (Cocchi et al., 2025):
World Ocean

EchoSight (Yan and Xie, 2024):
The red sea urchin (Mesocentrotus franciscanus) is found in the northeastern Pacific Ocean, specifically in the areas of Alaska to Baja California. ✗

IBA (Ours):
The purple sea urchin, Strongylocentrotus purpuratus, lives along the eastern edge of the Pacific Ocean, extending from Ensenada, Mexico, to British Columbia, Canada. ✗

Q: How big is an adult of this species typically?
Query Entity: Argiope bruennichi

ReflectiVA (Cocchi et al., 2025):
insects

EchoSight (Yan and Xie, 2024):
The answer is not explicitly stated in the provided text. It does not provide information about the size of the adult Argiope spider. ✗

IBA (Ours):
I did not find any information about the size of an adult Argiope species. ✓

Q:What did president bill clinton sign to create this memorial?
Query Entity: Oklahoma City National Memorial. Ground Truth Answer: Law

ReflectiVA (Cocchi et al., 2025):
Bill ✗
EchoSight (Yan and Xie, 2024):
The answer is: the Oklahoma City National Memorial Act of 1997. ✗
IBA (Ours):
President Bill Clinton signed the ""Oklahoma City National Memorial Act of 1997"" to create the Oklahoma City National Memorial. ✗

Figure 6: Error cases on image-question pairs from E-VQA, where we compare the answers provided by our propsed IBA with ReflectiVA (Cocchi et al., 2025) and EchoSight (Yan and Xie, 2024).

2. For the retained entities, obtain all wiki sections and send them to the BGE section reranker.

3. Obtain the section scores by comparing all sections with the query text to select the best section for downstream answer generation.

Empirically, our prepared metadata for E-VQA shows:

$$\text{avg. sections/example} \approx 24.7,$$
$$\text{avg. tokens/section} \approx 172.$$

Assuming questions contribute ≈ 20 tokens, the total BGE token budget per example is

$$24.7 \times (172 + 20) \approx 4.8\text{k tokens},$$

.

G.2 EchoSight Re-ranker

The EchoSight flow differs from our proposed IBA with larger re-ranking space.

1. Retrieve top 20 entities, which is same as our.

2. Expand all entities to sections, encode the image once, and encode *all* sections with BLIP-2 Q-Former.

3. Score by query(both image and text) and candidate sections relevance(similarity) and rerank.

Using the same section statistics as a proxy, the section pool grows to

$$20 \times \frac{24.7}{3} \approx 165 \text{ sections},$$

yielding a text load of $\approx 165 \times 172 \approx 28$k tokens, i.e., about $6-7\times$ more text than our proposed IBA.

G.3 Runtime Implications (full pipeline)

- **Ours**
  - Identification (Qwen2.5-VL-7B): ≈256 visual tokens (one image encode) + 70–100 text tokens input (question + 20 candidate titles) ⇒ 400 tokens.
  - Section rerank (BGE, text-only): ≈ 4.8k text tokens (24.7 sections $\times$ (172 per section + 20 for the question)).

  Overall cost is dominated by the BGE text forward; the image is encoded once in identification.

- **EchoSight reranker (no identification):**
  - One image encode (≈256 visual tokens).
  - Text side encodes around 28k text tokens (≈ 165 sections $\times$ 172 tokens per section) and fuses with the image. The model(*i.e.,* EchoSight Re-ranker based on BLIP-2) is heavier than the off-the-shelf textual re-ranker such as bge-v2-m3.

Thus, end-to-end, our identification + text rerank load is far smaller than EchoSight's multimodal re-rank; the gains come from entity pruning (section count reduced $6-7\times$) and using a lightweight text re-ranker.

H Computational Cost Analysis via FLOPs

We analyze computational cost using floating-point operations (FLOPs) rather than wall-clock latency. Wall-clock time is highly sensitive to implementation details, hardware configuration, batching strategy, and system-level optimizations, making it difficult to fairly compare methods with different architectures. In contrast, FLOPs provide a hardware-agnostic and reproducible proxy for inference complexity that reflects the intrinsic computational demand of a model. This metric has been widely adopted in prior work for comparing model efficiency across architectures and modalities.

FLOPs estimation protocol. To compare inference cost across different model architectures in a hardware-agnostic and reproducible manner, we estimate computational complexity using floating-point operations (FLOPs). Following prior work that analyzes the scaling and efficiency of Transformer models (Kaplan et al., 2020), we adopt a standard proxy for Transformer-based modules:

For a Transformer encoder with hidden dimension d and L layers, the forward FLOPs per token can be approximated as

$$\text{FLOPs per token} \approx 24 \times L \times d^2, \qquad (1)$$

which captures the dominant contributions of multi-head self-attention and feed-forward network operations across layers. The total compute is then obtained by multiplying this per-token cost by the number of tokens processed by the model. Using this unified protocol for all Transformer-only components (e.g., transformer layers in text or cross-

modal encoders) ensures a fair basis for comparison across methods.

Vision Transformers (ViT) (Dosovitskiy et al., 2021), however, require a different consideration because their self-attention operations compute pairwise interactions among image patch tokens, resulting in computation that scales quadratically with sequence length. Specifically, self-attention involves operations on $S \times S$ affinity matrices, where S is the number of visual tokens, dominating compute when S is large. For these vision encoders, we therefore estimate FLOPs by accounting explicitly for both the attention term ($O(S^2 \cdot d)$) and the feed-forward term ($O(S \cdot d \cdot d_{ff})$), rather than relying solely on the d^2 proxy. This exception reflects the inherent quadratic complexity of self-attention in visual processing and is consistent with practice in ViT analysis (Touvron et al., 2022; Marin et al., 2023; Zhang et al., 2023).

EchoSight. EchoSight performs multi-modal re-ranking using a BLIP-2-based (Li et al., 2023) architecture. During re-ranking, EchoSight does not invoke a large language model. Instead, it employs the BLIP-2 Querying Transformer (Q-Former), a BERT (Devlin et al., 2019) style Transformer (12 layers, hidden size 768), to encode both the multi-modal query (image + question) and all candidate wiki sections, where vision information will be processed with its Vision Transformer (Dosovitskiy et al., 2021) vision encoder [2] . Each candidate section is encoded independently using the same Q-Former in a text-only mode, and relevance scores are computed via embedding similarity. Given that EchoSight expands all retrieved entities into sections, the re-ranking stage processes a large number of text tokens, leading to substantial computational cost. We estimate the FLOPs of EchoSight re-ranking by accounting for: (i) the frozen vision encoder, (ii) the Q-Former multi-modal fusion, and (iii) the Q-Former-based text encoding of all candidate sections. For the vision encoder, EchoSight uses an EVA-CLIP Vision Transformer variant with a patch size of 14, hidden dimension $d_v = 1408$, and depth $L_v = 39$ transformer layers. The input image of size 224×224 is tokenized into $16^2 = 256$ patches plus one classification token, yielding $S = 257$ tokens.

To estimate FLOPs for Vision Transformers (ViT), we decompose the dominant contributions as:

- **Self-attention:** computing the attention score matrix QK^T involves $O(S^2 \cdot d)$ operations due to pairwise interactions among tokens.
- **Feed-forward network (FFN):** each token is transformed via two linear layers with intermediate dimension $d_{ff} \approx 4d$, contributing $O(S \cdot d \cdot d_{ff})$ operations.

This yields the per-layer FLOPs approximation:

$$\text{FLOPs per layer}_{\text{ViT}} \approx 2\, d_v\, S^2 \;+\; 4\, d_v\, d_{ff}\, S,$$

where the first term corresponds to self-attention and the second to FFN. Such decomposition reflects the quadratic dependence on token count intrinsic to self-attention in vision models.

Substituting $S = 257$, $d_v = 1408$, $L_v = 39$ and $d_{ff} \approx 4d_v$, the FLOPs for a single forward pass through the vision encoder can be approximated as:

$$\begin{aligned} \text{FLOPs}_{vision} &\approx L_v\big(2\, d_v\, S^2 + 4\, d_v\, d_{ff}\, S\big) \\ &\approx 3.3 \times 10^{11}. \end{aligned} \tag{2}$$

This indicates that a single forward pass through EVA-CLIP's Vision Transformer backbone incurs on the order of 10^{11} FLOPs, consistent with the understanding that self-attention costs grow quadratically with the number of tokens processed.

For the multi-modal fusion and text encoding, EchoSight employs the BLIP-2 Querying Transformer (Q-Former) (Li et al., 2023), a BERT-base style Transformer with hidden size $d_q = 768$ and $L_q = 12$ layers. Using the same transformer FLOPs proxy yields the following cost per token.,

$$\begin{aligned} \text{FLOPs}_{QFormer} &\approx 24 \times L_q \times d_q^2 \\ &\approx 24 \times 12 \times 768^2 \approx 1.7 \times 10^{10}. \end{aligned} \tag{3}$$

For the multi-modal fusion step, this cost is incurred once for the query and question tokens. For section text encoding, each candidate section is encoded with the same Q-Former in text-only mode. Given an approximate rerank pool of $\sim$ 28,000 tokens, the total FLOPs for text encoding becomes

$$28{,}000 \times 1.7 \times 10^{10} \;\approx\; 4.8 \times 10^{14}. \tag{4}$$

Overall, while the vision encoder and fusion contribute on the order of 10^{10}–10^{11} FLOPs, the text

[2] https://github.com/Go2Heart/EchoSight/blob/main/lavis/models/blip2_models/blip2_qformer_reranker.py

encoding FLOPs dominate the re-ranking cost at $\sim 4.8 \times 10^{14}$. These estimates use consistent transformer FLOPs proxies and model configuration data from the EchoSight implementation, supporting a concrete and fair comparison of computational demand across methods.

Our proposed IBA FLOPs. For our pipeline, the computational cost comprises two stages: (i) multi-modal identification using Qwen2.5-VL-7B Instruct and (ii) text-only re-ranking using the BAAI/bge-reranker-v2-m3 model.

Identification (Qwen2.5-VL-7B Instruct). Qwen2.5-VL-7B Instruct adopts a re-engineered vision–language architecture rather than a standard ViT encoder followed by a text-only transformer (Bai et al., 2025). In particular, the vision encoder is redesigned with windowed attention and multi-stage token merging, which significantly reduces the effective sequence length compared to a naive patch-based Vision Transformer. As a result, the quadratic S^2 self-attention cost characteristic of vanilla ViT is largely mitigated in practice.

For FLOPs estimation, we therefore treat the identification stage as a unified transformer-style forward pass and apply the standard transformer FLOPs proxy to the shared multi-modal backbone. According to the official configuration, Qwen2.5-VL-7B uses hidden size $d_{\text{Qwen}} = 3584$ and $L_{\text{Qwen}} = 28$ transformer layers (Bai et al., 2025). The per-token FLOPs is estimated as

$$\begin{aligned}\text{FLOPs}_{\text{Qwen2.5-VL}} &\approx 24 \times L_{\text{Qwen}} \times d_{\text{Qwen}}^2 \\ &\approx 24 \times 28 \times 3584^2 \approx 8.62 \times 10^9. \end{aligned} \tag{5}$$

During identification, the model processes one image together with the question and candidate entity names. Due to the internal vision token compression in Qwen2.5-VL, the effective number of tokens participating in the transformer layers is substantially smaller than the raw patch count. We conservatively approximate the overall forward pass as involving ~ 400 effective tokens, yielding

$$\text{FLOPs}_{\text{identification}} \approx 400 \times 8.62 \times 10^9 \approx 3.45 \times 10^{12}. \tag{6}$$

This estimate intentionally over-approximates the identification cost and provides a conservative upper bound for comparison.

Section re-ranking (bge-reranker-v2-m3). The off-shelf textual re-ranker(bge-reranker-v2-m3)(Chen et al., 2024) uses a RoBerta-based(Liu et al., 2019) cross-encoder with hidden size $d_{\text{BGE}} = 1024$ and $L_{\text{BGE}} = 24$ layers. Again applying the transformer FLOPs proxy yields following per token cost,

$$\begin{aligned}\text{FLOPs}_{\text{BGE}} &\approx 24 \times L_{\text{BGE}} \times d_{\text{BGE}}^2 \\ &\approx 24 \times 24 \times 1024^2 \approx 6.04 \times 10^8. \end{aligned} \tag{7}$$

Given a rerank pool of $\sim$ 4,800 text tokens per example, the total re-ranking FLOPs becomes

$$\text{FLOPs}_{\text{BGE rerank}} \approx 4800 \times 6.04 \times 10^8 \approx 2.90 \times 10^{12}. \tag{8}$$

Combining both stages gives

$$\text{FLOPs}_{\text{ours}} \approx 3.45 \times 10^{12} + 2.90 \times 10^{12} \approx 6.35 \times 10^{12},$$

indicating that our pipeline's compute proxy is dominated by identification and re-ranking FLOPs, and is orders of magnitude lower than the comparable EchoSight re-ranking cost.

Key efficiency advantage. A key source of efficiency in our pipeline stems from architectural decoupling, which reduces compute by *nearly two orders of magnitude* compared to EchoSight multi-modal re-ranker.

Under the same FLOPs proxy, EchoSight's re-ranking cost is dominated by the cost of encoding the entire candidate section pool with a high-capacity multi-modal encoder. Specifically, using an EVA-CLIP Vision Transformer backbone (4×10^{10} FLOPs per image) and a BLIP-2 Q-Former (1.7×10^{10} FLOPs per token), the re-ranking stage on a pool of $\sim 28{,}000$ text tokens incurs $\sim 4.8 \times 10^{14}$ FLOPs. By contrast, our method incurs only $\sim 3.45 \times 10^{12}$ FLOPs for the multi-modal identification stage with Qwen2.5-VL-7B and $\sim 2.90 \times 10^{12}$ FLOPs for the BGE reranker text stage, totaling 6.35×10^{12} FLOPs.

This means that even without considering the one-off vision encoding cost, our rerank-centric compute is *nearly two orders of magnitude lower* than the text encoding cost alone in EchoSight's re-ranker. The bulk of EchoSight's compute arises from repeatedly processing all candidate sections with a multi-modal model, whereas our approach confines expensive multi-modal processing to a single identification pass and relies on a lightweight text re-ranker for large-scale comparison. This structural decoupling yields **significantly lower computational demand** while preserving re-ranking effectiveness, demonstrating that effective knowledge-based VQA does not require heavy multi-modal encoding over the entire candidate space.