UniSONAR: Unified Source-Conditioned Attentive Retrieval for Knowledge-Based Visual Question Answering

Anonymous ACL submission

Abstract

Knowledge-Based Visual Question Answering (KB-VQA) requires retrieving entity knowledge from external sources to answer questions that cannot be resolved from visual content alone. However, existing RAG systems suffer from the Single-Source Retrieval Bottleneck and Source-Specific Reranker Degradation due to their reliance on individual retrieval sources. To address these challenges, we propose UniSONAR, a unified lightweight framework that effectively processes candidates from heterogeneous retrieval sources. By integrating dual-source coarse retrieval followed by a novel Source-Conditioned Attentive Fusion, UniSONAR facilitates robust cross-source generalization and enables both entity-level and section-level retrieval. Furthermore, we introduce a hybrid training strategy using contrastive learning and an auxiliary loss to enhance discriminative feature learning. Extensive experiments on E-VQA and InfoSeek demonstrate that UniSONAR achieves state-of-the-art performance. Code will be released.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) requires models to answer natural language questions about images. A more challenging variant, Knowledge-Based VQA (KB-VQA), further requires external entity knowledge beyond visual content from sources like Wikipedia. While Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Liu et al., 2024; Zhu et al., 2025) excel at general VQA, they often require external augmentation for knowledge-intensive queries. Consequently, Retrieval-Augmented Generation (RAG) (Caffagni et al., 2024; Zhang et al., 2024; Cocchi et al., 2025; Yan and Xie, 2024; Yang et al., 2025; Yuan et al., 2025) has become the dominant paradigm, retrieving query-relevant knowledge to contextualize MLLM generation. RAG frameworks typically adopt a three-stage pipeline: coarse retrieval extracts top-k candidate entities via vision-language encoders; fine-grained reranking optimizes the candidate order; and answer generation derives the final response based on the top-ranked entity's knowledge.

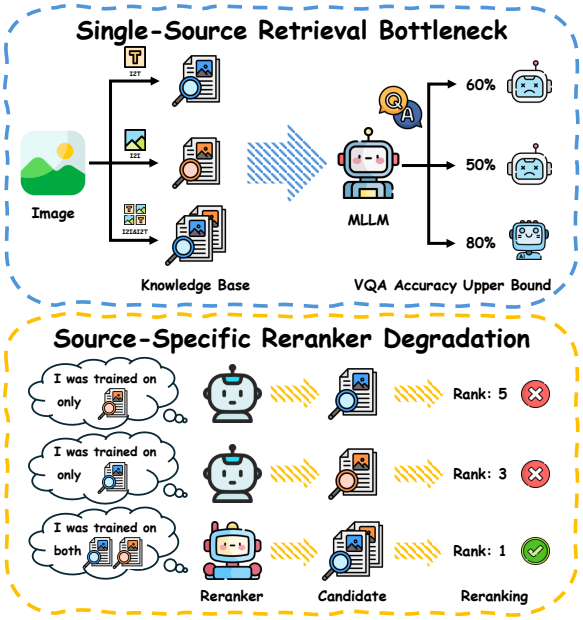

Figure 1: Limitations of existing KB-VQA systems. Top: Single-Source Retrieval Bottleneck, where ground-truth entities are frequently captured by only one modality, limiting downstream performance. Bottom: Source-Specific Reranker Degradation, where rerankers trained on single sources fail to generalize to non-native retrieval sources, misleading the answer generator.

As illustrated in Figure 1, current systems face two critical limitations. First, the reliance on individual retrieval sources, either image-to-image (I2I) or image-to-text (I2T), creates a Single-Source Retrieval Bottleneck. The bottleneck arises because these modalities diverge significantly: I2I focuses on morphological similarity (e.g., color, shape), while I2T emphasizes semantic consistency. As a result, they yield largely non-overlapping entity

sets under limited candidate budgets. Consequently, ground-truth entities are often retrieved by only one modality, rendering single-source approaches inherently incomplete. This bottleneck significantly hampers reranking performance, leading to hallucinated or incorrect answers. Second, current rerankers suffer from Source-Specific Reranker Degradation due to a lack of cross-modal generalization. Trained exclusively on single-source candidates, these models cannot effectively adapt when applied to non-native retrieval sources. This failure leads to the misranking of ground-truth entities, thereby propagating retrieval errors to the final answer generation.

To overcome these limitations, we propose **Uni**fied **S**ource-**Con**ditioned **A**ttentive **R**etrieval (**UniSONAR**), a lightweight multimodal RAG framework that effectively handles candidates from heterogeneous retrieval sources within a unified architecture. To fully utilize the distinct characteristics of dual source, we first perform Dual-Source Coarse-grained Retrieval by simultaneously matching the query image against entity images (I2I) and textual summaries (I2T). Subsequently, we introduce Source-Conditioned Attentive Fusion in the Multimodal Multisource Entity Reranking stage to enable robust cross-source generalization, culminating in the selection of the most reliable candidate via confidence-weighted fusion of top-ranked entities. Finally, we conduct Fine-grained Section Reranking within the selected entity to identify the most relevant knowledge for answer generation. Furthermore, we employ a training strategy that learns source-specific discriminative features via contrastive learning on fused representations, while preserving cross-modal matching capabilities through an auxiliary loss for downstream section-level retrieval. Extensive experiments demonstrate that our method outperforms existing state-of-the-art competitors. Our main contributions are as follows:

- We propose UniSONAR, a unified lightweight framework that efficiently synergizes I2I and I2T sources via our Source-Conditioned Multimodal Attentive Fusion module.
- We introduce a training strategy that combines contrastive learning on fused representations to learn source-specific features with an auxiliary loss that preserves cross-modal matching for section-level retrieval, enabling unified entity-to-section retrieval.
- Extensive experiments on E-VQA and InfoSeek demonstrate that our framework achieves state-of-the-art performance with high efficiency, establishing dual-source retrieval as a promising paradigm for knowledge-based VQA systems.

2 Related Work

2.1 KBVQA

KB-VQA extends traditional VQA (Antol et al., 2015) by incorporating external knowledge to answer questions requiring insights beyond the visual input. While early methods (Marino et al., 2021; Wu et al., 2022) achieved limited success, the field has advanced significantly through new benchmarks (Mensink et al., 2023; Chen et al., 2023) and LLM integration. PICa (Yang et al., 2022) pioneered GPT-3-based few-shot learning, leading to widespread adoption of In-Context Learning (ICL) (Hu et al., 2023; Khademi et al., 2023; Xenos et al., 2023; Shao et al., 2023) that relies on the parametric knowledge encoded in large language models (Achiam et al., 2023; Touvron et al., 2023). However, ICL methods rely entirely on parametric knowledge stored in model weights, lacking explicit knowledge sources or retrieval mechanisms to ground their predictions in verifiable evidence.

2.2 Multimodal RAG

Retrieval-Augmented Generation (RAG) (Lin and Byrne, 2022) has been expanded to KB-VQA to guide generation with external documents. Existing methods employ fine-grained encoding (Lin et al., 2024; Deng et al., 2025), hierarchical strategies (Caffagni et al., 2024; Yan and Xie, 2024), denoising (Jian et al., 2024; Qi et al., 2024), or reflective refinement (Zhang et al., 2024; Cocchi et al., 2025) to improve retrieval. Others utilize knowledge graphs (Yuan et al., 2025) or reinforcement learning (Hong et al., 2025). OMGM (Yang et al., 2025) specifically addresses granularity mismatch between coarse entity retrieval and fine-grained section matching. However, these approaches predominantly rely on a single retrieval modality: either image-to-image or image-to-text matching, which limits entity coverage when ground-truth entities are only captured by one source. In contrast, our method unifies both I2I and I2T retrieval within a single framework. By adaptively fusing complementary signals through source-conditioned attention, our approach exploits the synergistic poten-

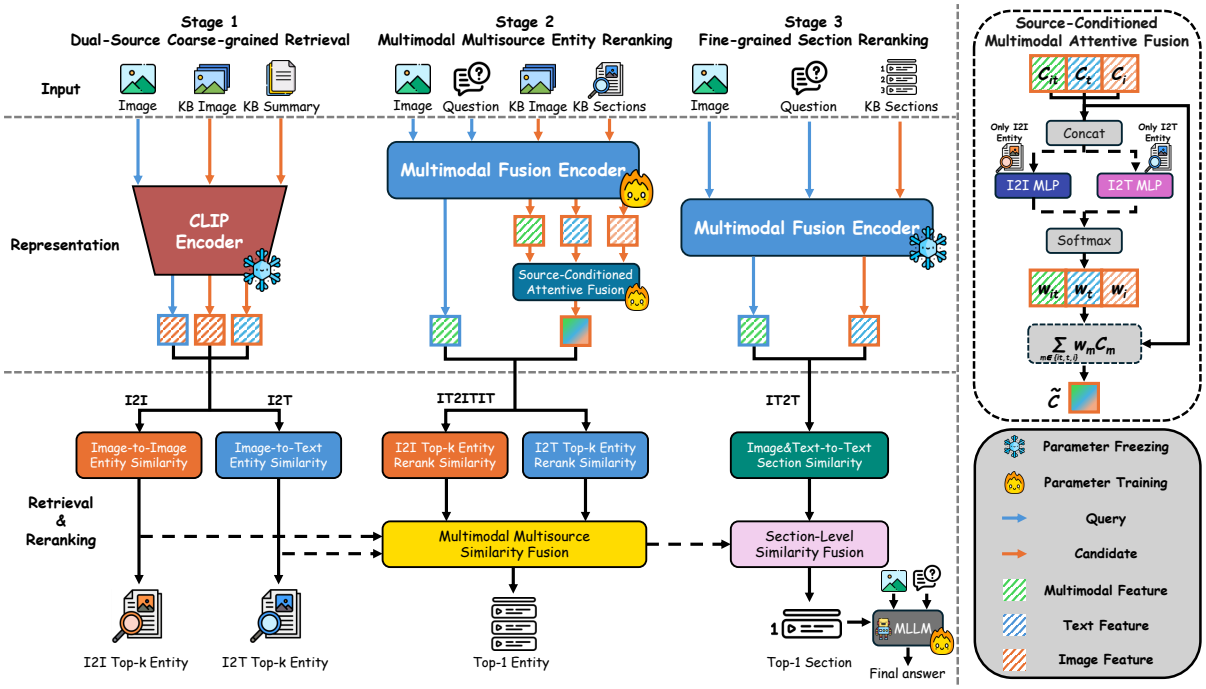

Figure 2: Overview of UniSONAR framework. Our approach operates in three stages: Stage 1: Dual-source coarse-grained retrieval obtains top-k candidates from both I2I and I2T sources; Stage 2: Multimodal multisource entity reranking employs source-conditioned attentive fusion to adaptively reweight multimodal features based on retrieval source, followed by confidence-weighted aggregation to select the top-1 entity; Stage 3: Fine-grained section reranking identifies the most relevant knowledge section for answer generation by MLLM.

tial of multi-source retrieval while maintaining a lightweight architecture.

3 Methodology

In this section, we present UniSONAR, a unified framework that jointly leverages image-to-image and image-to-text retrieval. Our approach operates in three stages: dual-source coarse retrieval obtains complementary candidate sets from both modalities, source-conditioned reranking adaptively fuses multimodal features to identify the most relevant entity, and section-level matching selects the optimal knowledge passage for answer generation. The complete pipeline is shown in Figure 2.

3.1 Dual-Source Coarse-Grained Retrieval

Coarse-grained entity retrieval aims to identify top-k candidates from a multimodal knowledge base. However, retrieval modalities exhibit distinct characteristics: visual-visual matching prioritizes morphological similarity such as color and shape, which can lead to visual ambiguity, whereas visual-text matching emphasizes semantic consistency, often resulting in conceptual over-generalization. Consequently, our dual-source retrieval strategy is designed to leverage the complementary strengths of these distinct retrieval behaviors.

Specifically, given a query image $\mathcal{I}_q$, candidate entities are retrieved by matching the query against two complementary knowledge-base representations: entity images $\mathcal{I}_e$ and textual summaries $\mathcal{S}_e$ of entity articles. This dual-branch process performs Image-to-Image (I2I) retrieval by aligning $\mathcal{I}_q$ with visual data, and Image-to-Text (I2T) retrieval by matching $\mathcal{I}_q$ against textual descriptions, formalized as,

$$\mathcal{C}_s^k = \mathcal{F}\big(\phi_v(\mathcal{I}_q), \mathbf{K}_s\big), \quad s \in \{\text{I2I}, \text{I2T}\} \qquad (1)$$

where $\mathbf{K}_s$ represents the set of knowledge-base embeddings encoded from raw data corresponding to source s, while $\phi_v(\cdot)$ and $\phi_t(\cdot)$ denote the visual and textual encoders of CLIP, respectively. Specifically, we define $\mathbf{K}_{\text{I2I}} = \phi_v(\mathcal{I}_e)$ and $\mathbf{K}_{\text{I2T}} = \phi_t(\mathcal{S}_e)$. The function $\mathcal{F}(\cdot)$ employs Faiss (Johnson et al., 2019) to retrieve the top-k nearest neighbors via inner-product similarity $sim_s^c(e)$ over pre-indexed embeddings. This yields two source-specific candidate sets $\mathcal{C}_{\text{I2I}}^k$ and $\mathcal{C}_{\text{I2T}}^k$ for subsequent reranking.

3.2 Multimodal Multisource Entity Reranking

With the coarse-grained candidate sets $\mathcal{C}^k_{\text{I2I}}$ and $\mathcal{C}^k_{\text{I2T}}$ obtained, the next critical step is to refine rankings through fine-grained feature interaction. While these sets provide high recall, they often lack sufficient discriminability due to visual homogeneity in I2I retrieval and semantic ambiguity in I2T retrieval. Therefore, we introduce Multimodal Multisource Entity Reranking, which explicitly exploits discriminative cues from the complementary modality via Source-Conditioned Multimodal Attentive Fusion module.

We employ VISTA (Zhou et al., 2024) as the multimodal encoder. Given an image-text pair $(\mathcal{I}, \mathcal{T})$, the encoder produces three modality-specific representations:

$$f_m = \text{BERT}(\psi_m(\mathcal{I}, \mathcal{T})), m \in \{it, t, i\} \quad (2)$$

where ψ_i extracts visual tokens via ViT (Dosovitskiy, 2020), ψ_t processes textual tokens, and ψ_{it} concatenates both. These token sequences are then encoded by BERT (Devlin et al., 2019).

For the query, we encode the image-question pair $(\mathcal{I}_q, \mathcal{T}_q)$ to obtain multimodal query representation $\mathbf{Q}_{it}$. For each candidate entity in $\mathcal{C}^k_s$ with knowledge base image $\mathcal{I}_e$ and article $\mathcal{T}_e$ comprising H sections $\mathcal{T}^h_e$ ($h \in [1, H]$), we encode each image-section pair to obtain section-level features $\mathbf{C}^h_{it}$, $\mathbf{C}^h_t$, $\mathbf{C}^h_i$ for multimodal, textual and visual modalities, respectively.

Source-Conditioned Multimodal Attentive Fusion. To learn discriminative modality preferences tailored to each retrieval source, we introduce Source-Conditioned Multimodal Attentive Fusion, which employs source-conditioned MLP_s to adaptively combine candidate features $\mathbf{C}^h_{it}$, $\mathbf{C}^h_t$, $\mathbf{C}^h_i$ from source s, producing section-level fused features $\tilde{\mathbf{C}}^h$:

$$\begin{cases} [w_{it}, w_t, w_i] = \text{Softmax}\big(\text{MLP}_s([\mathbf{C}^h_{it} \| \mathbf{C}^h_t \| \mathbf{C}^h_i])\big) \\ \tilde{\mathbf{C}}^h = w_{it}\mathbf{C}^h_{it} + w_t\mathbf{C}^h_t + w_i\mathbf{C}^h_i \end{cases} \quad (3)$$

where $\|$ denotes concatenation and MLP_s is the source-conditioned MLP for source s. To combine complementary strengths from coarse and fine-grained retrieval, we compute entity-level similarity $\text{sim}_s(e)$ by selecting the maximum query-section similarity across all H sections of entity e, then fusing with the coarse retrieval score:

$$sim_s(e) = \alpha_s \max_{h \in [1,H]} \mathbf{Q}^T_{it}\tilde{\mathbf{C}}^h + (1-\alpha_s) sim^c_s(e) \quad (4)$$

where $sim^c_s(e)$ is the coarse retrieval score from Eq. 1, and $\alpha_s \in [0, 1]$ balances the fine-grained and coarse signals.

Subsequently, to integrate the complementary signals from dual-source retrieval, we adopt a confidence-weighted fusion strategy, in which each source is assigned an adaptive weight according to its ranking reliability. Specifically, we quantify source confidence by the score margin between the top-ranked and second-ranked entities, where a larger margin indicates higher retrieval certainty:

$$\begin{cases} \Delta_s = sim_s(e^{(1)}) - sim_s(e^{(2)}) \\ \beta_s = \dfrac{\Delta_s}{\Delta_{\text{I2I}} + \Delta_{\text{I2T}}} \\ E^* = \underset{e \in (\mathcal{C}^k_{\text{I2I}} \cup \mathcal{C}^k_{\text{I2T}})}{\text{argmax}} \displaystyle\sum_s \beta_s \hat{sim}_s(e) \end{cases} \quad (5)$$

Here, $e^{(1)}$ and $e^{(2)}$ denote the top-1 and top-2 entities retrieved from source s, respectively, and $\hat{sim}_s(e)$ represents the z-score normalized similarity score for entity e under source s, with missing entities contributing zero.This confidence-aware aggregation yields Top-1 entity E^*.

3.3 Fine-grained Section Reranking

Given the selected entity E^*, we rerank its sections to identify the most relevant passage for answer generation. We combine section-level image-text-to-text (IT2T) similarity with entity-level retrieval scores to jointly leverage fine-grained and coarse-grained signals:

$$S^* = \max_{h \in [1,H]} [\gamma \mathbf{Q}^T_{it}\mathbf{C}^h_t + (1-\gamma) \max_s sim^h_s(E^*)] \quad (6)$$

where $\mathbf{C}^h_t$ denotes the textual feature of section h from E^*, and $\gamma \in [0, 1]$ balances section-level specificity and entity-level confidence. The predicted Top-1 section S^* is provided to an MLLM alongside $(\mathcal{I}_q, \mathcal{T}_q)$ for answer generation.

3.4 Reranker Training Objective

To train our reranker, we construct training pairs from the top-k candidates $\mathcal{C}^k_s$ retrieved in Stage 1. Positive pairs consist of the query with ground-truth entity images and their correct sections, while hard negatives combine the query with incorrect

entity images and random sections from those entities. We obtain N training pairs from both I2I and I2T sources. Pairs from each source are processed through their respective source-conditioned MLP$_s$ and fused into candidate features $\tilde{\mathbf{C}}$ via Eq. 3.

We optimize two complementary objectives, where λ balances their contributions. The primary contrastive loss trains the fused representation:

$$\mathcal{L}_{con} = -\log \frac{\exp(\mathbf{Q}_{it}^T \tilde{\mathbf{C}}^+ / \tau)}{\sum_{k=1}^{N} \exp(\mathbf{Q}_{it}^T \tilde{\mathbf{C}}^k / \tau)} \quad (7)$$

where $\tilde{\mathbf{C}}^+$ is the fused feature of the positive pair, τ is temperature. To prevent modality collapse and preserve IT2T matching capabilities required for downstream section-level reranking, we introduce an auxiliary loss that maintains alignment across multimodal, visual, and textual feature pairs:

$$\mathcal{L}_{aux} = -\sum_{m \in \{it,t,i\}} \log \frac{\exp(\mathbf{Q}_{it}^T \mathbf{C}_m^+ / \tau)}{\sum_{k=1}^{N} \exp(\mathbf{Q}_{it}^T \mathbf{C}_m^k / \tau)} \quad (8)$$

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda \mathcal{L}_{aux}, \quad (9)$$

4 Experiments

4.1 Datasets and Metrics

Datasets. We conduct experiments on two widely-adopted KB-VQA datasets: E-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023). E-VQA comprises 1M triplets $(\mathcal{I}_q, \mathcal{T}_q, y)$, generated by pairing each of the 221k unique QA pairs (spanning 16.7k entities) with up to five distinct entity images. Following EchoSight (Yan and Xie, 2024), we evaluate on single-hop questions, which are answerable from a single Wikipedia page, yielding 4.7k evaluation samples and a knowledge base of 2M Wikipedia articles. InfoSeek contains 1.3M triplets distributed across roughly 11k Wikipedia entities, partitioned into 934k training, 73k validation, and 348k test samples. Since ground-truth answers for the test split are not publicly available, we evaluate on the validation set, which contains both unseen entities (Unseen-E) and novel questions (Unseen-Q). We adopt the standard 100k knowledge base subset from the original 6M pages, consistent with recent work (Yang et al., 2025).

Metrics. For retrieval accuracy evaluation, we employ Recall@k to measure whether the GT appears within the top-k candidates, enforcing strict URL matching between the entities' Wikipedia URL exactly matches the GT. Moreover, we adopt dataset-specific metrics (BEM (Zhang et al., 2019) for E-VQA dataset and both VQA accuracy (Antol et al., 2015) and Relaxed accuracy (Methani et al., 2020) for InfoSeek) aligned with standard practices for answer quality assessment.

4.2 Implementation Details

Our Multimodal Multisource Reranker utilizes the VISTA (Zhou et al., 2024) architecture, combining a ViT-based visual encoder (EVA-CLIP-02-Base (Sun et al., 2023)) and a BERT-based text encoder (BGE-Base-v1.5 (Xiao et al., 2024)) for unified multimodal representation. Initialized with VISTA Stage 1 checkpoint, we freeze the visual encoder while fine-tuning the text encoder and source-conditioned fusion MLP$_s$. This facilitates IT2ITIT retrieval, supporting both entity and section-level reranking in a unified framework. Furthermore, we employ LoRA (Hu et al., 2022) to efficiently fine-tune the Qwen3-VL (Bai et al., 2025) model for answer generation, enabling it to leverage the triplets from the retrieval pipeline and align with our knowledge format. More training details are provided in Appendix B and C.

4.3 Main Results

The results of our method compared with other approaches are presented in Table 1 and Table 2 with additional efficiency comparison in Table 3.

Retrieval Performance. We compare our method against several state-of-the-art baselines, including Wiki-LLaVA (Caffagni et al., 2024), mR^{2}AG (Zhang et al., 2024), LLM-RA (Jian et al., 2024), VLM-PRF (Hong et al., 2025), ReflectiVA (Cocchi et al., 2025), EchoSight (Yan and Xie, 2024), OMGM (Yang et al., 2025). As illustrated in Table 1, methods focusing on inference-time optimization within the LLM generator (Wiki-LLaVA, ReflectiVA) suffer from poor performance, especially at low Recall@k on E-VQA. Likewise, multi-stage retrieval approaches restricted to single sources (EchoSight, OMGM) also achieve limited recall. In contrast, UniSONAR addresses these limitations, achieving state-of-the-art performance on E-VQA and InfoSeek with absolute Recall@1 gains of +4.7% and +2.4%, respectively. The pronounced improvement at higher k values (e.g., +11.2% on E-VQA R@20) highlights that our dual-source strategy substantially enhances candidate diversity compared to single-source baselines. This

Method	E-VQA				InfoSeek			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Wiki-LLaVA	3.3	-	9.9	13.2	36.9	-	66.1	71.9
mR^{2}AG	-	-	-	-	38.0	-	65.0	71.0
LLM-RA	-	-	-	-	47.3	53.8	-	-
VLM-PRF	-	-	-	-	54.9	-	-	-
ReflectiVA	15.6	36.1	-	49.8	56.1	77.6	-	86.4
Reranking on Image-to-Image coarse retrieval								
w/o Reranking	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
EchoSight	36.5	47.9	48.8	48.8	53.2	74.0	77.4	77.9
OMGM†	34.0	47.0	48.8	48.8	57.5	73.3	76.8	77.9
UniSONAR (Ours)	**39.3**	**48.4**	**48.8**	**48.8**	**57.7**	**75.8**	**77.9**	**77.9**
Reranking on Image-to-Text coarse retrieval								
w/o Reranking	19.1	41.2	49.8	58.7	52.6	73.9	80.0	84.8
EchoSight†	34.4	53.2	57.6	58.7	49.4	75.8	82.2	84.8
OMGM	42.8	55.7	58.1	58.7	64.0	80.8	83.6	84.8
UniSONAR (Ours)	**43.0**	**55.9**	**58.1**	**58.7**	**64.3**	**81.1**	**83.9**	**84.8**
Reranking on Dual-source coarse retrieval								
UniSONAR (Ours)	**47.5** (+4.7)	**60.7** (+5.0)	**67.0** (+8.9)	**69.9** (+11.2)	**66.4** (+2.4)	**82.3** (+1.5)	**86.4** (+2.8)	**88.5** (+2.1)

Table 1: Retrieval performance on E-VQA test split and InfoSeek validation split. w/o Reranking indicates coarse retrieval results. † denotes results evaluated using official checkpoints on non-native sources. Best in bold; second best underlined (under the same setting). Improvements over best baseline are shown in parentheses.

Model	Generator	E-VQA	InfoSeek		
			Unseen-Q	Unseen-E	Overall
Zero-shot MLLMs					
LLaVA-1.5-7B	-	14.5	8.9	8.8	8.8
InternVL3-8B	-	17.7	14.1	12.5	13.3
Qwen3-VL-8B	-	19.1	17.5	16.2	16.9
Retrieval-Augmented Models					
RoRA-VLM	LLaVA-1.5-7B	20.3	27.3	25.1	-
Wiki-LLaVA	LLaVA-1.5-7B	21.8	30.1	27.8	28.9
LLM-RA	BLIP2-Flan-T5XL	-	26.1	20.9	23.1
EchoSight	Mistral-7B | LLaMA3-8B	41.8	-	-	31.3
ReflectiVA	LLaVA-MORE-8B	35.5	40.4	39.8	40.1
mR^{2}AG	LLaVA-1.5-7B	-	40.6	39.8	40.2
mKG-RAG	LLaVA-MORE-8B	38.4	41.4	39.6	40.5
VLM-PRF	InternVL3-8B	43.5	40.4	42.1	42.5
OMGM	LLaVA-1.5-7B	50.2	43.5	43.5	43.5
UniSONAR (Ours)	Qwen3-VL-8B	**52.3**	**44.5**	**44.8**	**44.6**

Table 2: VQA performance on E-VQA test split and InfoSeek validation split. Best in bold; second best underlined.

advantage is particularly critical for E-VQA's 2M-article knowledge base, where I2I and I2T provide complementary coverage of ground-truth entities. Moreover, our unified model outperforms specialized baselines on their native sources, validating the effectiveness of the Source-Conditioned Fusion module in learning discriminative, source-specific representations.

VQA Performance. To evaluate end-to-end VQA accuracy, we compare our method against two categories of baselines: standard zero-shot MLLMs (LLaVA-1.5-7B (Liu et al., 2024), InternVL3-8B (Zhu et al., 2025), and Qwen3-VL-8B (Bai et al., 2025)) and retrieval-augmented approaches (RoRA-VLM (Qi et al., 2024), Wiki-LLaVA (Caffagni et al., 2024), LLM-RA (Jian et al., 2024), EchoSight (Yan and Xie, 2024), mR^{2}AG (Zhang et al., 2024), ReflectiVA (Cocchi et al., 2025), mKG-RAG (Yuan et al., 2025), VLM-PRF (Hong et al., 2025), and OMGM (Yang et al., 2025)). As shown in Table 2, zero-shot MLLMs yield low accuracy on both datasets, highlighting the necessity of external encyclopedic knowledge for answering knowledge-intensive visual questions. Notably, our approach outperforms the state-of-the-art baseline, achieving improve-

ments of +2.1% on E-VQA and +1.1% on InfoSeek. This demonstrates that incorporating dual-source retrieval with source-conditioned fusion significantly refines knowledge selection and improves generation quality.

Model	Backbone	Params	Ret. Type	Time(s)
EchoSight	BLIP-2	1.2B	IT2T	0.67
OMGM	BLIP-2	1.2B	IT2IT	3.50
UniSONAR	VISTA	**198M**	**IT2ITIT**	**0.62**

Table 3: Efficiency comparison. Retrieval time is averaged per sample over I2I and I2T sources on both benchmarks (on a single NVIDIA A6000).

Efficiency Analysis. Table 3 highlights our framework's efficiency. UniSONAR achieves competitive speed compared to single-source methods. The fusion modules are lightweight, adding only 1.18M parameters (0.6% overhead to the VISTA backbone), making this dual-source approach highly practical for large-scale KBVQA.

4.4 Ablation Study

Effect of Dual-Source Retrieval. Table 4 analyzes performance across varying candidate pool sizes k on E-VQA. The oracle results validate the single-source bottleneck shown in Fig. 1: at k=20, dual-source oracle (70.1%) substantially exceeds I2I (48.8%) or I2T (58.7%) alone, confirming that ground-truth entities are frequently captured by only one modality. Our dual-source reranker effectively exploits this complementarity via confidence-weighted fusion, achieving 60.7% Recall@5 at k=20. While larger k improves oracle performance, it increases reranking difficulty and computational cost with diminishing returns. We adopt k=20 to balance accuracy and efficiency.

Method	**Recall@5 on E-VQA**			
	k=10	k=20	k=50	k=100
I2I-Based Retrieval				
w/o Reranking		31.3		
Oracle	41.0	48.8	57.9	64.1
UniSONAR	40.9	48.4	55.0	58.1
Ret. Time(s)	*(0.44)*	*(0.62)*	*(2.13)*	*(6.52)*
I2T-Based Retrieval				
w/o Reranking		41.2		
Oracle	49.8	58.7	67.4	73.7
UniSONAR	49.0	55.9	61.3	64.3
Ret. Time(s)	*(0.47)*	*(0.62)*	*(1.97)*	*(5.98)*
Dual-Source Retrieval				
Oracle	62.1	70.1	77.3	81.6
UniSONAR	58.0	60.7	64.9	66.8
Ret. Time(s)	*(0.92)*	*(1.24)*	*(4.10)*	*(12.50)*

Table 4: Impact of candidate pool size k on retrieval performance and efficiency. Oracle indicates the coverage ceiling (Recall@k converted to Recall@5). w/o Reranking shows baseline Recall@5 from coarse retrieval.

Dataset	I2I Recall@K			I2T Recall@K		
	R@1	R@5	R@10	R@1	R@5	R@10
w/o I2I	31.4	43.2	46.5	38.9	53.7	57.5
w/o I2T	37.0	47.8	48.7	36.6	52.0	56.8
UniSONAR	37.5	48.2	48.8	39.7	55.1	57.9

Table 5: Impact of training data composition on cross-source generalization (E-VQA). We set α_s=1 for all experiments to focus on reranking performance without coarse score fusion.

Impact of Cross-Source Training Data. As shown in Table 1, source-specific rerankers exhibit degradation when applied to non-native sources. EchoSight and OMGM both show performance drops on their non-native retrieval sources. Table 5 ablates training data composition. Training exclusively on one source leads to substantial degradation on the other. More importantly, joint training on both sources outperforms single-source training even on individual sources. This stems from complementary learning: dual-source training provides both hard negatives from the native source and soft negatives from the complementary source, enabling more discriminative representations across modalities.

Ret. Modality	E-VQA		InfoSeek	
	I2I	I2T	I2I	I2T
$q_t \rightarrow c_t$	35.1	25.8	38.2	42.0
$q_{it} \rightarrow c_{it}$	21.6	28.2	37.9	46.3
$q_{it} \rightarrow (c_i, c_t)$	32.0	27.9	38.1	44.3
$q_{it} \rightarrow (c_{it}, c_i, c_t)$	37.5	39.7	39.5	45.8

Table 6: Modality ablation for entity-level reranking (Recall@1). We set α_s=1 to isolate reranking performance. Our three-modality fusion achieves optimal performance on both sources.

Impact of Reranking Modality. Table 6 ablates modality configurations for entity-level reranking. Results reveal distinct preferences across sources: for I2I retrieval, text matching substantially outperforms multimodal matching on E-VQA, as I2I candidates are already visually similar and require textual discrimination. Conversely, I2T retrieval benefits from multimodal

features since semantic alignment is captured by coarse retrieval. Our approach leveraging all three modalities ($q_{it} \rightarrow (c_{it}, c_i, c_t)$) achieves optimal performance on both sources, demonstrating the effectiveness of source-conditioned fusion.

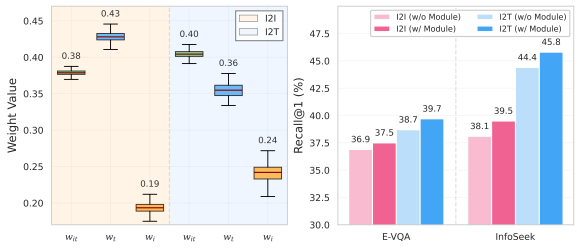

Figure 3: Analysis of source-conditioned fusion module. Left: Weight distribution across top-1 entities, showing source-specific modality preferences. Right: Ablation study demonstrating the module's impact on retrieval accuracy across both sources and datasets.

Effect of Source-Conditioned Multimodal Attentive Fusion. Figure 3 validates our fusion module's effectiveness. The learned weight distribution (left) reveals distinct modality preferences: I2I emphasizes textual and multimodal features while down-weighting visual cues, as candidates are already visually similar; conversely, I2T relies more on multimodal features for fine-grained discrimination. Ablation results (right) show that removing the module causes performance drops on both sources, particularly severe on I2T (-4.7% on E-VQA), confirming that adaptive fusion is essential for effective reranking.

Model	R@1	R@5	R@10
I2I-Based Retrieval			
EchoSight†	6.6	19.3	20.7
UniSONAR (Ours)	**31.3**	**38.0**	**38.5**
I2T-Based Retrieval			
OMGM	32.8	-	-
UniSONAR (Ours)	**34.1**	**41.2**	**41.7**
Dual-Source Retrieval			
UniSONAR (Ours)	**39.2**	**46.8**	**47.3**

Table 7: Section-level retrieval performance on E-VQA. † indicates reproduced results.

Impact of Section-Level Retrieval. Unlike OMGM, which requires a separate text-to-text reranker for section retrieval, our framework preserves cross-modal alignment through auxiliary training (Eq. 9), enabling natural IT2T section matching without additional modules. Table 7 shows substantial improvements over baselines on E-VQA: our unified architecture outperforms baselines even on individual sources. Most importantly, dual-source fusion achieves the best performance, demonstrating that modality complementarity extends to fine-grained section matching.

Stage 1	Stage 2	Stage 3	Stage 4	E-VQA	InfoSeek
✗	✗	✗	✗	19.1	16.9
✓	✗	✗	✗	28.4	31.6
✓	✓	✗	✗	43.2	34.3
✓	✓	✓	✗	45.2	35.7
✓	✓	✓	✓	**52.3**	**44.6**

Table 8: Progressive ablation on VQA accuracy. Stage 1: I2T coarse retrieval. Stage 2: Dual-source entity reranking. Stage 3: Section-level reranking. Stage 4: MLLM fine-tuning.

Impact of Pipeline Stages. Table 8 shows progressive improvements from each pipeline stage. Starting from a zero-shot MLLM baseline, coarse retrieval substantially improves accuracy by providing relevant knowledge. Entity-level reranking yields the largest gain by leveraging dual-source fusion to accurately select ground-truth entities. Section-level reranking further refines context quality, and MLLM fine-tuning adapts to dataset-specific distributions. Each stage proves essential, with the complete pipeline achieving strong performance on both benchmarks.

5 Conclusion

In this paper, we present UniSONAR, a unified lightweight framework specifically designed to address two critical limitations hindering current knowledge-based visual question answering systems: the Single-Source Retrieval Bottleneck and Source-Specific Reranker Degradation. By integrating dual-source retrieval with our novel Source-Conditioned Attentive Fusion module, our approach effectively synergizes the complementary strengths of heterogeneous retrieval sources. Consequently, UniSONAR achieves State-of-the-Art performance, surpassing the suboptimal baselines by significant margins of +4.7% and +2.4% Recall@1 on E-VQA and InfoSeek, respectively, while preserving superior VQA accuracy and high efficiency. Our work establishes adaptive fusion as a promising paradigm for advancing KB-VQA and multimodal RAG systems.

Limitations

Our framework exhibits several limitations that warrant future investigation. First, while multimodal encoding is unified across sources, source-specific MLPs are still required to generate fusion weights for each retrieval modality. A more intrinsic source perception mechanism that adaptively adjusts fusion strategies without explicit source indicators would reduce architectural complexity. Second, our approach requires encoding each candidate three times to extract multimodal, textual, and visual features independently, increasing inference cost proportionally. Exploring unified representations that preserve modality-specific discriminative power while reducing redundant encoding operations could improve computational efficiency. Third, our confidence-weighted fusion for cross-source entity aggregation operates without explicit supervision, relying solely on score margins. An end-to-end trained aggregation module that learns to dynamically weight and select entities based on query-source alignment could better exploit multi-source complementarity and further approach the oracle upper bound.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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/CVF International Conference on Computer Vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9209.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Lianghao Deng, Yuchong Sun, Shizhe Chen, Ning Yang, Yunfeng Wang, and Ruihua Song. 2025. Muka: Multimodal knowledge augmented visual information-seeking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9675–9686.

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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. 2025. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. *arXiv preprint arXiv:2510.14605*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. Mm-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*.

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.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

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.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

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*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Joseph A. Shaw and Edward A. Fox. 1994. Combination of multiple searches. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 105–108. National Institute of Standards and Technology (NIST).

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for listwise document reranking. *Preprint*, arXiv:2509.25085.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. *arXiv preprint arXiv:2310.13570*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

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

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

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3081–3089.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. 2025. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *arXiv preprint arXiv:2508.05318*.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024. mr ^{2} ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*.

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

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Prompt Templates

We design two prompt templates for answer generation: a zero-shot baseline and a retrieval-augmented version. Both prompts are tailored to dataset requirements, enforcing concise answers and standardized numerical formatting to align with evaluation protocols.

The zero-shot prompt instructs the model to answer from visual content and parametric knowledge, establishing a baseline. The retrieval-augmented prompt explicitly encourages the model to leverage its reasoning capabilities rather than over-relying on retrieved context. This design handles three scenarios: (1) correct entity with correct section, (2) correct entity with incorrect section, and (3) incorrect entity retrieval. By instructing the model to combine Wikipedia context with visual evidence and internal knowledge when appropriate, the prompt enables robust answer generation even when retrieval fails, preventing over-reliance on potentially irrelevant context.

Prompt for Retrieval-Augmented Models

System Prompt

```
You are an encyclopedic visual question
answering assistant.

Answer strategy:
1. If Wikipedia context contains the
answer → answer directly
2. If context has partial info → combine
with image details and your reasoning
3. If context is unrelated → answer from
image and your own internal knowledge

Format:
- Keep answers under 5 words.
- For numbers, dates, or counts, use
Arabic numerals (e.g., "2-8" not
"two-eight"). When context contains a
numerical answer (number or range),
preserve the format and units from
context.
- Do not output the explanations or
reasoning process. Perform all reasoning
internally.
```

User Prompt

```
Question: {question}
Wikipedia Context: {Wikipedia}
Answer:
```

Prompt for Zero-shot MLLMs

System Prompt

```
You are an encyclopedic visual question
answering assistant. Answer based solely
on the image and your internal
knowledge.

Rules:
- Do not describe image content
- Keep answers under 5 words
- For numbers, dates, or counts, use
Arabic numerals (e.g., "2-8" not
"two-eight")
- No explanations or reasoning
```

User Prompt

```
Question: {question}
Answer:
```

B More Training Details of Reranker

B.1 Model Architecture and Initialization.

We employ VISTA (Zhou et al., 2024) as our multimodal backbone, initializing from the publicly available VISTA Stage 1 checkpoint. During training, we freeze the visual encoder (EVA-CLIP-02-Base(Sun et al., 2023)) and fine-tune only the text encoder (BGE-Base-v1.5(Xiao et al., 2024)) along with our proposed Source-Conditioned Multimodal Attentive Fusion modules. This strategy preserves the visual encoder's pre-trained representations while enabling task-specific adaptation of textual and fusion components.

B.2 Training Data Construction.

To balance task difficulty across datasets, we construct 200k training samples: 160k from E-VQA and 40k from InfoSeek (Table 9). Since InfoSeek lacks ground-truth section annotations, we employ jina-reranker-v3 (Wang et al., 2025) to identify the most relevant sections for each ground-truth entity. We allocate 80k samples to I2I retrieval and 120k to I2T retrieval to facilitate training for both sources.

For each sample, we construct one positive pair

Datasets	Ret.Train	Gen. Train	Test	KB
E-VQA	160k	100k	71335	2M
InfoSeek	40k	100k	4750	100k

Table 9: Dataset statistics and knowledge base configurations. Ret. Train and Gen. Train denote retrieval and generation training samples, respectively.

and 15 hard negatives. The query consists of the input image $\mathcal{I}_q$ and question $\mathcal{T}_q$. Positive pairs combine the ground-truth entity's first image with its correct section, while negatives are sampled from other top-20 entities retrieved by the same source, pairing their images with random sections. During training, each batch contains samples from only one retrieval source (I2I or I2T), training the corresponding source-specific MLP_s in the fusion module.

B.3 Training Configuration.

We train with a learning rate of 2e-5, global batch size of 64, and employ DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) for memory-efficient training. Training on 4× NVIDIA A6000 GPUs takes approximately 5 hours for 1 epoch. The training objective is defined in Eq. 9, where λ controls the relative importance between the loss terms. We set $\lambda = 0.1$ for all experiments.

B.4 Inference Hyperparameters.

At inference time, we set α_{I2I}=0.45 and α_{I2T}=0.25 to balance coarse and fine-grained retrieval scores for entity-level reranking (Eq. 4). For section-level reranking, we set γ=0.45 (Eq. 6). Ablation studies on these hyperparameters are provided in Section D.2.

C More Training Details of MLLM

C.1 Training Data Construction.

We fine-tune the MLLM on 200k samples comprising 100k from E-VQA and 100k from InfoSeek(Table 9). Each sample consists of a four-tuple: query image $\mathcal{I}_q$, question $\mathcal{T}_q$, retrieved section S, and ground-truth answer y. To ensure the model maintains reasoning capabilities when retrieval errors occur, we construct three types of training instances with different section qualities:

- **Positive samples (60%)**: Correct entity with correct evidence section, representing successful retrieval.
- **Hard negatives (20%)**: Correct entity with incorrect section, simulating section-level retrieval errors.
- **Soft negatives (20%)**: Incorrect entity with random section, simulating entity-level retrieval failures.

For InfoSeek samples, ground-truth sections are obtained using jina-reranker-v3 as described in Section B. This distribution mirrors realistic retrieval scenarios and encourages the model to leverage both retrieved context and internal knowledge appropriately. Ablation studies validating this design are provided in Section D.4.

C.2 Training Configuration.

We employ Qwen3-VL-8B (Bai et al., 2025) as our multimodal backbone and apply LoRA (Hu et al., 2022) for parameter-efficient fine-tuning with rank 64 and α=128. Training uses a learning rate of 2e-5, global batch size of 32, and cosine schedule with 5% warmup ratio. Training for one epoch on 4× A6000 GPUs takes approximately 12 hours.

D More Ablation Experiments

D.1 Impact of Textual Granularity for Entity Reranking.

Granularity	I2I Based			I2T Based		
	R@1	R@5	R@10	R@1	R@5	R@10
article	24.1	46.3	48.6	33.9	53.5	57.6
summary	34.2	47.7	48.8	35.1	54.9	57.8
section	37.5	48.2	48.8	39.7	55.1	57.9

Table 10: Impact of textual granularity for entity representation in reranking on E-VQA.

At entity-level reranking, we must choose an appropriate textual representation for each candidate. Table 10 compares three granularity levels: full article, article summary, and individual sections. Full article encoding severely degrades performance. Aggregating multiple sections (ranging from 1-2 to dozens per entity) into a single representation dilutes fine-grained semantic signals and creates inconsistent features due to substantial length variance (hundreds to thousands of tokens). This also incurs memory overhead, limiting batch size. Article summaries improve both accuracy and efficiency by providing concise entity descriptions. However, they lack fine-grained details necessary to align with specific question intents.Section-level representation achieves optimal performance, balancing discriminative power and efficiency. By matching queries against individual sections, the model identifies which specific knowledge passage addresses the question while maintaining reasonable sequence lengths. This granularity also aligns with our subsequent section-level reranking stage.

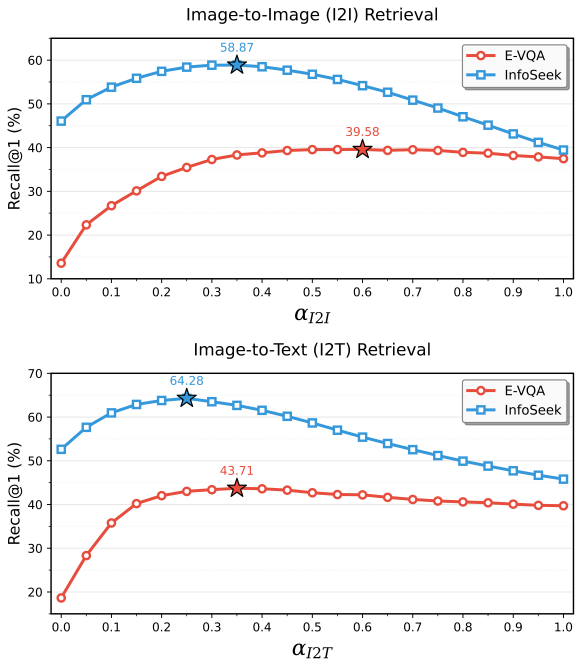

Figure 4: Impact of entity-level fusion weight α_s on Recall@1. The optimal values marked with stars.

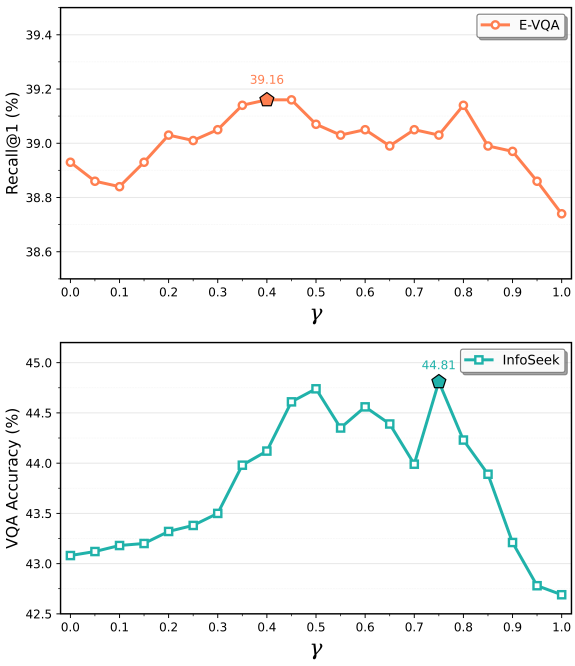

Figure 5: Impact of section-level fusion weight γ on E-VQA Recall@1 and InfoSeek VQA Accuracy. Best values marked with pentagons.

D.2 Impact of Hyperparameters.

Figure 4 examines α_s, which balances coarse retrieval and fine-grained reranking scores (Eq. 4). Both I2I and I2T retrieval exhibit smooth performance curves with optimal values around $\alpha \in [0.3, 0.5]$, demonstrating that fusing complementary signals from large-scale coarse encoders and lightweight rerankers consistently improves entity recall. Notably, I2I retrieval achieves peak performance at $\alpha_{\text{I2I}} = 0.45$, while I2T peaks at $\alpha_{\text{I2T}} = 0.25$, reflecting their distinct modality characteristics. The stable performance across a wide range validates the robustness of our fusion strategy. We adopt these optimal values for all experiments. For section-level fusion (Figure 5), γ combines entity-level confidence with IT2T retrieval capability. Due to the absence of ground-truth section annotations in InfoSeek, we evaluate on E-VQA Recall@1 and InfoSeek VQA Accuracy, adopting $\gamma = 0.45$ to balance performance across datasets.

D.3 Impact of Multi-Source Fusion Strategies.

Our framework requires fusing retrieval signals at multiple stages: (1) combining coarse and fine-grained scores within each source, (2)merging I2I and I2T sources for final entity selection, and (3) integrating IT2ITIT and IT2T modalities for section-level matching. We compare three fusion strategies on E-VQA using Recall@1: Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), CombSUM (Shaw and Fox, 1994), and confidence-weighted fusion (Eq. 5).

Table 11 presents results across different fusion stages. CombSUM achieves the best per-source performance, demonstrating effectiveness for within-source coarse-fine combination and section-level fusion. However, confidence-weighted fusion excels at dual-source entity selection (47.7%), leveraging score margins to identify the more reliable source. Based on these findings, we adopt CombSUM for coarse-fine fusion and section-level reranking, while using confidence weighting for final entity-level dual-source fusion (Eq. 5).

Strategy	Single-Source		Dual	Section
	I2I	I2T		
RRF	32.4	35.6	40.9	38.6
CombSUM	**39.6**	**43.6**	44.8	**39.2**
Conf-W	36.4	40.4	**47.7**	39.1

Table 11: Impact of fusion strategies on E-VQA (Recall@1). Single-Source: entity retrieval on I2I or I2T individually. Dual-Source: entity selection by fusing both sources. Section: section-level retrieval. Conf-W: Confidence-weighted fusion (Eq. 5).

D.4 Impact of Hard and Soft Negatives in MLLM Training.

To fine-tune the MLLM for answer generation with retrieved knowledge, we construct a training dataset that mirrors the retrieval pipeline's outputs. Each sample consists of: (1) a positive pair: the ground-truth entity's correct section, (2) a hard negative: an incorrect section from the same ground-truth entity, and (3) a soft negative: an incorrect section from a wrong entity. This design enables the model to learn fine-grained section discrimination while maintaining robustness when retrieval errors occur.

Training Data	E-VQA	InfoSeek		
		U-Q	U-E	All
w/o Hard Negatives	48.4	35.6	35.5	35.5
w/o Soft Negatives	49.7	41.1	41.3	41.2
UniSONAR	**52.3**	**44.5**	**44.8**	**44.6**

Table 12: Impact of hard and soft negatives in MLLM training on VQA accuracy.

Table 12 validates the necessity of both negative types. Removing hard negatives causes substantial degradation (e.g., -3.9% on E-VQA), as the model loses the ability to distinguish between sections within the correct entity. Removing soft negatives leads to moderate performance drops (e.g., -2.6% on E-VQA), as the model becomes less robust to entity-level retrieval errors. The full training strategy with both negative types achieves optimal performance, confirming that aligning training data distribution with actual retrieval outputs is essential for effective answer generation.

E Case Study

To provide intuitive insights into our framework's retrieval behavior, we visualize representative examples in Figure 6.

Success Cases (Rows 1-3). The first two rows demonstrate how dual-source retrieval overcomes single-source limitations on E-VQA. Despite poor coarse retrieval ranks from both I2I and I2T sources, our source-conditioned reranker successfully exploits cross-modal complementarity, achieving strong rankings on both individual sources and ultimately selecting the correct entity through confidence-weighted fusion.

The third row illustrates a particularly challenging InfoSeek case where visual similarity misleads both retrieval sources: when the query image prominently features a police vehicle, both I2I and I2T retrieval return police-related entities. However, the question asks about the owner of the *building behind the police car*. This highlights the critical advantage of our fine-grained multimodal fusion. By jointly modeling query text with candidate images and text, the reranker successfully identifies the ground-truth entity that single-source methods miss entirely.

Failure Cases (Rows 4-5). The failure cases reveal remaining limitations. Row 4 shows a scenario where extremely high I2I visual similarity causes the confidence-weighted fusion to over-rely on the I2I branch, despite I2T initially retrieving the correct entity. This results in the ground-truth being demoted, highlighting that our unsupervised cross-source aggregation strategy may not optimally balance source confidence in all cases.

Row 5 exposes another challenge: knowledge base entities often contain multiple images, but our framework uses only the first image during both training and inference. When this first image is unrepresentative or depicts irrelevant aspects of the entity, retrieval performance degrades. Developing methods to effectively leverage all available entity images is a promising avenue for future improvement.

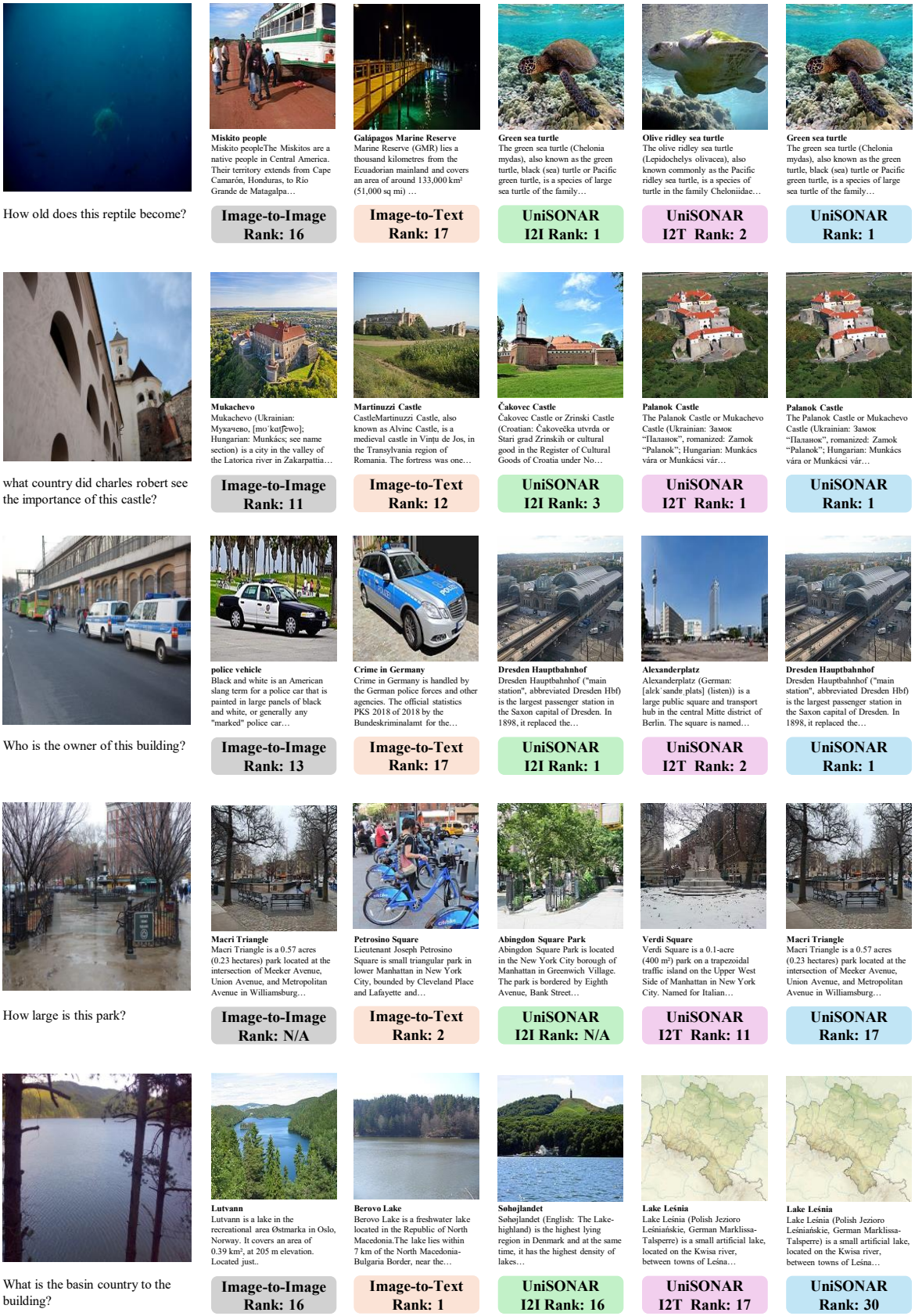

Figure 6: Qualitative retrieval examples on E-VQA and InfoSeek. Each sample shows the query (left) and retrieval results with ground-truth ranks. Rows 1-3: Success cases where UniSONAR achieves rank 1 despite poor single-source performance. Rows 4-5: Failure cases where dual-source fusion does not improve over coarse retrieval.