MaS-VQA: A Mask-and-Select Framework for Knowledge-Based Visual Question Answering

Xianwei Mao[1], **Kai Ye**[1], **Sheng Zhou**[1*]
Nan Zhang[2], **Haikuan Huang**[2], **Bin Li**[2], **Jiajun Bu**[1]
[1]Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems,
Zhejiang University [2]Alibaba Group
{maoxianwei, zhousheng_zju}@zju.edu.cn

Abstract

Knowledge-based Visual Question Answering (KB-VQA) requires models to answer questions by integrating visual information with external knowledge. However, retrieved knowledge is often noisy, partially irrelevant, or misaligned with the visual content, while internal model knowledge is difficult to control and interpret. Naïve aggregation of these sources limits reasoning effectiveness and reduces answer accuracy. To address this, we propose MaS-VQA, a selection-driven framework that tightly couples explicit knowledge filtering with implicit knowledge reasoning. MaS-VQA first retrieves candidate passages and applies a Mask-and-Select mechanism to jointly prune irrelevant image regions and weakly relevant knowledge fragments, producing compact, high-signal multimodal knowledge . This filtered knowledge then guides the activation of internal knowledge in a constrained semantic space, enabling complementary co-modeling of explicit and implicit knowledge for robust answer prediction. Experiments on Encyclopedic-VQA and InfoSeek demonstrate consistent performance gains across multiple MLLM backbones, and ablations verify that the selection mechanism effectively reduces noise and enhances knowledge utilization.

1 Introduction

Visual Question Answering (VQA) aims to answer questions by jointly modeling an image and its accompanying question (Antol et al., 2015; Huynh et al., 2025). While many VQA systems perform well when answers can be inferred directly from the image and question (Yang et al., 2025; Bai et al., 2025a), many real-world queries require additional background knowledge, such as commonsense or encyclopedic facts (Mensink et al., 2023; Chen et al., 2023; Deng et al., 2025). In such cases, the image-question pair alone is often insufficient. Knowledge-based VQA (KB-VQA) addresses this by incorporating external knowledge to support reasoning over visual content.

Existing KB-VQA methods can be broadly categorized into three paradigms: explicit, implicit, and hybrid approaches, as illustrated in Figure 1. Explicit methods align visual entities or regions with external knowledge sources (Ravi et al., 2023; Yan and Xie, 2024), while implicit methods rely on parametric knowledge embedded in large pretrained models for cross-modal reasoning (Yang et al., 2022; Yu et al., 2025). Hybrid approaches combine retrieved external knowledge with internal parametric knowledge to leverage their complementary strengths (Gui et al., 2022; Fang et al., 2025). Despite different implementations, these methods share the goal of enhancing VQA by incorporating external or latent knowledge.

However, a key challenge remains: how to effectively control which visual and retrieved knowledge should be exposed to the reasoning model under noisy inputs. In practice, both visual regions and retrieved passages may contain irrelevant or redundant information (Wang et al., 2024; Yan and Xie, 2024). More importantly, their usefulness is inherently cross-modally coupled: visual regions become meaningful only under textual knowledge guidance, while retrieved knowledge relevance depends on visual grounding. Existing evidence selection strategies mainly focus on individual modalities, such as retrieving relevant passages from external knowledge bases or identifying salient visual regions through attention-based attribution (Guo et al., 2023; Ma et al., 2024; Han et al., 2025). Such independent filtering fails to capture the interaction between visual and textual evidence, limiting effective knowledge utilization for downstream reasoning.

To address this limitation, we introduce a unified cross-modal selection paradigm that jointly regulates which visual regions and textual fragments

* Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 29796–29811
October 24-29, 2026 ©2026 Association for Computational Linguistics

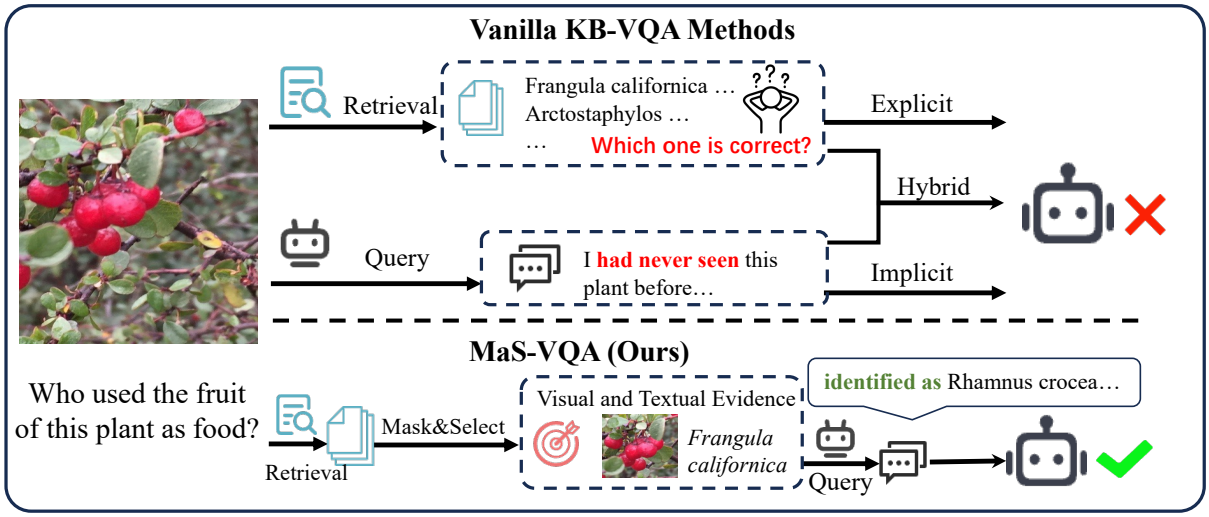

Figure 1: A comparison of vanilla KB-VQA and our proposed method. Compared to standard hybrid methods that separate explicit and implicit knowledge, MaS-VQA integrates their reasoning.

are provided to the reasoning model before answer generation. Unlike previous retrieval filtering or prompt compression approaches that operate on isolated modalities, our Mask-and-Select mechanism performs multimodal evidence selection within a unified relevance space, where visual masking and textual phrase selection serve as complementary operations for refining explicit knowledge. Furthermore, the selected explicit knowledge constrains implicit knowledge generation, transforming free-form knowledge elicitation into grounded reasoning within a more reliable semantic space.

Building on this insight, we propose **MaS-VQA**, a selection-based framework that tightly couples explicit knowledge filtering with implicit parametric reasoning, as illustrated in Figure 1. Given an image–question pair, MaS-VQA first retrieves top-k candidate passages via a multimodal retriever. It then performs explicit knowledge processing through a unified **Mask-and-Select** mechanism that jointly filters visual and textual knowledge: on the visual side, a knowledge-guided cross-attention module produces a knowledge-guided attention mask to suppress irrelevant regions; on the textual side, question-conditioned phrase selection preserves key knowledge fragments while masking noisy or weakly relevant content. Building upon the filtered multimodal knowledge, MaS-VQA further performs implicit knowledge processing conditioned on the refined explicit knowledge, guiding the model to activate and reason over relevant model-internal knowledge within a more constrained semantic space. In this way, explicit and implicit knowledge are co-modeled as complementary sources for answer prediction, yielding more robust KB-VQA under noisy retrieval and complex inputs.

We evaluate MaS-VQA on two challenging KB-VQA benchmarks, Encyclopedic-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023). Experimental results show that MaS-VQA consistently improves KB-VQA performance across different MLLM backbones, demonstrating strong robustness to noisy retrieval and complex inputs. Further analyses and ablations verify the effectiveness of each component in MaS-VQA, and qualitative visualizations illustrate that our selection mechanism yields more focused visual grounding and more reliable knowledge utilization. Our contributions can be summarized as follows:

- We propose MaS-VQA, a selection-based framework that tightly couples explicit knowledge filtering with implicit parametric reasoning for KB-VQA.
- We introduce a unified **Mask-and-Select** mechanism that performs fine-grained selection over both visual regions and retrieved knowledge, producing compact, high-signal explicit representations to mitigate noise accumulation.
- We conduct comprehensive experiments on Encyclopedic-VQA and InfoSeek, achieving consistent improvements and providing detailed ablations and qualitative analyses to validate the proposed design.

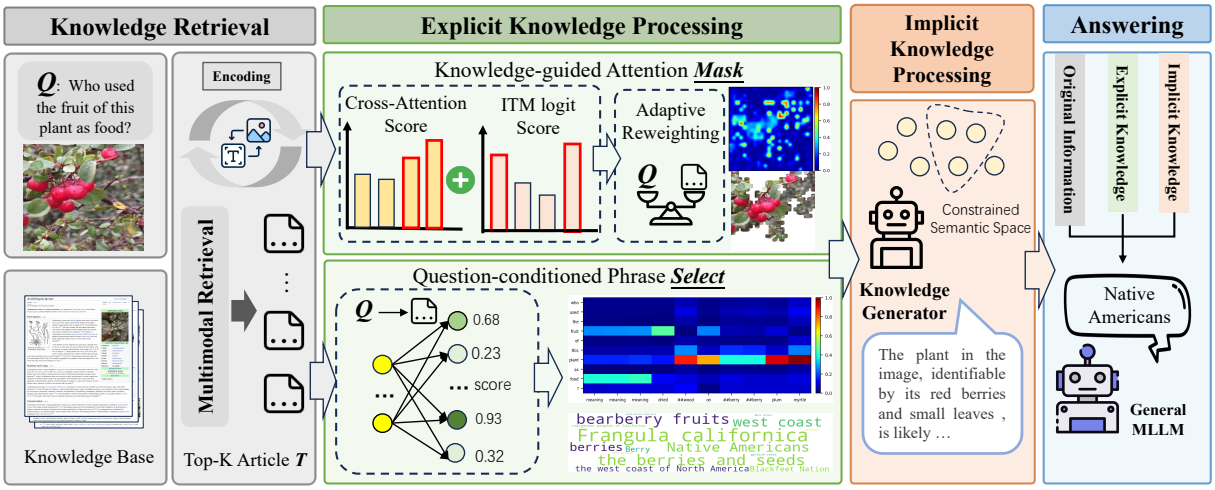

Figure 2: Overview of MaS-VQA. Given an image–question pair, MaS-VQA retrieves top-k passages from an external knowledge base and performs **Mask-and-Select** explicit knowledge processing, including a knowledge-guided attention mask for filtering irrelevant image regions and question-conditioned phrase selection for pruning noisy text. The filtered multimodal evidence is then used for implicit knowledge processing to elicit complementary model-internal knowledge, and both knowledge sources are co-modeled for final answer prediction.

2 Related Work

2.1 Explicit Knowledge Methods

Explicit knowledge methods retrieve facts from external knowledge bases and fuse them into answer prediction.

Early approaches construct KB queries from visual and textual cues, e.g., VLC-BERT (Ravi et al., 2023) concatenates detected object names with the question. Later work improves retrieval granularity and representations. KAT (Gui et al., 2022) aligns image patches with knowledge snippets, and RKVQA (Chen et al., 2025) jointly embeds global and local regions for retrieval. REVEAL (Hu et al., 2023b) further introduces a key–value memory to maintain updatable explicit knowledge.

These methods provide controllable and verifiable evidence but depend heavily on effective query formulation, retrieval granularity, and multimodal fusion.

2.2 Implicit Knowledge Methods

Implicit knowledge methods rely on knowledge stored in model parameters, typically elicited through prompting large language models.

Early work activates such knowledge via captioning, few-shot prompting, or pre-answers. PICa (Yang et al., 2022) prompts GPT-3 with image captions and examples, while ASB (Xenos et al., 2023) applies a similar strategy with LLaMA. Other methods enrich prompts: GeReA (Ma et al., 2024) uses Grad-CAM–guided captions, Img2LLM (Guo et al., 2023) adds related QA pairs, and PromptCap (Hu et al., 2023a) distills GPT-style question-conditioned captioning into a lightweight captioner. Prophet (Yu et al., 2025) and PCPA (Hu et al., 2024) introduce a lightweight VQA model to generate pre-answers that help unlock LLM reasoning.

More recent approaches improve knowledge selection and self-learning. ReflectiVA (Cocchi et al., 2025) adaptively chooses between internal knowledge and external retrieval, while HinD (Zhao et al., 2025) enhances MLLM self-learning for better knowledge generation and reasoning.

Implicit methods reduce dependence on KB coverage but remain prone to hallucination and outdated knowledge, motivating hybrid designs.

2.3 Hybrid Explicit-Implicit Knowledge Methods

Recent systems combine explicit retrieval with implicit knowledge from LLMs/MLLMs for knowledge-intensive reasoning. KAT (Gui et al., 2022) integrates patch-level visual retrieval with caption-based GPT-3 prompting to inject external evidence, while KRISP (Marino et al., 2021) jointly models explicit concept-graph reasoning and implicit answer priors. RKVQA (Chen et al., 2025) combines global and local retrieval signals with caption prompting to connect retrieved knowledge and QA. NoteMR (Fang et al., 2025) pairs an explicit retriever with direct MLLM QA, using the model itself as a complementary knowledge source.

Different from previous hybrid approaches that mainly combine retrieved knowledge with model-internal knowledge through direct fusion or prompting (Marino et al., 2021; Gui et al., 2022; Chen et al., 2025; Fang et al., 2025), MaS-VQA focuses on controlling the evidence presented to the reasoning model. Specifically, existing hybrid methods typically treat retrieved knowledge and visual information as complementary inputs, while MaS-VQA introduces an intermediate cross-modal selection stage to remove unreliable evidence before reasoning. By jointly selecting visual regions and textual fragments under a unified multimodal relevance space, MaS-VQA establishes a selection-driven paradigm for knowledge utilization rather than simple knowledge aggregation.

3 Method

In this section, we present the MaS-VQA framework. Sec. 3.1 defines the task and answer prediction objective with external knowledge. Sec. 3.2 describes explicit knowledge processing, including strategies for both image and text. Sec. 3.3 introduces implicit knowledge generation, where a frozen MLLM produces a concise grounded paragraph to complement the retrieved evidence. Fig. 2 shows the overall architecture.

3.1 Task Formulation

Given an image I and a question Q, KB-VQA aims to predict an answer $\hat{A} \in \mathcal{A}$ by leveraging an external knowledge source $\mathcal{K}$ in addition to visual evidence:

$$\hat{A} = \arg\max_{a \in \mathcal{A}} p(a \mid I, Q, \mathcal{K}). \tag{1}$$

As shown in Fig. 2, our framework constructs **explicit knowledge** from retrieval and **implicit knowledge** generated by a frozen multimodal large language model. We first retrieve a set of top-k candidate knowledge snippets as textual knowledge:

$$T = \{t_i\}_{i=1}^{k} = f_{\text{ret}}(I, Q; \mathcal{K}), \tag{2}$$

where each t_i is a retrieved passage (e.g., a Wikipedia segment).

To facilitate grounding of the retrieved knowledge to the image, we select a keyword set $\mathbf{k}$ from T and derive an attention map (mask) M that highlights question-relevant image regions under the guidance of retrieved text:

$$\mathbf{k} = f_{\text{key}}(T), \qquad M = f_{\text{mask}}(I, Q, T). \tag{3}$$

For notational convenience, we denote the explicit knowledge package as

$$E = \{T, \mathbf{k}, M\}. \tag{4}$$

Conditioned on the image, question, and explicit knowledge, we generate an *implicit knowledge* paragraph U (2–5 sentences) that captures concise grounded insights useful for answering the question:

$$U = f_{\text{imp}}(I, Q, E). \tag{5}$$

Finally, we query a frozen MLLM g_θ with all evidence and obtain the answer distribution:

$$p(a \mid I, Q, \mathcal{K}) = g_\theta(a \mid I, Q, E, U), \tag{6}$$

$$\hat{A} = \arg\max_{a \in \mathcal{A}} p(a \mid I, Q, \mathcal{K}). \tag{7}$$

Here, E corresponds to explicit knowledge processing (retrieval and grounding), while U represents implicit knowledge distilled under explicit guidance for robust final reasoning.

3.2 Explicit Knowledge Processing: Text and Image Signals

Motivation. A straightforward solution is to independently filter irrelevant image regions and textual knowledge fragments. However, such independent filtering ignores the inherent dependency between modalities: textual knowledge provides semantic guidance for visual grounding, while visual evidence determines the actual usefulness of retrieved knowledge. Therefore, MaS-VQA performs selection within a unified multimodal relevance space, where visual masking and phrase selection are two complementary outcomes of the same evidence selection process.

Given the retrieved knowledge passages $T = \{t_i\}_{i=1}^{k}$ and the question Q, we construct two explicit grounding signals: (i) an *image-side* attention mask M that localizes question-relevant visual regions under the guidance of the retrieved text, and (ii) a *text-side* keyword phrase set $\mathbf{k}$ that condenses the retrieved knowledge into a few high-salience spans. In practice, both signals are extracted from a pretrained image-text matching (ITM) encoder.

We tokenize the knowledge and question separately and concatenate them into a single sequence:

$$X = \texttt{[CLS]}\ T\ \texttt{[SEP]}\ Q\ \texttt{[SEP]}, \tag{8}$$

where we also keep the character-level offset mapping for knowledge tokens (enabled by using a fast

tokenizer). Let $\mathcal{I}_K = [k_s, k_e)$ and $\mathcal{I}_Q = [q_s, q_e)$ denote the token index ranges of knowledge and question in X, determined by the separator token positions. When the sequence exceeds the maximal text length, we truncate X and correspondingly clip the effective knowledge offsets.

Image-side: knowledge-guided attention mask generation. We compute an image patch mask that highlights regions supported by the joint evidence of the knowledge T and question Q. Let the visual encoder produce P patch embeddings ($P = g^2$ for a $g \times g$ patch grid), and let the text encoder attend to visual patches through cross-attention. At a chosen transformer block b, we extract the cross-attention weights $\mathbf{A}^{(b)} \in \mathbb{R}^{H \times L \times P}$ and their backpropagated sensitivity signals $\nabla \mathbf{A}^{(b)}$ with respect to the positive ITM logit. We construct token-to-patch relevance by head aggregation and non-negative filtering:

$$\mathbf{R}_{i,p} = \frac{1}{H} \sum_{h=1}^{H} \left(\mathbf{A}^{(b)}_{h,i,p} \cdot \mathrm{ReLU}(\nabla \mathbf{A}^{(b)}_{h,i,p}) \right), \quad (9)$$

where $i \in \{1, \ldots, L\}$ indexes text tokens and $p \in \{1, \ldots, P\}$ indexes visual patches. We reshape $\mathbf{R}$ into token-wise $g \times g$ maps. For numerical stability, each token map is min–max normalized to $[0, 1]$ independently.

Adaptive token reweighting. Different tokens in T and Q may contribute unequally to localization. We compute a token strength score by averaging its relevance over patches:

$$s_i = \frac{1}{P} \sum_{p=1}^{P} \mathbf{R}_{i,p}. \quad (10)$$

Let $\mathcal{I}_K$ and $\mathcal{I}_Q$ denote the token index sets corresponding to knowledge and question spans, respectively. We derive intra-group weights via a temperature-controlled softmax, applied separately within $\mathcal{I}_K$ and $\mathcal{I}_Q$:

$$\tilde{w}_i = \begin{cases} \mathrm{softmax}\left(\frac{s_i}{\tau}\right), & i \in \mathcal{I}_K, \\ \mathrm{softmax}\left(\frac{s_i}{\tau}\right), & i \in \mathcal{I}_Q, \\ 0, & \text{otherwise}. \end{cases} \quad (11)$$

To balance the overall contributions of knowledge and question tokens, we further compute a group-level adaptive factor:

$$\alpha = \mathrm{softmax}\left(\frac{[\bar{s}_K,\ \bar{s}_Q]}{\tau} \right), \quad (12)$$

$$\bar{s}_K = \frac{1}{|\mathcal{I}_K|} \sum_{i \in \mathcal{I}_K} s_i, \qquad \bar{s}_Q = \frac{1}{|\mathcal{I}_Q|} \sum_{i \in \mathcal{I}_Q} s_i. \quad (13)$$

We set $w_i = \alpha_K \tilde{w}_i$ for $i \in \mathcal{I}_K$ and $w_i = \alpha_Q \tilde{w}_i$ for $i \in \mathcal{I}_Q$, and normalize the weights such that $\sum_{i=1}^{L} w_i = 1$.

Token-wise thresholding and patch mask composition. We first form weighted token-to-patch scores

$$\hat{\mathbf{R}}_{i,p} = w_i \, \mathbf{R}_{i,p}. \quad (14)$$

Instead of thresholding a single aggregated patch score, we perform *token-wise* percentile thresholding. For each token i, we compute a token-specific threshold as the ρ-percentile of $\{\hat{\mathbf{R}}_{i,p}\}_{p=1}^{P}$ and obtain a binary token mask over patches:

$$\mathbf{B}_{i,p} = \mathbb{I}\left[\hat{\mathbf{R}}_{i,p} > \mathrm{Quantile}\left(\hat{\mathbf{R}}_{i,:}, \rho \right) \right]. \quad (15)$$

Finally, we compose the overall patch mask by combining token masks across all tokens using a logical OR operator:

$$M_p = \bigvee_{i=1}^{L} \mathbf{B}_{i,p}, \qquad p = 1, \ldots, P. \quad (16)$$

M is reshaped to a $g \times g$ patch mask, upsampled to the image resolution, and used for visualization by suppressing non-salient regions with a white background.

Text-side: question-conditioned phrase selection from retrieved knowledge. We additionally select a compact set of phrases from T to serve as high-precision textual hints. Using the same tokenized sequence X, we compute the self-attention probabilities at a selected text self-attention layer:

$$\mathbf{A} \in \mathbb{R}^{H \times L \times L}. \quad (17)$$

To quantify how strongly each knowledge token supports the question, we construct a question-to-knowledge interaction matrix by modulating $\mathbf{A}$ with its sensitivity signal and aggregating heads:

$$\mathbf{S} = \frac{1}{H} \sum_{h=1}^{H} \left(\mathbf{A}_h \odot \mathrm{ReLU}(\nabla \mathbf{A}_h) \right) \in \mathbb{R}^{L \times L}, \quad (18)$$

Table 1: Performance on the Encyclopedic-VQA test set and the InfoSeek validation set. Best results are highlighted in bold. * denotes methods that require additional training during the reasoning stage. † denotes results that are not directly comparable due to differences in the underlying knowledge bases. ◊ denotes our reproductions using different MLLM backbones.

Method	Model	Retrieval Mode	E-VQA		InfoSeek		
			Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot MLLMs							
BLIP-2 (Li et al., 2023)	Flan-T5$_{XL}$	-	12.6	12.4	12.7	12.3	12.5
InternVL3-8B (Zhu et al., 2025)	InternVL3-8B	-	15.7	15.9	13.5	11.7	12.5
Qwen2.5-VL-7B (Bai et al., 2025b)	Qwen2.5-VL-7B	-	16.3	16.3	16.8	15.8	16.3
Qwen3-VL-8B (Bai et al., 2025a)	Qwen3-VL-8B	-	19.5	19.5	19.7	17.4	18.5
GPT-4V (Achiam et al., 2023)	GPT-4V	-	26.9	28.1	15.0	14.3	14.6
Retrieval-Augmented Models							
DPR$_{V+T}$† (Lerner et al., 2024)	Multi-passage BERT	CLIP ViT-B/32	29.1	-	-	-	12.4
EchoSight† (Yan and Xie, 2024)	Mistral-7B/LLaMA-3-8B	EVA-CLIP-8B	19.4	-	-	-	27.7
Wiki-LLaVA (Caffagni et al., 2024)	Vicuna-7B	CLIP ViT-L/14+Contriever	17.7	20.3	30.1	27.8	28.9
Prophet++* ◊† (Yu et al., 2025)	Qwen3-VL-8B	EVA-CLIP-8B	22.5	20.0	13.2	11.6	12.3
NoteMR◊† (Fang et al., 2025)	Qwen3-VL-8B	EVA-CLIP-8B	25.6	23.6	28.7	29.8	29.2
ReflectiVA* (Cocchi et al., 2025)	LLaMA-3.1-8B	EVA-CLIP-8B	28.0	29.2	40.4	39.8	40.1
MMKB-RAG (Ling et al., 2025)	Qwen2-7B	EVA-CLIP-8B	39.7	35.9	36.4	36.3	36.4
VLM-PRF* (Hong et al., 2025)	InternVL3-8B	EVA-CLIP-8B	40.1	39.2	43.5	42.1	42.5
MaS-VQA (Ours)†	InternVL3-8B	EVA-CLIP-8B	40.8	40.1	42.9	42.6	42.8
MaS-VQA (Ours)†	Qwen3-VL-8B	EVA-CLIP-8B	**42.2**	**41.3**	**43.7**	**43.9**	**43.8**

and compute a score for each knowledge token $j \in \mathcal{I}_K$ by averaging over question tokens:

$$s_j = \frac{1}{|\mathcal{I}_Q|} \sum_{i \in \mathcal{I}_Q} \mathbf{S}_{i,j}. \tag{19}$$

We select the top-m knowledge tokens with the largest s_j, and map these token indices back to character spans in the original knowledge text using the offset mapping. Since selected tokens may be fragmented, we merge overlapping or adjacent spans within a small gap to form readable phrases. The final keyword set $\mathbf{k}$ is the collection of merged spans selected from T, and is used as the text-side explicit signal in the downstream prompting stage.

3.3 Implicit Knowledge Processing

Large multimodal language models (MLLMs) encode substantial world knowledge in their parameters. However, directly answering KB-VQA questions with a frozen MLLM can be unreliable when evidence is implicit or requires grounding. To better leverage internal knowledge while remaining faithful to retrieved evidence, we distill an *implicit knowledge* paragraph U conditioned on the retrieved explicit knowledge and its visual grounding.

Inputs and format. We generate U by conditioning a frozen MLLM on the image I, question Q, and explicit knowledge package $E = \{T, \mathbf{k}, M\}$. The prompt includes retrieved passages T, selected keywords $\mathbf{k}$, and an attention-guided image view induced by M. The model outputs a short paragraph (2–5 sentences) summarizing grounded clues useful for answering Q.

Role and usage. The generated U compresses retrieved passages into a question-focused summary and integrates grounded visual cues with textual knowledge. It is appended to (I, Q, E) in the final prompt to guide answer prediction.

4 Experiments

4.1 Experimental Setup

Dataset. E-VQA contains 2.21M VQA pairs linked to images from iNaturalist 2021 and GLDv2, including 1M train, 13.6k val, and 5.8k test samples. Questions involve single-hop or two-hop reasoning, and the knowledge base consists of 2M Wikipedia articles covering 16.7k fine-grained entities. We evaluate on the official test set.

InfoSeek is a large-scale IR-based VQA dataset with 1.3M image–question pairs built on about 11k visual entities from OVEN. It includes 934k train, 73k val, and 348k test samples. Following the official protocol, we report results on the full validation set with no overlap in questions or entities with the training data.

Baselines. We categorize the baselines into two groups: zero-shot MLLMs and retrieval-augmented models, including both current state-of-the-art methods and models with comparable parameter scales for fair comparison.

Table 2: Ablation study on Encyclopedic-VQA (test) assessing the impact of attention masking, phrase selection, and implicit knowledge. Best results are highlighted in bold.

Explicit Knowledge		Implicit Knowledge	E-VQA	
Attention Mask	Phrase Selection		Single-Hop	All
✓			39.8	36.8
	✓		38.5	36.6
✓	✓		40.9	38.4
		✓	40	35.9
✓	✓	✓	**42.2**	**41.3**

Table 3: Effect of retrieval breadth on Encyclopedic-VQA. MaS-VQA performance as the number of retrieved passages k varies. The retriever and MLLM backbone are fixed; only k is changed.

k	E-VQA	
	Single-Hop	All
1	35.1	31.6
3	39.5	36.7
5	**42.2**	**41.3**
7	41.8	41.1

Evaluation Metrics. Following the official protocols, we use dataset-specific metrics. For E-VQA, we adopt BERT Matching (BEM) (Bulian et al., 2022), which evaluates predicted answers conditioned on the question and ground-truth answer and is robust to paraphrases. For InfoSeek, STRING and TIME questions are evaluated with VQA accuracy (Goyal et al., 2017) (exact/normalized match), while NUMERICAL questions use Relaxed Accuracy (Methani et al., 2020) with type-specific tolerances for small numeric deviations.

Implementation Details. We adopt InternVL3-8B (Zhu et al., 2025) and Qwen3-VL-8B (Bai et al., 2025a) as MLLM backbones, using the same model for both knowledge generation and final answering. We employ EchoSight (Yan and Xie, 2024) as the multimodal retriever and retrieve the top-k passages with $k = 5$. For image–text encoding, we use a pretrained BLIP model (Li et al., 2022). Experiments are conducted on NVIDIA H20 GPUs with FlashAttention-2 (Dao, 2023). Unless otherwise specified, inference uses a temperature of 0.7, a maximum generation length of 512 tokens, and a batch size of 16. More details are provided in Appendix C.

4.2 Main Results

Table 1 summarizes the main results on Encyclopedic-VQA and InfoSeek, with the best numbers highlighted in bold. Overall, **MaS-VQA** delivers the strongest performance across both benchmarks and all reported splits.

Results across benchmarks. On Encyclopedic-VQA, MaS-VQA significantly outperforms zero-shot MLLMs that rely solely on parametric knowledge. With Qwen3-VL-8B, performance improves from 19.5 (zero-shot) to **42.2** (Single-Hop) and **41.3** (All), highlighting the importance of incorporating external knowledge and effectively selecting relevant evidence from noisy retrieval. MaS-VQA also achieves the best overall accuracy compared with strong retrieval-augmented baselines such as MMKB-RAG and VLM-PRF.

On InfoSeek, MaS-VQA achieves the best results on Unseen-Q, Unseen-E, and All, reaching **43.7**, **43.9**, and **43.8**, respectively. These results indicate that selecting compact, question-relevant evidence helps the model generalize better to novel questions and entities.

Effect of backbone models. MaS-VQA consistently improves performance with both InternVL3-8B and Qwen3-VL-8B, suggesting that the gains stem from our MaS-VQA framework rather than a specific backbone.

Discussion on methods requiring additional training. In Table 1, * denotes methods that require additional training during the reasoning stage. While such approaches typically increase computation and may reduce fairness when compared to purely inference-time methods, MaS-VQA improves reasoning *without* additional training at inference time by selecting and compressing retrieved knowledge and question-relevant visual cues to better elicit implicit (parametric) knowledge from a frozen backbone.

4.3 Ablation Study

Table 2 studies the contribution of each component in MaS-VQA (Qwen3-VL-8B) on Encyclopedic-VQA (test). Under the explicit-knowledge setting, enabling the question-guided attention mask alone yields 39.8/36.8 (Single-Hop/All), while using phrase selection alone achieves 38.5/36.6. Combining the two explicit grounding modules further improves performance to 40.9/38.4, indicating that visual-region localization and question-conditioned phrase selection provide complementary signals for constructing higher-quality explicit evidence. When only implicit knowledge

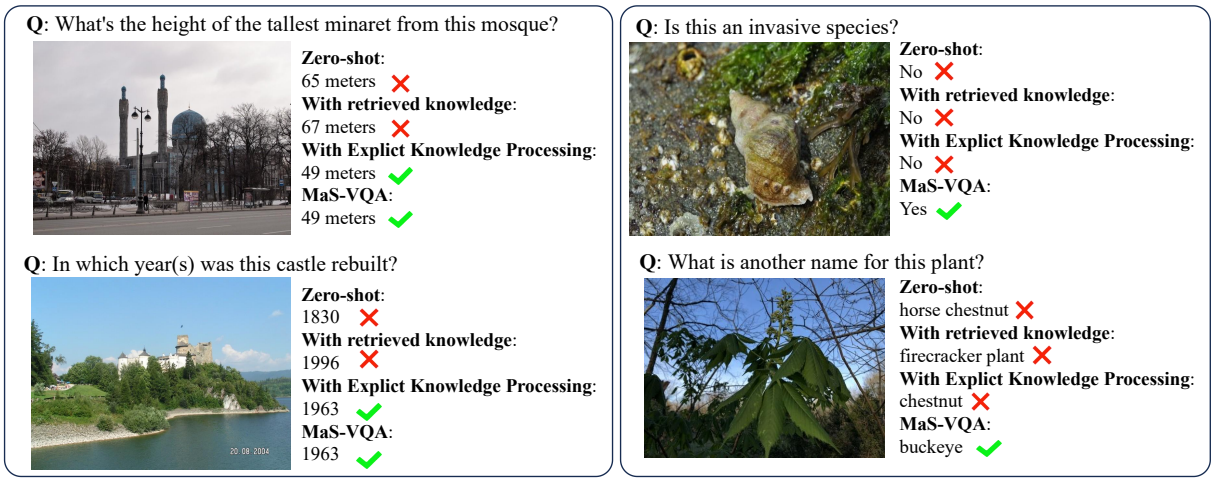

Figure 3: Qualitative case studies. Left: explicit external knowledge helps bridge missing factual gaps and corrects errors made without retrieval. Right: implicit knowledge complements retrieved evidence when the final decision requires commonsense/domain priors beyond the retrieved text.

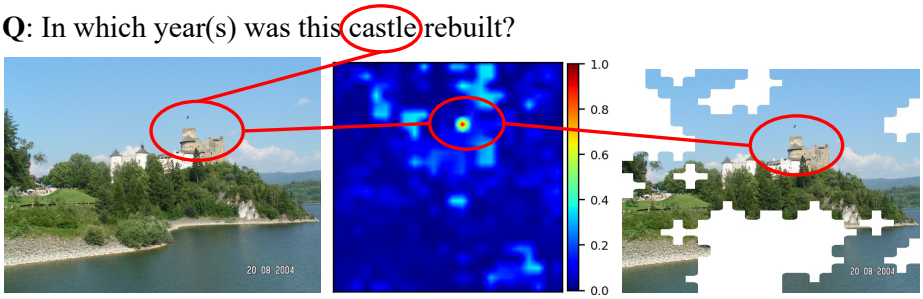

Figure 4: Specific Case Analysis of the Effectiveness of Explicit Knowledge Processing. Explicit knowledge processing allows the model to focus more on information relevant to the answer.

is used (i.e., without explicit grounding), the performance drops to 40.0/35.9, suggesting that parametric knowledge alone is less reliable and can be distracted by unfiltered or redundant inputs. Finally, integrating implicit knowledge with both explicit modules achieves the best results, reaching **42.2/41.3**, which confirms that MaS-VQA benefits from jointly modeling explicit retrieval-grounding and implicit knowledge distillation.

We further analyze the effect of retrieval breadth in Table 3. Increasing the number of retrieved passages from $k = 1$ to $k = 5$ consistently improves performance (35.1/31.6 $\rightarrow$ 42.2/41.3), showing the importance of having sufficient evidence coverage. However, further increasing k to 7 leads to a slight drop (41.8/41.1), indicating that overly broad retrieval may introduce additional noise that offsets the benefit of more passages. Based on this trade-off, we set $k = 5$ in all experiments.

4.4 Case Study

Effectiveness of Explicit Knowledge Processing. As shown in Figure 3 (left), directly appending retrieved passages often introduces redundant or weakly relevant evidence, which may distract the model with spurious associations and ultimately mislead answer selection. With our *Explicit Knowledge Processing*, MaS-VQA performs joint modality filtering: a knowledge-guided attention mask suppresses irrelevant image regions and encourages the model to attend to the question-critical visual cues, while question-conditioned phrase selection prunes noisy or off-topic textual fragments and keeps only the most informative spans. Consequently, the final context becomes more compact, better aligned across vision and language, and easier for the MLLM to reason over. Figure 4 further visualizes this effect—the model concentrates on the correct visual region and retains the key phrase from retrieved knowledge—providing an intuitive

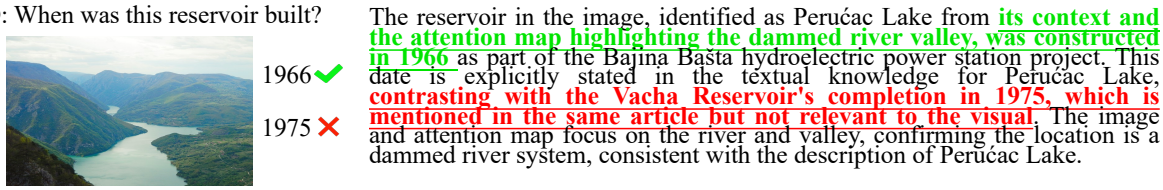

Figure 5: Specific Case Analysis of the Effectiveness of Implicit Knowledge Processing. Even with filtered evidence, some questions need commonsense or domain priors, which implicit knowledge processing elicits for correct reasoning.

explanation for the consistent gains over raw retrieval prompting.

Effectiveness of Implicit Knowledge Processing. Figure 3 (right) presents challenging cases where explicit filtering alone remains insufficient, e.g., when answering requires additional commonsense/domain priors or implicit relational reasoning beyond the retrieved snippets. In these cases, *Implicit Knowledge Processing* further elicits and organizes parametric knowledge conditioned on the filtered evidence, enabling MaS-VQA to complete the missing reasoning step and correct the prediction. A representative example is provided in Figure 5, where implicit knowledge complements the cleaned explicit evidence to reach the correct answer.

5 Conclusion

We presented **MaS-VQA**, a selection-based framework for KB-VQA that addresses the noise and distraction introduced by raw retrieval prompting. By jointly filtering question-relevant visual regions and retrieved knowledge via a **Mask-and-Select** mechanism, MaS-VQA constructs low-noise explicit evidence aligned with the question. Building on this filtered context, MaS-VQA further elicits complementary *implicit* knowledge conditioned on the selected evidence, enabling more reliable reasoning. Experiments on E-VQA and InfoSeek show consistent improvements over strong baselines across different MLLM backbones, while ablations and case studies verify our method's effectiveness.

Limitations

This work has several limitations. First, the framework relies on the quality and coverage of the external knowledge base. If relevant information is missing or poorly retrieved, the model may still produce incorrect answers. Second, the design of the Mask-and-Select mechanism introduces several hyperparameters (e.g., thresholds and the number of selected phrases), and the performance may vary under different settings. Finally, our experiments are conducted in an English-only setting, as both the datasets and the knowledge sources are primarily English-based, and its applicability to other languages remains to be studied.

Ethics Statement

This work focuses on improving knowledge-based visual question answering by enhancing how external knowledge and model-internal knowledge are selected and integrated. The proposed framework operates on publicly available datasets and knowledge sources and does not involve the collection of new personal or sensitive data.

However, KB-VQA systems may inherit biases or inaccuracies from the underlying knowledge bases and pretrained models. As a result, the system may occasionally produce incorrect or misleading answers. Therefore, such systems should not be used as the sole source of information in high-stakes domains, and further research on bias analysis and factual reliability is encouraged.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No: 62372408), Ningbo Natural Science Foundation (Grant No: 2025J001).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 260 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 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. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1818–1826.

Weipeng Chen, Xu Huang, Zifeng Liu, Jin Liu, and Lan Yo. 2025. Rk-vqa: Rational knowledge-aware fusion-in-decoder for knowledge-based visual question answering. *Information Fusion*, 118:102969.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9199–9209. IEEE.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

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*.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. 2025. Notes-guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19597–19607. IEEE.

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.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 956–968.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877.

Seunghoon Han, Mingyu Choi, Hyewon Lee, SoYoung Park, Jong-Ryul Lee, Sungsu Lim, and Tae-Ho Kim. 2025. Diverse knowledge selection for enhanced zero-shot visual question answering. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2161–2169.

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. In *Advances in Neural Information Processing Systems*.

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

Zhongjian Hu, Peng Yang, Yuanshuang Jiang, and Zijian Bai. 2024. Prompting large language model with context and pre-answer for knowledge-based vqa. *Pattern Recognition*, 151:110399.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023b. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil Aryal, Imran Razzak, and Hakim Hacid. 2025. Visual question answering: from early developments to recent advances–a survey. *arXiv preprint arXiv:2501.03939*.

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

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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

Ziyu Ma, Shutao Li, Bin Sun, Jianfei Cai, Zuxiang Long, and Fuyan Ma. 2024. Gerea: Question-aware prompt captions for knowledge-based visual question answering. *arXiv preprint arXiv:2402.02503*.

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.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. Vlc-bert: Visual question answering with contextualized commonsense knowledge. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1155–1165.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011.

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

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

Shuo Yang, Siwen Luo, and Soyeon Caren Han. 2025. Multimodal commonsense knowledge distillation for visual question answering (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 29545–29547.

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.

Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. 2025. Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6797–6808.

Yu Zhao, Ying Zhang, Xuhui Sui, Baohang Zhou, Li Shen, and Dacheng Tao. 2025. Hindsight distillation reasoning with knowledge encouragement preference for knowledge-based visual question answering. *arXiv preprint arXiv:2511.11132v1*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Test Dataset Statistics

Table 4 reports the dataset statistics for our evaluation splits.

Templated. Templated questions are manually created by domain experts based on predefined super-categories, resulting in controlled and standardized question patterns. **Automatic.** Automatic questions are generated by feeding Wikipedia articles into a question generation model, producing more diverse question formulations. **Multi-answer.** For multi-answer questions, the authors first use automatically generated questions to produce an initial candidate answer list, and then ask human annotators to complete and verify all possible correct answers. **Two-Hop.** Two-hop questions are constructed using a *bridge entity*: the answer to the first single-hop question is used as the subject of a second single-hop question, forming a two-step reasoning chain.

Split	Question Type		#Questions	Total
E-VQA Test	Single-Hop	Templated	1000	5750
		Automatic	2750	
		Multi-answer	1000	
	Two-Hop	–	1000	
InfoSeek Val	–	–	71335	71335

Table 4: Dataset statistics of the Encyclopedic-VQA test set and the InfoSeek validation set.

B Prompts

Implicit Knowledge Synthesis. Given the explicit evidence (textual knowledge and extracted keywords) and the attention-guided visual signals (white regions are masked), the model produces a concise paragraph (2–5 sentences) that summarizes grounded insights useful for answering the question later. We explicitly instruct the model **not** to output the final answer in this stage.

> Synthesize implicit knowledge by integrating the four inputs below. Do not answer the question. Produce one concise paragraph (2–5 sentences) capturing the key grounded insights that would be useful for answering the question later.
> 1. Evidence: Each Keyword paragraph is extracted from the Knowledge paragraph above it.
> Textual Knowledge: Textual information relevant to answering the question. You need to identify information that helps you answer the question.
> Keywords: Keywords selected based on the textual knowledge to assist your answer.
> 2. Original Image: An image relevant to the question.
> 3. Attention Map: Visualization of key areas based on the original image. White areas are masked; you only need to focus on the non-white areas.
>
> Evidence: {*evidence*}
> Original Image: <image>
> Attention Map: <image>
> Question: {*question*}
> Implicit Knowledge:

Final Answer Prediction. The model answers the question using (i) the selected explicit evidence, (ii) the synthesized implicit knowledge, and (iii) the visual inputs. The attention map is treated as supplementary grounding information to reduce distractions from irrelevant regions. The output is constrained to a single word or short phrase to match the evaluation format.

> Your task is to perform visual question answering based on the given text and image. I will provide you with the following information:
> 1. Evidence: Each Keyword paragraph is extracted from the Knowledge paragraph above it.
> Textual Knowledge: Textual information relevant to answering the question. You need to identify information that helps you answer the question.
> Keywords: Keywords selected based on the textual knowledge to assist your answer.
> 2. Original Image: An image relevant to the question.
> 3. Attention Map: Visualization of key areas based on the original image. White areas are masked; you only need to focus on the non-white areas.
> Please use the evidence and original image as global information, and the attention map as supplementary information to answer the question.

Evidence: {*evidence*}
Implicit Knowledge: {*imknowledge*}
Original Image: <image>
Attention Map: <image>
Question: {*question*}
Answer the question using a single word or phrase.
Answer:

C Experimental Setup

Implementation Details. We adopt InternVL3-8B (Zhu et al., 2025) and Qwen3VL-8B (Bai et al., 2025a) as the frozen MLLM backbones. For consistency, the same backbone is used for both implicit knowledge generation (f_{imp}) and final answer inference (g_θ) throughout the pipeline, and we do not update MLLM parameters in any stage.

Retrieval and explicit knowledge construction. For explicit textual evidence retrieval (f_{ret}), we use EchoSight (Yan and Xie, 2024) as the multimodal retriever to query the external knowledge base $\mathcal{K}$, and keep the top-k passages with $k = 5$, forming $T = \{t_i\}_{i=1}^{k}$. Retrieved passages are concatenated with the question into a single sequence $X =$ `[CLS]` T `[SEP]` Q `[SEP]`. When the input exceeds the maximal text length of the encoder (`max_txt_len` $= 512$), we truncate X accordingly. We enable fast tokenization and store character-level offset mappings to support phrase span recovery.

Grounding encoder and attention-based signals. To derive explicit grounding signals, we employ a pretrained BLIP image–text matching (ITM) encoder (Li et al., 2022) (`blip_image_text_matching_large`) to implement f_{mask} and f_{key}. For the image-side attention mask M, we extract cross-attention weights from transformer block $b = 7$ and compute gradient-weighted relevance maps (attention $\times$ ReLU(gradient)) with respect to the positive ITM logit, followed by head averaging and per-token min–max normalization. We adopt adaptive token reweighting with softmax temperature $\tau = 1.0$, including intra-group token normalization within the knowledge span and the question span, as well as group-level balancing between knowledge and question tokens. Patch-level binarization is performed via token-wise percentile thresholding with percentile $\rho = 90$. Token masks are combined using a logical OR operation across tokens (`combine_strategy=any`) to obtain the final $g \times g$ patch mask, which is then upsampled to the image resolution and visualized by suppressing non-salient regions with a white background.

For the text-side keyword set $\mathbf{k}$, we compute gradient-modulated text self-attention interactions (question-to-knowledge) at the same transformer block ($b = 7$) with `output_attentions=True`. We score each knowledge token by averaging interaction scores over all question tokens and select the top-m tokens with $m = 30$. Selected token indices are mapped back to character spans in the retrieved passages using offset mappings and merged into readable phrases with a span merge gap of 3 character. We keep the first 10 merged phrases as the final keyword set $\mathbf{k}$.

Inference configuration and hardware. All experiments are conducted on NVIDIA H20 GPUs with FlashAttention-2 (Dao, 2023) enabled. Unless otherwise specified, we use a temperature of 0.7, a maximum generation length of 512 tokens, and a batch size of 16 for decoding acceleration.

D Hyperparameter Sensitivity Analysis

To evaluate the robustness of the proposed selection mechanism, we analyze two key hyperparameters: the percentile threshold ρ for visual mask generation and the number of selected tokens m for textual phrase selection.

D.1 Sensitivity of Visual Selection Threshold

Table 5 reports the performance under different percentile thresholds. The results show that MaS-VQA remains stable across a broad range of ρ. A smaller threshold retains excessive visual regions and introduces noise, while an overly large threshold removes useful evidence.

Table 5: Sensitivity analysis of percentile threshold ρ on E-VQA.

ρ	70	80	90	95
Accuracy	39.8	40.5	**41.3**	38.1

D.2 Sensitivity of Selected Tokens

Table 6 studies different numbers of selected tokens m. Too few tokens may remove necessary evidence, whereas retaining too many tokens introduces irrelevant information.

Table 6: Sensitivity analysis of selected token number m on E-VQA.

m	10	30	50	70
Accuracy	37.9	41.3	**42.1**	40.8

E Additional Case Study

To provide a deeper understanding of MaS-VQA, we include additional and more detailed qualitative case studies in this section. These examples further illustrate the behavior of the model. As shown in Figs. 6 and 7, MaS-VQA effectively grounds retrieved knowledge to question-relevant image regions and leverages the generated implicit knowledge to support robust multi-step reasoning.

Question:
What is the maximum length of this bird in centimeters?

Explicit Knowledge

Attention Mask

Phrase Select

white underparts
Little bunting
Savannah sparrow
birds
13 to 14 centimeters in length
about 16 cm long
Brown
11 to 17 cm (4.3 to 6.7 in)
Cape Sable seaside sparrow
wings are brown

Ground Truth:
17

Zero-shot:
15
With retrieved knowledge:
14
With Explict Knowledge Processing:
17
MaS-VQA:
17

Implicit Knowledge

The bird in the image, a small bunting with streaked brown plumage and a whitish underbelly, is most likely a Savannah sparrow or a similar small sparrow species, as it lacks the chestnut face or black crown stripes of a Little bunting. Given the attention map highlights the bird's head, chest, and back, and its overall size, it aligns with the Savannah sparrow's range of 11–17 cm.

Question:
What is another name for this plant?

Explicit Knowledge

Attention Mask

Phrase Select

Aesculus pavia
scarlet buckeye
flowers
Carya myristiciformis
synoecious tree
chestnut
shagbark hickory
mockernut hickory
willow oak
hippocastanum

Ground Truth:
buckeye

Zero-shot:
horse chestnut
With retrieved knowledge:
firecracker plant
With Explict Knowledge Processing:
chestnut
MaS-VQA:
buckeye

Implicit Knowledge

The plant shown, identifiable by its large, compound leaves and upright flower cluster, is a species of horse chestnut (Aesculus), likely Aesculus hippocastanum, which is commonly known by several names including buckeye and conker tree, though "horse chestnut" is its most widely recognized common name. The attention map highlights the leaves and flowers, which are characteristic of this genus, and the textual knowledge confirms that "buckeye" is a common name for New World Aesculus species, while "horse chestnut" is the standard name for A. hippocastanum, which matches the visual evidence.

Figure 6: Case studies of MaS-VQA.

Question:
Is this plant abundant or scarce where there is chalk bedrock?

Explicit Knowledge

Attention Mask

Phrase Select

south-central United States
Euphorbia tuberosa
dry summer
grey-green
over the Winter
Digitaria sanguinalis
grain
fertility and drought
prominent midrib
sometimes

Ground Truth:
abundant

Zero-shot:
scarce
With retrieved knowledge:
scarce
With Explict Knowledge Processing:
scarce
MaS-VQA:
abundant

Implicit Knowledge

The plant in the image, likely a member of the aster family and flowering in late summer/fall, is native to prairie habitats with calcareous (chalky) soil, as seen in the attention map focusing on the plant amidst grassy terrain. It is an annual species that produces pink to white flowers, matching the visual cues. While Palafoxia callosa is explicitly noted as thriving in rocky, calcareous prairie soils, the image's context and the plant's morphology suggest it is adapted to such environments.

Question:
Along with greece, to what ancient civilization can the culture of the continent where this plant is found be traced?

Explicit Knowledge

Attention Mask

Phrase Select

dune-slacks
moorland
Anacamptis coriophora
East to Iran
grassland sites
temperate climates
Neotinea maculata
United Kingdom
Hook Moor SSSI
Neottia ovata

Ground Truth:
Rome

Zero-shot:
egypt
With retrieved knowledge:
Roman
With Explict Knowledge Processing:
Minoan
MaS-VQA:
Rome

Implicit Knowledge

The plant, identified as Neottia ovata (eggleaf twayblade) from its morphology and habitat context, is native to Europe, where it grows in varied habitats including meadows. This aligns with the cultural and geographic scope of ancient civilizations such as Greece and Rome, which both flourished across the continent where this orchid is found.

Figure 7: Case studies of MaS-VQA.