EchoSight: Advancing Visual-Language Models with Wiki Knowledge

Yibin Yan **Weidi Xie**

School of Artificial Intelligence, Shanghai Jiao Tong University
https://go2heart.github.io/echosight

Abstract

Knowledge-based Visual Question Answering (KVQA) tasks require answering questions about images using extensive background knowledge. Despite significant advancements, the large generative visual-language models often struggle with these tasks due to the limited integration of external knowledge. In this paper, we introduce **EchoSight**, a novel multimodal Retrieval-Augmented Generation (RAG) framework that enables to answer visual questions requiring fine-grained encyclopedic knowledge. To strive for high-performing retrieval, EchoSight first searches wiki articles by using visual-only information, subsequently, these candidate articles are further reranked according to their relevance to the combined text-image query. This approach significantly improves the integration of multimodal knowledge, leading to enhanced retrieval outcomes and more accurate VQA responses. Our experimental results on the Encyclopedic VQA and InfoSeek datasets demonstrate that EchoSight establishes new state-of-the-art results in knowledge-based VQA, achieving an accuracy of 41.8% on Encyclopedic VQA and 31.3% on InfoSeek.

1 Introduction

Visual Question Answering (VQA) addresses the challenge of enabling machines to understand and respond to questions about visual content, typically images or videos. Broadly, this task can be divided into two categories: standard VQA (Antol et al., 2015; Goyal et al., 2017) with questions that can be answered directly from the visual content, for example, counting objects, identifying colors, or recognizing simple actions, which rely solely on commonsense and information present in the image; and knowledge-based VQA (Marino et al., 2019; Schwenk et al., 2022; Mensink et al., 2023; Chen et al., 2023) requiring additional context or external knowledge, such as historical facts, detailed object properties, or specific situational contexts not evident in the visual content.

Addressing these two types of questions presents different challenges for VQA systems. Questions that draw answers directly from visual content demand robust image understanding capabilities, encompassing tasks such as object detection, scene recognition, and spatial reasoning. Conversely, questions requiring external knowledge call for additional mechanisms to access and integrate information from external sources. In this paper, we focus on the latter type of visual question answering, by building a retrieval-augmented multimodal system, that enables searching an external knowledge base for more nuanced understanding and accurate responses.

Despite the recent accomplishments in developing Visual-language Models (VLMs) (Achiam et al., 2023; Gemini Team et al., 2023; Abdin et al., 2024; Liu et al., 2024), knowledge-based VQA remains challenging. This complexity primarily stems from two aspects. (i) Existing VLMs struggle to adequately encode all essential knowledge, due to its limited model capacity, and infrequent inclusion of encyclopedic, long-tail information in their training data (Mensink et al., 2023). (ii) The visual component of the questions often provides limited help in addressing the queries, as establishing a meaningful connection between entity knowledge and visual attributes can be difficult. For example, an image of a church alone barely reveal information about its construction date.

In this paper, we introduce **EchoSight**, a novel retrieval-augmented vision-language system designed for knowledge-based visual question answering. EchoSight employs a dual-stage search mechanism that integrates a retrieval-and-reranking process with the Retrieval Augmented Generation (RAG) paradigm. Initially, the system performs a visual-only retrieval from an external knowledge base, to effectively narrow the knowledge search

Findings of the Association for Computational Linguistics: EACL 2024, pages 1538–1551

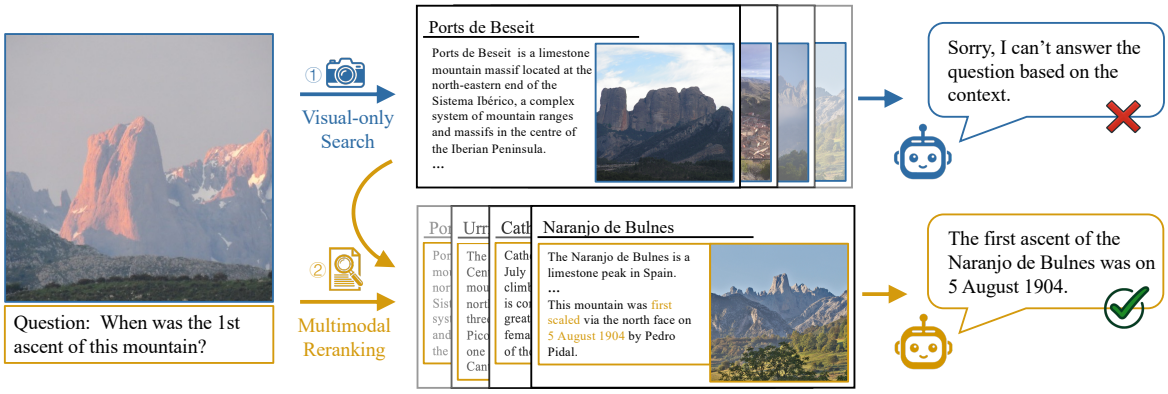

Figure 1: For visual questions such as *"When was the 1st ascent of this mountain?"*, visual-only search methods consider image similarity only, ignoring the textual details of the accompanying article. By incorporating multimodal reranking, the correct entry, accounting for both visual and textual information, can be accurately identified.

space, only focusing on candidates that are closely align with the visual context of the reference image. In the subsequent multimodal reranking stage, the system refines the candidates ranking by incorporating both the reference image and the textual query. This approach guarantees that the selected results are pertinent not only visually, but also contextually to the multimodal query. After acquiring the most relevant information through this coarse-to-fine grained search, our model generates the precise answer to the posed question.

Overall, we present three contributions: *First*, we propose a multimodal retrieval-augmented generation framework, termed as **EchoSight**, that enables to answer visual questions that require fine-grained encyclopedic knowledge; *Second*, we adopt a retrieval-and-reranking scheme to improve retrieval performance, specifically, it first searches images with visual-only information, and then conduct a fine-grained multimodal reranking on the candidates; *Third*, we conduct thorough experiments on both Encyclopedic VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023) benchmarks, **EchoSight** demonstrates state-of-the-art performance on both benchmarks, significantly outperforming existing VLMs or other retrieval-augmented architectures.

2 Method

This section starts with the problem formulation of retrieval-augmented VQA (Sec. 2.1), followed by detailing the retrieval-and-reranking module in EchoSight (Sec. 2.2), and finally the answer generation module (Sec. 2.3).

2.1 Problem Formulation

Given a reference image, and question of free-form texts, our goal is to construct a visual question answering system, that can benefit from the access of an external knowledge base. In our case, this is a million-scale dataset of entity articles and their corresponding images from Wikipedia webpage, *i.e.*, $\mathcal{B} = \{(a_1, I_1), \ldots, (a_n, I_n)\}$.

The overall architecture of our proposed method, EchoSight, is illustrated in Figure 2. It consists of four main components: an external knowledge base (KB), a retriever, a reranker, and an answer generator. (i) The process begins with the retriever, which utilizes the reference image to filter and extract relevant KB entries with similar images; (ii) Next, the reranker takes these candidate entries and employs their textual contents to have them reranked, based on their relevance to both the reference image and the textual question; (iii) Finally, the reranked KB entries are fed into the answer generator to produce the final answer.

2.2 Retrieval and Reranking

The goal of this stage is to identify relevant entries from a large-scale external knowledge base using the given reference image and question. We employ a two-stage procedure: first, a visual-only search identifies candidates that are visually similar to the query image. Subsequently, a multimodal reranking process evaluates both visual and textual information to reorder the retrieved entries. This ensures that the most pertinent article entry can be ranked at the top, facilitating efficient and accurate answer generation.

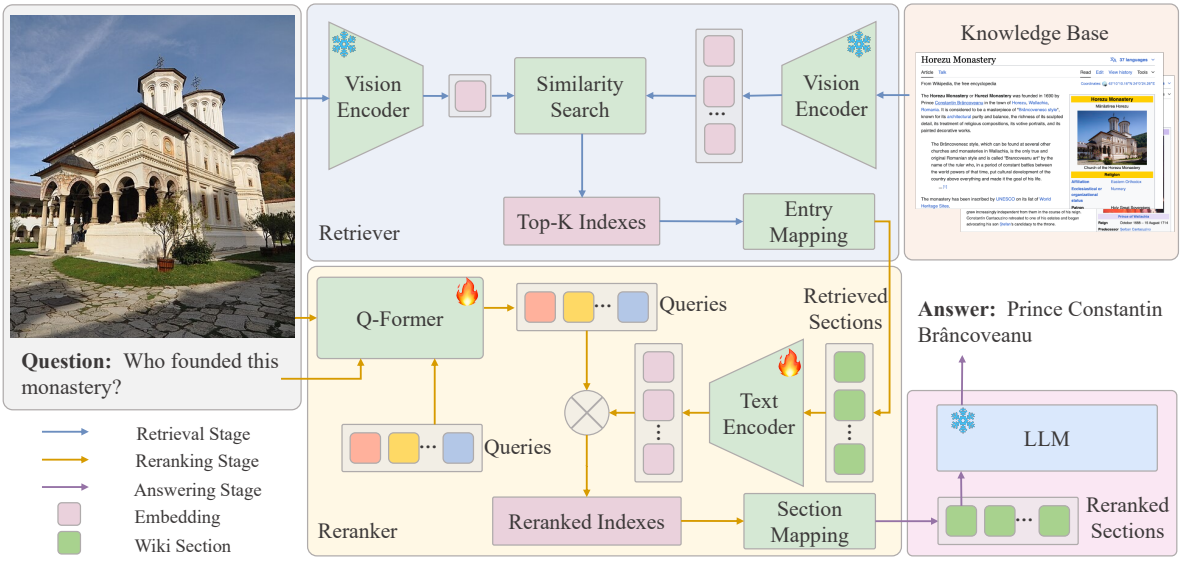

Figure 2: **The overall view of our proposed EchoSight**. **(i)** Given a visual question with an image, the retriever searches the reference image in the knowledge base for top k similar images to get their corresponding Wikipedia Entries. **(ii)** After changing the granularity to sections, all the sections of retrieved entries are then reranked with the maximum pairwise similarity of their textual embeddings and the reference image+question's Q-Former query tokens. **(iii)** The top reranked section will be utilized as RAG prompt for the LLM to generate the ultimate answer.

Visual-only Search. Given the extensive size of the knowledge base, potentially encompassing millions of image-article pairs, optimizing the efficiency of the image search process is critical. To achieve this, we transform all images into vectors and utilize the cosine similarity metric to assess their proximity to a reference image.

$$S_\Omega = \left\{ s_i = \left\langle \frac{v_r}{||v_r||} \cdot \frac{v_i}{||v_i||} \right\rangle, i = 1, \ldots, n \right\},$$

where $v_r = \Phi_{\text{vis}}(I_{\text{ref}})$ and $v_i = \Phi_{\text{vis}}(I_i)$ denote the visual embedding for reference image and database image, respectively, computed by a pre-trained visual encoder. We employ the FAISS library(Douze et al., 2024) for vector search, and keep the top k best-matched images and their corresponding wiki article entries from the knowledge base, *i.e.*, $\mathcal{E}_v = \{(a_1, I_1), \ldots, (a_k, I_k)\}, k \ll n$.

Multimodal Reranking. After initially filtering the candidates based on visual similarities, the reranker module integrates both textual and visual inputs from the multimodal query and the top k retrieved Wikipedia article entries. This stage aims to prioritize entries that are most pertinent to the question, ensuring the articles with highest relevancy are ranked at the top.

Specifically, we employ the Q-Former (Li et al., 2023b) architecture to extract multimodal information from the reference image and textual question, resulting 32 query tokens.

$$z_m^i = \text{Q-Former}\left(I_{\text{ref}}, Q\right)^i,$$

where z_m^i denotes the ith query token embedding of the reference image I_{ref} and textual question Q.

On the candidates side, we break each of the wiki articles into sections, with each section prefixed by the article's title, for example, $a_i = \{\text{sec}_1^i, \text{sec}_2^i, \ldots, \text{sec}_p^i\}$, and further encode them with Q-Former's text encoder. We initialize the Q-Former with BLIP-2's weights and fine-tune all parameters except the visual encoder.

The reranking score for each section is calculated as follows:

$$S_r^{\text{sec}} = \max_{1 \le i \le N_q} \left(\text{sim}(z_m^i, z_s^{\text{sec}})\right),$$

where S_r^{sec} is the reranking score for section "sec", determined using the Q-Former's Image-to-Text Correspondence (ITC) method. This method computes the highest pairwise similarity between each multimodal query token embedding z_m^i from the reference image and question pair, and the `[CLS]` token embedding of a Wikipedia article section z_s^{sec}. N_q denotes the number of query tokens.

In the final step of multimodal reranking, the reranker combines the visual similarity score from the previous stage and the reranking score into a weighted sum:

$$\text{sec}_{vl} = \arg\max_{\text{sec} \in a} \left(\alpha \cdot S_v^{\text{sec}} + (1 - \alpha) \cdot S_r^{\text{sec}}\right),$$

where sec_{vl} refers to the highest-ranked entry section produced by the reranker, α is a weight parameter that balances the visual similarity score S_v^{sec} and the reranking score S_r^{sec}. Note that, S_v^{sec} is calculated in the visual-only search stage using the best-matched image from the wiki entry to which sec belongs.

Reranker Training. Here, we implement hard negative sampling within a contrastive learning framework. Specifically, the negative samples are specifically selected from examples that are visually similar yet contextually distinct, *i.e.*, the initial visual-only retrieval efforts were unsuccessful. With such training, the reranker is thus forced to select the most relevant articles for the multimodal queries, enhancing the overall accuracy and effectiveness of the system (Robinson et al., 2021).

The training objective of the reranker is given as follows:

$$\mathcal{L} = -\log \frac{\exp(\max_{1\leq i\leq N_q} \text{sim}(z_m^i, z_s)/\mathcal{T})}{\sum_{j=1}^{N} \exp(\max_{1\leq i\leq N_q} \text{sim}(z_m^i, z_s^j)/\mathcal{T})},$$

where z_s denotes the positive section embedding, N is the total number of samples including both positive and negatives and $\mathcal{T}$ is the temperature parameter that controls the smoothness of the softmax distribution.

2.3 Answer Generation with LLMs

Once the relevant entries are identified from the knowledge base, large language models (LLMs) will integrate such information to answer the questions, *i.e.*, $A = \text{LLM}(\text{sec}_{vl}, Q)$, where the off-the-shelf LLM acts as an answer generator, sec_{vl} denotes the retrieved wiki article section, and Q refers to the target question. Comparing to existing generative VLMs, such retrieval-augmented generation (RAG) (Lewis et al., 2020), enables the model with the essential contextual knowledge, improving the system's ability to handle complex questions that demand precise and detailed knowledge.

3 Experiments

3.1 Datasets

Encyclopedic VQA (Mensink et al., 2023) contains 221k unique question and answer pairs each matched with (up to) 5 images, resulting in a total of 1M VQA samples. These images are derived from iNaturalist 2021 (iNat21) (Van Horn et al., 2021) and Google Landmarks Dataset V2 (GLDv2) (Weyand et al., 2020). The visual questions focus on the fine-grained categories and instances. There are single-hop and two-hop questions that require different reasoning steps in the dataset. Notably, the dataset provides a controlled knowledge base with 2M Wikipedia articles with images, ensuring all the questions can be answered if correct Wikipedia article is given. For our experiments on E-VQA, we consider the single-hop questions using the provided 2M knowledge base.

InfoSeek (Chen et al., 2023) comprises 1.3M visual information-seeking questions, covering more than 11K visual entities from OVEN (Hu et al., 2023a). InfoSeek provides a knowledge base with 100K Wikipedia articles with images. The questions of the dataset are diverse and the answers can be referenced from Wikipedia. There are a human-labeled 8.9K collection and an automated generated 1.3M collection in InfoSeek. Due to the unavailability of groundtruth for test split, we report evaluation results on the validation split. We note that, the original authors did not publicly release their knowledge base, we therefore filter a 100K knowledge base from E-VQA instead. We will release ours to the community for reproduction and future comparison.

3.2 Metrics

To evaluate the performance of our proposed retrieval-augmented QA model, we focus on two aspects, namely, retrieval and question answering. The retrieval results gauge the system's capability to accurately retrieve relevant articles from a large-scale multimodal knowledge base, while the question answering results assess its holistic effectiveness in providing precise and correct answers to visual questions

Metrics for Retrieval. We utilize the standard metric Recall@K. Recall@K assesses whether the correct article entries appear among the top k retrieved results. An article is considered correct only if its URL exactly matches the target URL, making our retrieval evaluation more stringent and precise compared to methods that only match the content of answers to the retrieved articles.

Metrics for Question Answering. Here, we follow the conventional practise, use different metrics depending on the considered datasets. For E-VQA dataset (Mensink et al., 2023), we use the BEM (Balanced Evaluation Metric) score (Zhang et al., 2019), while for the InfoSeek dataset (Chen

Method	Recall@K			
	K=1	K=5	K=10	K=20
Google Lens	47.4	62.5	64.7	65.2
CLIP I-T	3.3	7.7	12.1	16.5
EchoSight				
w/o. Reranking	13.3	31.3	41.0	48.8
w. Reranking	**36.5**	**47.9**	**48.8**	48.8

Table 1: **E-VQA retrieval experiments.** While Google Lens can be recognized as a *upperbound* in E-VQA, CLIP I-T indicates the retrieval from the reference image to Wikipedia entry texts with CLIP (Radford et al., 2021).

Method	Recall@K			
	K=1	K=5	K=10	K=20
DPR^*_{V+T}	29.6	-	-	-
CLIP I-T	32.0	54.0	61.6	68.2
EchoSight				
w/o. Reranking	45.6	67.1	73.0	77.9
w. Reranking	**53.2**	**74.0**	**77.4**	77.9

Table 2: **InfoSeek retrieval experiments.** Note that, DPR^*_{V+T} (Lerner et al., 2024) actually used an in-house 1.5M knowledge base. Its recall is calculated by answer matching (if the answer appeared in the retrieved text) instead of the absolute article matching we used.

et al., 2023), we employ the VQA accuracy (Goyal et al., 2017; Marino et al., 2019) and *Relaxed Accuracy* (Methani et al., 2020; Masry et al., 2022). These metrics are chosen to align with the evaluation settings specific to each dataset.

3.3 Implementation Details

The Retriever. We compute the visual embedding for the reference images and images from database with a frozen Eva-CLIP vision encoder (Eva-CLIP-8B) (Sun et al., 2024). The pooled last-layer embedding are used as the features for computing cosine similarity between images, with FAISS library.

The Reranker. The reranking module is initialized with pre-trained BLIP-2 (Li et al., 2023b) weights using the LAVIS Library (Li et al., 2023a). The number of query tokens N_q is 32 and weighting parameter α is 0.5. Instead of using in-batch contrastive learning, we employ hard negative sampling, where each positive sample is paired with $N = 24$ negative samples.

In practise, a positive sample is constructed using the evidence section text from the corresponding Wikipedia article. While for negative samples, we perform a visual-only search on the reference images. Knowledge base entries with images that fail to match the reference images ranked within the top k are selected as negative samples. During training, we randomly sample sections from these negative entries as well as from the non-evidence sections of the positive entries. Note that, as only E-VQA dataset provides labeled evidence sections for all its training data, we train the reranker on this dataset, and directly use it on InfoSeek in a zero-shot manner.

We adopt OneCycleLR (Smith and Topin, 2019) scheduler, with AdamW (Loshchilov and Hutter, 2018) optimizer. We use learning rate 10^{-4}, batch size 6, and the negative samples per example being 24. For training the reranker module with 900K examples in Encyclopedic VQA, 150K steps require 40 hours on 1 Nvidia A100 (80G).

The Answer Generator. We use Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) as the question generator for E-VQA and LLaMA-8B-Instruct (AI@Meta, 2024) for InfoSeek.

3.4 Results

In this section, we present experimental results on the E-VQA and InfoSeek benchmarks.

On Retrieval. The experiment results for the retrieval tasks across different configurations are detailed in Table 1 and Table 2. The CLIP I-T setting involves using CLIP for cross-modal similarity search, from the reference image to the Wikipedia article. The articles are represented as CLIP embedding of their title and descriptions. The 'Google Lens' refers to the approach used in Encyclopedic VQA (Mensink et al., 2023), where Google Lens indexes billions of images from the Internet, not limited to Wikipedia, to find and return the most closely matching images along with an entity prediction. The best corresponding knowledge base entry identified by Google Lens is considered as its retrieval results. Given its vast image index, which potentially includes the image from the test set and capability to associate images with relevant entities, Google's retrieval can be viewed as a *top performer* in E-VQA retrieval.

From both tables, we can draw the observation that, our proposed reranking module has shown to significantly improve the retrieval performance, for example, it improves Recall@1 from 13.3% to 36.5% on E-VQA benchmark, 45.6% to 53.2% on InfoSeek, largely bridging the gap towards the 'Google Lens' top performer.

Method	LLM	Retrieval	E-VQA	InfoSeek
Google Lens	PaLM	KB Article	48.0	-
Google Lens	PaLM	KB Section	48.8	-
	PaLM	-	19.7	1.0
Vanilla	Mistral-7B	-	21.0	0.4
	LLaMA3-8B	-	18.7	2.4
BLIP-2	Flan-T5XXL	-	12.6	12.5
LLaVA-1.5	Vicuna-7B	-	16.3	9.5
Wiki-LLaVA	Vicuna-7B	KB Section	21.8	28.9
DPR^{*}_{V+T}	Multi-passage BERT	KB Section	29.1	12.4
EchoSight				
w/o. Reranking	Mistral-7B | LLaMA3-8B[1]	KB Article	19.4	27.7
w. Reranking	Mistral-7B | LLaMA3-8B[1]	KB Section	**41.8**	**31.3**

Table 3: VQA Accuracy comparison with the SOTA methods. Google Lens method is the closed source top performer. Vanilla method indicates the LLM directly generate answers with textual questions only. BLIP-2 (Li et al., 2023b) and LLaVA(Liu et al., 2024) are strong vision language models yet with no retrieval augmented. Wiki-LLaVA(Caffagni et al., 2024) and DPR^{*}_{V+T}(Lerner et al., 2024) are recent works focusing on retrieval-augmented answer generation. Our proposed **EchoSight** is reported without and with multimodal reranking.

VQA Results. As shown in Table 3, we present the comparison to state-of-the-art approaches on final VQA results. For methods that do not utilize an external knowledge base or retrieval system, we present the results of large language models (LLMs), and multimodal large language models (MLLMs). The vanilla method refers to scenarios where only the textual question of the multimodal query is provided. The performance of multimodal-LLMs, including BLIP2 (Li et al., 2023b) and LLaVA (Liu et al., 2024), are reported in Wiki-LLaVA (Caffagni et al., 2024), where both the reference image and question are simultaneously processed. For methods with external knowledge bases, we compare with Wiki-LLaVA (Caffagni et al., 2024) and DPR^{*}_{V+T} (Lerner et al., 2024).

It is clear that our proposed EchoSight (w. reranking) has outperform the prior works by a significant margin, even approaching the upperbound results reported by original E-VQA (Mensink et al., 2023) benchmark, where two giant models are adopted, *i.e.*, 'Google Lens' for knowledge retrieval, and PaLM for answer generation.

3.5 Ablation Study

For all ablation studies, we use the E-VQA dataset. On the retrieval side, we conduct the following ablations: (i) to compare different vision backbones in retrieval, (ii) to study the impact of reranking scope, and (iii) to investigate the importance of hard negative sampling. On final answer generation, we carry out ablation studies on: (i) the impact of various language models and (ii) to experiment with the answer generator under oracle settings.

Backbone	Recall@K			
	K=1	K=5	K=10	K=20
OpenAI-CLIP				
w/o. Reranking	10.1	19.5	25.8	32.2
w. Reranking	**23.8**	**31.4**	**32.1**	32.2
Eva-CLIP				
w/o. Reranking	13.3	31.3	41.0	48.8
w. Reranking	**36.5**	**47.9**	**48.8**	48.8

Table 4: Retrieval performance analysis on different vision backbones.OpenAI-CLIP is CLIP-ViT-Large (Radford et al., 2021) and Eva-CLIP is Eva-CLIP-8B (Sun et al., 2024) from BAAI. We both take the visual encoder's last layer output as the image feature.

Impact of vision backbones. We assess the effect of different vision backbones on the retrieval stage, as detailed in Table 4. We compare the Vision Transformer (ViT) from EvaCLIP-8B (Sun et al., 2024) with OpenAI's CLIP-ViT-Large (Radford et al., 2021). The EvaCLIP-8B's ViT achieves a recall@20 of 48.8%, outperforming the CLIP-ViT-Large, which scored 32.2%. This substantial improvement is likely due to EvaCLIP-8B's larger parameter size and more extensive training dataset, allowing it to develop more robust representations.

While the initial Recall@1 shows a modest difference between the two models (10% for CLIP-ViT-Large and 13% for EvaCLIP-8B), adopting our multimodal reranking significantly boosts performance, increasing Recall@1 to 23.8% and 36.5% for CLIP-ViT-Large and EvaCLIP-8B, respectively.

[1]The E-VQA accuracy is tested with Mistral-7B and InfoSeek accuracy is tested with LLaMA3-8B.

This results in a marked 13% difference, underscoring the effectiveness of our approach, especially when combined with a more capable backbone.

Impact of reranking scope. The reranking scope refers to the number of candidates considered by the reranker module. Involving a higher reranking scope means calculating more embeddings during the reranking process. The reranking scope, which can be any number up to k, *i.e.*, the total number of candidates returned by the retriever. As shown in Table 5, our reranker can consistently improve the results with increasing scope from Top-5 to Top-500. As the thoughput experiment showed in Table 6, considering the balance of efficiency and quality, the scope of 20 candidate entries is used when reporting our final VQA accuracy on E-VQA and InfoSeek.

Scope	Recall@K			
	K=1	K=5	K=10	K=20
Top 5	29.4	32.2	-	-
Top 10	34.3	40.7	40.9	-
Top 20	36.5	47.9	48.8	48.8
Top 50	38.3	53.6	56.9	57.9
Top 100	38.8	55.9	60.8	63.0
Top 500	39.8	58.5	65.3	70.3

Table 5: The ablation study on impact of the reranking scope. Our reranker can consistently improve the results with increasing scope from Top-5 to Top-500.

Scope	Total Retrieval Time	Throughput
Top 10	0.602	1.66
Top 20	1.171	0.85
Top 50	2.720	0.37
Top 100	5.082	0.20
Top 500	21.591	0.05

Table 6: Throughput Study. This study uses a NVIDIA A100 GPU with 80GB memory limiting the batch size to 1. For visual-only retrieval, we use the Faiss library with an exhaustive search. The throughput is calculated as the number of queries processed per second (QPS).

Impact of hard negative sampling. The training strategy of the reranker module is critical for its performance. Rather than using randomly selected, irrelevant article entries, we employ a hard negative sampling during training, *i.e.*, top negative candidates returned by the retriever. This approach ensures the reranker to be trained on more demanding examples, thereby improving its performance and robustness. The effects of different training strategies on reranking performance are detailed in Table 7.

Sampling	Recall@K			
	K=1	K=5	K=10	K=20
EchoSight				
w/o. Hard Neg	31.4	46.0	48.5	48.8
w. Hard Neg	**36.5**	**47.9**	**48.8**	48.8

Table 7: The ablation study of how sampling methods affect the overall retrieval-and-reranking performance.

LLMs	GPT-4	PaLM	Mistral	LLaMA3
Accuracy	44.4	39.0	41.8	38.9

Table 8: The ablation study of impact of language models. The results are generated with the retrieval results of EchoSight with reranking scope 20.

Consistency of EchoSight across LLMs. The choice of LLMs influences the RAG paradigm greatly (Shao et al., 2023; Hu et al., 2022). We compare PaLM (Chowdhery et al., 2023), GPT-4 (Achiam et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and LLaMA3-8B-Instruct (AI@Meta, 2024) as answer generators. Specifically, we provide them with same reranking results (KB entries). As shown in Table 8, the accuracy results are calculated with BEM (Zhang et al., 2019) following (Mensink et al., 2023). The results indicate that though better language models yield better scores, the overall performance across all tested language models is quite stable. This validates our method adapts well across modern language models.

Effect of oracle retrieval. Oracle retrieval indicates that the correct Wikipedia entry is always provided for generating the answer. As shown in Table 9, LLMs can *almost* answer the question if oracle retrieval is provided.

LLM	Retrieval	Accuracy
PaLM	KB Title	31.0
Mistral-7B	KB Title	29.4
LLaMA3-8B	KB Title	32.0
PaLM	KB Article	78.4
Mistral-7B	KB Article	84.8
LLaMA3-8B	KB Article	84.6
PaLM	KB Section	87.0
Mistral-7B	KB Section	89.9
LLaMA3-8B	KB Section	90.7

Table 9: The ablation study with VQA results on the effect of oracle retrieval.

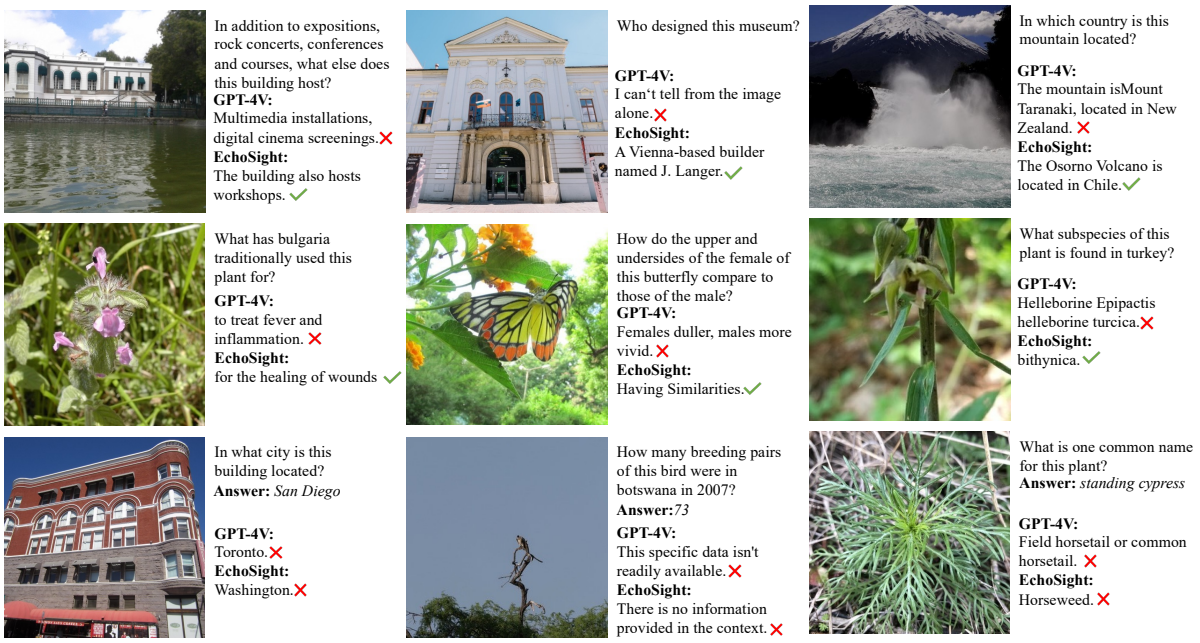

Figure 3: Qualitative VQA results from Encyclopedic VQA comparing to GPT-4V. The first row shows results in landmarks and the second row in natural species. Some failure cases are shown in the third row altogether with ground-truth.

3.6 Qualitative Results

As shown in Figure 3, our EchoSight demonstrates significant improvements in multimodal understanding and generation tasks compared to the state-of-the-art GPT-4V (Achiam et al., 2023).

4 Related Work

4.1 Visual Question Answering

Visual Question Answering (VQA) is the task of answering open-ended questions based on an image with natural language response. VQA tasks can be divided into two types: standard VQA and knowledge-based VQA.

Standard VQA. Datasets such as VQAv1 (Antol et al., 2015), VQAv2 (Goyal et al., 2017), and VizWiz (Gurari et al., 2018) focus on questions that can be answered by analyzing the image content alone, without external information. These datasets typically cover questions about objects in the image, their attributes and other perceptual details that can be inferred from the visual input.

Knowledge-based VQA. The task involves questions that require information not present in the image. Pioneering datasets like OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), which include questions needing knowledge beyond what is visually depicted, necessitate the integration of external world knowledge and commonsense reasoning. Uni-modal knowledge bases like GS112K (Luo et al., 2021) and Wiki21M (Karpukhin et al., 2020) are adopted in prior works (Lin et al., 2023; Lin and Byrne, 2022; Gao et al., 2022; Luo et al., 2021, 2023). However, uni-modal knowledge bases are text-only, which limits their applicability in scenarios where visual context is paramount. To better utilize multimodal information, multiple previous attempts have been made (Ding et al., 2022; Zhu et al., 2020; Wu et al., 2022; Chen et al., 2022).

Datasets such as Encyclopedic VQA (E-VQA) (Mensink et al., 2023) and InfoSeek (Chen et al., 2023) have been developed with multimodal knowledge bases. These datasets utilize Wikipedia as a multimodal knowledge base to provide detailed and specific information on various topics. E-VQA covers a wide range of topics like animals, plants, and landmarks, while InfoSeek focuses on info-seeking questions about various visual entities. These datasets require models to recognize visual entities and accurately retrieve and use relevant information from external sources (Lerner et al., 2024; Caffagni et al., 2024; Lin et al., 2024).

4.2 Vison Language Models for VQA

Advances in Vision Language Models (VLMs) such as GPT-4V (Achiam et al., 2023), Gem-

ini (Gemini Team et al., 2023), LLaVA (Liu et al., 2024), and Phi-3-Vision (Abdin et al., 2024) have demonstrated impressive capabilities in standard Visual Question Answering (VQA) tasks, exhibiting strong image analysis and accurate response generation (Li et al., 2023d). However, these models encounter difficulties with knowledge-based VQA due to issues such as hallucination, where responses are generated based on nonexistent content and internal biases (Li et al., 2023c), and the lack of efficient knowledge retrieval mechanisms which hampers the integration of external knowledge bases for reasoning (Caffagni et al., 2024).

Recently, research has shifted towards retrieval-augmented generative systems. While Retrieval-Augmented Generation (RAG) has been well-established in Large Language Models (LLMs), its application in VLMs remains underexplored. Systems like KAT (Gui et al., 2021), REVIVE (Lin et al., 2022), and REVEAL (Hu et al., 2023b) show promise for questions involving commonsense reasoning, yet they struggle with complex, knowledge-intensive tasks like Encyclopedic VQA (E-VQA) and Infoseek. These limitations stem from their restricted ability to fetch and incorporate precise information from extensive encyclopedic knowledge bases (Mensink et al., 2023).

EchoSight addresses these issues through a novel two-stage process combining visual-only retrieval and multimodal reranking. This approach significantly enhances the alignment between retrieved textual knowledge and visual content, leading to improved performance on benchmarks such as Encyclopedic VQA and InfoSeek.

5 Conclusion

In this paper, we introduced EchoSight, a novel retrieval-augmented vision language system designed to address the challenges of knowledge-based Visual Question Answering (VQA). Our approach enhances the retrieval capabilities of multimodal models through a two-stage process: initial visual-only retrieval followed by a multimodal reranking stage. This methodology significantly improves the alignment between visual and textual information, leading to more accurate and contextually relevant answers. Experimentally, we have conducted thorough ablation studies to demonstrate the effectiveness of our proposed components. While comparing to existing state-of-the-art approaches on the Encyclopedic VQA and InfoSeek datasets, EchoSight demonstrates significant performance improvement, with an accuracy of 41.8% on E-VQA and 31.3% on InfoSeek. The success of EchoSight highlights the importance of efficient retrieval processes and the integration of multimodal information in enhancing the performance of large language models (LLMs) in knowledge-based VQA tasks.

Limitations

Although our proposed EchoSight demonstrates impressive performance on Knowledge-based VQA like Encyclopedic-VQA and InfoSeek, several limitations must be acknowledged. EchoSight's performance is heavily dependent on the quality and comprehensiveness of the underlying knowledge base used for retrieval. Domain-specific knowledge not covered in these databases may lead to sub-optimal performance in specialized queries. In addition, the retrieval process, especially when involving multimodal reranking of candidates, introduces significant computational overheads, making it less suitable for real-time applications. These overheads can impact the efficiency and response time of the system. Future work focusing on improving the quality of knowledge bases and mitigating computational overheads remains to be explored.

Acknowledgements

This work is funded by National Key R&D Program of China (No.2022ZD0161400).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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

AI@Meta. 2024. Llama 3 model card.

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 International Conference on Computer Vision*.

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 Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

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*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5089–5098.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

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

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

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 Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023d. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

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*.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

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

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

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 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.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *icml*, pages 8748–8763. PMLR.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision*, pages 146–162. Springer.

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 Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks

using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.

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

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12884–12893.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2575–2584.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.

Dataset	Question Type	Number of IQA pairs		
		Train	Val	Test
E-VQA	Templated	66,535	1,827	1,000
	Automatic	737,114	8,025	2,750
	Multi Answer	112,736	1,844	1,000
	Total	**916,385**	**11,696**	**4,750**
InfoSeek	**Total**	**902,509**	-	**71,335**

Table 10: Dataset details used in our EchoSight's traning and testing.

A Dataset Details

In this section, we provide more details of in the Dataset we used. We summarize the statistics of in Table 10.

A.1 E-VQA

We focus only on Single-hop questions of E-VQA (Mensink et al., 2023), namely Templated, Automatic, and Multi Answer questions in the table.

A.2 InfoSeek

And for Infoseek (Chen et al., 2023), due to the missing entities in the knowledge-base we use, we remove the examples in the dataset. Specifically, 916,385 examples in training split out of 934,048 are kept (98.1%), and 71,335 examples of validation split out of 73,620 are kept (96.9%). Therefore, the results we obtain with our knowledge base are consistent with the dataset's original setting while considering for the limitations of our knowledge base.

B Vision backbones

In addition to CLIP, there are other robust vision backbones available, such as Dino (Caron et al., 2021; Oquab et al., 2023). Unlike CLIP, which employs a visual-language training method, Dino leverages a self-supervised, visual-focused training approach. To evaluate its performance, we benchmarked DinoV2 as our visual-only retriever, presenting the results in Tables 11 and 12. Despite observing a notable performance improvement in the Encyclopedic VQA task, there was a significant drop in performance on the InfoSeek task. Therefore, to maintain the consistency and overall performance of EchoSight, we have decided to continue using Eva-CLIP as our vision backbone.

Backbone	Recall@K			
	K=1	K=5	K=10	K=20
Eva-CLIP				
w/o. Reranking	13.3	31.3	41.0	48.8
w. Reranking	**36.5**	**47.9**	**48.8**	48.8
DINOv2				
w/o. Reranking	17.3	38.6	46.0	51.4
w. Reranking	**40.8**	**50.7**	**51.3**	51.4

Table 11: DINOv2 comparison to CLIP in E-VQA.

Backbone	Recall@K			
	K=1	K=5	K=10	K=20
Eva-CLIP				
w/o. Reranking	45.6	67.1	73.0	77.9
w. Reranking	**53.2**	**74.0**	**77.4**	77.9
DINOv2				
w/o. Reranking	34.7	53.4	60.0	64.5
w. Reranking	**38.2**	**60.0**	**64.1**	64.5

Table 12: DINOv2 comparison to CLIP in InfoSeek.

C Prompt Template

C.1 E-VQA

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

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

C.2 InfoSeek

Due to the strict metrics of exact match are used by InfoSeek (Chen et al., 2023), we have to consider the format of the prompt so that the generated answer is comparable with the ground truth. Thereby, by using a one-shot example to keep the format correct, our prompt we use for InfoSeek is:

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

USER:Context: The sounthern side of the
Alps is next to Lake Como.
```

```
Question: Which body of water is this
mountain located in or next to?
Just answer the questions, no
explanations needed.
Short answer is: Lake Como

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