

# Seeing Beyond: Enhancing Visual Question Answering with Multi-Modal Retrieval

Boqi Chen<sup>1\*</sup>, Anuj Khare<sup>2</sup>, Gaurav Kumar<sup>2</sup>, Arjun Akula<sup>2</sup>, Pradyumna Narayana<sup>2</sup>

<sup>1</sup>University of North Carolina at Chapel Hill, <sup>2</sup>Google

## Abstract

Multi-modal Large language models (MLLMs) have made significant strides in complex content understanding and reasoning. However, they still suffer from model hallucination and lack of specific knowledge when facing challenging questions. To address these limitations, retrieval augmented generation (RAG) has emerged as an effective solution. While incorporating knowledge has led to improvements, it also highlights the need for a more robust knowledge selection strategy. For multi-modal tasks, such as visual question answering (VQA), integrating all modalities is crucial in providing comprehensive information for accurate answers. Therefore, we propose to construct an encoder model for extracting joint embedding from all modalities, enabling alignment between the corresponding query and knowledge through contrastive learning. To further improve performance, we introduce an additional MLLM re-selection step, which selects the best matching knowledge from the top- $k$  retrieved results of our alignment model. We evaluated our method, SeBe-VQA, on the Encyclopedic VQA dataset. Our knowledge retrieval results demonstrate the benefit of our multi-modal framework. By incorporating the retrieved knowledge along with the question, we achieve a significant performance improvement compared with the previous method and scenarios without knowledge provision.

## 1 Introduction

Multi-modal large language models (MLLMs) are rapidly becoming widely used tools for everyday queries. Despite their impressive capabilities in content understanding and reasoning, MLLMs still face significant challenges when answering questions that require specific domain knowledge, as shown in Fig. 1. In such situations, the MLLM may



**Question:** What organization estimates the population of this bird in Europe?  
**Correct answer:** BirdLife International

**Gemini-1.5:** The question cannot be answered from the given context. The image shows a bird, but there is no information about the bird's species or any population estimates.

**LLaVa-1.5:** The European Bird Census is an organization that estimates the population of various bird species in Europe. They conduct regular bird counts and surveys to monitor and assess the health of bird populations in the region. The bird in the image is a small brown bird, which could be a part of the bird population that the European Bird Census is monitoring.

**SeBe-VQA:** BirdLife International.

Figure 1: An example showcasing where existing MLLMs either refuse to respond or provide incorrect answers to a query. Our multi-modal retrieval augmented method is able to select the relevant knowledge and guide the MLLM for correct responding.

decline to answer due to insufficient information or provide a related but incorrect response.

Researchers are actively exploring ways to enhance MLLM performance by incorporating extensive amounts of training samples (Hoffmann et al., 2022; Team et al., 2023; Touvron et al., 2023). However, this approach can be computationally demanding. Retrieval augmented generation (RAG) has recently proven effective for various scenarios (Lewis et al., 2020; Gao et al., 2023) by incorporating retrieved knowledge into the LLM along with the query. This method not only facilitates better answer generation but also provides the source of the generated result. Despite its effectiveness, developing an accurate retrieval method remains challenging given the vast amount of information in the knowledge base. Furthermore, most existing RAG solutions are based on a single modality and cannot effectively address multi-modal scenarios. A particularly relevant study (Caffagni et al., 2024) applies a 2-step retrieval process: selecting candidates by query-knowledge image matching using the CLIP model (Radford et al., 2021), and then filtering using the query text. While this method uses both image and text for retrieval, the two-stage pipeline is suboptimal compared to combining modalities in the same stage. Therefore, there

\*Work done during an internship at Google

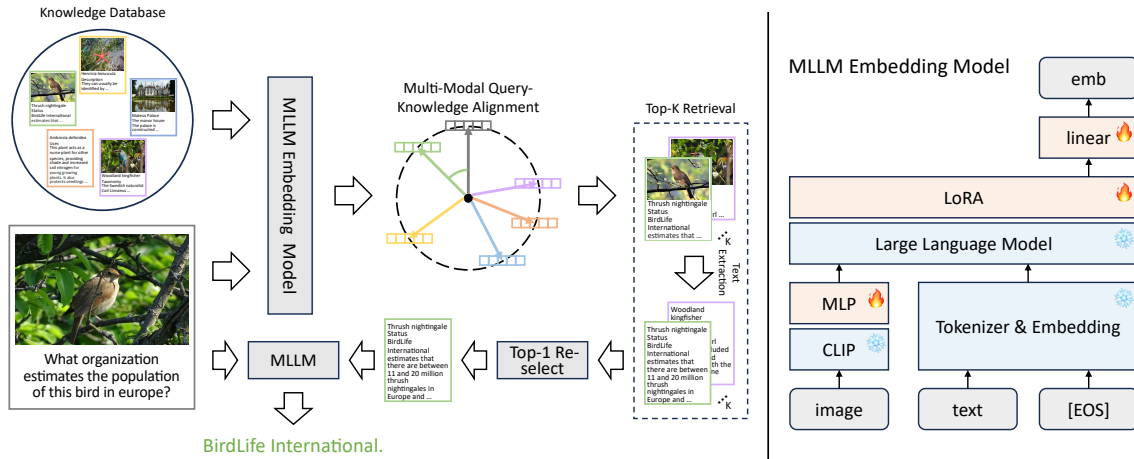


Figure 2: Our multi-modal retrieval augmented visual question answering framework. **Left:** The model independently encodes both the multi-modal query and all Wikipedia sections. Feature distances between the query and each Wikipedia section are calculated, and the top- $k$  knowledge is selected to guide the question answering process. **Right:** The model extracts features by treating image embeddings as tokens for input. A [EOS] token is added at the end, and its output embedding is used as the multi-modal joint feature embedding.

is a clear need for a multi-modal RAG approach that can effectively leverage the interplay between different modalities for knowledge selection.

To effectively connect a multi-modal query with its relevant knowledge in VQA, we need to create a joint embedding space for accurate alignment. However, existing methods (Radford et al., 2021; Girdhar et al., 2023) face two issues: 1) only semantically similar data are aligned, which may not perfectly fit between the query and its knowledge, and 2) joint feature extraction of multiple modalities is not supported, given the semantic gap between different modalities. As illustrated in Fig. 1, aligning based on image similarity might retrieve information about a visually similar bird that is irrelevant to the question, while relying solely on the text might yield results about birds in general without addressing the specific species in the image. Therefore, straightforward methods such as averaging individual features in the joint embedding space can be misleading. To address these limitations, we propose a novel approach that leverages contrastive learning to achieve robust query-knowledge alignment across both single and multiple modalities.

For query-knowledge alignment, constructing a multi-modal embedding is crucial. Typical approaches involve concatenation or element-wise multiplication of individually extracted features for each modality. With the success of MLLMs, LLaVa (Liu et al., 2024) demonstrates the strong image-text understanding capability by treating im-

age embeddings as tokens for generation tasks. However, its decoder-only architecture and autoregressive objective make it suboptimal for feature extraction. To overcome this, we propose modifying existing MLLMs for feature encoding and using contrastive learning to align the query with its knowledge.

To answer visual questions accurately, RAG systems typically rely on retrieving the single best piece of knowledge. Our multi-modal alignment model effectively identifies relevant knowledge, but we further refine this process by considering the top- $k$  retrieved candidates. An MLLM then re-selects the most suitable knowledge from this refined set, leading to improved performance in VQA. This re-selection step is inspired by work highlighting the benefits of re-ranking in RAG (Glass et al., 2022; Song et al., 2024).

In this work, we proposed SeBe-VQA to improve MLLM’s generation accuracy on challenging VQA tasks. Our contributions are as follows:

- We develop a multi-modal feature encoder that extends from a pre-trained MLLM, which supports both single and multiple modalities.
- We construct a joint embedding space that aligns the multi-modal query with its knowledge, enabling accurate knowledge retrieval.
- We further improve retrieval performance by re-selecting from the top- $k$  retrieved knowledge using an existing MLLM.

- By incorporating the extracted knowledge with the query, SeBe-VQA significantly improves MLLM’s responses on the Encyclopedic VQA dataset (Mensink et al., 2023).

## 2 Related Works

Extracting joint embeddings from multi-modal data is crucial for comprehensive understanding. While early works focused on concatenation or element-wise multiplication (Antol et al., 2015; Anderson et al., 2018) of independently extracted features, recent advances in LLMs have prompted exploration of their potential for feature extraction. To encode information from a decoder-only LLM architecture, several approaches propose appending a special token to the input sequence and utilizing its output representation as the feature embedding (Wang et al., 2023; Ma et al., 2024). Alternatively, LLM2Vec (BehnamGhader et al., 2024) extends the decoder-only architecture to a bi-directional framework, requiring additional training. However, these methods primarily focus on encoding single modalities, limiting their applicability to tasks like VQA. To address this gap, we extend LLM encoding techniques to MLLMs, leveraging their strong understanding capabilities in a multi-modal setting.

Data retrieval plays a vital role in various applications. A common approach involves constructing an embedding space where semantically similar data are clustered together. With the growing abundance of multi-modal data, cross-modal retrieval has become increasingly important. CLIP (Radford et al., 2021) employs contrastive learning to align image and text representations, while ImageBind (Girdhar et al., 2023) extends this approach to align six different modalities using image as a common anchor. However, these methods primarily focus on retrieval between single modalities, which may be suboptimal for VQA tasks that require joint reasoning over multiple modalities. Therefore, we propose a multi-modal alignment framework that encodes multiple, potentially semantically irrelevant modalities together.

Direct retrieval from a large database can be challenging and inaccurate. To address this, many works (Nogueira and Cho, 2019; Ren et al., 2021; Shen et al., 2021) incorporate a re-ranking step to refine retrieval results. Conventional methods include pseudo-relevance feedback, graph-based, clustering-based approaches, etc. (Arun et al., 2017) Recently, several works (Sun et al., 2023;

Pradeep et al., 2023) have explored using LLM for re-ranking, which turned out to be effective. In this work, we focus on data re-selection, specifically utilizing the top-1 result after re-ranking.

## 3 Method

To enhance the performance of MLLMs for visual question answering, we propose a 2-step approach: (1) multi-modal query-knowledge alignment, which constructs an embedding space for effective retrieval, and (2) retrieval-augmented visual question answering, which leverages the retrieved knowledge to generate accurate answers.

### 3.1 Multi-modal Query-Knowledge Alignment

To extract the joint embedding of multiple modalities, various approaches have been explored. Early approaches (Antol et al., 2015; Anderson et al., 2018) combine individual features through concatenation or element-wise multiplication. More recent works (Lu et al., 2019; Yu et al., 2022; Li et al., 2022) proposed treating image embeddings as tokens, enabling explicit alignment between image and text. LLaVa (Liu et al., 2024, 2023) further simplified this process by employing direct auto-regression. Specifically, images are encoded into feature embeddings using a CLIP (Radford et al., 2021) encoder with an additional MLP for feature transformation and dimensionality matching. These image features, along with tokenized text embeddings, are then fed into an LLM for text generation. While effective for generating multi-modal conversations, the decoder-only architecture is not inherently designed for feature extraction.

To address this limitation, we follow previous works (Wang et al., 2023; Ma et al., 2024) and append a special [EOS] token to the end of the input sequence, as shown in Fig. 2 right. Due to the decoder-only architecture, only the last token can attend to the entire input sequence. Therefore, we utilize the output representation of this [EOS] token as the joint embedding for the multi-modal input. For computational efficiency, we avoid fine-tuning the entire LLM and instead incorporate LoRA layers (Hu et al., 2021) to the pre-trained LLM.

To align multi-modal queries with their corresponding knowledge in the embedding space, we independently extract their feature embeddings ( $z$ ) using the aforementioned feature encoder. Both the query and knowledge share the same network

Model	section-wise				article-wise			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Wiki-LLaVa*	-	-	-	-	3.3	-	9.9	13.2
SeBe-VQA-Text	20.0	37.5	45.6	54.0	24.7	43.3	51.8	60.3
SeBe-VQA-TextImg	16.5	33.0	40.9	49.8	22.4	40.7	49.2	57.7

(a) Recall value of our multi-modal query-knowledge alignment method, presented both section-wise and article-wise. Values marked with \* are taken from the previous paper.

2-Step R@1	section-wise				article-wise			
	top-1	top-5	top-10	top-20	top-1	top-5	top-10	top-20
SeBe-VQA-Text	20.0	30.2	33.1	33.0	24.7	35.3	38.9	40.1
SeBe-VQA-TextImg	16.5	27.2	30.5	32.7	22.4	33.4	37.4	40.5

(b) R@1 value from our 2-step method re-selected by Gemini-1.5, where top- $k$  represents re-selecting the best matching knowledge from the closest  $k$  retrieved candidates using our multi-modal alignment method.

Table 1: Recall value from the WikiWeb2M (Burns et al., 2023) dataset. **Top:** Direct retrieval results from our multi-modal query-knowledge alignment. **Bottom:** Our R@1 results after a 2-step re-selection from MLLM.

weights, and we align positive feature pairs using the following contrastive loss (Chen et al., 2020):

$$loss = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)}, \quad (1)$$

where  $\tau$  is the temperature,  $N$  is the total number of pairs in a batch,  $z_i$  and  $z_j$  are positive features pairs,  $z_k$  is all other features except for  $z_i$ , and  $sim(\cdot, \cdot)$  represents cosine similarity between the features.

### 3.2 Retrieval Augmented Visual Question Answering

To retrieve the knowledge needed for visual question answering, we first encode all queries and knowledge using the aforementioned feature encoder. For each query, we compute the cosine similarity between its feature and those from the entire knowledge database. We then retrieve the top- $k$  nearest knowledge entries. When  $k > 1$ , we apply an additional re-selection step using an existing MLLM. This re-selection step utilizes the following prompt as input to the MLLM:

```
From the below k contexts, select the most
related one for answering the following
question. Your response should be of the
following format: 'Answer: $NUMBER' (without
quotes) where NUMBER is one of 012...k-1.
0 <OPTION 0>
1 <OPTION 1>
...
k-1 <OPTION k-1>
<IMAGE>
Question: <QUESTION>
```

With the final knowledge being retrieved from the knowledge database, we use the following

prompt to generate answers from existing MLLM for a given query:

```
<IMAGE>
Context: <KNOWLEDGE>
Question: <QUESTION>
The answer is:
```

## 4 Experiments

### 4.1 Dataset

We use the Encyclopedic VQA (Mensink et al., 2023) dataset, which comprises 221k unique question-answer pairs. Each question is associated with up to 5 images during training, and we randomly sample 1 image per epoch. Questions are categorized by type and labeled with at least one corresponding Wikipedia section from the WikiWeb2M dataset (Burns et al., 2023; Srinivasan et al., 2021). Following the approach in (Caffagni et al., 2024), we exclude all 2-hop questions during evaluation, resulting in a testing set of 4,750 questions.

Our knowledge database is derived from the entire WikiWeb2M dataset (Burns et al., 2023; Srinivasan et al., 2021), containing approximately 2 million Wikipedia articles with an average of 8 sections per article (see the Appendix for an example article corresponding to the question in Fig. 1). Each section includes the article title, section title, and section text. The section image is selected as the first image appearing in the main section of the article. If no image is present in the main section, only the text is used for feature encoding.

To ensure the retrieval of accurate and concise knowledge, we define the corresponding section for each query as positive and sections from all other

Wikipedia articles as negative. Sections within the same article, excluding the corresponding section, are treated as neither explicitly positive nor negative. This is because these sections may be partially related to the query but not directly answer it. While one solution is to prevent multiple sections from the same article appearing in the same training batch, the large knowledge base size and the small batch size make this occurrence unlikely. Therefore, we simplify the training process by contrasting only the positive section for each query.

## 4.2 Training

Our MLLM embedding model utilizes the 7B parameter LLaVa-1.5 (Liu et al., 2023) as its backbone. This architecture employs a CLIP-ViT-Large (Radford et al., 2021) with 2-layers of MLP for image encoding, and vicuna-7b-v1.5 (Zheng et al., 2024) as the LLM. We initialize the model weights from pre-trained LLaVa 1.5 and incorporate LoRA (Hu et al., 2021) layers to the LLM for computational efficiency. An additional linear layer is added to project the output features to a dimension of 2048. During training, both the CLIP and LLM weights are frozen.

We train with DeepSpeed (Rajbhandari et al., 2020) for distributed training with a batch size of 64 and employ the AdamW optimizer (Loshchilov and Hutter, 2017) with a cosine scheduler. The learning rate is set to  $2e-5$  for MLP and  $2e-4$  for all other parameters. All images are resized to  $336 \times 336$  pixels and then divided into patches of  $14 \times 14$ . The model is trained for 3 epoch on the training set using  $4 \times 40G$  NVIDIA A100 GPUs.

To retrieve the top- $k$  most relevant knowledge entries for a given query, we use Faiss (Douze et al., 2024) for efficient nearest-neighbor lookup in the embedding space. We compare two retrieval strategies: (1) directly using the closest knowledge entry retrieved by our query-knowledge alignment model, and (2) a 2-step method where an additional re-selection step is performed using Gemini-1.5-flash (Team et al., 2023; Reid et al., 2024) on the top- $k$  retrieved entries. The selected knowledge is then used along with the query for VQA.

We develop two model variations: (1) SeBe-VQA-Text, which encodes Wikipedia sections using only textual data, and (2) SeBe-VQA-TextImg, which encodes sections using both text and the first image in the main section of the corresponding article. In both cases, the query consists of both the question image and text.

Method	MLLM	2-step	Accuracy
LLaVa-1.5*	Vicuna-7B	-	16.9
LLaVa-1.5*	Vicuna-7B-Finetuned	-	28.5
Wiki-LLaVa*			26.4
Vanilla Oracle	Gemini-1.5	-	18.7
		-	87.4
		-	35.2
SeBe-VQA-Text	Gemini-1.5	top-5	43.2
		top-10	45.4
		top-20	46.5
		-	32.5
SeBe-VQA-TextImg	Gemini-1.5	top-5	40.4
		top-10	43.6
		top-20	45.0

Table 2: Visual question answering accuracy on the Encyclopedic VQA dataset. Our SeBe-VQA enhances MLLM’s generation capability and outperforms the previous method. We also present the vanilla and oracle result from Gemini-1.5.

## 4.3 Results

### 4.3.1 Knowledge Retrieval

To evaluate the effectiveness of our knowledge retrieval method, we show the recall values in Tab. 1, where  $R@k$  represents the percentage of questions for which the corresponding knowledge is retrieved within the  $k$ -nearest neighbors. We report results using two metrics: section-wise and article-wise recall. Section-wise recall requires that both the correct Wikipedia article and the correct section within that article are selected. Article-wise recall is a more relaxed metric, considering retrieval successful if any section from the correct Wikipedia article is selected. Consequently, article-wise recall is generally higher than section-wise recall.

Tab. 1a presents the recall values achieved by our methods and demonstrates their superior performance compared to Wiki-LLaVa (Caffagni et al., 2024), which utilizes CLIP (Radford et al., 2021) for Wikipedia article retrieval. Comparing the article-wise and section-wise recall of our methods reveals that the article-wise results are only marginally better than the section-wise results. Given that each article contains an average of 8 sections, this finding highlights our model’s ability to directly retrieve the correct section. Interestingly, SeBe-VQA-TextImg exhibits slightly lower accuracy than SeBe-VQA-Text, despite incorporating additional image features. We attribute this to the presence of mismatched images between queries and their corresponding Wikipedia articles (see Appendix for examples).





Image				
Text	Who currently lives at this building?	What is the wingspan of this bird in cm?	What state lists this plant as threatened?	When was this park created?
Ground Truth	Denise Anne Carey	16–21 cm (6.3–8.3 in)	Michigan	1820s
Vanilla	This building is currently vacant and owned by the state of Rhode Island. It is currently <b>not lived in by anyone</b> .	It is <b>not possible to determine</b> the wingspan of the bird from the image provided.	This plant is <b>not listed as threatened in any state</b> .	This question <b>cannot be answered</b> from the given image.
SeBe-VQA-Text	<u>Denise Anne Carey</u> , an architect from New York.	<u>16–21 cm</u>	<u>Kentucky</u> .	<u>1852</u>
SeBe-VQA-Text-Top5	<u>Denise Anne Carey</u> , an architect from New York, currently lives at Seaview Terrace.	<u>16–21 cm</u>	<u>Michigan</u> .	The gardens were created in the <u>1820s</u> .

Figure 3: Visual question answering example from the Encyclopedic VQA test set. The top half shows the query’s image, question, and ground truth answer. The bottom half presents Gemini-1.5 without additional knowledge, with the top-1 retrieved knowledge, and our 2-step method re-selected from the top-5 retrieved knowledge. For the left two examples, SeBe-VQA-Text is able to directly retrieve the correct section from all Wikipedia sections, and for the right two examples, our 2-step method can correctly refine from the top-5 retrieved knowledge.

Tab. 1b shows the R@1 values achieved by our 2-step method. As no re-selection is necessary for top-1 retrieval, these values are identical to those in Tab. 1a. Re-selecting from the top-5 retrieved knowledge entries significantly improves performance ( $\sim 50\%$ ) compared to using only the top-1 entry. This improvement likely stems from the extensive knowledge base, which makes direct retrieval more challenging. However, the performance gain diminishes with a larger candidate pool. This can be attributed to both the plateauing of R@k with increasing  $k$ , as shown in Tab. 1a, and the limitations of existing MLLMs in effectively selecting from a large context window.

#### 4.3.2 Visual Question Answering

For visual question answering using MLLM, we provide the model with the best matched knowledge selected as described in sec. 4.3.1, along with the query image and text, for answer generation. As shown in Tab. 2, all our proposed methods achieve higher accuracy than both the previous method and the vanilla Gemini model. Incorporating knowledge retrieved via our 2-step method further boosts performance, with a particularly significant improvement observed when moving from top-1 (1-step) to top-5 (2-step) retrieval. This trend mirrors the recall value improvements shown in Tab. 1b. We also present the oracle result for Gemini, where the ground-truth knowledge section is provided for answer generation. This result repre-

sents the upper bound of retrieval augmented visual question answering.

Fig. 3 provides illustrative examples. The left two columns showcase instances where both our multi-modal query-knowledge alignment model and the 2-step method successfully select the correct Wikipedia section. In contrast, the right two columns demonstrate how our 2-step method can rectify incorrect top-1 retrieval by re-selecting the corresponding knowledge from the top-5 retrieved sections. Additional examples including the retrieved Wikipedia sections are in the Appendix.

## 5 Conclusion

In this work, we presented a multi-modal retrieval augmented visual question answering method, where both the queries and knowledge can encompass multiple modalities. To retrieve relevant knowledge from the database, we employed contrastive learning to align the multi-modal queries with their corresponding knowledge in the embedding space. These embeddings are derived from our proposed MLLM embedding model. We further enhanced retrieval performance by incorporating an additional re-selection step, which also improved the visual question answering capabilities of existing MLLMs. Evaluation on the Encyclopedic VQA dataset demonstrated that our multi-modal retrieval framework outperforms previous method and that the retrieved knowledge effectively guides the MLLM toward more accurate responses.

In the future, we plan to dynamically incorporate relevant Wikipedia images for enhanced knowledge retrieval. We also aim to refine our multi-modal alignment model by considering different sections within the relevant article and dynamically weighting them during contrastive learning.

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- 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.
- KS Arun, VK Govindan, and SD Madhu Kumar. 2017. On integrating re-ranking and rank list fusion techniques for image retrieval. *International Journal of Data Science and Analytics*, 4:53–81.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.
- Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. *arXiv preprint arXiv:2305.03668*.
- Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- 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.
- 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.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.
- 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. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.
- Thomas Mensink, Jasper Uijlings, Lluís 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.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.
- Ronak Pradeep, Sahel Sharifmoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.
- 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 *International conference on machine learning*, pages 8748–8763. PMLR.

- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.
- Xi Shen, Yang Xiao, Shell Xu Hu, Othman Sbai, and Mathieu Aubry. 2021. Re-ranking for image retrieval and transductive few-shot classification. *Advances in Neural Information Processing Systems*, 34:25932–25943.
- EuiYul Song, Sangryul Kim, Haeju Lee, Joonkee Kim, and James Thorne. 2024. Re3val: Reinforced and reranked generative retrieval. *arXiv preprint arXiv:2401.16979*.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.
- 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*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.



## A Additional Examples



**Question:** Whose statue is on the facade of this museum?

**Ground truth:** Dewitt Clinton and Alexander Hamilton

**Vanilla:** The statue on the facade of this museum is of the [city's founder](#).

**SeBe-VQA:** The statues on the facade of the Museum of the City of New York are of [Alexander Hamilton and DeWitt Clinton](#).

**SeBe-VQA retrieved Wikipedia section:**

**Article title:** Museum of the City of New York

**Section title:** Museum of the City of New York

**Section Text:** The Museum of the City of New York (MCNY) is a history and art museum in Manhattan, New York City, New York. It was founded by Henry Collins Brown, in 1923 to preserve and present the history of New York City, and its people. It is located at 1220–1227 Fifth Avenue between East 103rd to 104th Streets, across from Central Park on Manhattan's Upper East Side, at the northern end of the Museum Mile section of Fifth Avenue. The red brick with marble trim museum was built in 1929–30 and was designed by Joseph H. Freedlander in the neo-Georgian style, with [statues of Alexander Hamilton and DeWitt Clinton by sculptor Adolph Alexander Weinman facing Central Park from niches in the facade](#). The museum is a private non-profit organization which receives government support as a member of New York City's Cultural Institutions Group, commonly known as "CIG"s. Its other sources of income are endowments, admission fees, and contributions. The museum is hosting the "New York At Its Core" permanent exhibition, the first-ever museum presentation of New York City's full history.



**Question:** In addition to the olympic games, what other games were held at this sports venue in 2012?

**Ground truth:** Paralympic Games

**Vanilla:** The answer is [BMX Supercross](#). The image includes a banner that says "uci) BMX Supercross".

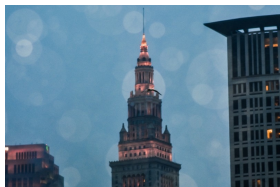
**SeBe-VQA:** [Paralympic games](#).

**SeBe-VQA retrieved Wikipedia section:**

**Article title:** Lee Valley VeloPark

**Section title:** London 2012

**Section text:** The venue was used for the [2012 Olympic and Paralympic track cycling competition](#) was held in the Velodrome with the adjoining BMX track hosting the Olympic BMX competition. Team GB dominated the track cycling competition winning seven out of a possible ten gold medals plus one silver and one bronze. The GB Paralympic track cycling team won a total of 15 medals, comprising five golds, seven silver and three bronze.



**Question:** How many floors are there in this building?

**Ground truth:** 52

**Vanilla:** [18](#)

**SeBe-VQA:** The Terminal Tower has [52](#) floors.

**SeBe-VQA retrieved Wikipedia section:**

**Article title:** Terminal Tower

**Section title:** Architecture

**Section text:** Built for \$179 million by the Van Sweringen brothers, the tower was to serve as an office building atop the city's new rail station, the Cleveland Union Terminal. Originally planned to be 14 stories, the structure was expanded to [52 floors](#) with a height of 708 feet (216m). It rests on 280-foot (85m) caissons. Designed by the firm of Graham, Anderson, Probst & White, the tower was modeled after the Beaux-Arts New York Municipal Building by McKim, Mead, and White. The Terminal Tower was completed in 1927 and opened for tenants in 1928, though the Union Terminal complex wasn't dedicated until 1930. It remained the tallest building in the world outside of New York City until the completion of the main building of Moscow State University in Moscow in 1953; it was the tallest building in North America outside of New York until the Prudential Center in Boston, was completed in 1964. In the late 1960s through 1980s, radio station WCLV-FM (Cleveland's fine arts station) maintained studios on the 15th floor. The transmitter was on the 43rd floor with transmission line running up the outside of the building to the antenna assembly attached to the flagpole atop the building. In the tiny cupola, which is technically the 52nd floor, there was equipment for the antenna heaters and a small telephone. Radio station WHK placed antennas on the building to increase the range of its 1420kHz signal. In the 1980s, developers sought approval to make their proposed BP Building taller than the Terminal Tower, but city officials forced them to scale it down. The Terminal Tower remained the tallest building in Ohio until the 1991 completion of Society Center, now Key Tower.



**Question:** Who currently lives at this building?

**Ground truth:** Denise Anne Carey

**Vanilla:** This building is currently vacant and owned by the state of Rhode Island. It is currently **not lived in by anyone**.

**SeBe-VQA:** [Denise Anne Carey](#), an architect from New York, currently lives at Seaview Terrace.

**SeBe-VQA retrieved Wikipedia section:**

**Article title:** Seaview Terrace

**Section title:** Carey Mansion

**Section text:** In 1974, Millicent and Martin T. Carey of New York, purchased the mansion. Being one of the largest of Newport's mansions, the Careys were faced with large costs for upkeep. The main house and gatehouse were leased to Salve Regina University, which renamed it the Carey Mansion. The former stables were also leased as a dormitory for Salve Regina, which renamed it "Seaview". The mansion's Drawing Room, used by the university for performances and practice, was renamed Cecilia Hall, for the patron saint of music (Saint Cecilia). During the 1980s Seaview Terrace housed the American syndicate of the America's Cup. On August 31, 2009, Salve Regina University terminated the lease with the Carey family. The Careys' daughter, [Denise Anne Carey, an architect from New York, currently lives there](#). The Syfy network featured the mansion in the first season, second episode of its paranormal reality show Stranded on March 6, 2013. Seaview Terrace was featured on Travel Channel's Ghost Nation, on October 31, 2020. Seaview Terrace is privately owned and is not open for tours or tourist visits.



**Question:** What is the wingspan of this bird in cm?

**Ground truth:** 16–21 cm (6.3–8.3 in)

**Vanilla:** It is **not possible to determine** the wingspan of the bird from the image provided.

**SeBe-VQA:** [16–21 cm](#).

**SeBe-VQA retrieved Wikipedia section:**

**Article title:** Black-capped chickadee

**Section title:** Description

**Section Text:** The black-capped chickadee has a black cap and "bib" with white sides to the face. Its underparts are white with rusty brown on the flanks. Its back is gray and the tail is normally slate gray. This bird has a short dark beak of 8–9.5 mm (0.31–0.37 in), short, rounded wings 63.5–67.5 mm (2.50–2.66 in), a tarsus of 16–17 mm (0.63–0.67 in), and a long tail at 58–63 mm (2.3–2.5 in). Its total body length is 12–15 cm (4.7–5.9 in), [wingspan is 16–21 cm \(6.3–8.3 in\)](#), and body mass is 9–14 g (0.32–0.49 oz). Sexes look alike, but males are slightly larger and longer than females. Although range can generally be used to separate them, the black-capped chickadee is very similar in appearance to the Carolina chickadee. The black-capped is larger on average, but this cannot be used reliably for identification. The most obvious difference between the two is in the wing feathers. In the black-capped chickadee, the wing feathers have white edges that are larger and more conspicuous than those of the Carolina chickadee. The latter is often mistaken for black-capped chickadees with feather dystrophy, which sometimes affects the appearance of the primary feathers making them look slimmer, a phenomenon caused by illnesses such as fatty liver disease in malnourished birds. Overall, the Carolina appears slightly paler colored, whereas the flanks of the black-capped can appear to have a trace of off-yellow or rusty coloration. Also, the black-capped generally has a more "ragged" looking black bib, whereas the bib of the Carolina has a more smooth-edged look. These subtle features are often even more vague in populations around where the black-capped and Carolina overlap in range (possibly the result of hybrids) and the two cannot always be distinguished as two species. The two species were formerly thought to be easily distinguished by call, but they often learn each other's vocalizations where their ranges overlap (their point of overlap is a narrow band that runs along the east-central United States, with the black-capped chickadee to the north). A bird located near the zone of overlap that sings both songs, or sings "odd-sounding" songs, cannot be positively identified solely by voice in the field.

Figure 4: Additional examples when the vanilla model fails to answer the question, while our proposed method is able to select the corresponding knowledge for answering.

## B Failure Cases

Image				
Text	Which river flows near this place?	In which city is this monastery located?	What are the seeds of this plant used to make?	Where is this plant native to?
Ground Truth	Vardar River	Zhengding	jewellery	Mexico
Vanilla	The image <b>does not provide any information</b> about the river near this place.	This is the Foguang Temple in <b>Wutaishan</b> , Shanxi Province, China.	The seeds of this plant are used to make <b>mesquite flour</b> .	This plant is native to <b>Mexico</b> .
SeBe-VQA-Text	The river that flows near Ulpiana is the <b>Graqanica</b> .	<b>Beijing</b>	The seeds of this plant are used to make <b>mesquite flour</b> .	<b>India and Southern Asia</b> .
SeBe-VQA-Text-Top5	<b>Graqanica</b> river.	<b>Suzhou</b> .	Mesquite flour for making traditional <b>horno bread</b> .	The plant is native to the warmer and moister parts of <b>North America</b> .
Top5 Retrieved Article, Section	<ol style="list-style-type: none"> <li><a href="#">Ulpiana, Geography</a></li> <li>Roman heritage in Kosovo, Municipium Dardanorum</li> <li><a href="#">Scupi, Scupi</a></li> <li>Roman heritage in Kosovo, Vendenis</li> <li>Koviljkin grad, Koviljkin grad</li> </ol>	<ol style="list-style-type: none"> <li>Bailin Temple (Beijing), Bailin Temple (Beijing)</li> <li>Shuxiang Temple, Shuxiang Temple</li> <li>Wenshu Temple (Chengdu), Wenshu Temple (Chengdu)</li> <li><a href="#">Hanshan Temple, Hanshan Temple</a></li> <li><a href="#">Longxing Temple, Longxing Temple</a></li> </ol>	<ol style="list-style-type: none"> <li><a href="#">Parkinsonia microphylla, Uses</a></li> <li>Senegalia greggii, Ethnobotany</li> <li>Saguaro, Ethnobotany</li> <li><a href="#">Prosopis glandulosa, Indigenous peoples</a></li> <li>Pachycereus pecten-aboriginum, Food</li> </ol>	<ol style="list-style-type: none"> <li>Thunbergia fragrans, Distribution</li> <li><a href="#">Tithonia rotundifolia, Tithonia rotundifolia</a></li> <li>Gerbera, Distribution</li> <li>Thunbergia alata, Thunbergia alata</li> <li>Tagetes patula, Tagetes patula</li> </ol>

Figure 5: Failure cases of our model. **Red** represent the section selected by Gemini from our top-5 retrieved sections. **Green** represent the correct section corresponds to the query. The first two examples show when SeBe-VQA-Text correctly selects the knowledge between top-2 ~ 5, but Gemini fails to identify the correct one given the query image and text. The third example shows when SeBe-VQA-Text successfully retrieves the correct knowledge for top-1, but Gemini instead re-selected another. The last example shows when none of the top-5 retrieved knowledge are correct.

## C Query Knowledge Image Comparison

Query Text	Where did this bird end up after it was transferred to the brookfield zoo?	How tall is this lighthouse?	What cultivar is considered a form of this tree?	In kilometers, how far away from havana is this building?	What was an educational film made at this village called working in rural new england?
Query Image					
Wiki Image					

Figure 6: Image comparison between the query and the Wikipedia. For the left examples, the Wikipedia images match well with the query image. For the right examples, it's hard to align between the query and the Wikipedia image from a human perspective. Therefore, we think this misalignment provided noise to the SeBe-VQA-TextImg model.

## D Wikipedia Example

**Thrush nightingale** article title

From Wikipedia, the free encyclopedia

The **thrush nightingale** (*Luscinia luscinia*), also known as the **sprosser**, is a small **passerine** bird that was formerly classed as a member of the **thrush** family Turdidae, but is now more generally considered to be an **Old World flycatcher**, Muscicapidae.<sup>[c]</sup> It, and similar small European species, are often called **chats**.

It is a **migratory** insectivorous species breeding in forests in Europe and the **Palaearctic** and overwintering in Africa. The distribution is more northerly than the very closely related **common nightingale**, *Luscinia megarhynchos*, which it closely resembles in appearance. It nests near the ground in dense undergrowth.

The thrush nightingale is similar in size to the **European robin**. It is plain greyish-brown above and white and greyish-brown below. Its greyer tones, giving a cloudy appearance to the underside, and lack of the **common nightingale**'s obvious **rufous** tail side patches are the clearest plumage differences from that species. Sexes are similar: It has a similar but more powerful song than that of the nightingale.

■  
■  
■

**Behaviour** section title

The thrush nightingale feeds chiefly on the ground taking **earthworms**, **spiders** and the adults, **larvae** and **pupae** of insects such as **beetles**, small **moths**, **ants** and **flies**. In the autumn, the berries of **currants** (*Ribes* spp.) and **elders** (*Sambucus* spp.) are also eaten.<sup>[8]</sup> Before crossing the **Sahara** on its migration, thrush nightingales build up their fat reserves. It has been found experimentally that **magnetic cues** may stimulate the birds to do this. A simulation of the magnetic field found in northern Egypt encouraged birds preparing to migrate from Sweden to further build up their body fat.<sup>[9]</sup>

The thrush nightingale breeds in damp forests, nesting on the ground, often in the middle of a bed of **singing nettles** (*Urtica dioica*). The nest rests on a platform of dead leaves and is composed of dead grass stalks, bents (*Agrostis* spp.), sedges and stems, lined with finer material. It is built by the female which lays four or five (occasionally six) eggs. These are a milky-blue colour, usually plain but sometimes with a slight speckling of rusty-brown and measure an average of 21.7 by 16.2 millimetres (0.85 in × 0.64 in). The hen **incubates** the eggs which hatch in about thirteen days. The young are fed by both parents and **fledge** when about eleven days old, but are not fully independent for another twelve days or so.<sup>[9]</sup>

**Status** section title positive section


**BirdLife International** estimates that there are between 11 and 20 million thrush nightingales in Europe and that, as Europe forms somewhere between 50% and 74% of the bird's global range, the total world population may be between 15 and 41 million individuals. In Europe, the population seems to be increasing slightly. The bird is considered to be of **Least Concern** by the International Union for Conservation of Nature **IUCN**.<sup>[10]</sup>

57 languages

Read Edit View history Tools

**1st image in the main section**

**Thrush nightingale**



At Uglich, Russia

**Conservation status**


Extinct  Threatened  Least Concern

EW CR EN VU NT LC

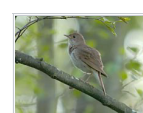
**Least Concern** (IUCN 3.1)<sup>[1]</sup>

**Scientific classification**

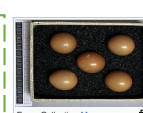
Domain: Eukaryota



The sonograms of *Luscinia luscinia* and *Luscinia megarhynchos* singing help to distinguish these two species by voice definitely.



In Poland



Eggs, Collection Museum Wiesbaden, Germany

**Thrush nightingale**  
**Status**  
 BirdLife International estimates that there are between 11 and 20 million thrush nightingales in Europe and that, as Europe forms somewhere between 50% and 74% of the bird's global range, the total world population may be between 15 and 41 million individuals. In Europe, the population seems to be increasing slightly. The bird is considered to be of Least Concern by the International Union for Conservation of Nature IUCN.

Figure 7: A Wikipedia article example corresponding to the question in Fig. 1. Each article contains multiple sections, and only the relevant section is used as positive data. The article title and section title are prepended to the section during training.