Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering

Omar Adjali and **Olivier Ferret**
Université Paris-Saclay
CEA, List
F-91120, Palaiseau, France

Sahar Ghannay
Université Paris-Saclay
CNRS, LISN
Orsay, France

Hervé Le Borgne
Université Paris-Saclay
CEA, List
F-91120, Palaiseau, France

Abstract

The Knowledge-Aware Visual Question Answering about Entity task aims to disambiguate entities using textual and visual information, as well as knowledge. It usually relies on two independent steps, information retrieval then reading comprehension, that do not benefit each other. Retrieval Augmented Generation (RAG) offers a solution by using generated answers as feedback for retrieval training. RAG usually relies solely on pseudo-relevant passages retrieved from external knowledge bases which can lead to ineffective answer generation. In this work, we propose a multi-level information RAG approach that enhances answer generation through entity retrieval and query expansion. We formulate a joint-training RAG loss such that answer generation is conditioned on both entity and passage retrievals. We show through experiments new state-of-the-art performance on the VIQuAE KB-VQA benchmark and demonstrate that our approach can help retrieve more actual relevant knowledge to generate accurate answers.

1 Introduction

Knowledge-based Visual Question Answering (KB-VQA) has recently gained significant attention as a challenging yet promising task for evaluating systems' capabilities to deeply understand both visual and textual information in order to answer multimodal queries. In contrast to standard VQA tasks, KB-VQA further extends the challenge by requiring access to external knowledge sources (e.g., knowledge bases, knowledge graphs), as image and text queries do not explicitly carry the required knowledge for accurate answers. Recently, Large Language Models (LLMs) e.g., Palm (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and Large Multimodal Models (LMMs) e.g., GIT (Wang et al., 2022), BLIP-2 (Li et al., 2023), PaLi (Chen et al., 2023a), and BEiT-3 (Wang et al., 2023) achieved impressive results in vision-language tasks thanks to their billions of parameters trained on large-scale corpora, endowing them with huge memorization capabilities for solving any downstream tasks (Izacard et al., 2023). However, they still struggle to successfully address the challenges related to knowledge-intensive tasks (e.g. KB-VQA), facing problems such as hallucinations and outdated parametric knowledge (Kandpal et al., 2023; Gao et al., 2023). Indeed, solving KB-VQA requires addressing two tasks, namely Retrieval and Reading Comprehension: relevant information (passages, documents) is first retrieved from a Knowledge Base (KB) and then answers are extracted/generated. Decoupling these two steps during training accounts for the main drawback as the retriever lacks feedback from the reader. Retrieval training is often performed using pseudo-labels indicating if the retrieved items are relevant for answering queries. For example, a common approach for passage retrieval consists in training dual encoders using pseudo-relevant passages where relevance is assessed based on string matching between passages and answers. Obviously, such heuristics yield noisy supervision retrieval signals that do not necessarily help answer questions. Retrieval Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a new paradigm where generated answers serve as training signals for the retriever, circumventing the need for exhaustive memorization within model parameters and addressing the limitation of training the retriever and the reader in a decoupled way. In this work, we rely on the RAG framework to tackle the Knowledge-Aware Visual Question Answering about Entities (KVQAE) task (Shah et al., 2019; Lerner et al., 2022), which involves answering visual questions related to named entities specified in a knowledge base. Applying RAG to solve the KVQAE task involves retrieving relevant passages used as query context during answer generation. However, the single-step retrieval em-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16499–16513

ployed in standard RAG may produce irrelevant passages misleading the generation process. In such a context, we found that determining which entity a question is related to plays a crucial role in generating accurate answers. We thus propose a multi-level information RAG (MiRAG) approach that performs retrieval at different levels of granularity as follows: 1) Given an initial question, we perform entity retrieval obtaining coarser-grain information which might enhance the subsequent finer-grain retrieval. 2) Using entity information, we perform query entity expansion on the initial question before applying passage retrieval offering more relevant context that improves the generator's ability to provide accurate and informative answers. To avoid propagating irrelevant information during the iterative retrieval process, we formulate a RAG loss to train all the components in a seamless end-to-end way, such that answer generation is conditioned on both entity and passage retrieval.

2 Related Work

In recent years, retrieval-augmented models have made substantial progress, driven by the necessity to exploit world knowledge for tasks like visual question answering (VQA) and open-domain question answering. Beyond the implicit knowledge captured in pretrained language models (PLMs) parameters, the retrieval augmented (RA) learning paradigm has shown how PLMs can greatly benefit from external knowledge augmentations, thus alleviating problems such as *hallucinations* (Lewis et al., 2020; Izacard et al., 2023). RA learning principle consists in optimizing through backpropagation standard language modeling objectives combined with a retrieval step that learns providing useful knowledge during training, inference, and fine-tuning on downstream tasks. Regarding the Open-domain Question Answering (Open-QA) task, early work such as ORCA (Lee et al., 2019) proposed an end-to-end approach that jointly trains retriever and reader models using Inverse Cloze Task (ICT). Similarly, REALM (Guu et al., 2020) is pretrained on masked language modeling (MLM) including a differentiable retrieval step that augments an encoder with external knowledge. These examples have paved the way for more recent retrieval augmented generation approaches that employ generative models for reading comprehension (answer generation) instead of span-based answer extractors. For example, Izacard and Grave (2021) proposed an iterative process where the cross-attention scores computed for answer generation are distilled as feedback to train the retriever. Singh et al. (2021) introduced a differentiable training method for RA-Open-QA allowing information fusion from multiple retrieved documents during answer generation. Paranjape et al. (2022) addressed equally pseudo-relevant passages using an additional guide retriever during training that leverages the target outputs to retrieve the actual relevant passages. In contrast, Lee et al. (2022) proposed a single model trained end-to-end which integrates retriever and reranker models as internal passage-wise attention mechanisms within a transformer architecture. While neural retrievers require large training datasets to achieve good performance, Ram et al. (2022) explored self-supervised pretraining by recurring spans across passages in a document to create pseudo examples. In contrast, Hofstätter et al. (2022) proposed to filter noisy training examples using confidence scores on the relevance labels, by measuring the connection between query-answer examples and items in a knowledge base. Furthermore, Izacard et al. (2023) explored with ATLAS, a pre-trained RA language model, several objectives (e.g. likelihood distillation, attention distillation) on QA and Fact-checking downstream tasks in the few-shot setting. Other approaches dynamically perform retrieval showing that the retrieved items are not systematically beneficial for the generator. Self-RAG (Asai et al., 2024) for example generates *reflection* special tokens used during inference for retrieval decision making while FLARE (Jiang et al., 2023) employs next sentence generation token probability as a threshold to trigger the retrieval process.

Multimodal RAG. Lin and Byrne (2022) tackled the KB-VQA task, which requires retrieval of external knowledge to answer, and proposed RA-VQA, a joint training scheme that integrates differentiable multimodal Passage Retrieval with answer generation, enabling end-to-end training. Alternatively, REVEAL (Hu et al., 2023b), extended the RAG architecture with a memory module to encode multimodal world knowledge, which helps retrieve relevant entries to answer queries. Additionally, they proposed an attentive fusion layer enabling seamless training of the retriever and the generator. Finally, Lin et al. (2024) extended RA-VQA (Lin and Byrne, 2022) with Fine-grained Late-interaction Multi-modal Retrieval (FLMR), an efficient retrieval method that increases the in-

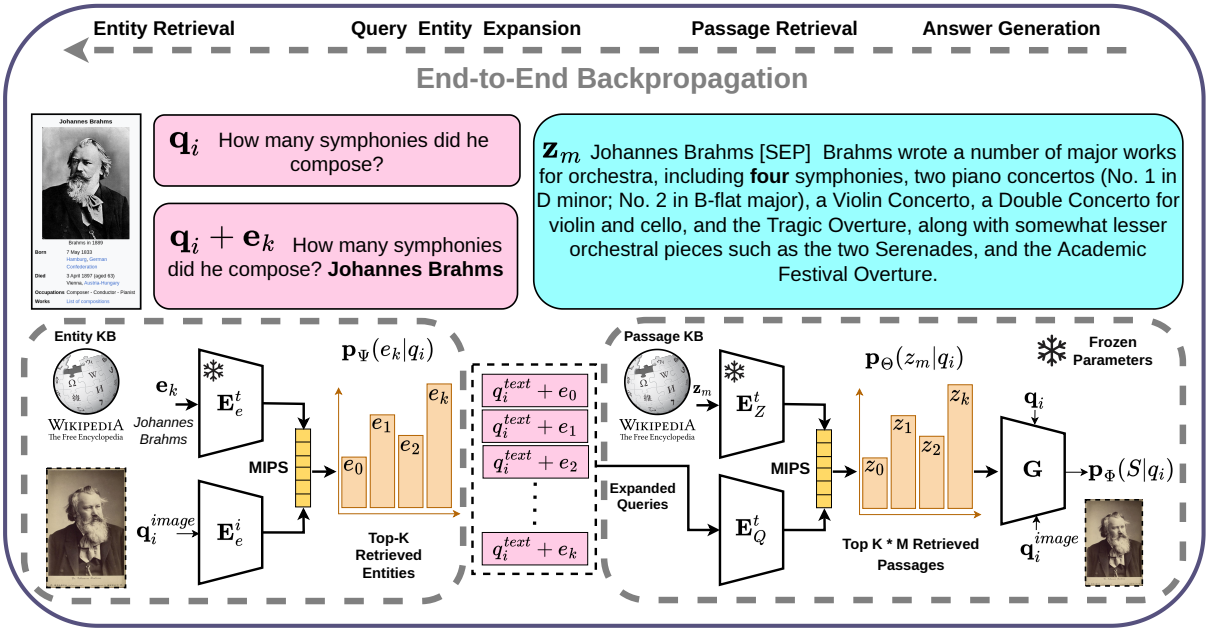


Figure 1: MiRAG approach overview.

teraction between query and document representations using multi-dimensional embeddings. These approaches perform knowledge retrieval at the passage level only. We explore in this work how leveraging entity-level information through retrieval and query expansion may enhance RAG training for KB-VQA tasks.

3 Method

Our MiRAG proposed approach relies on the RAG framework to solve the KVQAE task, which involves end-to-end training of the retriever and reader components. Figure 1 illustrates the main steps of our approach. The core idea is to perform retrieval at different levels of granularity (entity and passage) given a multimodal query with both textual and visual inputs. This multi-level retrieval strategy progressively refines the query before generating answers using a generator model. Retrieval is initially conducted at a coarse "entity" level to identify a set of candidate entities relevant to the query. Subsequently, these retrieved entities are prepended to the query before conducting the retrieval again at a finer "passage" level. Adding entities to queries provides a better understanding of the context of the query, making it easier to select more relevant passages and providing additional context for the answer generator. Formally, we aim to learn the probability $p(a|q, z, e)$ of generating an answer a conditioned on the query q, a retrieved passage z, and its corresponding retrieved entity e.

3.1 Entity retriever

Knowledge-intensive tasks such as KVQAE rely on an external unstructured KB (e.g. Wikipedia) from which relevant information is retrieved. Given a general-domain knowledge base represented as a set of multimodal documents $\mathcal{D}$, we define the corresponding set of entities $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$, where each entity e_i is related to a document $d_i \in \mathcal{D}$. The objective is to retrieve the most relevant entity given an input visual query. Although KVQAE questions are multimodal, we leverage only their visual content as mapping question texts to entities remains inefficient. We therefore perform cross-modal entity retrieval using a pretrained CLIP-based dual encoder (Radford et al., 2021), which excels at aligning image-text pairs. More formally, we encode the image of a question q and the title description of an entity into dense vector representations[1] using respectively the visual encoder $\mathbf{E}_\mathbf{e}^\mathbf{i}(\cdot)$ and the textual encoder $\mathbf{E}_\mathbf{e}^\mathbf{t}(\cdot)$. We then build a Faiss index (Johnson et al., 2019) that maps each entity $e_i \in \mathcal{E}$ to a dense vector, allowing us to perform fast exact maximum inner product search (MIPS). Formally, we compute the inner product between the dense vectors of q and all $e_i \in \mathcal{E}$ as follows :

$$s(q, e_i) = \mathbf{E}_\mathbf{e}^\mathbf{i}(q)^T \mathbf{E}_\mathbf{e}^\mathbf{t}(e_i) \quad (1)$$

During training, we compute the retrieval joint probability over the retrieved entities using the *soft-*

[1]We take the 512-d $L2$-normalized hidden state vector of the start special token.

max function allowing the selection of the K most relevant entities:

$$p_\Psi(e_k|q) = \frac{\exp(s(q, e_k))}{\sum_{j=1}^{K} \exp(s(q, e_j))} \quad (2)$$

where Ψ denotes the model parameters of the CLIP visual encoder $\mathbf{E}^{\mathbf{i}}_{\mathbf{e}}(\cdot)$. In contrast, the CLIP textual encoder parameters are frozen to avoid recomputing embeddings of all entities in $\mathcal{E}$ at each training step. As demonstrated in (Lewis et al., 2020), combining a pre-computed index and a trainable query encoder is sufficient for RAG learning.

Achieving decent cross-modal entity retrieval performance requires strengthening alignments between visual and textual entity descriptions. Thus, we pretrain the aforementioned CLIP encoders on the cross-modal retrieval task (see Appendix C for details).

3.2 Query entity expansion

Before retrieving relevant passages useful for answer generation, we propose to expand the textual query with the top-K candidate entities retrieved in the previous stage. Query expansion techniques enrich original queries with additional context that helps improve retrieval performance, especially in settings where only pseudo-relevant supervision signals are available. Specifically, during training, given a question text q, a set of candidate entities $\{e_k\}_{i=1}^{K}$, we expand q by appending its text with the corresponding title of each entity e_k yielding K expanded queries ready for the subsequent retrieval.

3.3 Passage retriever

Starting from the set of documents $\mathcal{D}$, we split each document into passages of 100 words and each passage is headed with its corresponding document title. We perform passage-level retrieval using a BERT-base (Devlin et al., 2019) dual encoder similar to the Dense Passage Retriever (DPR) (Karpukhin et al., 2020). Questions and passages are separately encoded into dense vector representations[2] using respectively a question encoder $\mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(\cdot)$ with trainable parameters Θ and a passage encoder $\mathbf{E}^{\mathbf{t}}_{\mathbf{Z}}(\cdot)$ with fixed parameters. Given a set of expanded questions $\{q_k\}_{i=1}^{K}$ and a collection of passages $\{z_i\}_{i=1}^{N_z}$, the objective is to retrieve passages that are relevant for question answering. We perform maximum inner product search for each expanded question q_k after encoding and indexing all the passages as follows:

$$s(q_k, z_i) = \mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(q_k)^T \mathbf{E}^{\mathbf{t}}_{\mathbf{Z}}(z_i) \quad (3)$$

Similar to the entity retrieval stage, we select the *M* highest relevance scores for each q_k, and compute the joint probability distribution of the $K \times M$ retrieved passages:

$$p_\Theta(z_i|q_k) = \frac{\exp(s(q_k, z_i))}{\sum_{j=1}^{K\times M} \exp(s(q_k, z_j))} \quad (4)$$

3.4 Answer generator

Answer generation is performed using an encoder-decoder based model denoted $\mathbf{G}(\cdot)$ with parameters Φ. Questions are augmented with the retrieved passages using concatenation.

3.5 MiRAG joint training

During training, each original question q is expanded with K entities obtained after entity retrieval; then M passages are retrieved for each expanded question, yielding the sets of candidate entities and passages $\{e_k\}_{k=1}^{K}$, and $\{z_m\}_{m=1}^{K\times M}$. Answers are generated for each z_m and the best candidate answer $\hat{a}$ is selected according to the joint probability of entity retrieval, passage retrieval, and answer generation such that:

$$\hat{a}, \hat{e}, \hat{z} = \arg\max_{a, z_m, e_k} p_{\Phi,\Psi,\Theta}(a, e_k, z_m|q) = \arg\max_{a, z_m, e_k} \left[p_\Psi(e_k|q) \cdot p_\Theta(z_m|q, e_k) \cdot p_\Phi(a|q, z_m)\right] \quad (5)$$

$\hat{e}$ and $\hat{z}$ are respectively the best-retrieved entity and passage candidates. The answer generator is trained on the following cross-entropy loss:

$$\mathcal{L}_{\text{MiRAG}} = -\sum_{i=1}^{N_q}\sum_{k=1}^{K}\sum_{m=1}^{K\times M} \log(p_{\Psi,\Theta,\Phi}(s_i^*|q_i, e_k, z_m)) = -\sum_{i=1}^{N_q}\sum_{k=1}^{K}\sum_{m=1}^{K\times M} \log(p_\Psi(e_k|q_i) p_\Theta(z_m|q_i, e_k) p_\Phi(s_i^*|q_i, z_m)) \quad (6)$$

where N_q is the number of questions in a batch, s_i^* is the ground truth answer string, and $p_\Phi(s_i^*|q_i, z_m)$ = $\mathbf{G}(q_i, z_m)$.

Following the answer generation loss in Equation 6, our end-to-end learning objective guides the

[2]We used the 768-dimensional vector of the `[CLS]` token from the last hidden layer of each encoder.

generator in providing answers conditioned on both the retrieved entities and passages. Thus, gradients are propagated through $\mathbf{G}(\cdot)$, $\mathbf{E}^{\mathbf{i}}_{\mathbf{e}}(\cdot)$, and $\mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(\cdot)$, allowing to jointly adapt entity and passage retrievals to select the most relevant knowledge for question answering.

4 Experimental Setup

4.1 Datasets

We conduct experiments on the ViQuAE (Lerner et al., 2022) KVQAE benchmark, whose questions are specifically related to named entities (2,397 unique entities) and cover a broad range of entity types beyond named persons only, such as in (Shah et al., 2019). ViQuAE provides manually annotated challenging questions, 95.2% of which (vs. 29.2% for OK-VQA benchmark (Marino et al., 2019)) require external knowledge to answer (Chen et al., 2023b). ViQuAE also relies on an external knowledge base built from Wikipedia dumps comprising 1.5M entities that act as distractors during retrieval. Table 1 reports the ViQuAE dataset and KB statistics.

Table 1: ViQuAE dataset and KB statistics.

KB-VQA Datasets	External KB		Splits		
	#Entities	#Passages	Train	Val	Test
ViQuAE	1.49M	11.8M	1,190	1,257	1,250

4.2 Evaluation

Systems are evaluated on their ability to answer multimodal questions. For fair comparison and reproducibility purposes, we evaluate our approach using the standard metrics used for assessing the ViQuAE benchmark: F1-score and Exact Match metrics. We further evaluate our approach for passage and entity retrieval using Precision@1 (P@1), Precision@20 (P@20), and Mean Reciprocal Rank at 100 (MRR) metrics. Following (Lerner et al., 2024; Lin et al., 2024), we use pseudo-relevance to assess passages retrieval i.e., passages are considered relevant if they contain ground truth answers.

4.3 Baselines

To demonstrate the effectiveness of our proposed approach on KB-VQA, we compare its performance against published baselines in the literature and two RAG baseline settings we carefully implemented and categorized as follows. To assess the model size effect, we experimented with two pre-trained generators: T5-large (Raffel et al., 2020) with 738M parameters and BLIP2-Flan-T5-XL with 3.9B parameters. Experimental training details are given in Appendix A.

Literature baselines. Several works addressed the KB-VQA task following the two-step retrieve and read principle. Lerner et al. (2023) proposed ECA (Early Cross-Attention) and ILF (Intermediate Linear Fusion) early fusion retrieval approaches using multimodal dense representations. Lerner et al. (2024) proposed DPR$_{v+t}$, a Score-Based Fusion (SBF) approach that achieved strong KB-VQA performance, where scores from text retrieval, image retrieval, and cross-modal retrieval are fused using linear interpolation combined with span-based answer extraction. Another baseline includes the PaLM LLM (Chowdhery et al., 2023) (540B parameters) using only textual queries. We do not include in our study related work that are not directly comparable. Specifically, Lin and Byrne (2022) rely on image-to-text transformations to leverage query visual content during retrieval, which necessitates advanced preprocessing such as objects/attributes detection and OCR, while Lin et al. (2024) focus on improving retrieval using multi-embedding retrieval. Our approach is thus complementary to these latter.

RAG. Similar to (Lewis et al., 2020), we jointly train the answer generator and a DPR retriever following Equation 6. These baselines leverage only textual information during retrieval.

SBF-RAG. Since our approach requires visual and textual query information during entity and passage retrievals, we implemented a strong baseline, referred to as SBF-RAG, which leverages, during retrieval, multi-modal information for both questions and passages. During training, the answer generator is fed with passages obtained after fusing the scores from DPR, CLIP mono-modal image, and CLIP cross-modal retrievals. Before fusion, scores are normalized to zero mean and unit variance to have comparable distributions. We used the pretrained CLIP model (Radford et al., 2021) to encode the texts and images associated with questions and passages. For text retrieval, we used the same DPR model across all experiments and settings.

5 Results

Table 2 shows the performance results of our approach and baseline systems under different set-

Method	Joint Fine-tuning	M_{train}	K_{train}	Metrics	
				EM	F1
Literature Baselines					
PaLM (few-shot) (Chen et al., 2023b)	✗	✗	✗	31.5	-
ECA (Lerner et al., 2023)	✗	✗	✗	20.6	24.4
ILF (Lerner et al., 2023)	✗	✗	✗	21.3	25.4
DPR_{V+T} (Lerner et al., 2024)	✗	✗	✗	30.9	34.3
Systems *w/o* KB retrieval augmentation					
T5-large	✗	✗	✗	18.7	22.5
BLIP-2	✗	✗	✗	14.5	21.6
RAG Systems					
RAG (T5-large)	✓	5	✗	22.1	26.1
RAG (T5-large)	✓	15	✗	22.5	26.6
RAG (BLIP-2)	✓	5	✗	30.6	34.4
Score-based Multimodal Fusion RAG Systems					
SBF+RAG (T5-large)	✓	5	✗	25.4	29.9
SBF+RAG (T5-large)	✓	15	✗	25.6	30.3
SBF+RAG (BLIP-2)	✓	5	✗	31.1	37.4
Ours					
MiRAG (T5-large)	✓	5	3	29.8	34.1
MiRAG (BLIP-2)	✓	2	2	**36.6**	**41.2**

Table 2: KB-VQA performance results of our approach and baseline systems evaluated on ViQuAE test sets. At inference time, the number of retrieved passages M_{test} = 5 for T5-large models and M_{test} = 3 for BLIP2 models to fit in one GPU. For systems with entity retrieval, we select the best-retrieved entity (K_{test} = 1) for query expansion.

tings evaluated on the ViQuAE test set. Experimental settings include whether joint training is enabled i.e., answer generator and retriever are trained end-to-end. It also includes the number of training retrieved passages M_{train} and which type LLM/LMM is used for answer generation. In the MiRAG setting, it mentions the number of training retrieved entities K_{train}, so that the total number of retrieved passages is $K_{train} \times M_{train}$. Note that we systematically evaluate the different approaches in settings where the total number of retrieved passages is comparable to ensure a fair comparison. Overall, in comparable settings, we can observe that our approach consistently outperforms all baselines regardless of model size, which validates the ability of MiRAG to better augment answer generators with actual relevant information that boosts the performance of KB-VQA. In particular, MiRAG (BLIP-2) with 3.94B parameters achieves the best overall performance reaching an EM of 36.6 and an F1 of 41.2, surpassing all literature baselines, including PaLM (540B). In contrast to T5 models, BLIP-2 is a generator that encodes text-image input pairs for answering multimodal questions, thus allowing MiRAG (BLIP-2) to benefit from text-image information at both the retrieval and answer generation steps. This suggests that even such a large multimodal model benefits from entity-level knowledge not provided by visual/textual features.

5.1 RAG joint training

We observe that RAG methods consistently enhance the performance across different models and settings compared to fine-tuned generators. RAG (T5-large) with M_{train} = 5 improves EM from 18.7 to 22.1 (+3.8% gain) and F1 from 22.5 to 26.1 (+3.6% gain). Increasing the number of retrieved passages M_{train} to 15 naturally provides a slight gain of 0.5% in EM (22.5) and F1 (26.6) since the answer generator benefits from more retrieved passages with potentially useful information during training. This confirms the importance of retrieval augmentation for answer generation.

5.2 Text vs. multimodal retrieval

The SBF+RAG (T5-large) baseline with M_{train}=5 significantly improves over RAG (T5-large), achieving an EM of 25.6 (+3.5%) and an F1 of 29.9 (+4.2%) while increasing M_{train} to 15 shows only slight gains over M_{train}=5. Compared to the DPR-based RAG approach, the SBF+RAG approach processes multimodal information allowing better retrieval performance which in turn benefits answer generation. MiRAG (T5-large) with M_{train}=5 and K_{train}=3 further improves the performance, yielding an EM of 29.8 (+4.2%) and

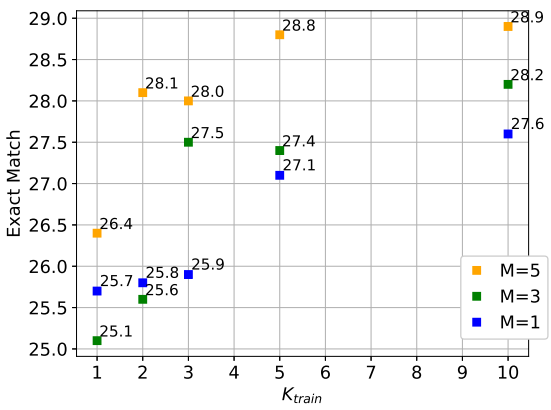


Figure 2: Exact match on ViQuAE validation set vs. number of retrieved entities K_{train} and passages M_{train} using MiRAG (T5-large). At inference time, we select the best entity (K_{test} = 1) and feed the generator with the concatenated top-5 passages (M_{test} = 5). Passage concatenation showed in experiments marginally better results than selecting the passage with the best answer probability score.

an F1 of 34.1 (+3.4%). Similarly, our approach achieves performance gains of the same order using BLIP-2 (3.9B params). In contrast to SBF+RAG approaches, which require leveraging the image-text pairs associated with both questions and passages and combining text retrieval, image retrieval, and cross-modal retrieval, our MiRAG approach leverages only the visual query to perform cross-modal entity retrieval and textual information for passage retrieval. This suggests that expanding queries with entities helps retrieve more actual relevant passages, thus validating our statement that entity-level information carries additional signals that complement visual and textual features.

5.3 Effect of increasing M_{train} and K_{train}

Figure 2 shows the impact of increasing the number of retrieved entities and passages during training. In general, more retrieved entities and passages increase the EM score up to some extent but at an additional memory and computational cost. For example, we found in our experiments that $K_{train} \times M_{train}$ = 15 is a good trade-off between the computational cost and answer generation performance, particularly seeing that increasing the number of retrieved passages yields only marginal improvement or hurts the performance. Another key observation includes that retrieving a few entities (e.g. K_{train}=3) is sufficient to boost KB-VQA performance. This is confirmed in Table 3, where given a question, its correct answer can be found in pseudo-relevant passages that are related to 4 entities on average. This highlights the importance of entity retrieval pretraining to leverage the most useful entity-level knowledge during RAG training.

Mean	Median	StdDev	Min	Max
4.00	4.00	5.45	2	60

Table 3: ViQuAE dataset statistics about the number of entities covered by the pseudo-relevant passages of a question.

Method	**MRR@100**	**P@1**	**P@20**
BM25	19.0	13.1	6.0
DPR	31.9	22.6	15.8
CLIP$_{i-i}$	15.7	12.7	6.4
CLIP$_{i-t}$	18.0	11.9	9.6
SBF-RAG	36.0	25.9	17.9
ILF	37.3	26.8	19.1
ECA	37.8	26.7	19.5
RAG (T5-large)	32.2	23.0	16.3
MiRAG-No-Grad	**38.9**	**30.0**	22.3
MiRAG (T5-large)	38.3	30.0	**22.5**
MiRAG (BLIP-2)	37.6	28.9	21.3

Table 4: Pseudo-relevance passage retrieval performance results. The best performing model settings are: RAG (T5-large) with M$_{train}$=15, MiRAG (T5-large) with M$_{train}$=3 and K$_{train}$=3, MiRAG (BLIP-2) with M$_{train}$ = 2 and K$_{train}$=2.

Further analyses about the impact of K_{train} and M_{train} during training on passage retrieval are provided in Appendix D.

5.4 Pseudo-relevance retrieval performance

Table 4 reports the pseudo-relevance passage retrieval performance, which was evaluated using Precision@20 (P@20) and Mean Reciprocal Rank at 100 (MRR@100) metrics. While only pseudo-relevant supervision signals are available for pretraining DPR and CLIP encoders, these allow assessing the effect of answer generation feedback on pseudo-relevance retrieval performance during RAG joint training. We compare retrieval systems involved in our KB-VQA experiments and baseline systems found in the literature. Those latter include two early fusion baselines, ECA (Early Cross-Attention) and ILF (Intermediate, Linear Fusion) (Lerner et al., 2023), which rely on multimodal dense representations to perform retrieval. We also report the retrieval performance of CLIP-based image and cross-modal retrieval (CLIP$i-i$

Method	Freeze $\mathbf{E}^{\mathbf{i}}_{\mathbf{e}}(\cdot)$	Freeze $\mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(\cdot)$	Entity retrieval			Passage retrieval			Answer generation	
			MRR	P@1	P@20	MRR	P@1	P@20	EM	F1
MiRAG (T5-large)	✓	✓	40.3	30.3	13.8	38.9	30.0	22.3	26.2	31.3
MiRAG (T5-large)	✓	✗	40.3	30.3	13.8	38.7	30.7	21.8	28.7	33.1
MiRAG (T5-large)	✗	✗	43.9	32.9	15.0	38.3	30.0	22.5	**29.8**	**34.1**
MiRAG (BLIP-2)	✓	✓	40.3	30.3	13.8	38.9	30.0	22.3	35.3	40.8
MiRAG (BLIP-2)	✓	✗	40.3	30.3	13.8	39.6	30.4	22.4	35.1	40.0
MiRAG (BLIP-2)	✗	✗	41.9	30.5	14.7	37.6	28.9	21.3	**36.6**	**41.2**

Table 5: Ablation study on the effect of entity and passage retrievals during MiRAG joint training by stopping parameters update of the entity and question encoders $\mathbf{E}^{\mathbf{i}}_{\mathbf{e}}(\cdot)$ and $\mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(\cdot)$. When both retrievers are frozen, the answer generator is trained using the standard cross-entropy loss of the generated answers i.e., without the entity and passage retrieval terms $p_\Psi(e_k|q_i)$ and $p_\Theta(z_m|q_i, e_k)$ of the MiRAG loss in Equation 6. For better readability, the performance of frozen components is highlighted in bold gray.

and CLIP$i-t$) used to compute the score-based multimodal fusion system (SBF). In MiRAG-No-Grad, we evaluate the effect of query entity expansion on passage retrieval performance by stopping the gradient propagation of entity and passage retrievers during training. We also report the effect of RAG training using our MiRAG approach. DPR achieves good retrieval performance (31.9 MRR@100, 22.6 P@1) slightly enhanced by RAG (T5-large) joint training (+0.3% MRR@100, +0.4%P@1). In general, leveraging visual features drastically improves the retrieval performance for all multimodal systems, indicating that the image feature complements the text feature with additional useful information. Moreover, we see that query entity expansion (MiRAG-No-Grad) shows the highest pseudo-relevant retrieval performance in MRR@100 and P@1, which indicates the potential benefits of incorporating entity information in the retrieval process. We observe that MiRAG approaches do not improve pseudo-relevance retrieval performance compared to MiRAG-No-Grad. Indeed, MiRAG's end-to-end training objective is to improve answer generation by optimizing entity and passage encoders to retrieve the most relevant knowledge items to answer a question. Thus, pseudo-relevance retrieval metrics indicate only potential improvements for answer generation, as many pseudo-relevant passages are not truly relevant and can mislead answer generation training. Similarly, while MiRAG (BLIP-2) achieves superior performance on the KB-VQA task compared to MiRAG (T5-large), we note that MiRAG (T5-large) and MiRAG (BLIP-2) achieve near-similar pseudo-relevant passage retrieval performance although T5 answer generator has 5 times fewer parameters. This clearly demonstrates the contribution of our approach in providing answer generation with more pseudo-relevant passages and at the same time, it learns to retrieve more actual relevant ones.

5.5 Joint performance analysis

We evaluate in Table 5 the effect of gradient propagation through the entity and question encoders $\mathbf{E}^{\mathbf{i}}_{\mathbf{e}}(\cdot)$ and $\mathbf{E}^{\mathbf{t}}_{\mathbf{Q}}(\cdot)$ during MiRAG training and the interaction between entity retrieval, passage retrieval, and answer generation performances. Overall, we observe that the best answer generation performance is obtained when enabling gradient propagation for both T5-large (29.8 vs. 28.7 vs. 26.2 EM) and BLIP-2 (36.6 vs. 35.1 vs. 35.3 EM), which validates the contribution of our MiRAG approach. Disabling gradient propagation for both retrievers achieves worse but strong answer generation performance. This demonstrates, on the one hand, the positive impact of query entity expansion on retrieving more relevant passages for answer generation and on the other hand, the benefit of the MiRAG joint training loss. Furthermore, when the entity encoder parameters are fixed, answer generation feedback is only provided to the passage retrieval encoder, yielding slightly better pseudo-relevance retrieval performance. In contrast, when gradient propagation is enabled all over the components, it benefits the entity retrieval performance, which in turn improves answer generation. As with passage retrieval, MiRAG (T5-large) entity retrieval performance boost is more significant than MiRAG (BLIP-2) (43.9 vs. 41.9 MRR) despite the superior performance of this latter on the target task. We assume that the raw BLIP-2 ability to generate answers allows to better guide

the retrieval components towards discarding more pseudo-relevant items in favor of passages and entities that are relevant for answering questions. Appendix F shows some qualitative examples.

6 Conclusion

We proposed in this paper Multi-level information Retrieval Augmented Generation (MiRAG), a fully integrated approach for end-to-end training of RAG systems augmented with retrieved items at different levels of granularity. MiRAG combines entity and passage retrieval using query expansion that helps provide useful knowledge for answer generation. Our experiments demonstrate that our training scheme, which conditions answer generation training on both entity and passage retrievals, increases its ability to retrieve more relevant knowledge to solve the KB-VQA task.

7 Limitations

Our approach focuses on cross-modal entity retrieval to perform query expansion. A straightforward way to improve entity retrieval is to leverage visual and textual content associated with both the questions and entities. Experiments on the benchmark proposed in (Hu et al., 2023a) will help recognize millions of entities. In the same way, performing multimodal dense passage retrieval may contribute positively to our approach as shown in (Lin et al., 2024). While we considered one prominent entity in the image during entity retrieval, addressing the presence of multiple irrelevant entities in query images poses additional challenges for KB-VQA systems. As shown in (Wu and Mooney, 2022), determining the critical entity in the image for answering a question may help in retrieving relevant documents and generating more accurate answers. We conducted experiments on the ViQuAE benchmark, which is relevant for KB-VQA tasks that cover a large number of entity types. Experimenting with benchmarks more limited in terms of entity types might also yield valuable insights. Finally, this work is limited to integrating entity-level information in the RAG framework. Considering entity type knowledge for MiRAG might be a good start for future investigations.

Acknowledgment

This work was supported by the ANR-19-CE23-0028 MEERQAT project and the ANR-23-PEIA-0008 SHARP project, supported by France 2030. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013719 made by GENCI. It also relied on the use of the FactoryIA cluster, financially supported by the Ile-de-France Regional Council.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.

Elias Bassani and Luca Romelli. 2022. ranx.fuse: A Python Library for Metasearch. In *CIKM*, pages 4808–4812. ACM.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023a. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023b. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. ACL.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented

language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. Multi-Task Retrieval-Augmented Text Generation with Relevance Sampling. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

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 IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

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

Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. ACL.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, Vancouver, Canada. ACL.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

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. ACL.

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher Manning, and Kyoung-Gu Woo. 2022. You Only Need One Model for Open-domain Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3047–3060, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2023. Multimodal Inverse Cloze Task for Knowledge-based Visual Question Answering. In *European Conference on Information Retrieval*, pages 569–587. Springer.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal Retrieval for Knowledge-based Visual Question Answering. In *46th European Conference on Information Retrieval (ECIR 2024): Advances in Information Retrieval*, pages 421–438, Glasgow, Scotland. Springer Nature Switzerland.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, Madrid, Spain. ACM.

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

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.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2024. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2022. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, Long Beach, CA, USA. MIT Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, Virtual Event. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to Retrieve Passages without Supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2687–2700, Seattle, United States. ACL.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 8876–8884, Honolulu, Hawaii, USA. AAAI Press.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

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*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. ACL.

Jialin Wu and Raymond Mooney. 2022. Entity-Focused Dense Passage Retrieval for Outside-Knowledge Visual Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8061–8072.

A Training Details

We initialized our entity retriever with a pre-trained ViT-B/32-base CLIP dual encoder (Radford et al., 2021) and continued pre-training following section 3.1 on cross-modal entity retrieval using 90% of the entities in the ViQuAE KB and leaving the rest for validation. We pre-trained the CLIP encoder for 10 epochs using a learning rate of 2e-6. Using gradient checkpointing, we adopted a large batch size of 1,000 to improve contrastive learning (Equation 7). The best model for entity retrieval was selected according to the in-batch mean reciprocal rank on the validation set. For passage retrieval, we pre-trained the BERT-base DPR (Karpukhin et al., 2020) dual encoder on the TriviaQA (Joshi et al., 2017) dataset to ensure state-of-the-art retrieval performance and then

we fine-tuned it on each KB-VQA dataset using pseudo-relevant contrastive learning allowing a warm start initialization for RAG end-to-end joint training. For answer generation, we experimented with two pre-trained generators: T5-large (Raffel et al., 2020) and BLIP2-Flan-T5-XL, which we finetuned on ViQUAE before retrieval augmentation. Following (Lin et al., 2024), we adopted LoRA (Hu et al., 2022), a low-rank matrix decomposition approach that greatly reduces the number of LLMs trainable parameters to train BLIP2-Flan-T5-XL on a single GPU. Our implementation relies on the huggingface-PEFT package with the following configuration: $r = 8$, $lora\_alpha = 32$, $lora\_dropout = 0.1$. The batch size was set to 2 due to GPU memory limit. We trained end-to-end all the components using a linear-decay schedule that reduces the learning rate from 2e-5 to 0 after 20 epochs for the answer generator, and a fixed learning rate of 1e-6 for the question and image entity encoders. At inference time, we decoded using beam-search with 5 beams. Results on test sets were reported after averaging the scores of 2 different runs initialized with random seeds. Model checkpoints were selected based on validation performance. All experiments needed only one Nvidia A100 (80G) GPU. Our implementation is based on PyTorch (Paszke et al., 2017). Pretrained models were obtained using Hugginface and Transformers (Wolf et al., 2020). Retrieval performance evaluation was done using the ranx library (Bassani and Romelli, 2022). Faiss (Johnson et al., 2019) allowed MIPS search and vector indexing. Our code will be released at: `https://github.com/OA256864/MiRAG`.

B Computational Cost

The computational cost in terms of GPU time on a single Nvidia A100 (80G) is the following: it required 40 mins per epoch (training + inference on the validation split) for training MiRAG (BLIP-2) with frozen retrievers. In contrast, joint training required 2h30 per epoch for MiRAG (T5-Large) with 934M trainable parameters and $K_{train} \times M_{train}$ = 15 against 1h50 for MiRAG (BLIP-2) with 201M trainable parameters using LORA and $K_{train} \times M_{train}$ = 4.

C Entity Retriever Pretraining

We pretrained the CLIP dual encoders on the cross-modal retrieval task using the title-image pairs in the KB, which enables projecting the title and image representations of an entity closer in the embedding space. Specifically, we followed a standard in-batch negative sampling strategy and optimized the following contrastive loss:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(q,e_i)/\tau)}{\sum_{j=1}^{N}\exp(s(q,e_j)/\tau)} \quad (7)$$

where τ is the temperature hyperparameter controlling the level of penalties on hard negative pairs. $\mathcal{L}_{CL}$ maximizes the embedding similarity between matching pairs and minimizes the similarity otherwise.

D Additional Analyses about K_{train} and M_{train}

Figure 3 shows the impact of increasing the number of retrieved entities and passages during training on passage retrieval. We note that varying the retrieved entities and passages marginally affects pseudo-relevance retrieval metrics. MiRAG end-to-end training continuously selects and discards retrieved passages according to answer generation following Equation 5.

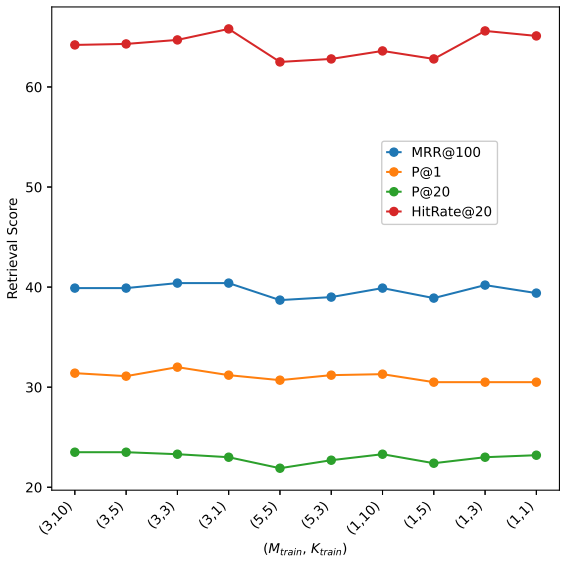


Figure 3: Pseudo-relevance passage retrieval performance vs. number of retrieved entities K_{train} and passages M_{train} using MiRAG (T5-large).

E Baseline Settings

Table 6 gives the settings of all our baselines.

Method	Joint Fine-tuning	Entity Retriever	Passage Retriever	Reader	# Params
Literature Baselines					
PaLM (few-shot) (Chen et al., 2023b)	✗	✗	✗	Text	540B
ECA (Lerner et al., 2023)	✗	✗	Multimodal	Text	432M
ILF (Lerner et al., 2023)	✗	✗	Multimodal	Text	432M
DPR_{V+T} (Lerner et al., 2024)	✗	✗	Multimodal	Text	481M
Systems *w/o* KB retrieval augmentation					
T5-large	✗	✗	✗	Text	738M
BLIP-2	✗	✗	✗	Multimodal	3.94B
RAG Systems					
RAG (T5-large)	✓	✗	Text	Text	955M
RAG (BLIP-2)	✓	✗	Text	Multimodal	4.16B
Score-based Multimodal fusion RAG Systems					
SBF+RAG (T5-large)	✓	✗	Multimodal	Text	1.16B
SBF+RAG (BLIP-2)	✓	✗	Multimodal	Multimodal	4.37B
Ours					
MiRAG (T5-large)	✓	Cross-modal	Text	Text	1.1B
MiRAG (BLIP-2)	✓	Cross-modal	Text	Multimodal	4.3B

Table 6: System settings.

F Qualitative Examples

Figures 4 and 5 show some qualitative examples.

Question

He was killed when competing in which Grand Prix?

Answer / Reference entity

SAN MARINO / Formula One

SBF-RAG (BLIP-2) — **Generated Answer**: Monaco Grand Prix

Retrieved Passage

Sauber Motorsport [SEP] The season took a turn for the worse after a 4th place by Wendlinger following the tragic deaths of Ayrton Senna and Roland Ratzenberger at the San Marino Grand Prix. Just two weeks later, Wendlinger was seriously injured after crashing in practice for the Monaco Grand Prix; losing control of his car under braking for the Nouvelle Chicane. He suffered serious head injuries, which left him in a coma for weeks, and he was sidelined for the rest of the season.

MiRAG (BLIP-2) + frozen entity retriever — **Generated Answer:** The Belgian Grand Prix — **Predicted Entity**: Ayrton Senna

Retrieved Passage

Sauber Motorsport [SEP] The season took a turn for the worse after a 4th place by Wendlinger following the tragic deaths of Ayrton Senna and Roland Ratzenberger at the San Marino Grand Prix. Just two weeks later, Wendlinger was seriously injured after crashing in practice for the Monaco Grand Prix; losing control of his car under braking for the Nouvelle Chicane. He suffered serious head injuries, which left him in a coma for weeks, and he was sidelined for the rest of the season.

MiRAG (BLIP-2) — **Generated Answer**: San Marino — **Predicted Entity**: Ayrton Senna

Retrieved Passage

Sauber Motorsport [SEP] The season took a turn for the worse after a 4th place by Wendlinger following the tragic deaths of Ayrton Senna and Roland Ratzenberger at the San Marino Grand Prix. Just two weeks later, Wendlinger was seriously injured after crashing in practice for the Monaco Grand Prix; losing control of his car under braking for the Nouvelle Chicane. He suffered serious head injuries, which left him in a coma for weeks, and he was sidelined for the rest of the season.

Prediction rationales

All systems successfully retrieved the passage containing the answer, however only MiRAG (BLIP-2) generated the correct answer

Question

Who was the first US President to speak in this palace?

Answer / Reference entity

Barack Obama / Palace of Westminster

SBF-RAG (BLIP-2) — **Generated Answer**: Franklin Pierce

Retrieved Passage

George Peabody [SEP] In 1851, when the US Congress refused to support the American section at the Great Exhibition at the Crystal Palace, Peabody advanced £3000 to improve the exhibit and uphold the reputation of the United States. In 1854, he offended many of his American guests at a Fourth of July dinner when he chose to toast Queen Victoria before US President Franklin Pierce; Pierce's future successor, James Buchanan, then Ambassador to London, left in a huff

MiRAG (BLIP-2) + frozen entity retriever — **Generated Answer:** Barack Obama — **Predicted Entity**: Palace of Westminster

Retrieved Passage

Barack Obama [SEP] On May 25, 2011, Obama became the first President of the United States to address both houses of the UK Parliament in Westminster Hall, London. This was only the fifth occurrence since the start of the 20th century of a head of state being extended this invitation, following Charles de Gaulle in 1960, Nelson Mandela in 1996, Queen Elizabeth II in 2002 and Pope Benedict XVI in 2010.

MiRAG (BLIP-2) — **Generated Answer**: Barack Obama — **Predicted Entity**: Palace of Westminster

Retrieved Passage

Barack Obama [SEP] On May 25, 2011, Obama became the first President of the United States to address both houses of the UK Parliament in Westminster Hall, London. This was only the fifth occurrence since the start of the 20th century of a head of state being extended this invitation, following Charles de Gaulle in 1960, Nelson Mandela in 1996, Queen Elizabeth II in 2002 and Pope Benedict XVI in 2010.

Prediction rationales

Predicting the correct entity helps MiRAG (BLIP-2) retrieving the actual relevant passage

Figure 4: Qualitative examples outputs comparison of: SBF-RAG (BLIP-2), MiRAG (BLIP-2) with frozen entity retriever and MiRAG (BLIP-2). These show the benefit of leveraging entity-level information to retrieve actual relevant passages.

Question

this man died in which month of 1959?

Answer / Reference entity

February / Buddy Holly

SBF-RAG (BLIP-2) — **Generated Answer**: September

Retrieved Passage

Johnny O'Keefe [SEP] On 14 February 1975 (St Valentine's Day) at the Masonic Hall, Waverley, O'Keefe married for the second time to Maureen Joan Maricic, a 29-year-old fashion consultant. They opened a boutique, J. O'K Creations, at Paddington in 1978. O'Keefe's last public appearance was on Seven Network's "Sounds" program, taped on 30 September 1978, six days prior to his death. Johnny O'Keefe died from a drug overdose on 6 October 1978.

MiRAG (BLIP-2) + frozen entity retriever — **Generated Answer:** March — **Predicted Entity**: Paul O'Grady

Retrieved Passage

George Formby [SEP] On Valentine's Day 1961, seven weeks after Beryl's death, Formby and Howson announced their engagement. Eight days later he suffered a further heart attack which was so severe that he was given the last rites of the Catholic Church on his arrival at hospital. He was revived and, from his hospital bed, he and Howson planned their wedding, which was due to take place in May. He was still there when, on 6 March, he had a further heart attack and died at the age of 56.

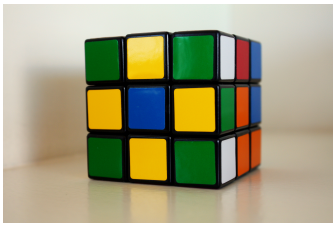
MiRAG (BLIP-2) — **Generated Answer**: February — **Predicted Entity**: Buddy Holly

Retrieved Passage

Buddy Holly [SEP] Shortly after 12:55 am on February 3, 1959, Holly, Valens, Richardson, and Peterson were killed instantly when the aircraft crashed into a frozen cornfield five miles northwest of Mason City, Iowa, airport shortly after takeoff. The three musicians, who were ejected from the fuselage upon impact, suffered severe head and chest injuries. Holly's funeral was held on February 7, 1959, at the Tabernacle Baptist Church in Lubbock.

Prediction rationales

Training jointly the entity retriever allowed MiRAG (BLIP-2) to correctly predict the target entity, thus providing the actual relevant passage for answering.



Question

How many squares does this toy have on each side?

Answer / Reference entity

Nine / Rubik's Cube

SBF-RAG (BLIP-2) — **Generated Answer**: 4

Retrieved Passage

Rubik's Revenge [SEP] However, there also exist Cubes with alternative colour arrangements (yellow opposite green, blue opposite white and red opposite orange). The Eastsheen version has purple (opposite red) instead of orange. There are 8 corners, 24 edges and 24 centres.

MiRAG (BLIP-2) + frozen entity retriever — **Generated Answer:** 6 — **Predicted Entity**: Rubik's Cube

Retrieved Passage

Rubik's Cube [SEP] It is widely considered to be the world's best-selling toy. On the original classic Rubik's Cube, each of the six faces was covered by nine stickers, each of one of six solid colours: white, red, blue, orange, green, and yellow. The current version of the cube has been updated to coloured plastic panels instead, which prevents peeling and fading.

MiRAG (BLIP-2) — **Generated Answer**: 9 — **Predicted Entity**: Rubik's Cube

Retrieved Passage

Rubik's Cube [SEP] It is widely considered to be the world's best-selling toy. On the original classic Rubik's Cube, each of the six faces was covered by nine stickers, each of one of six solid colours: white, red, blue, orange, green, and yellow. The current version of the cube has been updated to coloured plastic panels instead, which prevents peeling and fading.

Prediction rationales

Use case where MiRAG (BLIP-2)-Frozen does not answer correctly even though it has retrieved the correct entity and the actual relevant passage

Figure 5: Qualitative examples outputs comparison of: SBF-RAG (BLIP-2), MiRAG (BLIP-2) with frozen entity retriever and MiRAG (BLIP-2). These show the benefit of leveraging entity-level information to retrieve actual relevant passages.