Entity-Focused Dense Passage Retrieval for Outside-Knowledge Visual Question Answering

Jialin Wu
Department of Computer Science
The University of Texas at Austin
jialinwu@utexas.edu

Raymond J. Mooney
Department of Computer Science
The University of Texas at Austin
mooney@cs.utexas.edu

Abstract

Most Outside-Knowledge Visual Question Answering (OK-VQA) systems employ a two-stage framework that first retrieves external knowledge given the visual question and then predicts the answer based on the retrieved content. However, the retrieved knowledge is often inadequate. Retrievals are frequently too general and fail to cover specific knowledge needed to answer the question. Also, the naturally available supervision (whether the passage contains the correct answer) is weak and does not guarantee question relevancy. To address these issues, we propose an **En**tity-**Fo**cused **Re**trieval (EnFoRe) model that provides stronger supervision during training and recognizes question-relevant entities to help retrieve more specific knowledge. Experiments show that our EnFoRe model achieves superior retrieval performance on OK-VQA, the currently largest outside-knowledge VQA dataset. We also combine the retrieved knowledge with state-of-the-art VQA models, and achieve a new state-of-the-art performance on OK-VQA.

1 Introduction

Passage retrieval under a multi-modal setting is a critical prerequisite for applications such as outside-knowledge visual question answering (OK-VQA) (Marino et al., 2019), which requires effectively utilizing knowledge external to the image. Recently, dense passage retrievers with deep semantic representations powered by large transformer models have shown superior performance to traditional sparse retrievers such as BM25 (Robertson and Zaragoza, 2009) and TF-IDF under both textual (Karpukhin et al., 2020; Chen et al., 2021; Lewis et al., 2022) and multi-modal settings (Luo et al., 2021; Qu et al., 2021; Gui et al., 2021).

In this work, we investigate two main drawbacks of recent dense retrievers (Karpukhin et al., 2020; Chen et al., 2021; Lewis et al., 2022; Luo et al., 2021; Qu et al., 2021; Gui et al., 2021), which are typically trained to produce similar representations for input queries and passages containing ground-truth answers.

Figure 1: Top: Examples of critical entities upon which retrieval models should focus; Bottom: Example of improved passage retrieval using critical entities.

First, as most retrieval models encode the query and passages as a whole, they fail to explicitly discover entities critical to answering the question (Chen et al., 2021). This frequently leads to retrieving overly-general knowledge lacking a specific focus. Ideally, a retrieval model should identify the critical entities for the query and then retrieve question-relevant knowledge specifically about them. For example, as shown in the top half of Figure 1, retrieval models should realize that the entities "turkey" and "teddy bear" are critical.

Second, on the supervision side, the positive signals are often passages containing the right answers with top sparse-retrieval scores such as BM 25 (Robertson and Zaragoza, 2009) and TF-IDF. However, this criterion is inadequate to guarantee question relevancy, since good positive passages should reveal facts that actually support the correct answer using the critical entities depicted in the image. For example, as shown in the bottom of Figure

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8061 - 8072
December 7-11, 2022

1, both passages mention the correct answer "vegetable" but only the second one which focuses on the critical entity "bell pepper" is question-relevant.

In order to address these shortcomings, we propose an **En**tity-**Fo**cused **Re**trieval (EnFoRe) model that improves the quality of the positive passages for stronger supervision. EnFoRe automatically identifies critical entities for the question and then retrieves knowledge focused on them. We focus on entities that improve a sparse retriever's performance if emphasized during retrieval as critical entities. We use the top passages containing *both* critical entities and the correct answer as positive supervision. Then, our EnFoRe model learns two scores to indicate (1) the importance of each entity given the question and the image and (2) a score that measured how well each entity fits the context of each candidate passage.

We evaluate EnFoRe on OK-VQA (Marino et al., 2019), currently the largest knowledge-based VQA dataset. Our approach achieves state-of-the-art (SOTA) knowledge retrieval results, indicating the effectiveness of explicitly recognizing key entities during retrieval. We also combine this retreived knowledge with SOTA OK-VQA models and achieve a new SOTA OK-VQA performance. Our code is available at https://github.com/jialinwu17/EnFoRe.git.

2 Background and Related Work

2.1 OK-VQA

Visual Question Answering (VQA) has witnessed remarkable progress over the past few years, in terms of both the scope of the questions (Antol et al., 2015; Hudson and Manning, 2019; Wang et al., 2018; Gurari et al., 2018; Singh et al., 2019), and the sophistication of the model design (Antol et al., 2015; Lu et al., 2016; Anderson et al., 2018; Kim et al., 2018, 2020; Wu et al., 2019; Wu and Mooney, 2019; Jiang et al., 2018; Lu et al., 2019; Nguyen et al., 2021). There is a recent trend towards outside knowledge visual question answering (OK-VQA) (Marino et al., 2019), where open domain external knowledge outside the image is necessary. Most OK-VQA models (Marino et al., 2019; Gardères et al., 2020; Zhu et al., 2020; Li et al., 2020; Narasimhan et al., 2018; Marino et al., 2021; Wu et al., 2022; Gui et al., 2021; Gao et al., 2022) incorporate a retriever-reader framework that first retrieves textual knowledge relevant to the question and image and then "reads" this text to predicts the answer. As an online free encyclopedia, Wikipedia is often used as the knowledge source for OK-VQA. While most previous works focused more on the answer prediction stage, the performance is still lacking because of the imperfect quality of the retrieved knowledge. This work focuses on knowledge retrieval and aims at retrieving question-relevant knowledge that focuses explicitly on the critical entities for the visual question.

2.2 Passage Retrieval

Sparse Retrieval: Before the recent proliferation of transformer-based dense passage retrieval models (Karpukhin et al., 2020), previous work mainly explored sparse retrievers, such as TF-IDF and BM25 (Robertson and Zaragoza, 2009), that measure the similarity between the search query and candidate passage using weighted term matching. These sparse retrievers require no training signals on the relevancy of the passage and show solid baseline performances. However, exact term matching prevents them from capturing synonyms and paraphrases and understanding the semantic meanings of the query and the passages.

Dense Retrieval: To better represent semantics, dense retrievers (Karpukhin et al., 2020; Chen et al., 2021; Lewis et al., 2022; Lee et al., 2021) extract deep representations for the query and the candidate passages using large pretrained transformer models. Most dense retrievers are trained using a contrastive objective that encourages the representation of the query to be more similar to the relevant passages than other irrelevant passages. During training, the passage with a high sparse retrieval score containing the answer is often regarded as a positive sample for the question-answering task. However, these positive passages may not fit the question's context and only serve as very weak supervision. Moreover, the query and passages are often encoded as single vectors. Therefore most dense retrievers fail to explicitly discover and utilize critical entities for the question (Chen et al., 2021). This often leads to overly general knowledge without a specific focus.

2.3 Dense Passage Retrieval for VQA

Motivated by the trend toward dense retrievers, previous work has also applied them to OK-VQA. Qu et al. (2021); Gao et al. (2022) utilize Wikipedia as a knowledge source. Luo et al. (2021) crawl Google search results on the training set as a knowledge source. However, the weak training signals

for passage retrieval become more problematic for VQA as the visual context of the question makes it more complex. Therefore, a "positive passage" becomes less likely to fit the visual context and actually provide suitable supervision. In order to better incorporate visual content, Gui et al. (2021) adopt an image-based knowledge retriever that employs the CLIP model (Radford et al., 2021) pretrained on large-scale multi-modal pairs as the backbone. However, question relevancy is not considered, so the retriever has to retrieve knowledge on every aspect of the image for different possible questions.

This work proposes an **En**tity-**Fo**cused **Re**trieval (EnFoRe) model that recognizes key entities for the visual question and retrieves question-relevant knowledge specifically focused on them. Our approach also benefits from stronger passage-retrieval supervision with the help of those key entities.

2.4 Phrase-Based Dense Passage Retrieval

The most relevant work to ours is phrase-based dense passage retrieval. Chen et al. (2021) employ a separate lexical model that is trained to mimic the performance of a sparse retriever that is better at matching phrases. Lee et al. (2021) propose DensePhrase model that extracts each possible phrase feature in the passage and only uses the most relevant phrase to measure the similarity between the query and passage. However, the training signals still come from exactly matching ground truth answers, and the phrases are parsed from the candidate passage, limiting the scope of the search. In contrast, our approach collects entities from many aspects of the question and image, including object recognition, attribute detection, OCR, brands, captioning, etc., building a rich unified intermediate representation.

3 Entity Set Construction

Our EnFoRe model is empowered by a comprehensive set of extracted entities. Entities are not limited to phrases from the question and passages as in (Lee et al., 2021). We collect entities from the sources below. Most entity extraction steps are independent and can execute in parallel, except for answering sub-questions, which first requires parsing the questions. Parallelizing these steps can significantly reduce run time.

3.1 Question-Based Entities

Entities from Questions: First, the noun phrases in questions usually reveal critical entities. Following Wu et al. (2022), we parse the question using a constituency parser (Gardner et al., 2018) and extract noun phrases at the leaves of the parse tree. Then, we link each phrase to the image and extract the referred object with its attributes. We use a pretrained ViLBERT model (Lu et al., 2020) as the object linker.

Entities from Sub-Questions: OK-VQA often requires systems to solve visual reference problems as well as comprehend relevant outside knowledge. Therefore, we employ a general VQA model to find answers to the visual aspects of the question. In particular, we collect a set of sub-questions by appending each noun phrase in the parse tree to the common question phrases "What is..." and "How is..." When the confidence for an answer from a pretrained VilBERT model (Lu et al., 2020) exceeds 0.5, it is added to the entity set. For the example in Fig. 2, the noun phrases "plush toy" and "president" generate the sub-questions: "What is plush toy?", "How is plush toy?", "What is president?", "How is president?". The answer confidence for "teddy bear" exceeds 0.5 for the first question, so we include it in the entity set.

Entities from Answer Candidates: Standard state-of-the-art VQA models are surprisingly effective at generating a small set of promising answer candidates for OK-VQA (Wu et al., 2022, 2020). Therefore, we finetune a ViLBERT model (Lu et al., 2019) on the OK-VQA data set and extract the top 5 answer candidates and add them to entity set.

3.2 Image-Based Entities

Question-based entities are high precision and narrow down the search space for knowledge retrievers. To complement this, we also collect image-based entities to help achieve higher recall.

Entities from Azure tagging: Following Yang et al. (2022), we use Azure OCR and brand tagging to annotate the detected objects in the images using a Mask R-CNN detector (He et al., 2017).

Entities from Wikidata: As suggested by Gui et al. (2021), common image and object tags can be generic with a limited vocabulary, leading to noise or irrelevant knowledge. Therefore, we also leverage recent advanced visual-semantic matching approaches, i.e. CLIP (Radford et al., 2021), to extract image-relevant entities from Wikidata. In

particular, the entities with their descriptions in Wikidata and sliding windows of the images are used as inputs. Then, at most 18 entities with top maximum CLIP scores over these sliding windows are preserved. We follow the released code for KAT (Gui et al., 2021) and resize the image such that the size of the shorter edge is 384. The sliding window size is set to 256 with a stride of 128.

Entities from Captions: Captions provide a natural source of salient objects in the image, and do not suffer from the limited vocabulary of object detectors (Wu et al., 2018). Similar to extracting entities from the question, we parse captions and extract noun phrases from the parse tree. During training, we use the human captions provided by the COCO dataset to provide richer entities, and during testing, we use generated captions from the OFA captioning model (Wang et al., 2022).

3.3 Oracle Critical Entity Detection

Given the comprehensive set of entities $\mathcal{E}$ covering different aspects of the question and image, we introduce an approach to automatically find critical entities and passages containing them. Then, those entities and passages are used during training to provide more substantial supervision. The intuition is that a good passage that fits the visual question's context should mention *both* the key entities and the correct answer. Also, emphasizing critical entities should improve retrieval performance.

Given a question q, we use BM25[1] (Robertson and Zaragoza, 2009) as the sparse retriever to retrieve an initial set of passages $\mathcal{P}_{init} = \{p_1, ..., p_K\}$. We calculate a baseline score $\texttt{SRR}_{init}$ for these K passages using summed reciprocal ranking (SRR) as shown in Eq. 1.

$$\texttt{SRR}(\mathcal{P}) = \sum_{i=1}^{K} \frac{\mathbb{1}[\texttt{ans} \in p_i]}{i} \tag{1}$$

We use summed reciprocal ranking instead of reciprocal ranking since it provides more stable scores for evaluating the set of retrieved passages and does not overweight the highest ranked document.

Then, for each entity $e \in \mathcal{E}$, we retrieve another set of passages $\mathcal{P}_e$ using an entity-emphasizing query where the entity is appended to the end of the question. Note that the BM25 retriever does not take word order into account, so simply appending entities will not lead to undesired results due to the linguistic disfluency of the query.

[1]https://github.com/castorini/pyserini.git

The final score for an entity $S(e)$ is computed as the difference between the SRR of these two sets of retrieved passages, i.e. $S(e) = \texttt{SRR}(\mathcal{P}_e) - \texttt{SRR}(\mathcal{P}_{init})$. We regard entities with $S(e)$ over a threshold θ as critical entities, i.e. $\mathcal{E}_{oracle} = \{e \in \mathcal{E} | S(e) > \theta\}$.

Qu et al. (2021) extract the top-k passages containing the correct answer from $\mathcal{P}_{init}$ to construct the positive passage set $\mathcal{P}^+_{init}$. As we have identified oracle entities, the passage that contains both the right answer and the oracle entity is more likely to fit in the context of the question. Therefore, we augmented the positive passage set to include those passages for each oracle entity, *i.e.* $\mathcal{P}^+_{\mathcal{E}} = \bigcup_{e \in \mathcal{E}_{oracle}} (\{p^+_e\})$, where p^+_e denotes the first passage that contains both the right answer and the oracle entity. On average, there are 3.4 new positive passages per question. The negative passages are the same as those in (Qu et al., 2021), and the number of training instances (positive-negative pairs) is not changed.

4 Entity-Focused Retrieval

Entity-**Fo**cused **Re**trieval (EnFoRe) automatically recognizes critical entities and retrieves question-relevant knowledge specifically focused on them. "proj" denotes a projection function that consists of an MLP layer with layer-norm as normalization.

4.1 Encoders

Query encoder: As observed by Qu et al. (2021) and Luo et al. (2021), multi-modal transformers encode questions and visual content better than uni-modal transformers, so we adopt LXMERT (Tan and Bansal, 2019) for query encoding. In particular, we project the "pooled_output" at the last layer from LXMERT as the feature vector $f_q \in R^d$ given the query q that contains a visual question Q and the set of detected objects $\mathcal{V}$ in the image as shown in Eq. 2. See the LXMERT paper for further details.

$$f_q = \texttt{proj}(\texttt{LXMERT}(Q, \mathcal{V})) \tag{2}$$

Passage encoder: Following Qu et al. (2021), we use BERT (Devlin et al., 2019) as the passage encoder and project the "[CLS]" representation to compute the vector features for each passage p.

$$f_p = \texttt{proj}(\texttt{BERT}(p)) \tag{3}$$

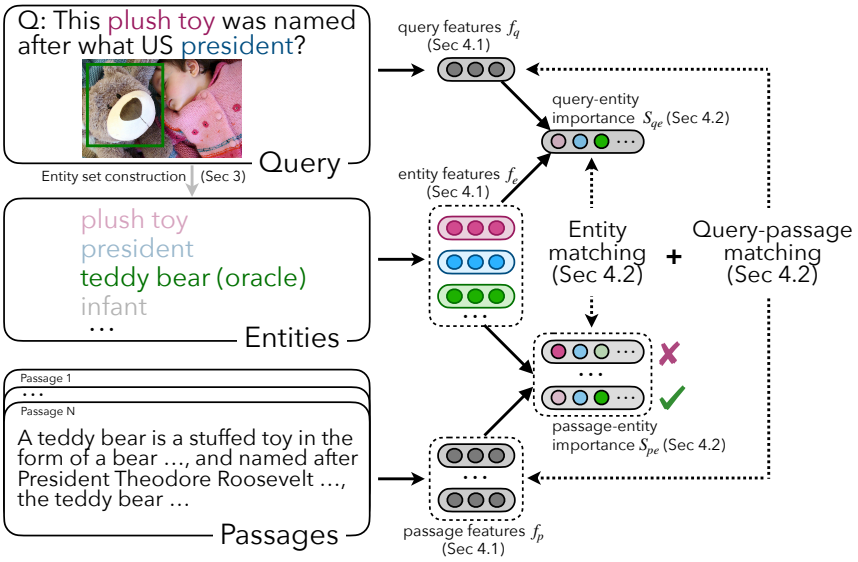

Figure 2: EnFoRe model overview. We first extract a set of entities from the query consisting of a question and an image (Sec. 3). Then, the EnFoRe model computes the features for the query, the entities, and the passages (Sec. 4.1). Query features and passage features, together with entity features, are used to compute a query-entity score and a passage-entity score to indicate the importance of the entities given the query and the passages, respectively (Sec. 4.2). These two importance scores are combined to produce an entity-matching score, and the features of the query and the passages are used to predict a query-passage matching score.

Entity encoder: In order to provide query context for each entity, we append the question and a generated image caption (Wang et al., 2022) after each entity. The input to the Entity encoder is "`[CLS] entity [SEP] question [SEP] caption`". Similar to the passage encoder, we use BERT (Devlin et al., 2019) as the entity encoder and project the "`[CLS]`" representation to compute the features for each entity.

$$f_e = \texttt{proj}(\texttt{BERT}(e)) \tag{4}$$

4.2 Retrieval Scores

EnFoRe aims to retrieve question-relevant knowledge that focuses on critical entities. Therefore, the similarity metric consists of two parts: a question relevancy term and an entity focus term.

Modeling question relevancy: We model the question relevancy term S_{qp} as the inner-product of the query and passage features, i.e. $S_{qp}(q,p) = f_q^T f_p$. During inference, as the query and passage features are decomposable, maximum inner product search (MIPS) can be applied to efficiently retrieve top passages for the query.

Modeling entity focus: The entity focus term consists of two parts, where query features are used to identify critical entities from the set of entities in Sec. 3, and passage features are used to determine whether it contains these key entities. For each entity, we compute the query-entity score $S_{qe}(q,e)$ as the inner-product of the projected query and entity feature, i.e. $S_{qe}(q,e) = \texttt{proj}(f_q)^T \texttt{proj}(f_e)$, and we compute the passage-entity score as $S_{pe}(p,e) = \texttt{proj}(f_p)^T \texttt{proj}(f_e)$. Then, we combine all of the entities and compute the entity-focused score S_{qpe} per Eq. 5:

$$S_{qpe}(q,p,\mathcal{E}) = \frac{\sum_{e\in\mathcal{E}} \sigma(S_{qe}(q,e)) \times S_{pe}(p,e)}{\sum_{e\in\mathcal{E}} \sigma(S_{qe}(q,e))} \tag{5}$$

where σ denotes the sigmoid function. Another way to interpret Eq. 5 is to treat it as modeling the conditional distribution $\Pr(p \mid q)$ and consider the entities as hidden variables.

The final score $S(q,p)$ for the query q and passage p linearly combines both terms, i.e. $S(q,p) = S_{qp}(q,p) + \lambda S_{qpe}(q,p,\mathcal{E})$, where the weight λ controls the balance between the these two terms.

4.3 Training

We train our EnFoRe model with a set of training instances consisting of a query containing the visual question with an image, a positive passage, a retrieved negative passage, and the set of entities. We present more details on constructing the training data in Sec. 6.1. We adopt the "R-Neg+IB-All" setting introduced by Qu et al. (2021) that regards the retrieved negatives, along with all

Methods	Val		Test	
	MRR@5	P@5	MRR@5	P@5
BM25-Obj	0.3772	0.2667	0.3686	0.2541
BM25-Cap	0.4727	0.3483	0.4622	0.3367
BM25 w. entities	0.3620	0.2558	0.3732	0.2620
BM25 w. oracle entities	0.6591	0.4548	0.6401	0.4345
DPR-LXMERT (Qu et al., 2021)	0.4704	0.3364	0.4526	0.3329
EnFoRe-LXMERT	**0.4881**	**0.3488**	**0.4800**	**0.3444**
EnFoRe-LXMERT w. oracle entities	0.4898	0.3533	0.4853	0.3451

Table 1: MRR and precision retreival results on OK-VQA. The first four rows present sparse retrieval results and the others are dense retrieval results.

other in-batch passages, as negative samplings. Following previous work (Karpukhin et al., 2020), we use cross-entropy loss to maximize the relevancy score $S_{qp}(q, p)$ and the entity focusing score $S_{qpe}(q, p, \mathcal{E})$ of the positive passage given the negatives identified above. In addition, we regard the oracle entities, defined in Sec. 3.3, as positive entities and others as negative entities. We use binary cross-entropy loss to supervise the importance score $S_{qe}(q, e)$. We use AdamW (Loshchilov and Hutter, 2018) with a learning rate of 1e-5 to train the EnFoRe model for 8 epochs where 10% of the iterations are used to warm up the model linearly. The batch size is set to 6 per GPU, and we use 4 GPUs (Tesla V100) for each experiment. The training process takes about 45 hours for each model. We save the parameters every 5000 steps and present the best results (MRR@5) on the validation set. The hidden states size is set to 768 following Qu et al. (2021) for fair comparison. The threshold θ for recognizing critical entities is set to 0.8. As our model consists of a BERT encoder and a LXMERT encoder, resulting in 430M parameters in total.

Inference: As the question relevancy term is decomposable, we again adopt MIPS to retrieve the top-80 passages. Then, we evaluate the entity focus term for each passage and use the combined score $S(q, p)$ to rerank the retrieved passages.

5 Reader

We employ the current state-of-the-art KAT model (Gui et al., 2021) as our VQA reader. KAT is a generation-based reader that learns to generate the answer given the retrieved knowledge. It adopts an FiD (Izacard and Grave, 2021) architecture to incorporate both implicit knowledge, generated by a frozen GPT-3 model, and explicit knowledge. For implicit GPT-3 knowledge, the input format is "`question:ques?candidate:cand. evidence:expl.`", where the `ques`, `cand` and `expl.` denotes the question, answer and its explanation generated by the GPT-3 model (Brown et al., 2020). For the explicit knowledge, the input format is "`question:ques? entity:ent. description:desc.`", where `ent`, `desc` denote the retrieved entity and its description. See (Gui et al., 2021) for further details.

We change the original explicit knowledge to the knowledge retrieved by our EnFoRe model. As the retrieved passage contains multiple sentences, and usually not all are relevant, we select the most relevant sentence for each passage. Specifically, following Wu et al. (2022), we convert the question and the candidate answers to a set of statements. Then, we decontextualize each sentence for each passage and compute the BertScore (Zhang* et al., 2020) between the decontextualized sentences and each statement. The sentence with the highest BertScore across these statements is extracted for each passage. The input format is "`question:ques?entity:ents. description:desc.`", where the `ents`, `desc` denote the top-10 entities judged by the query-entity importance score $S_{qe}(q, e)$ and the extracted sentence.

Following Gui et al. (2021), we perform experiments for two KAT settings: (1) "KAT-base + EnFoRe" setting is a single model that employs T5-base (Raffel et al., 2020) as the backbone encoder and decoder. (2) "KAT-full + EnFoRe" is an ensemble model, where each model employs T5-large as the backbone encoder and decoder. As our knowledge is question-aware, we only encode the top 10 retrieved sentences in contrast to the 40 sentences in the original KAT. We adopt the same training scheme as KAT.

	Image-based entities			Question-Image-based entities			MRR@5	P@5
	Tags	Wikidata	Cap.	Ques.	Sub-Ques.	Cand.		
DPR-LXMERT							0.4526	0.3329
EnFoRe (Backbone)							0.4632	0.3317
EnFoRe (Image)	✓	✓	✓				0.4688	0.3351
EnFoRe (Question)				✓	✓	✓	0.4750	0.3409
EnFoRe (Full)	✓	✓	✓	✓	✓	✓	0.4800	0.3444

Table 2: Ablation study on the entity sources used during re-ranking.

6 Experimental Results

6.1 Dataset

We use the OK-VQA dataset[2] (Marino et al., 2019) (version 1.1), the largest open-domain English knowledge-based VQA dataset at present, to evaluate the EnFoRe model. The questions were crowdsourced on Amazon Mechanical Turk (AMT) and are guaranteed to require external knowledge beyond the images. The dataset contains 14,031 images and 14,055 questions covering a variety of knowledge categories (i.e. 9,009 for training and 5046 for test). For knowledge retrieval, we adopt the same data configuration as Qu et al. (2021) that evenly splits the test set of the OK-VQA dataset into a validation set and a test set, and we refer to these as RetVal and RetTest, respectively.

Following Qu et al. (2021), we take the Wikipedia passage collection with 11 million passages created by previous work as our knowledge source, where each passage contains at most 384 "word pieces" with intact sentence boundaries. We extract 25 passages with the highest BM 25 scores (CombSum setting in (Qu et al., 2021)) that do not contain the correct answers as our retrieved negative samples, and the top 5 passages that contain the correct answer as retrieved positive samples for training. In addition, we also consider the most relevant passage that contains each of the oracle entities and the correct answer as positive passages. The positive and negative passages are randomly paired up to form the training instances. During evaluation, any passages containing at least one of the correct answers are considered as gold passages. For VQA models, we adopt the same model architecture and training scheme and only switch the external knowledge for the KAT models. Due to limits on computational resources, we adopt 10 retrieved sentences for the KAT model. The models are evaluated every 500 steps. We normalize the predictions by lowercasing, lemmatizing, and removing articles, punctuation and duplicated whitespace. We follow the standard evaluation metric recommended by the VQA challenge.[3] The results for "KAT-base + EnFoRe" are obtained by averaging three runs with different random seeds.

6.2 Passage Retrieval Results

We present our passage retriever results in Table 1, comparing them with the current state-of-the-art systems. We adopt MRR and Precision at a cut-off of 5 as our automatic evaluation metric. The first four rows present sparse retrieval results. The BM25 approach using our oracle entities achieves an MRR@5 of 0.6401, and a precision@5 of 0.4345 on the OK-VQA RetTest set, indicating the comprehensiveness and the potential helpfulness of the extracted entities. With the help of these entities, EnFoRe-LXMERT outperforms the previous SOTA DPR-LXMERT (with the same architecture for visual and textual embedding) by 2.74% MRR@5 and 1.15% precision@5. We perform a student's paired t-tests with a p-value of 5% to test the significance of our results. In particular, we found that the MRR and the precision gap between our EnFoRe (Full) model and (1) the DPR-LXMERT and (2) the EnFoRe (Backbone) are statistically significant.

Ablation study on entity sources: We also performed an ablation study on entity-based re-ranking shown in Table 2. The EnFoRe backbone without re-ranking achieves an MRR of 0.4632, outperforming DPR (Qu et al., 2021) by 1.06%. This indicates that using our entities during training helps the retriever build better representations. It is because (1) we add additional supervision that tells the retriever which entities are more likely to lead to the correct answers, and (2) we add additional training passages that contain both the oracle entities and the right answers. Image-based and Question-based entities help our EnFoRe model achieve MRR of 0.4688 and 0.4750, respectively.

[2]https://okvqa.allenai.org/

[3]https://github.com/GT-Vision-Lab/VQA

Method	Knowledge Resources	VQA Scores
Q-only (Marino et al., 2019)	—	14.9
BAN (Kim et al., 2018)	—	25.2
MUTAN (Ben-Younes et al., 2017)	—	26.4
Mucko (Zhu et al., 2020)	Dense Caption	29.2
ConceptBert (Gardères et al., 2020)	ConceptNet	33.7
KRISP (Marino et al., 2021)	Wikipedia + ConceptNet	38.9
MAVEx (Wu et al., 2022)	Wikipedia +ConceptNet + Google Image	39.4
RVL (Shevchenko et al., 2021)	Wikipedia + ConceptNet	39.0
VRR (Luo et al., 2021)	Google Search	39.2
PICa (Yang et al., 2022)	Frozen GPT-3	48.0
KAT-base	Frozen GPT3 + Wikidata	(50.58)
KAT-base + EnFoRe	Frozen GPT3 + Wikipedia	51.34 (52.24)
KAT-full	Frozen GPT3 + Wikidata	(54.41)
KAT-full + EnFoRe	Frozen GPT3 + Wikipedia	54.35 (55.23)

Table 3: EnFoRe knowledge boosts the current state-of-the-art approaches on OK-VQA. The middle column lists the external knowledge sources if any, used in each system. The additional result shown in parentheses is computed by an unofficial evaluation metric that takes the max over 1.0 and number of annotators agreements divided by 3.

Our full model, taking advantage of both image- and question-based entities, achieves an MRR of 0.4800, showing that these two types of entities are complementary.

We also present an ablation study on individual entity sources in Table 4. We introduce a particularly challenging "RetTest Hard" split that collects all of the examples in "RetTest" where none of the correct answers is in the entity set. Our EnFoRe model consistently achieves better retrieval performance (i.e. MRR@5 and P@5) by incorporating entities extracted from each source. On the normal RetTest set, removing entities from candidate answers yields the largest decrease in MRR@5. This is due to the fact that the candidate answers cover plenty of correct answers in the OK-VQA test split and therefore provide direct hints to the desired content. On the RetTest Hard set, image-based entities generally help improve the retrieval performance more, indicating the need for explicitly discovering critical visual clues.

6.3 Visual Question Answering Results

We present the VQA performance of incorporating our EnFoRe knowledge in the state-of-the-art KAT model in Table 3. While a plain KAT-base model, which uses GPT-3 and CLIP (Radford et al., 2021) to retrieve image-based knowledge, achieves a score of (50.58)[4], switching to our EnFoRe knowledge brings a 1.7 point improvements, achieving a score of 51.34 (52.24). Our ensemble model (KAT-full + EnFoRe) achieves a new SOTA score of 54.35 (55.23).

Qualitative results: We present sample results in Figure 3 where (a)–(d) show cases where our EnFoRe model correctly identifies the critical entities ($i.e.$ the orange, the kite, the calico cat, and the teddy bear) and retrieved question-relevant knowledge focused on them. Case (e) shows an example where the retrieved sentence misleads the reader, because the reader currently only receives the textual input, and it fails to verify whether the pizza actually has a thin crust. Case (f) shows an example where the retriever properly focuses on the critical entity "NORWOOD" but fails to understand that this is the destination for the bus.

Human evaluation: We also conducted a human evaluation on AMT of the retrieved entities and sentences to demonstrate that the knowledge retrieved by EnFoRe better supports the correct answers. We first randomly sampled 1,000 test questions that are correctly answered by both the original KAT-base model and our "KAT-base + EnFoRe" model. Next, we extracted the top-3 sentences with the highest attention score averaged over all attention heads from the last decoder layer for both models. We also extracted the top-3 visual entities. For EnFoRe, the top-3 entities with the highest attention scores in the input prompts are selected. For the original KAT model, we use the three entities from the three top retrieved sentences. Next, we show AMT workers the question, the predicted answer, the image with bounding boxes for the top entities, and the three retrieved sentences, for both systems randomly ordered. We present an example in the Appendix. Finally, workers are asked to judge which system's set of highlighted entities and sentences

[4]The additional result shown in parentheses is computed by an unofficial evaluation metric that takes the max over 1.0 and number of annotators agreements divided by 3.

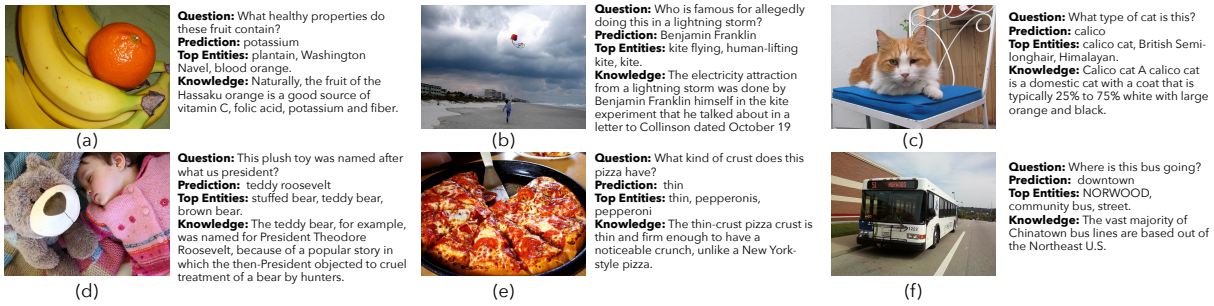

Figure 3: Qualitative results on EnFoRe; (a)-(d) present cases where EnFoRe correctly identifies the critical entities and retrieved question-relevant knowledge properly focuses on them; (e) and (f) present two failure cases.

Sources	RetTest		RetTest Hard	
	MRR@5	P@5	MRR@5	P@5
None	0.4632	0.3329	0.2525	0.1553
Image-based entities	0.4688	0.3351	**0.2709**	**0.1637**
Question-based entities	0.4750	0.3409	0.2594	0.1612
Full	**0.4800**	0.3444	0.2643	0.1632
w/o. Tags	0.4788	0.3410	0.2624	0.1606
w/o. Wikidata	0.4775	0.3429	0.2617	0.1574
w/o. Caption	0.4794	**0.3449**	0.2626	0.1611
w/o. Question	0.4786	0.3442	0.2647	0.1627
w/o. Sub-Question	0.4784	0.3411	0.2625	0.1605
w/o. Candidate	0.4693	0.3332	0.2664	0.1622

Table 4: Ablation study on entity sources.

best supports the given answer. Experimental results show that judges pick our EnFoRe knowledge 61.8% of the time, indicating a clear preference over the original KAT knowledge. Such information can be considered an explanation or rationale for the system's answer, and improved explanations can engender greater trust and acceptance from users and provide additional transparency of the system's operation.

7 Conclusion

In this work, we presented an Entity-Focused Retrieval (EnFoRe) model for retrieving knowledge for outside-knowledge visual questions. The goal is to retrieve question-relevant knowledge focused on critical entities. We first construct an entity set by parsing the question and the image. Then, EnFoRe predicts a query-entity score, predicting how likely it will lead to finding a correct answer, and a passage-entity score showing how likely the entity fits in the context of the passage. These two scores are combined to re-rank the conventional query-passage relevancy score. EnFoRe demonstrates the clear advantages of improved multi-modal knowledge retrieval and helps improve VQA performance with its improved retrieved knowledge.

8 Limitations

Our EnFoRe model is empowered by a comprehensive set of parsed entities from the question and the image. However, as shown in the failure cases in the experiment section, those entities may contain detection errors that lead to undesired results. In addition, during training, we adopt a fully automatic scheme for annotating critical entities assuming they can help a sparse retriever achieve better SRR results; however, explicit human annotation could potentially improve the quality of the critical entities identified. While we have explored collecting both question-based and image-based entities in our current approach, they are not fully adequate in that ideally it could be beneficial to include not only the relevant objects for the visual question but other kinds of descriptors that may act as useful clues for knowledge retrieval. Another limitation of the current approach is that we encode each entity separately, ignoring the relationships between entities, which could be helpful for knowledge retrieval.

9 Acknowledgements

The research was supported by the NSF-funded Institute for Foundations of Machine Learning (IFML) at UT Austin.

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 VQA. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One? *arXiv preprint arXiv:2110.06918*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. A Thousand Words Are Worth More Than a Picture: Natural Language-Centric Outside-Knowledge Visual Question Answering. In *CVPR*.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. In *EMNLP*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. KAT: A Knowledge Augmented Transformer for Vision-and-Language. *arXiv preprint arXiv:2112.08614*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz Grand Challenge: Answering Visual Questions from Blind People. In *ICCV*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *ICCV*.

Drew A Hudson and Christopher D Manning. 2019. GQA: a New Dataset for Compositional Question Answering over Real-World Images. In *CVPR*.

Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *ICLR*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the Winning Entry to the VQA Challenge 2018. *arXiv*.

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 *EMNLP*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *NeurIPS*.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Jinhyuk Lee, Alexander Wettig, and Danqi Chen. 2021. Phrase Retrieval Learns Passage Retrieval, Too. In *EMNLP*.

Patrick Lewis, Barlas Oğuz, Wenhan Xiong, Fabio Petroni, Wen-tau Yih, and Sebastian Riedel. 2022. Boosted Dense Retriever. In *NAACL*.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *ACMMM*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-Task Vision and Language Representation Learning. In *CVPR*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-attention for Visual Question Answering. In *NeurIPS*.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering. In *EMNLP*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. In *CVPR*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out-of-The-Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *NeurIPS*.

Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. 2021. Movie: Revisiting modulated convolutions for visual counting and beyond. In *ICLR*.

Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. In *SIGIR*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*.

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. *JMLR*.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Information Retrieval*.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Rread. In *CVPR*.

Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. 2018. Attention on Attention: Architectures for Visual Question Answering (VQA). *arXiv preprint arXiv:1803.07724*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-based Visual Question Answering. *TPAMI*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities through a Simple Sequence-to-Sequence Learning Framework. *arXiv preprint arXiv:2202.03052*.

Jialin Wu, Liyan Chen, and Raymond J Mooney. 2020. Improving VQA and its Explanations by Comparing Competing Explanations. *arXiv preprint arXiv:2006.15631*.

Jialin Wu, Zeyuan Hu, and Raymond J Mooney. 2018. Joint Image Captioning and Question Answering. *arXiv preprint arXiv:1805.08389*.

Jialin Wu, Zeyuan Hu, and Raymond J Mooney. 2019. Generating Question Relevant Captions to Aid Visual Question Answering. In *ACL*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-Modal Answer Validation for Knowledge-Based VQA. In *AAAI*.

Jialin Wu and Raymond J Mooney. 2019. Self-Critical Reasoning for Robust Visual Question Answering. *NeurIPS*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. In *AAAI*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In *IJCAI*.

A Appendix

A.1 Varying Weights during Re-Ranking

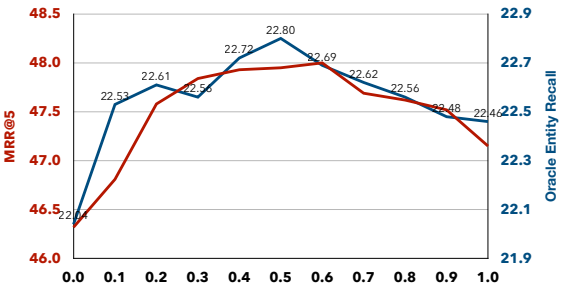

Figure 4: MRR and oracle-entity recall with different reranking weights.

In Figure 4, we present the MRR (red line), and the oracle-entity recall (blue line) at a cut-off of 5, which is defined as the fraction of oracle entities appearing in the top-5 retrieved passages over the total number of the oracle entities. Our EnFoRe model not only improves the MRR results but also retrieves more oracle entities in the top passages, making the retrieved content more relevant. Also, the EnFoRe model is robust to the re-ranking weight, yielding consistent improvements for a broad range of weights.

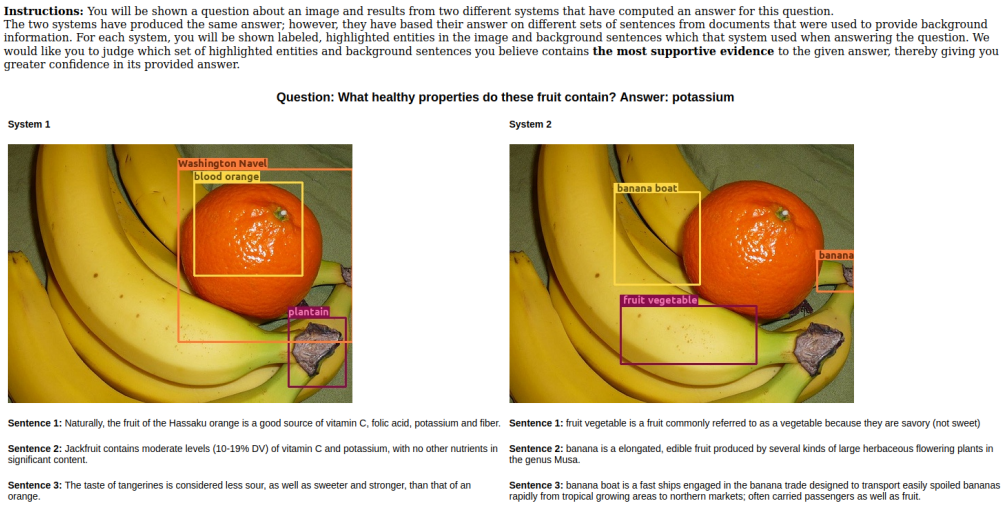

Figure 5: Sample question for the human evaluation.

A.2 Human Evaluation Details

We use Amazon Mechanical Turk (AMT) as our platform to perform human evaluation. We randomly sample 1,000 test questions that are correctly answered by both the original KAT-base model and our "KAT-base + EnFoRe" model in order to focus on evaluating the explanations for their answers rather than their correctness. In each HIT (Human Inference Task), we include four questions together with a quality control example, where the preference should be clear. We eliminate data where the quality control is not passed, but pay the workers 80 cents for finishing the HIT regardless of passing the quality control example. The average time workers spent on each HIT is 2 min and 33 sec. Figure 5 shows a sample question from a HIT.

A.3 Hyperparameters

We present details of the searching hyperparameters for the EnFoRe model in Table 5 . While most of the hyperparameters are set to the same as in (Qu et al., 2021), we tune the threshold θ for recognizing critical entities (0.6, 0.8, 1.0), batch size (2, 4, 6 per GPU), and the number of training epochs (2, 4, 6). We use a greedy approach (Singh et al., 2018) to search hyperparameters in the order of θ, batch size, and training epochs. Maximizing MRR@5 is used as the objective.

Hyperparameters	Value
BM25 Retriever k	1.1
BM25 Retriever b	0.4
CLIP	ViT-B/16
Learning rate	1e-5
Optimizer	AdamW
Batch size	6 per GPU
#Gpus	4
Retriever hidden states	768
Critical entity threshold	0.8
#Epochs	8
Learning rate in KAT	3e-5
Optimizer in KAT	AdamW
#Sentences in KAT	10
Batch size in KAT	24

Table 5: Configurations for best-performing models.