End-to-end Knowledge Retrieval with Multi-modal Queries

Man Luo[1] **Zhiyuan Fang**[2] **Tejas Gokhale**[1] **Yezhou Yang**[1] **Chitta Baral**[1]
[1] Arizona State University   [2] Amazon Alexa
{mluo26, tgokhale, yz.yang, chitta}@asu.edu, zyfang@amazon.com

Abstract

We investigate knowledge retrieval with multi-modal queries, *i.e.* queries containing information split across image and text inputs, a challenging task that differs from previous work on cross-modal retrieval. We curate a new dataset called ReMuQ [1] for benchmarking progress on this task. ReMuQ requires a system to retrieve knowledge from a large corpus by integrating contents from both text and image queries. We introduce a retriever model "ReViz" that can directly process input text and images to retrieve relevant knowledge in an end-to-end fashion without being dependent on intermediate modules such as object detectors or caption generators. We introduce a new pretraining task that is effective for learning knowledge retrieval with multimodal queries and also improves performance on downstream tasks. We demonstrate superior performance in retrieval on two datasets (ReMuQ and OK-VQA) under zeroshot settings as well as further improvements when finetuned on these datasets.

1 Introduction

Humans recall, retrieve, and communicate information using many indirect hints and cues. For instance, if we want to explain the concept of a "leopard" but have forgotten the name, we can relate the concept to a picture of a tiger and say "it is an animal that looks like this, but has spots instead of stripes". Similarly, when children learn to draw a new shape like an *oval*, teachers often prompt them by showing a circle, but saying "make the circle stretched-out". This method of learning new concepts from visual aids and language descriptions is a common way of reinforcing existing knowledge and allowing learners to explore and retrieve new concepts (Kinder, 1942).

We propose a task for vision-language models to retrieve knowledge with multi-modal queries,

i.e. queries in which hints about the information to be retrieved are split across image and text inputs. Figure 1 contains an example of this task, where the image shows the Empire State Building in New York City. If we retrieve knowledge using only the image, is it likely that the retrieved information (K1) will be related to the Empire State Building. However, K1 is insufficient to answer the question. On the other hand, if we retrieve knowledge using only the question, then the information retrieved (K2) is likely to be related to the tallest building in all cities (and not restricted to New York City). K2 by itself is also insufficient to answer the question. This example shows that the combined query containing both image and text (question) is necessary for retrieving relevant knowledge (K3).

We introduce a new benchmark and dataset called ReMuQ (**Re**trieval with **Mu**ltimodal **Q**ueries) to train and evaluate models to retrieve the answer from a corpus given multimodal (vision + language) queries. To create multimodal queries, we start with the WebQA (Chang et al., 2022) dataset as a source – WebQA contains images annotated with questions and answers. We select questions from WebQA where the answer includes both an image and text. We then remove any image information from text and combine the image and the augmented text to form a new multimodal query. We also construct a large retrieval corpus consisting of answer options of all questions as the source of knowledge for this task.

This task requires integrating the contents from both modalities and retrieve knowledge – in this paper we denote such a system as a "VL-Retriever". Existing VL-Retrievers (Qu et al., 2021; Luo et al., 2021; Gao et al., 2022) typically follow a two-step process to retrieve knowledge: (1) converting the image into captions or keywords, appending them to the text query, and (2) using a text-retriever system to retrieve the knowledge. However, this approach can result in a loss of important information

---

Question: Is this the tallest building in the city?

External Knowledge

K1: The Empire State Building is a 102-story Art Deco skyscraper in Midtown Manhattan, New York City

K2: The 828 metre (2,717 ft) tall Burj Khalifa in Dubai has been the tallest building since 2010. The Burj Khalifa has been classified as megatall.

K3: The tallest building in New York is One World Trade Center which rise 1,776 feet (541 m).

Figure 1: An illustration of the multimodal retrieval task from the ReMuQ dataset. The image shows the Empire State Building and the question asks if it is the tallest building in "the city". Neither the image nor the question explicitly mentions that "the city" is New York. The challenge therefore is to use the cues in the image and question to retrieve relevant information and answer the question. In this illustration we show the retrieved knowledge using only the image (K1), only the question (K2), or both image and question (K3). Only K3 can be used to answer the question correctly.

from the image, such as context and background. Additionally, using a caption generation model trained on a particular domain does not transfer well to other domains in real-world applications.

To address these issues, we propose an end-to-end VL-Retriever that has the potential to leverage the entire image, rather than just object categories, keywords, and captions. We call this model *ReViz*, a retriever model for "**Re**ading and **Viz**ualizing" the query. As part of ReViz, we use a vision transformer-based model, ViLT (Kim et al., 2021), to directly encode the image from raw pixels with context inputs, and we employ BERT (Devlin et al., 2019) as the knowledge encoder to represent the long, free-form text as a knowledge embedding. ReViz differs from previous retrieval models in two main ways. First, it does not require an extra cross-modal translator (e.g., a captioning model) or object detector to represent the images. Second, its end-to-end design allows for the flexible retraining of each submodule of the model, which can mitigate potential issues caused by domain gaps.

Unlike neural text-retrievers (Karpukhin et al., 2020; Luo et al., 2022), the query and knowledge encoders in ReViz are of different types of modality (*i.e.* multimodal transformer and language transformer). The different semantic spaces of the query and knowledge embeddings make alignment between them difficult. To address this, we propose a novel multimodal retrieval pretraining task. To create training data, we construct triplets of (input-image, input-text, output-knowledge) from the WiT (Srinivasan et al., 2021) dataset which contains encyclopedia-type knowledge from Wikipedia. We process the data such

that the input image and text have mutually exclusive information.

Our contributions and findings are listed below.

- We introduce a new dataset *ReMuQ* to facilitate research on retrieval with multimodal queries.
- We propose an end-to-end VL-Retriever, ReViz, that directly acquires knowledge given multimodal query. ReViz is not dependent on any cross-modal translator, such as an image captioning model or an object detector.
- We pretrain ReViz on a novel multimodal retrieval pretraining task, VL-ICT. We observe that with the proposed pretraining on the WiT dataset, our VL-Retriever is a powerful zero-shot multimodal retriever that surpasses existing single-modal knowledge retrieval methods.

2 Related Work

Cross-Modal Retrieval aims to find information from a different modality than the query; for instance retrieving images from text (text-to-image), text from images (image-to-text) (Young et al., 2014; Lin et al., 2014), text-to-video and video-to-text (Rohrbach et al., 2015; Xu et al., 2016; Zhou et al., 2018). In contrast, we consider retrieval of knowledge for queries comprised of both modalities (i.e. image and text) together.

Knowledge-based Question Answering. Retrievers are important for finding relevant knowledge to aid knowledge-based question-answering models for tasks such as FVQA (Wang et al., 2017) (commonsense knowledge), Text-KVQA (Singh et al., 2019) which requires knowledge of the text
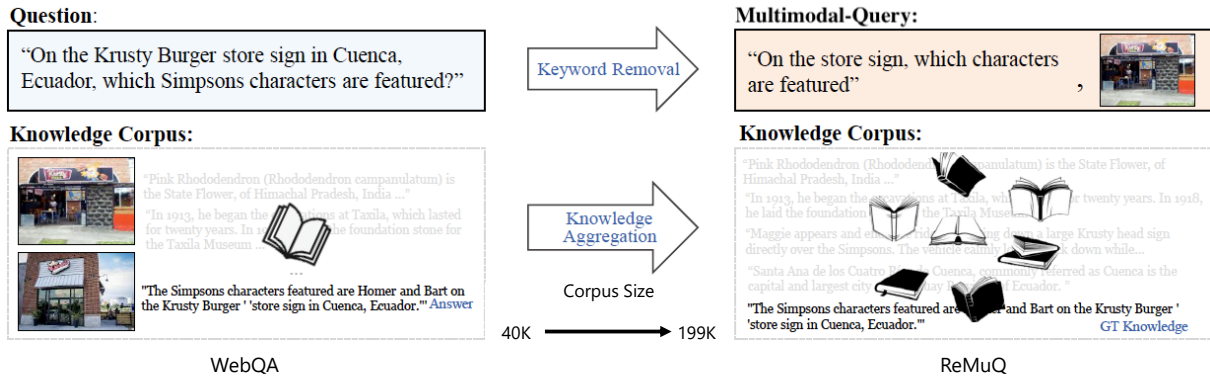
Figure 2: Dataset creation procedure for ReMuQ using WebQA as the source of the raw data. The multimodal-Query in ReMuQ is the combination of an image and the question from WebQA where the overlapped information with the image is removed. The ground truth knowledge of ReMuQ is the answer from WebQA. The corpus consists of all answers and the distracted knowledge candidates given in ReMuQ.

in the image, and KVQA (Shah et al., 2019)(world knowledge about named entities). Both FVQA and KVQA are equipped with knowledge graph as external corpus. In OKVQA (Marino et al., 2019) and its augmented versions S3VQA (Jain et al., 2021) and A-OKVQA (Schwenk et al., 2022), models are free to use any existing knowledge bases to retrieve relevant knowledge. WebQA (Chang et al., 2022) is a multi-hop reasoning dataset that requires a system to aggregate multiple sources to answer a question, where the answers can be found either via image search or general web search. Fang et al. (2020) introduce a video question answering dataset that requires a system to answer questions using commonsense knowledge about intentions and effects of people's actions in videos.

Knowledge-Retrieval with Multimodal Queries While there are methods for retrieving knowledge from knowledge graphs (Narasimhan et al., 2018; Li et al., 2020; Marino et al., 2021), in this work, we focus on systems that retrieve knowledge from free-form text, which is more readily available and comprehensive. Previous methods involve converting images into language representations such as captions (Qu et al., 2021; Gao et al., 2022) or object tags (Gui et al., 2022; Yang et al., 2022), and then using a text-based retriever such as BM25 (Robertson and Zaragoza, 2009) or DPR (Karpukhin et al., 2020) to find relevant knowledge. Gao et al. (2022) leverage GPT-3 (Brown et al., 2020) to generate the knowledge. Qu et al. (2021); Luo et al. (2021) use a vision and language model to obtain cross-modal representations. CLIP (Radford et al., 2021) has also been applied to retrieval tasks; however it has limitations due to its separate encoding of text and

image without a multi-modal fusion module.

3 Retrieval with Multimodal Queries

In this section, we define the problem statement for knowledge retrieval with multimodal queries and describe the construction of the ReMuQ dataset to assess models performing this task.

3.1 Problem Statement

Given a query $Q = (I, T)$ containing as image I and text T, we wish to learn a mapping to relevant textual knowledge K from a corpus C. Note that the two modalities I and T are such that each contains partial information about K. Both I and T are necessary for successful retrieval of K and Only using one of the two modalities is inadequate.

3.2 ReMuQ Dataset Creation

In ReMuQ each query has exactly one ground truth knowledge associated with it. To create such queries, we augment WebQA questions (Chang et al., 2022), and collect a large corpus to serve as the knowledge source for any retrieval systems. WebQA is a multihop and multimodalQA dataset including text questions of different types such as Yes/No, Open-ended (e.g. shape, color, etc.), and multi choice (MC) questions. The images are crawled from Wikimedia Commons, both questions and text answers are created by annotators.

To create multimodal queries, we utilize the MC questions in WebQA, which are associated with multiple choices as knowledge sources in the form of text or images. The ground truth answers of the questions include text-only, image-only, or both text and image. We adapt important steps to create
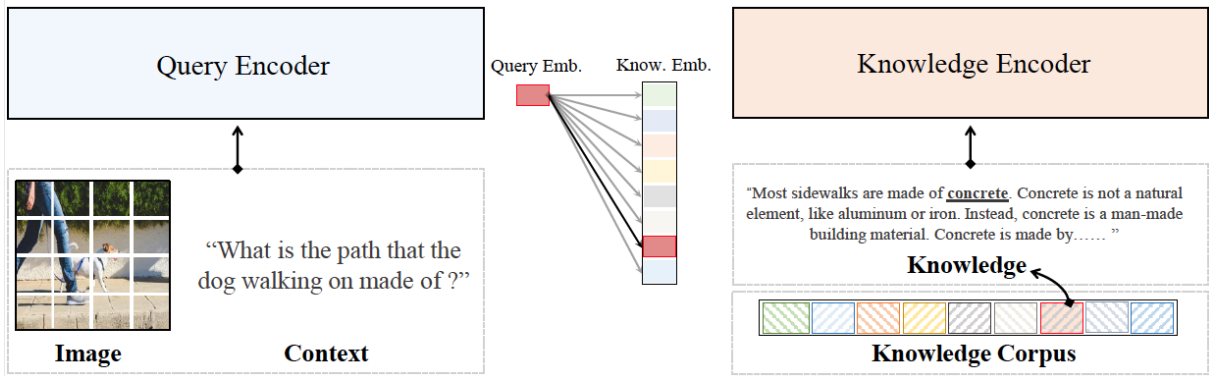
Figure 3: Overall architecture of ReViz. ReViz consists of a Vision-Language Transformer that encodes the image and text and a knowledge encoder that projects the knowledge into knowledge embedding. During inference, ReViz selects the knowledge from the corpus that has the largest relevance score with the image-text embedding.

multimodal queries and explain the pipeline of the curation procedure below and in Figure 2 (more examples are given in Appendix).

(1) Question Filtering. We select multiple-choice questions which have answer choices containing both image and text.

(2) Multimodal Query Construction. The initial multimodal query is the combination of the question and the corresponding image. In order to enforce a system to integrate information from both text and images, we use *tf-idf* to select keywords and then remove them in the question. Our new multimodal-query is then the concatenation of the augmented question and the image, with the text-answer to be the ground-truth knowledge.

(3) Retrieval Corpus Construction. We aggregate the textual knowledge from all samples as the common knowledge corpus for multimodal retrieval, resulting in a large corpus of $\sim 199k$ knowledge descriptions.

(4) Dataset Train-Test Split. We divide ReMuQ into 70% for training and 30% as testing split. The new curated dataset contains 8418 training samples and 3609 testing samples, together with a knowledge corpus with $195,837$ knowledge descriptions. More statistic of ReMuQ is given in Table 1.

4 Method

Prior work on Vision-Language (VL)-Retrievers has focused on two-stage methods where the first stage involves feature-extraction using pretrained visual and textual encoders and the second stage learns retrieval using these features. A typical VL-

Retriever can be expressed as:

$$K = \text{VL-RETRIEVER}(T, F; C), \qquad (1)$$

where C is the knowledge corpus, T is the text component of the query, and F denotes the extracted features of image I. This feature extraction can be done in two ways; (1) by converting the visual inputs into a human-readable textual description via an image captioning model or a series of object tags by object detector, (2) by extracting object features using an object detector.

End-to-End VL-Retriever. Instead, in this work, we are interested in building an end-to-end VL-Retriever, that encodes and selects the knowledge from the corpus using a VL model:

$$K = \text{VL-RETRIEVER}(T, I; C). \qquad (2)$$

We propose ReViz, an end-to-end VL-RETRIEVER that learns to maximize the multimodal query and knowledge similarity for knowledge retrieval tasks. We introduce its architecture below.

4.1 ReViz Model Architecture

ReViz can read and visualize the input query, consists of two components, the multimodal query encoder and the knowledge encoder. Figure 3 illustrates the pipeline of our model.

Multimodal Query Encoder. We use ViLT (Kim et al., 2021) to jointly encode the text input T and the image I. In ViLT, an image is first partitioned into a set of a fixed size of patches – these patches are encoded as continuous visual tokens through a linear projection layer (Dosovitskiy et al., 2020). These visual tokens are concatenated with the text
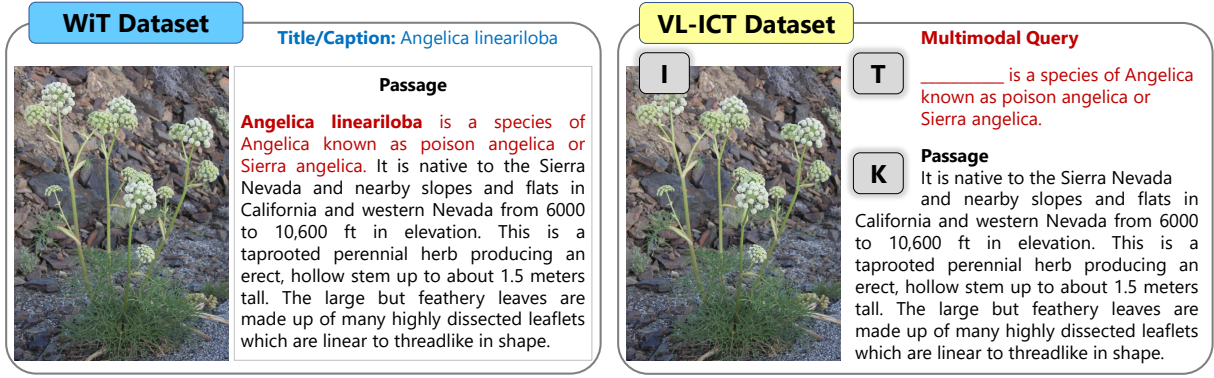
8576

Figure 4: Figure on the left shows an example of the WIT dataset (Srinivasan et al., 2021), crawled from Wikipedia. Figure on the right shows our constructed (T, I, K) triplet: T is a sentence from the passage and the words overlapped with the title/caption is masked; K is the remaining passage after removing the sentence.

tokens and summed with the position embeddings and fed into a stack of several self-attention blocks. The final multimodal representation is obtained by applying linear projection and hyperbolic tangent upon the first index token embedding.

$$\mathbf{Z}_q = \text{ViLT}(I, T) \tag{3}$$

Knowledge Encoder. To encode knowledge, we use a pre-trained BERT (Devlin et al., 2019) model, which produces a list of dense vectors $(h_1, \ldots, h_n)$ for each input token, and the final representation is the vector representation of special token [CLS].

$$\mathbf{Z}_k = \text{BERT}(K) \tag{4}$$

After the embeddings of query and knowledge are computed by the encoders, inner-dot product of the embeddings is considered as the relevancy score.

$$\text{Score}(I, T, K) = \mathbf{Z}_k^\top \cdot \mathbf{Z}_q \tag{5}$$

4.2 Training

The training objective of ReViz draws inspiration from the instance discrimination principle based on contrastive learning. The loss function to be minimized is given below:

$$\mathcal{L} = -\log \frac{\exp(\mathbf{z}_q \cdot \mathbf{z}_k)}{\exp(\mathbf{z}_q \cdot \mathbf{z}_k) + \sum\limits_{\hat{\mathbf{k}} \in \mathbf{B_k}, \hat{k} \neq k} \exp(\mathbf{z}_q \cdot \mathbf{z}_{\hat{\mathbf{k}}})}, \tag{6}$$

where $\mathbf{z}_q$ denotes the query embedding, $\mathbf{z}_k$ denotes the relevant knowledge embedding, and $\mathbf{z}_{\hat{k}}$ is the irrelevant knowledge embedding which serves as negative instances. We use all in-batch samples ($\mathbf{B_k}$) as the negative instances.

Training with Hard Negatives. Adopting random samples as negative instances may cause sub-optimal metric space. Existing work shows that mining with hard negative samples leads to discriminative representations and has been applied to a broad series of tasks like face recognition (Zhang et al., 2017), object detector (Shrivastava et al., 2016), and metric learning for retrieval tasks (Faghri et al., 2018; Harwood et al., 2017). Inspired by this, we also experiment with the hard negative technique to further boost the retrieval performance. To obtain the meaningful hard negative samples, we first train ReViz with the supervisions in *eq.* 6. With that, for each training question, we retrieve the top-100 knowledge instances (excluding the ground-truth) as the hard negative samples. Note that we only apply hard negative mining to fine-tuning on downstream task but not the pretraining task (introduced in the next section).

5 Pretraining Task for VL Retriever

Previous work (Chang et al., 2020; Lee et al., 2019; Guu et al., 2020) suggests that pretraining a retriever on unsupervised task that closely resembles retrieval can greatly improve the downstream tasks performance. We propose a pretraining task called VL-ICT, which is inspired by ICT (Lee et al., 2019) task in NLP domain.

ICT aims to train text-based information retrieval (IR) system for the open-domain question answering task. To train a model without annotated data, Lee et al. (2019) propose to construct pseudo $(question, context)$ pairs as the training data for IR system. In particular, given a passage P, a random sentence S in the passage is selected as the pseudo question, and the remaining passage

Datasets	Source		Average Length		Size		
	Image	Knowledge	Question	Knowledge	Train-D	Test-D	Knowledge
VL-ICT	Wiki	Wiki	24.15	111.79	10,783,957	-	-
OKVQA	COCO	GS/Wiki	9.15	67.05/100.00	8,958	5,046	112,724/21M
ReMuQ	Wiki	Wiki	14.97	48.60	8,418	3,609	195,837

Table 1: A comparison of the datasets used in our experiment in terms of the sources of images and knowledge, average length of question and knowledge, and the sizes of each dataset.

P' is considered as the relevant context. Such a weakly-supervised setting enables large-scale ICT pre-training, leveraging any available knowledge base as the training corpus.

VL-ICT. We propose VL-ICT task to pre-train ReViz, which can be applied to multi-modal scenarios when both language and vision inputs exist in the query. In VL-ICT, a (I, T, K) triplet is used for training. Importantly, I and T, contain mutually exclusive information and are both necessary for knowledge retrieval. However, such condition is not naturally existing, thus, we propose an automatic procedure to construct triplet satisfying this condition in the following.

VL-ICT Training Data. Figure 4 shows a snapshot of our data construction process where we use the WiT dataset (Srinivasan et al., 2021) as the source. Each WiT entry provides a title of the page or an image caption, a passage, and an image. We use the image from this WiT entry as the image I in our VL-ICT triplet. We observe that the title or caption is usually entities, it allows us to simply use word matching to find the sentences in the page passage that include the title/caption. We take such sentences as the text (T), then we remove this sentence from the passage and use the remaining passage as the knowledge (K). To enforce that (T) and (I) have mutually exclusive but important information, we mask keywords in T that appear in both T as well as the caption. In our experiments, we only select the English entities in WiT and execute the above process, and this results in 3.2 million (I, T, K) training triplets.

6 Experiments and Results

Datasets. In addition to ReMuQ, we conduct experiments on OKVQA to obtain stronger evidence for the efficacy of our method. Here, instead of QA task, we use OKVQA as a testbed for retrieval task, i.e. to retrieve a relevant knowledge to a question such that it contains the answer span. Furthermore,

we use two corpora, a small corpus collected from Google search API introduced in Luo et al. (2021), and a large corpus which contains 21M Wikipedia knowledge used in Gao et al. (2022). The statistic of each dataset is given in Table 1.

Evaluation Metrics. Following Gao et al. (2022); Luo et al. (2021), we evaluate the performances of models by Precision@K (P@K), Recall@K (R@K), and MRR@5. We use similar metrics to evaluate the ReMuQ challenge except that P@1 is used instead of P@5 since ReMuQ has exactly one correct knowledge per query.

6.1 Zero-shot Retrieval

We first introduce three zero-shot baselines and then present the results.

CLIP Baseline. CLIP (Radford et al., 2021) is a vision-language model pre-trained on over 400M image-text pairs. We encode all knowledge descriptions via CLIP's textual encoder **K**. Then, given an image-text pair as the query, we use the image encoder to get the visual representations (**I**) and use the textual encoder to get the embedding of **Q**. We compute the inner-dot products between all encoded visual representations (**I**) and **K** to get the top-100 knowledge for evaluation, similarly for **Q**. Finally we sum the scores and re-rank the top-100 knowledge. We find this performs the best than using individual modality (see Appendix).

BM25 Baseline. BM25 (Robertson and Zaragoza, 2009) is a well-known efficient retrieval algorithm for text-based retrieval task based on the sparse representation. We use the caption of the image to represent the information of the image and thus we convert the multi-modal knowledge retrieval task into a pure text-based retrieval task.

DPR Baseline. We adopt DPR (Karpukhin et al., 2020) trained on NaturalQuestions (Kwiatkowski et al., 2019) dataset as a baseline, to retrieve the knowledge given an input image-text pair. First,

Model	Dataset	KB-Size	Metric						
			MRR@5	P@5	R@5	R@10	R@20	R@50	R@100
CLIP-IMG+Q	OKVQA	GS-112K	19.08	11.13	34.54	50.48	65.08	80.62	88.11
BM25 (GenCap)	OKVQA	GS-112K	36.36	27.54	51.35	63.04	73.37	84.21	90.39
DPR (GenCap)	OKVQA	GS-112K	39.15	27.72	55.56	66.44	75.59	87.17	92.42
ReViz+VL-ICT	OKVQA	GS-112K	**45.77**	**33.18**	**64.05**	**75.39**	**84.21**	**91.64**	**94.59**
TRiG (Gao et al., 2022)	OKVQA	Wiki-21M	-	-	45.83	57.88	72.11	80.49	86.56
CLIP-IMG+Q	OKVQA	Wiki-21M	16.45	9.66	29.81	43.00	55.73	72.73	82.26
BM25 (GenCap)	OKVQA	Wiki-21M	36.43	27.89	50.16	60.92	71.62	82.82	88.74
DPR (GenCap)	OKVQA	Wiki-21M	41.15	28.10	59.41	71.13	81.73	89.90	93.39
ReViz+VL-ICT	OKVQA	Wiki-21M	**44.03**	**32.94**	**62.43**	**73.44**	**82.28**	**89.93**	**93.76**
CLIP-IMG+Q	ReMuQ	199K	0.34	0.17	0.78	1.36	2.41	7.34	47.88
BM25 (GenCap)	ReMuQ	199K	3.80	5.59	8.78	10.75	12.88	15.88	17.98
DPR (GenCap)	ReMuQ	199K	**31.23**	**35.79**	**43.42**	**48.77**	**54.47**	61.40	67.30
ReViz+VL-ICT	ReMuQ	199K	23.61	29.52	39.43	46.77	53.56	**63.70**	**71.13**

Table 2: Zero-shot performance of ReViz and baselines on two datasets: OKVQA and ReMuQ. OKVQA is evaluated on two knowledge sources. ReViz shows superior zero-shot performance in majority of the cases.

Model	Dataset	KB-Size	Metric						
			MRR@5	P@5	R@5	R@10	R@20	R@50	R@100
ReViz	OKVQA	GS-112K	46.92	34.51	66.05	77.80	86.33	93.34	95.90
ReViz+VL-ICT	OKVQA	GS-112K	**54.47**	**41.74**	**73.35**	**83.17**	**89.56**	**94.73**	**96.81**
ReViz	OKVQA	Wiki-21M	41.66	30.08	60.88	72.20	81.07	89.16	93.10
ReViz+VL-ICT	OKVQA	Wiki-21M	**43.68**	**31.36**	**61.91**	**72.63**	81.05	**89.28**	**93.44**
ReViz	ReMuQ	199K	41.03	49.08	62.40	71.63	78.92	86.60	92.17
ReViz+VL-ICT	ReMuQ	199K	**53.39**	**62.11**	**76.23**	**83.32**	**88.56**	**93.41**	**96.12**

Table 3: Comparison of ReViz when it is fine-tuned on downstream tasks. We compare ReViz and ReViz+VL-ICT (our pretraining task). VL-ICT enables ReViz to be a stronger multimodal-query retrieval model.

we use the contextual encoder of DPR to index the corpus, then we concatenate the question and the caption of the image as a joint textual query. With that, the question encoder of the DPR extracts the dense representation of the query for later computation. Lastly, we retain the most relevant knowledge pieces by calculating the inner-dot product between the query and the knowledge embedding.

Results. Table 2 shows the performances of baselines as well as ReViz pretrained on VL-ICT task. Among the baselines, we see that DPR is the strongest baselines. Surprisingly, although CLIP has shown strong performance on many classification and cross-modality pretraining task, it does not perform well on multimodal query retrieval task, this suggests that multimodal query retrieval is a challenging task for VL model. More importantly, we observe clearly that ReViz outperforms the baselines in terms of all metrics on OKVQA task on corpus of small and large size. On the ReMuQ dataset, ReViz wins CLIP and BM25 on all metrics, and DPR on two metrics. This demonstrates

the effectiveness of our proposed pretraining task and the model design.

6.2 Fine-tuning on Downstream Tasks

To further demonstrate the effectiveness of VL-ICT pretraining task, we finetune models on downstream tasks and compare performance. We compare two versions of ReViz: (1) ReViz directly trained on the downstream task and (2) ReViz first pretrained on VL-ICT and then finetuned the downstream task. In addition, We study two senarios: in-domain, where a model is trained on the training set of X domain and evaluated on the testing set of X; out-of-domain, where a model is trained on the training set of X domain and evaluated on the testing set of Y domain.

In-Domain Results. Table 3 shows the in-domain performance. On both datasets, pretrained ReViz consistently outperform vanilla ReViz, suggesting that the pretraining task equips ReViz better alignment between the multimodal queries and the relevant knowledge.

Model	FT	KB-Size	Metric						
			MRR@5	P@5	R@5	R@10	R@20	R@50	R@100
VRR-IMG (Luo et al., 2021)	✓	GS-112K	-	31.80	62.52	73.96	83.04	90.84	94.67
VRR-CAP (Luo et al., 2021)	✓	GS-112K	-	39.42	71.52	81.51	88.57	94.13	96.95
ReViz+VL-ICT	✓	GS-112K	**54.47**	**41.74**	**73.35**	**83.17**	**89.56**	**94.73**	96.81
TRiG (Gao et al., 2022)	✗	Wiki-21M	-	-	45.83	57.88	72.11	80.49	86.56
ReViz+VL-ICT	✗	Wiki-21M	**44.03**	**32.94**	**62.43**	**73.44**	**82.28**	**89.93**	**93.76**

Table 4: Comparison of our best model with existing models on OKVQA. "FT" denotes fine-tuning. Our model surpasses existing methods by significant margins with or without fine-tuning and with different knowledge corpus.
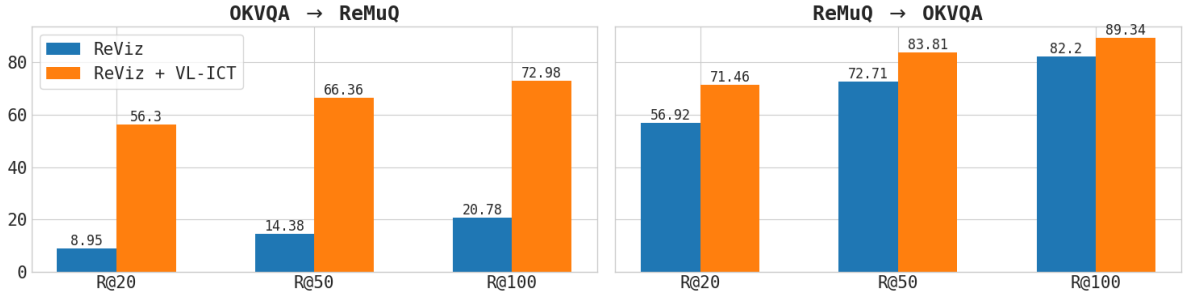
Figure 5: Evaluation of out-of-domain performances of ReViz and ReViz+VL-ICT. For OKVQA, we retrieve knowledge from GS-112K corpus. VL-ICT substantially improves the generalization of ReViz. Other metrics are given in Appendix. X->Y denotes using X as the training domain and Y as the testing domain.

Out-of-Domain Results. We investigate if the VL-ICT pretraining task can improve the generalization of ReViz. We study the performances of ReViz under two settings: train on OKVQA (domain **X**) and test on ReMuQ (domain **Y**); and the inverse. Table 5 shows that ReViz+VL-ICT+**X** shows obviously better results than ReViz+**X** on **Y**, especially when **X** is OKVQA and **Y** is ReMuQ. This suggests that models pre-trained with VL-ICT tasks are more robust than models without VL-ICT. We also see that the generalization performance still has a large gap with the fine-tuning, which suggests that OKVQA and ReMuQ are quite different tasks, and ReMuQ can be a good complement to OKVQA to study multimodal query retrieval task.

6.3 Comparison with Existing Methods

We compare ReViz with existing retrieval methods for the OKVQA task. Note that most of the models on the leaderboard of OKVQA only report the final question answering accuracy but not the retrieval performance. In our experiments we include systems which report the retrieval performance.

Baselines. Luo et al. (2021) present two fine-tuned multimodal retrievers: VRR-IMG which uses LXMERT (Tan and Bansal, 2019) and VRR-CAP to convert the image into captions for knowledge re-

trieval. Both retrievers use GS-112K as the knowledge corpus. TriG (Gao et al., 2022) uses zeroshot retriever and Wikipedia 21M as the knowledge corpus. Since these systems use either fine-tuned retriever or zero-shot retrievers, for fair comparison, we compare the best fine-tuned model and zeroshot model with the corresponding corpus.

Results. In the fine-tuning scenario, in majority of the cases (only one exception, R@100), our models consistently shows better performance than previous methods overall metrics. Similarly, in the zero-shot case, our model is better than previous model on all metrics by large margins.

6.4 Effects of Mask Ratio in VL-ICT Task

In VL-ICT, we mask the keywords in the sentence to prevent information leakage. Despite this, we find that the certain masked sentences still somehow overlap with the retrieved knowledge. We conjecture that this overlapping makes the VL-ICL task inevitably easy, and thus impairs the effects of pre-training. To study the optimal mask ratio, we conduct experiments to randomly mask the words in the sentence by different ratios. This study is performed on a smaller corpus of 1 million VL-ICT training triplets and models are trained for one epoch. Figure 7 shows the results. We observe that
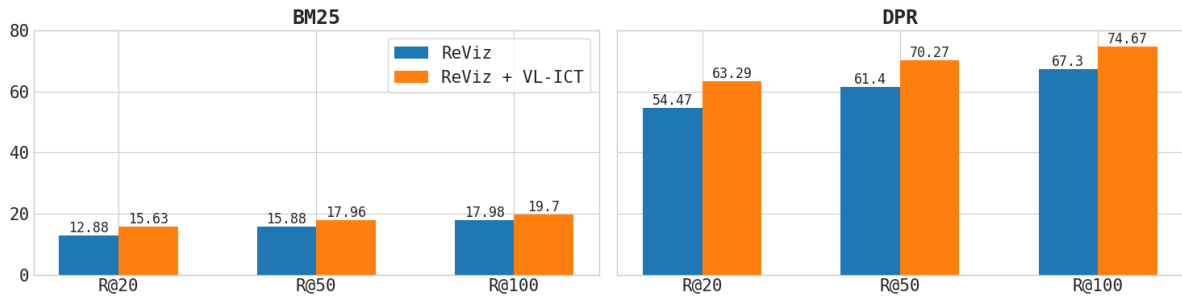
Figure 6: Comparison of captioning-dependent retrievers using generated captions and ground truth captions. The ground truth captions always lead to better performance than generated caption.
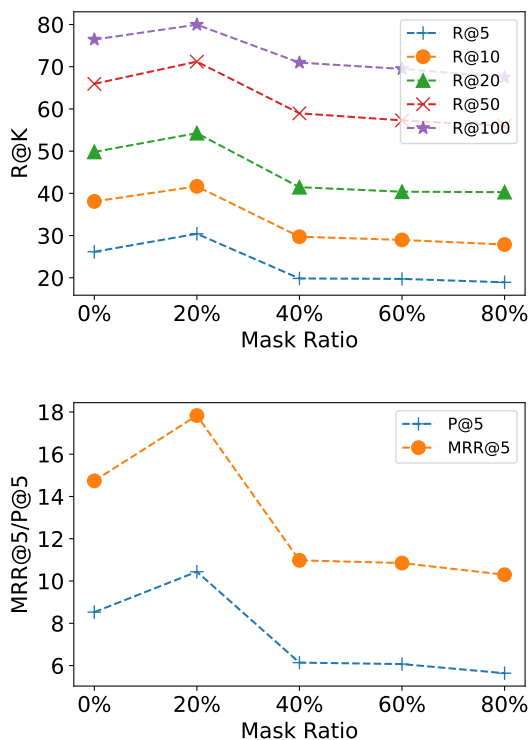

Figure 7: Effect of the masking ratio of sentences in VL-ICT task on ReViz's performance on OKVQA Task. We use GS112K as the knowledge corpus.

removing 20% of the keywords yields the best performance amongst all ratios and is also better than maintaining the sentences intact (0% masking).

6.5 Effect of Generated Captions

Previous systems which rely on the caption generation model are affected by the quality of generated captions. This may hamper the retrieval performance when the caption generation model is not trained on the same domain as the downstream task. In our ReMuQ dataset, the images are from Wikipedia, but the caption generator is trained on

MS-COCO (Lin et al., 2014). We compare our two baselines, BM25 and DPR, using ground-truth image captions and the generated captions. Table 6 shows that using the ground truth caption is much better than the generated caption in all cases. This suggests that the caption generator is the bottleneck of the retrieval methods to convert the image information to image captioning. This demonstrates the limitations of previous methods and justifies our exploration of end-to-end training.

7 Conclusion

We study knowledge retrieval with multimodal (vision and language) queries, which, compared with existing retrieval tasks, is more challenging and under-explored. In addition, multimodal-query information retrieval has numerous potential applications, not only in retrieval tasks such as image, text, and video retrieval, but also in question answering, recommendation systems, and personal assistant. The proposed dataset (ReMuQ) is ideally positioned to support the development of such functionalities. We propose an end-to-end VL-retriever model, ReViz, which does not rely on any intermediate image to text translation modules. A novel weakly-supervised task (VL-ICT) is proposed to enable large-scale pre-training. Extensive evaluations on ReMuQ and OK-VQA datasets demonstrate that ReViz exhibits strong performance amongst all retrieval models in both zero-shot and fine-tuning scenarios. Our proposed dataset and model provide a foundation for future work which could potentially lead to new findings and innovative applications in multimodal-query information retrieval.

Limitations

During the creation of the ReMuQ dataset, we simply remove the words in the question that are duplicated in the image caption – in some cases, this may result in grammatical errors in the text query. We performed the experiments for studying optimal masking ratio on a subset of the pretraining data, due to resource constraints.

Acknowledgments

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860.

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. *arXiv preprint arXiv:2201.05299*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Ben Harwood, Vijay Kumar B. G, Gustavo Carneiro, Ian D. Reid, and Tom Drummond. 2017. Smart mining for deep metric learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2840–2848. IEEE Computer Society.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

James S. Kinder. 1942. Chapter viii: Visual aids in education. *Review of Educational Research*, 12(3):336–344.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022. Improving biomedical information retrieval with neural retrievers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11038–11046.

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

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121.

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.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2659–2670.

Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1753–1757.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Information Retrieval*, 3(4):333–389.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.

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, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8876–8884. AAAI Press.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. 2016. Training region-based object detectors with online hard example mining. In *2016 IEEE Conference on Computer Vision and Pattern Recognition,*

CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 761–769. IEEE Computer Society.

Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From strings to things: Knowledge-enabled VQA model that can read and reason. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4601–4611. IEEE.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. 2017. Range loss for deep face recognition with long-tailed training data. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5419–5428. IEEE Computer Society.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press.

Appendix

A Experimental Setup

All ReViz models consist of a ViLT query encoder and a BERT context encoder, both with 12 transformer blocks with 12 attention heads each. For pretraining, we use Adam optimizer with 100 warm-up steps, learning rate at 1e-6, a dropout probability of 0.1, and pre-train the model in 5 epochs. For down-stream task fine-tuning, we use Adam optimizer with 10 warm-up steps in 30 epochs. Learning rate 1e-6 is applied to fine-tune a pretrained ReViz on the down-stream task, and learning rate 1e-5 is used if fine-tune a vanilla ReViz. All models use 64 batch-size in the training on a machine with eight Quadro RTX 8000 GPUs.

B Effect of Hard Negative Training

We show the effectiveness of hard negative training in Table 6. We experiment with both OkVQA and our ReMuQ dataset and the pretrained models on VL-ICT. We see that using the hard negative examples to train the model is much better than without this training step.

C Additional Visualizations

Examples of VL-ICT Pretraining Task. Figure 8 presents more examples of VL-ICT pretraining task.
More Examples of ReMuQ Task. We present some examples of ReMuQ in Figure 9, consisting of an image, an input context and the corresponding knowledge.

D Examples of Retrieval Results

In Table 7, we present some examples of ReViz+VL-ICT+OKVQA, the best model performing on the GS-112K corpus for OKVQA dataset. In Table 8, we present some examples of ReViz+VL-ICT, the best model performing on the Wiki-21M corpus for OKVQA dataset.

E CLIP Performance

As we mention in the experiment section that CLIP is one of the baselines. We compare three methods to retrieve knowledge using CLIP. First one is
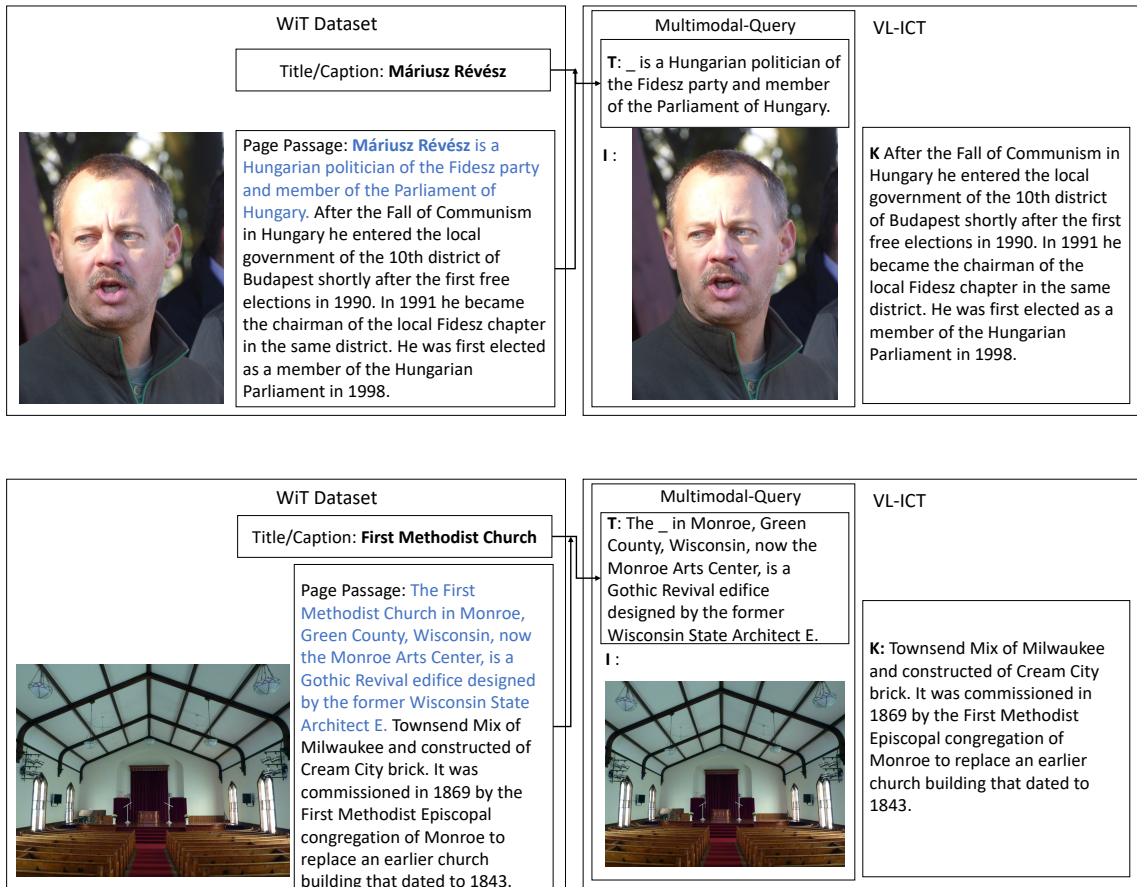
Figure 8: More examples of VL-ICT pretraining Task.

Question: does the flower have petals in a cup shape?
knowledge: No, a Minnetonka Rhododendron flower does not have petals in a cup shape.

Question: on what is the clock in ceiling of hung on?
Knowledge: The clock on the ceiling of Dijon Halles intérieur is hung on a black railing structure.

Figure 9: Examples of ReMuQ datasets.

Model	Dataset	Metric						
		MRR@5	P@K	R@5	R@10	R@20	R@50	R@100
CLIP-IMG	OKVQA	18.96	11.03	34.50	50.48	65.12	80.60	88.11
CLIP-Q	OKVQA	5.06	4.46	10.15	13.77	20.61	35.63	44.03
CLIP-IMG+Q	OKVQA	19.08	11.13	34.54	50.48	65.08	80.62	88.11
CLIP-IMG	ReMuQ	0.28	0.11	0.69	1.27	2.33	7.29	47.88
CLIP-Q	ReMuQ	0.00	0.00	0.00	0.03	0.03	0.11	0.17
CLIP-IMG+Q	ReMuQ	0.34	0.17	0.78	1.36	2.41	7.34	47.88

Table 5: CLIP performance on two datasets using three approaches to retrieve knowledge. For OKVQA, GS-112K corpus is used. P@5 is used for OKVQA and P@1 is used for ReMuQ as shown in the main paper.

Model	KB-Size	Metric						
		MRR@5	P@5	R@5	R@10	R@20	R@50	R@100
ReViz+VL-ICT+OKVQA⁻	GS-112K	47.82	36.50	66.49	77.35	86.23	95.14	95.70
ReViz+VL-ICT+OKVQA	GS-112K	**54.47**	**41.74**	**73.35**	**83.17**	**89.56**	**94.73**	96.81
ReViz+VL-ICT+ReMuQ⁻	ReMuQ	50.93	42.67	64.17	72.10	79.27	86.81	90.58
ReViz+VL-ICT+ReMuQ	ReMuQ	**62.11**	**53.39**	**76.23**	**83.32**	**88.56**	**93.41**	**96.12**

Table 6: Compare the performance of without using hard negative (-) and with hard negative.

Image	Question	Retrieved Knowledge	Answer
	This bear is what kind of bear?	**brown** bears are found in asia, europe, and north america, giving them the widest ranges of bear species.they also inhabited north africa and the middle east. in north america, **grizzly** bears previously ranged from alaska down to mexico and as far east as the western shores of hudson bay...	grizzly; brown
	What do you call the device that keeps boats in place at sea?	an **anchor** is a device, normally made of metal, used to connect a vessel to the bed of a body of water to prevent the craft from drifting due to wind or current. they have the reputation of not breaking out with tide or wind changes, instead slowly turning in the bottom to align with the force...	anchor; locator
	How long do these animals live?	we usually consider 1 year for dog equals to 7 years for you. similar to dogs, birds' age should be viewed relative to its life span. different bird species have different life span.	15 years; fifteen years

Table 7: Examples of retrieved knowledge for OKVQA under GS-112K corpus by ReViz+VL-ICT+OKVQA.

only using the image, the second one is only by question, and the last one is by both image and question. In the last method, we firstly use the image embeddings and the knowledge embeddings to obtain the top-100 relevant knowledge, then we use the question embeddings to obtain the top-100 relevant knowledge. Lastly, we obtain the final top-100 knowledge by the sum of the scores given by the image and question embeddings. Table 5 shows the performance of CLIP using three methods. Using both image and question achieves the best performance.
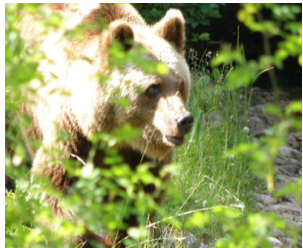
Image	Question	Retrieved Knowledge	Answer
	This bear is what kind of bear?	a 1,600-square mile island. There are currently about 55,000 wild **grizzly** bears total located throughout North America, most of which reside in Alaska. Only about 1,500 grizzlies are left in the lower 48 states of the US. Of these, about 800 live in Montana...	grizzly; brown
	What do you call the device that keeps boats in place at sea?	the vessels are not controlled when they arrive in harbor, it depends of many characteristics (old, flag state, cargo, quality of classification society ...). The Memorandum of understanding (MOU) is a group (area) to avoid this restrictive investigation...	anchor; locator
	How long do these animals live?	in the morning and their dog cries at the window, it exhibits sadness. A growling dog who doesn't like it when someone touches its favorite toy is showing anger. Animals can feel love as well as other basic emotions humans feel. Dogs that grow up with siblings create strong bonds to their sibling.	15 years; fifteen years

Table 8: Examples of retrieved knowledge for OKVQA under Wiki-21M corpus by ReViz+VL-ICT model.

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
all public datasets

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. using previously published open-source data

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.