REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory

Ziniu Hu[1]*, Ahmet Iscen[2], Chen Sun[2], Zirui Wang[2], Kai-Wei Chang[1], Yizhou Sun[1]
Cordelia Schmid[2], David A. Ross[2], Alireza Fathi[2]
[1]University of California, Los Angeles, [2]Google Research

Figure 1. We augment a visual-language model with the ability to retrieve multiple knowledge entries from a diverse set of knowledge sources, which helps generation. Both retriever and generator are trained jointly, end-to-end, by optimizing a language modeling objective.

Abstract

*In this paper, we propose an end-to-end Retrieval-Augmented Visual Language Model (REVEAL) that learns to encode world knowledge into a large-scale memory, and to retrieve from it to answer knowledge-intensive queries. REVEAL consists of four key components: the memory, the encoder, the retriever and the generator. The large-scale memory encodes various sources of multimodal world knowledge (*e.g. *image-text pairs, question answering pairs, knowledge graph triplets,* etc.*) via a unified encoder. The retriever finds the most relevant knowledge entries in the memory, and the generator fuses the retrieved knowledge with the input query to produce the output. A key novelty in our approach is that the memory, encoder, retriever and generator are all pre-trained end-to-end on a massive amount of data. Furthermore, our approach can use a diverse set of multimodal knowledge sources, which is shown to result in significant gains. We show that REVEAL achieves state-of-the-art results on visual question answering and image captioning. The project page of this work is* reveal.github.io.

1. Introduction

Recent large-scale models such as T5 [33], GPT-3 [4], PaLM [9], CoCa [49], Flamingo [2], BEIT-3 [43] and PaLI [7] have demonstrated the ability to store substantial amounts of world knowledge, when scaled to tens of billions of parameters and trained on vast text and image corpora. These models achieve state-of-the-art results in downstream tasks such as image captioning, visual question answering and open vocabulary recognition. Yet, these models have a number of drawbacks: (i) they require massive scale, of parameters, data and computation, and (ii) they need to be re-trained every time the world knowledge is updated.

To address these issues, we adopt a different approach. Instead of statically compiling world knowledge into model weights, we transform the knowledge into a key-value memory through neural representation learning. Our model learns to utilize the memory for answering knowledge-intensive queries. By decoupling the knowledge memorization from reasoning, we enable our model to leverage various external sources of knowledge (*e.g.*, Wikipedia passages and images [37], the WikiData knowledge graph [40], Web image-text pairs [5] and visual question answering data [12]). This enables the model parameters to focus on understanding the query and conducting reasoning, rather than being dedicated to memorization.

Retrieval-augmented models have attracted a fair amount of attention in the fields of NLP [14, 18] and computer vision [13,25]. Typically, these models often use a pre-existing single-modality backbone to encode and retrieve information from the knowledge corpus. Such approaches do not leverage all available modalities in the query and knowl-

*This work was done when Ziniu was an intern at Google.

edge corpora, and hence they might not find the information that is most helpful for generating the model output. A key novelty in our approach is that we encode and store various sources of multimodal world knowledge into a unified memory, which the retriever can access via multimodal query encodings, to find the most relevant information from across complementary sources. Our multimodal memory and retriever are pre-trained end-to-end together together with the rest of the model, on a massive amount of data and using diverse knowledge sources.

A key challenge of pre-training the multimodal retriever end-to-end is the lack of direct supervision. There is no ground-truth indicating which knowledge entries are most helpful for answering knowledge-intensive queries. Some of the existing works in NLP [14, 23, 34] propose to acquire training signal by assessing the usefulness of each retrieved knowledge entry independently for helping language modelling. This approach is inefficient, as it involves estimating hundreds of retrieved knowledge entries independently, and also inaccurate as it discards the dependency between different knowledge entries in the retrieval set. In contrast, we propose to get this training signal while simultaneously considering multiple retrieved knowledge entries, by introducing an attentive fusion layer that injects retrieval score into the attention calculation procedure. This enables the retrieval module to be differentiable and jointly pre-trained with the rest of the model.

In summary, our key contributions are as follows:

- We are the first to propose an end-to-end pre-training paradigm that learns to index into a large-scale memory to solve knowledge-intensive visual-language tasks.
- Our method can construct a large-scale memory by encoding various sources of multimodal world knowledge, including Wikipedia passage, web images with alt-text captions, and knowledge graph triplets.
- REVEAL achieves state-of-the-art performance on several knowledge-intensive visual question answering and image captioning datasets. Notably on the OKVQA benchmark, REVEAL achieves a new state-of-the-art, 59.1% accuracy, while using order of magnitude fewer parameters than previous works.

2. Related Work and Background

Knowledge-based Visual Question Answering. To evaluate a model's ability to comprehend multimodal world knowledge not easily inferred from input data, several knowledge-based Visual Question Answering (VQA) datasets have been introduced. KB-VQA [41] and FVQA [42] design questions that can be answered by retrieving relevant triplets from domain-specific structured knowledge graphs. OK-VQA [29] improves these datasets by necessitating the use of external knowledge, which goes beyond what can be directly observed in the input images. More recently, A-OKVQA [35] offers further improvements to OK-VQA by exclusively selecting questions that demand both external knowledge and commonsense reasoning about the image scenes. To tackle knowledge-based VQA tasks, many approaches have been proposed to incorporate external knowledge into visual-language models. One line of research uses explicit knowledge from structured knowledge graphs [11,15,31,44] or unstructured text corpora [27,28,46]. The key component for these works is the knowledge retriever. Some works [11, 28, 31, 46] utilize off-the-shelf vision detection models to generate image tags for knowledge retrieval, while others train the retrieval model via distant supervision [27] or auxiliary tasks (*e.g.* entity linking) [13]. Another research direction aims to incorporate implicit knowledge from pre-trained Large Language Models, such as GPT-3 [4] or PaLM [9]. These approaches utilize off-the-shelf image caption models to convert images into text, feed them into a language model, and use the generated text output as augmented knowledge [13, 24, 48]. Our work follows the first direction, augmenting a vision-language model with an explicit knowledge retriever. The main distinction is that we propose an end-to-end training framework to jointly learn the answer generator and retriever, rather than using a fixed or predefined knowledge retrieval.

End-to-End Training of Retrieval-Augmented Models. Given the advantage of knowledge retrieval, a key question is how to get learning signal to train the retrieval model. For tasks with annotated retrieval ground-truth, retrieval training can be conducted via standard contrastive learning [22]. However, most tasks do not provide clear indications of which knowledge entries are relevant for generating answers. To this end, a series of studies have investigated retrieval training using supervision derived from downstream tasks. REALM [14] trains a single-document retriever by concatenating each retrieved result with the query, to calculate the final loss independently. A similar approach has been used by EMDR^2 [34] for multi-document retrieval training. FID-KD [17] proposes to use the aggregated attention score calculated by the generator as a distillation signal to train the retriever. Atlas [18] further introduces a perplexity distillation loss and a leave-one-out variant. Our REVEAL proposes to inject the retrieval scores directly into an attentive fusion module, enabling to train the retriever to directly optimize downstream tasks as well as pre-training objectives.

3. Method

We propose a Retrieval-Augmented Visual Language Model (REVEAL), which learns to use knowledge from different sources for solving knowledge-intensive tasks. For both pre-training and fine-tuning, our goal is to learn the distribution $P(y \mid x)$ to generate a textual output y conditioned on a multimodal input query x. REVEAL contains

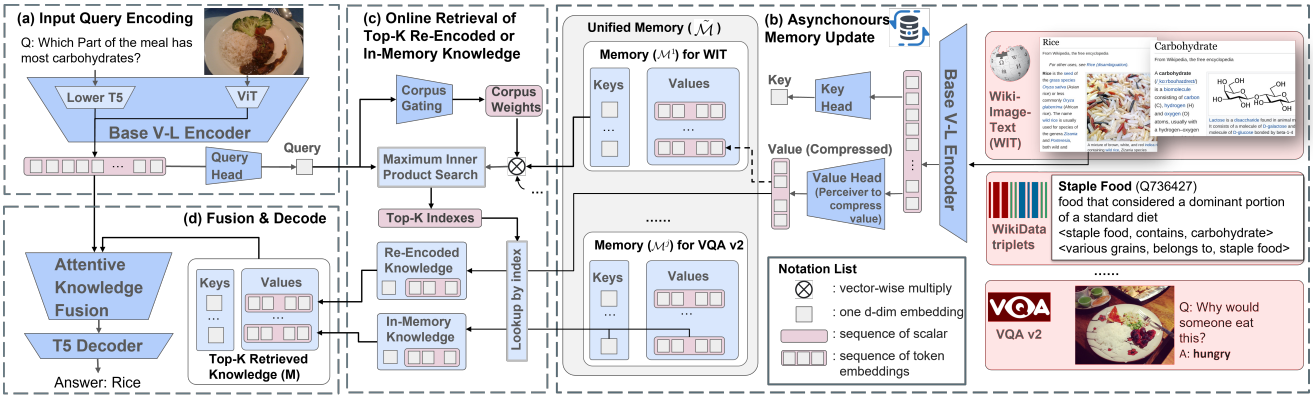

Figure 2. **The overall workflow of REVEAL** consists of four main steps: **(a)** encode a multimodal input into a sequence of token embeddings and a summarized query embedding; **(b)** encode each knowledge entry from different corpus into unified key and value embedding pairs, where key is used to index the memory and value contains full information of the knowledge; **(c)** retrieve top-K most similar knowledge items from different knowledge sources, and return the pre-computed in-memory value embeddings and re-encoded value; and **(d)** fuse the top-K knowledge items via attentive knowledge fusion layer by injecting the retrieval score as a prior during attention calculation. This facilitates REVEAL's key novelty: the memory, encoder, retriever and the generator can be jointly trained in an end-to-end manner.

four components: knowledge encoding, memory, retrieval and generation. Given an input query x, we first retrieve K possibly helpful entries $M = \{m_1, \cdots, m_K\}$ from the memory corpora $\mathcal{M}$. Each m is a memory entry containing the encoded single key embedding and a sequence of value embeddings (we will describe how to encode knowledge items into memory entries in Sec. 3.2). With it, the retriever can use embedding similarity to find relevant memory entries. We model this retrieval process as sampling from distribution $p(M \mid x)$. Then, we condition on both the retrieved set M and the original input query x to generate the output y, modeled as $p(y \mid x, M)$. To obtain the overall likelihood of generating y, we treat M as a latent variable from the entire memory $\tilde{\mathcal{M}}$ and marginalize over it yielding:

$$p(y \mid x) = \sum_{M \subset \tilde{\mathcal{M}}} \underbrace{p(M \mid x)}_{retrieval} \cdot \underbrace{p(y \mid x, M)}_{generation}. \quad (1)$$

However, this marginal probability involves an intractable summation over all size-K subsets of the memory corpora $\tilde{\mathcal{M}}$. We approximate this instead by using the top-K entries in memory with the highest probability under $p(M \mid x)$. This is reasonable if most of the unrelated memory entries do not contribute to the generation. Note that we use an online memory that is updated as the knowledge encoder is trained end-to-end with the rest of the model.

Figure 2 illustrates the overall workflow of REVEAL, and we describe each component in this section. In particular, in Sec. 3.1 we describe how the query is encoded. In Sec. 3.2 we go over how the multimodal knowledge memory is constructed and updated during pre-training. Next, we describe how we retrieve the memory entries that are most relevant to the input query in Sec. 3.3. Finally, in Sec. 3.4 we describe the generator that fuses the query and retrieved knowledge and decodes them into the generated text.

3.1. Query Encoding

Figure 2 (a) depicts how the input image-text query is encoded. We use a base visual-language encoder $b(\cdot)$ to turn the query input and each knowledge item (with potentially different modalities *e.g.* text-only, image-only or image-text pairs) into a sequence of embeddings (tokens). We adopt a Vision Transformer (ViT) [10] to encode the images and we use a lower-layer[1] T5 encoder [33] to encode the texts. We add a projection layer on top of the ViT model to map the image tokens into the same space as the text tokens. We then concatenate the two modalities together. We use an upper-layer T5 module as both the query Head $\phi_{\texttt{Query}}(\cdot)$ and the key Head $\phi_{\texttt{Key}}(\cdot)$ to compute the query embedding and memory keys. We take the output of the first `[CLS]` tokens followed by a linear projection and L2-normalization to summarize the input into a d-dimensional embedding.

3.2. Memory

Figure 2 (b) shows how memory is constructed and updated by encoding knowledge items. Our approach differs from previous works primarily by leveraging a diverse set of multimodal knowledge corpora (WikiData knowledge graph, Wikimedia passages and images, Web image-text pairs). Throughout the paper, we denote each corpus as $\mathcal{C}^j = \{z_1^j, \ldots, z_N^j\}$, in which each $z_i^j \in \mathcal{C}^j$ is a knowledge item that could be an image-text pair, text only, image only, or a knowledge graph triplet. We denote the unified knowledge corpus as $\tilde{\mathcal{C}} = \mathcal{C}^1 \cup \mathcal{C}^2 \cdots \cup \mathcal{C}^S$ that combines $|\tilde{\mathcal{C}}| = S$ different knowledge corpora. We encode the external knowledge corpora into a unified memory $\tilde{\mathcal{M}} = [\mathcal{M}^1, \ldots, \mathcal{M}^{|\tilde{\mathcal{C}}|}]$. Each knowledge item z_i is encoded into a key/value pair $m_i = (\texttt{Emb}_{\texttt{Key}}(z_i), \texttt{Emb}_{\texttt{Value}}(z_i))$ in memory. Each key

[1]We denote the last l layers of a T5 encoder as 'upper-layer', and the remaining ones including the token embedding layer as 'lower-layer'.

$\texttt{Emb}_{\texttt{Key}}(z) = \phi_{\texttt{Key}}\big(b(z)\big) \in \mathbb{R}^d$ is a d-dimensional embedding vector encoded via Key Head. Each value is a sequence of token embeddings representing the full information of knowledge item z. We follow a similar procedure as in [14] to precompute key/value embeddings of knowledge items from different sources and index them in a unified knowledge memory. We continuously re-compute the memory key/value embeddings as the model parameters get updated during the pre-training phase. We update the memory $\tilde{\mathcal{M}}$ asynchronously at every 1000 training steps.

Scaling Memory by Compression A naive solution for encoding the memory value is to keep the whole sequence of tokens for each knowledge item. Then, the generator could fuse the input query and the top-K retrieved memory values by concatenating all their tokens together and feeding them into a Transformer Encoder-Decoder pipeline [23]. This approach has two issues: (1) storing hundreds of millions of knowledge items in memory is impractical given that each memory value would consist of hundreds of tokens; (2) transformer encoder has quadratic complexity with respect to the total number of tokens times K for self-attention.

Therefore, we propose to use the Perceiver architecture [19] as the Value Head to encode and compress knowledge items. The Perceiver model uses a transformer decoder $\psi(\cdot)$ with learnable c-length latent embeddings to compress the full token sequence into an arbitrary length c, such that $\texttt{Emb}_{\texttt{Value}}(z) = \psi(b(z)) \in \mathbb{R}^{c\times d}$ (In our experiments we use $c = 32$). This lets us retrieve top-K memory entries for K as large as a hundred. To make the compressed embeddings generated by Perceiver more expressive, we add two additional regularizations. The first one is a disentangled regularization [16] that forces every two output tokens to be linearly de-correlated $\mathcal{L}_{\text{decor}} = \sum_{i,j=1}^{K} \Big\|\text{Covariance}\big(\psi\big(b(z_i)\big), \psi\big(b(z_j)\big)\big)\Big\|_F^2$, and the second one is an alignment regularization that minimizes the distance of L2-Norm between the query and compressed knowledge embedding: $\mathcal{L}_{\text{align}} = \Big|1 - \frac{\sum_z \|\psi\big(b(z)\big)\|_2}{\sum_x \|b(x)\|_2}\Big|$.

3.3. Retriever

Figure 2 (c) shows REVEAL's retrieval procedure. Given the input query x, the retriever's task is to find top-K memory entries M with the highest probability $p(M \mid x)$ which we approximate as $p(M \mid x) = \prod_{m\in M} p(m \mid x)$ by retrieving each entry independently. Note that we retrieve from a large-scale unified memory $\tilde{\mathcal{M}} = [\mathcal{M}^1, \ldots, \mathcal{M}^{|\tilde{\mathcal{C}}|}]$ that is constructed from a diverse set of knowledge sources. To help the query to better choose the most appropriate knowledge sources, we learn a gating function that models the probability of retrieving from each memory corpus. With the corpus gating, for $m_i^j \in \mathcal{M}^j$ we re-weight $p(m^j \mid x)$ by

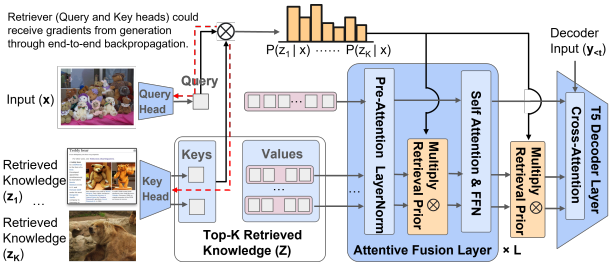

Figure 3. Detailed procedure of attentive knowledge fusion module. We inject retrieval probability as a prior to knowledge token embeddings, so the retriever can receive gradients via back-propagating over {self/cross}-attention part.

the computed corpus gating score:

$$p(m_i^j \mid x) = p(\mathcal{M}^j \mid x) \cdot p(m_i^j \mid x; \mathcal{M}^j) \tag{2}$$

$$= Gate_{\mathcal{M}^j}(x) \cdot \frac{\exp\big(Rel(x, m_i^j)/\tau\big)}{\sum_{m_k^j \in \mathcal{M}^j} \exp\big(Rel(x, m_k^j)/\tau\big)} \tag{3}$$

where $Gate_{\mathcal{M}^j}(x) = \text{Softmax}(W \cdot \texttt{Emb}_{\texttt{Query}}(x) + b)[j]$ is a softmax gating that assigns a score to each memory corpus M^j, with W and b as function parameters. $Rel(x, m_i^j)$ models relevance score between query x and each memory entry via embedding dot product, such that $Rel(x, m_i^j) = \texttt{Emb}_{\texttt{Query}}(x)^T \cdot \texttt{Emb}_{\texttt{Key}}(z_i^j)$. where z_i is the knowledge item corresponding to the memory entry m_i and τ is the temperature parameter.

After identifying the top-K memory entries, the retriever passes the pre-computed in-memory key and value embeddings to the generator. In the meantime, to support end-to-end training of the encoders, we also re-encode a small portion (i.e., 10%) of the retrieved knowledge items z_i from scratch. In this way, the memory encoders could be updated with moderate computational cost. We concatenate the re-encoded knowledge with in-memory ones to construct the final top-K retrieved key/value embeddings.

3.4. Generator

Figure 2 (d) shows how the query and the retrieved knowledge items are fused to generate the output answer. All K retrieved memory values are concatenated with the query embedding, which is feasible due to the Perceiver module utilized as the value head $\psi(\cdot)$, compressing each knowledge item into a short sequence. We denote the concatenated query embedding and memory values as $X = [b(x), \psi(b(z_1)), \ldots, \psi(b(z_K))] \in \mathbb{R}^{(I+c\cdot K)\times d}$, where I is the number of tokens of the input query x and c is the number of compressed tokens. To guide the generator towards attending to the most important items in X and facilitate backpropagation of gradients to the retriever, we propose an attentive fusion module $f(\cdot)$ capable of incorporating the retriever score as a prior for calculating

Knowledge Source	Corpus Size	Type of Text	Avg. Text Length
WIT [37]	5,233,186	Wikipedia Passage	258
CC12M [5]	10,009,901	Alt-Text Caption	37
VQA-V2 [12]	123,287	Question Answer	111
WikiData [40]	4,947,397	Linearlized Triplets	326

Table 1. Statistics of the knowledge sources used.

Model Name	T5 Variant	Image Encoder	# params.	GFLOPs
REVEAL-Base	T5-Base	ViT-B/16	0.4B	120
REVEAL-Large	T5-Large	ViT-L/16	1.4B	528
REVEAL	T5-Large	ViT-g/14	2.1B	795

Table 2. Model configuration of different REVEAL variants.

cross-knowledge attention. The detailed procedure is illustrated in Figure 3. We firstly compute a latent soft attention mask over X as $\texttt{Mask}_{\texttt{att}} = [1, p(z_1|x), \ldots, p(z_K|x)]$. Finally, we pass the fused representation $f(X, \texttt{Mask}_{\texttt{att}})$ into a T5 decoder module $g(\cdot)$ to generate the textual output.

4. Generative Pre-Training

The existing VQA datasets are not large enough for training a complex multi-component model like ours from scratch. Therefore, we pre-train our model on a massive image-text corpus. In Sec. 4.1 we go over the details of our pre-training data and objective. Then in Sec. 4.2 we introduce the various sources of knowledge used in our experiments. Finally, in Sec. 4.3 we describe the pre-training implementation details.

4.1. Pre-Training Objective

We pre-train our model on the Web-Image-Text dataset [51], a large-scale corpus containing 3 billion image alt-text caption pairs collected from the public Web. Since the dataset is noisy, we add a filter to remove data points whose captions are shorter than 50 characters. This yields roughly 1.3 billion image caption pairs for pre-training.

We denote the pre-training Web-Image-Text dataset [51] as $\mathcal{D}$. We use the text generation objective used in Wang et al. [45]) to pre-train our model on $\mathcal{D}$. Given an image-text example $x = (\texttt{img}, \texttt{txt})$ from $\mathcal{D}$, we randomly sample a prefix length T_p. We feed $x_{<T_p}$ that contains the text prefix and image to the model as input and our objective is to generate $x_{\geq T_p}$ containing the rest of the text as output. The training goal is to condition on $x_{<T_p}$ and autoregressively generate the remaining text sequence $x_{\geq T_p}$:

$$\begin{aligned} \mathcal{L}_{\text{PrefixLM}} &= -\,\mathbb{E}_{x\sim\mathcal{D}}\big[\log p(x_{\geq T_p} \mid x_{<T_p})\big] \\ &= -\,\mathbb{E}_{x\sim\mathcal{D}}\Big[\sum_{i\geq T_p} \log p(x_i \mid x_{<i})\Big]. \end{aligned} \tag{4}$$

Warm Starting the Model In order to pre-train all components of our model end-to-end, we need to warm start the retriever at a good state. Otherwise, if starting with random weights, the retriever would often return irrelevant memory items that would never generate useful training signals.

To avoid this cold-start problem, we propose to construct an initial retrieval dataset with pseudo ground-truth knowledge to give the pre-training a reasonable head start. We create a modified version of the Wikipedia-Image-Text (WIT) [37] dataset for this purpose. Each image-caption pair in WIT also comes with a corresponding Wikipedia passage (words surrounding the text). We put together the surrounding passage with the query image and use it as the pseudo ground-truth knowledge that corresponds to the input query. As the passage provides rich information about the image and caption, it definitely is useful for initializing the model. To avoid the model from relying on low-level image features for retrieval, we apply random data augmentation to the input query image. Given this modified dataset that contains pseudo retrieval ground-truth, we train the query and memory key embeddings by optimizing the following contrastive loss:

$$\mathcal{L}_{contra} = -\text{logSoftmax}(\texttt{Emb}_{\texttt{Query}}(x)^T \texttt{Emb}_{\texttt{Key}}(\hat{z}))$$

where $\hat{z}$ represents the pseudo ground-truth knowledge entry corresponding to the input query x.

4.2. Knowledge Sources

We use the following four sources of knowledge in our experiments: **Wikipedia-Image-Text (WIT)** [37] consists of the images in Wikipedia, as well as their alt-text captions and contextualized text passages. **Conceptual (CC12M)** [5] contains web images paired with alt-text captions. It includes many long-tail entities. **VQA-v2** [12] is a visual question answering dataset. We merge all question-answer pairs per image into a single passage. **WikiData** [40] is a structural knowledge graph encoding relations between Wikipedia entities. We linearize all relational triplets per entity into a textual passage following the procedure of [32]. We have listed the statistical details of these knowledge sources in Table 1.

4.3. Implementation Details

Incorporating all the components introduced above, REVEAL can be directly pre-trained over large-scale image caption datasets after proper initialization. As our model architecture is based on T5 and ViT, we use pre-trained ViT checkpoints from [50] and pre-trained T5 checkpoints from [33] to initialize the encoder parameters. The query head, key head and attentive fusion layers are initialized from upper T5, while the base text encoder is initialized from lower T5. The combination of these modules can be found in Table 2 for three three model variants, REVEAL-Base, REVEAL-Large and REVEAL, of which the largest REVEAL model has around 2 billion parameters.

VQA Model Name	Knowledge Sources	Accuracy (%)	Memory (GB)
MUTAN+AN [29]	Wikipedia + ConceptNet	27.8	-
ConceptBERT [11]	Wikipedia	33.7	-
KRISP [28]	Wikipedia + ConceptNet	38.4	-
Visual Retriever-Reader [27]	Google Search	39.2	-
MAVEx [46]	Wikipedia+ConceptNet+Google Images	39.4	-
KAT-Explicit [13]	Wikidata	44.3	1.5
PICa-Base [48]	Frozen GPT-3	43.3	350
PICa-Full [48]	Frozen GPT-3	48.0	350
KAT [13] (Single)	Wikidata + Frozen GPT-3	53.1	1.5 + 352+ 500
KAT [13] (Ensemble)	Wikidata + Frozen GPT-3	54.4	4.6 + 352 + 500
ReVIVE [24] (Single)	Wikidata + Frozen GPT-3	56.6	1.5 + 354 + 500
ReVIVE [24] (Ensemble)	Wikidata+Frozen GPT-3	58.0	4.6 + 354 + 500
REVEAL-Base	WIT + CC12M + Wikidata + VQA-2	55.2	0.8 + 7.5 + 744
REVEAL-Large	WIT + CC12M + Wikidata + VQA-2	58.0	2.8 + 10 + 993
REVEAL	WIT + CC12M + Wikidata + VQA-2	**59.1**	4.2 + 10 + 993

Table 3. **Visual Question Answering** results on OK-VQA, compared with existing methods that use different knowledge sources. For the memory cost, we assume all models use bfloat16. Green means on-device model parameters that are learnable, Blue means on-device memory of frozen model parameters, and Red means CPU/disk storage cost that are not involved in computation.

Distributed Online Retrieval. Finding the top-k most-relevant knowledge entries is a standard Maximum Inner Product Search (MIPS) problem. There are approximate search algorithms [8, 36] that scale sub-linearly with the size of the knowledge corpus $|C|$. We use TPU-KNN [8] to conduct distributed MIPS search, by splitting and storing the memory embeddings across all training devices. The query is synced to each device, which retrieves approximate top-K results from its own memory. Then these results are combined to compute the global top-K retrieved items.

Pre-Training Pipeline. We first train the multimodal retriever on our modified version of the Wikipedia Image Text (WIT) dataset via $\mathcal{L}_{contra}$. We use the Adafactor optimizer without momentum ($\beta_1 = 0$, $\beta_2 = 0.999$), with weight decay of 0.001[2], and with a peak learning rate of $6e4$, to train for 10 epochs. We use this checkpoint to warm-start our generative pre-training. We set the number of retrieved knowledge entries as $K = 10$ during pre-training, and use adafactor with a peak learning rate of $1e{-}3$ and inverse squared root learning rate scheduler with 10,000 linear warm-up steps. We use $\mathcal{L}_{PrefixLM}$ as the main objective, adding $\mathcal{L}_{contra}$, $\mathcal{L}_{decor}$ and $\mathcal{L}_{align}$ weighted by 0.01. We use a batch size of 4096 across 256 CloudTPUv4 chips and train for about 5 days.

5. Experimental Results

We evaluate our proposed method on knowledge-based VQA in Sec. 5.1 and image captioning in Sec. 5.2. We then conduct ablation studies in Sec. 5.3 to analyze the impact of each model component on overall performance.

[2]The remaining experiments use the same optimizer configuration.

VQA Model Name	Accuracy (%)
ViLBERT [26]	30.6
LXMERT [38]	30.7
ClipCap [30]	30.9
KRISP [28]	33.7
GPV-2 [21]	48.6
REVEAL-Base	50.4
REVEAL-Large	51.5
REVEAL	**52.2**

Table 4. **Visual Question Answering** results on A-OKVQA.

5.1. Evaluating on Knowledge-Based VQA

One of the most knowledge intensive visual-language tasks is knowledge-based visual question answering (VQA), exemplified by the OK-VQA [29] and A-OKVQA [35] benchmarks. To finetune our pre-trained model on these VQA tasks, we use the same generative objective where the model takes in an image question pair as input and generates the text answer as output. There are a few differences between the fine-tuning and the pre-training stages: 1) we set the number of retrieved knowledge entries to $K = 50$, so the model is able to retrieve sufficient supporting evidence; 2) we freeze the whole base V-L encoder to stabilize training; and 3) we use a batch size of 128, with the Adafactor optimizier, a peak learning rate of 1e-4. We use the soft VQA accuracy metric [3] to evaluate the model's generated answer.

Our results on OKVQA and A-OKVQA datasets are shown in Table 3 and Table 4 respectively. For OKVQA, earlier attempts that incorporate a fixed knowledge retriever report results that are below 45%. Recently a series of works utilize large language models (*e.g.* GPT-3) as implicit knowledge sources, which achieve much better performance

Model Name	MSCOCO	NoCaps	# params.
Flamingo [2]	138.1	-	80B
VinVL [52]	140.9	105.1	0.4B
SimVLM [45]	143.3	112.2	1.5B
CoCa [49]	143.6	122.4	2.1B
REVEAL-Base	141.1	115.8	0.4B
REVEAL-Large	144.5	121.3	1.4B
REVEAL	**145.4**	**123.0**	2.1B

Table 5. **Image Captioning** results on MSCOCO (Karpathy-test split) and NoCaps (val set). Evaluated using the CIDEr metric.

with the trade-off of a huge computational cost. REVEAL achieves higher performance than those methods without relying on such large language models[3]. Compared with the previous state-of-the-art, KAT and ReVIVE, which also utilizes T5-Large as a generator, REVEAL achieves accuracy of 59.1%, which is +6.0% higher than the single KAT [13] model and +2.5% higher than ReVIVE [24].

On A-OKVQA, REVEAL achieves 52.2% accuracy, which is +3.6% higher than the previous best, GPV-2 [21]. We also show two examples of these datasets in Figure 4. All these results show that, with proper end-to-end retrieval training and a diverse set of knowledge sources, REVEAL can learn to retrieve meaningful knowledge entries, and achieve promising results without relying on a large language model.

5.2. Evaluating on Image Captioning

We also evaluate REVEAL on image captioning benchmarks: MSCOCO Captions [6] and NoCaps [1]. We follow the evaluation protocol used in [49]. We directly fine-tune our generator model on the MSCOCO training split via cross-entropy generative objective. We measure our performance on the MSCOCO test split and NoCaps val set with the CIDEr metric [39]. The results of these two datasets are shown in Table 5. Note that REVEAL achieves better results than strong recent baselines such as SimVLM [45] and CoCa [49] on both benchmarks. Notably, REVEAL −Large with 1.4B parameters outperforms the 2.1B-parameter CoCa model and is significantly better than 80B-parameter Flamingo model [2].

5.3. Analyzing Effects of Key Model Components

In the following we study which design choices contribute most to the model's performance. We focus on three research questions: (1) Does utilizing multiple knowledge sources enhance performance? (2) Does the proposed attentive fusion surpass existing end-to-end retrieval training methods?

[3]As shown in the last column of Table 3, REVEAL stores external knowledge as value embeddings on disk, occupying 993GB of space. The key embeddings consume 10GB space and are kept in TPU memory for fast lookup. On the other hand, KAT and REVIVE need to load the entire 350GB GPT-3 model in the GPU/TPU memory. Furthermore, storing WikiData on disk consumes 500GB of disk memory.

Figure 4. **VQA Examples**. REVEAL is able to use knowledge from different sources to correctly answer the question. We show more examples in Figure 1-3 of Supplementary Material, indicating that our model can retrieve and use items from diverse knowledge sources to correctly solve different input query.

(3) Can we add knowledge by only updating the memory without modifying model parameters?

Analyzing multiple knowledge sources. A major distinction of REVEAL compared to previous retrieval-augmented approaches is its capacity to utilize a diverse set of knowledge sources during inference. To assess the relative importance of each data source and the efficacy of retrieving from various corpora, we conduct two ablation studies: 1) **Only-One-Left**: employing a single knowledge source to evaluate the outcomes; and 2) **Leave-One-Out**: excluding one knowledge source from the complete set $\mathcal{C}$. These ablation studies are executed using the REVEAL $_{\text{Base}}$, evaluated on the OKVQA validation set under the aforementioned conditions. As shown in Figure 5, among the four knowledge sources utilized in this paper, WIT is the most informative, with the highest accuracy when used in isolation (53.1The remaining three corpora, CC12M, VQA-v2, and WikiData, do not offer the same level of informativeness as WIT when utilized independently. However, excluding any of these corpora from the complete dataset results in performance decreases of 1.3%, 0.6%, and 1.1%, respectively. This observation implies that these knowledge sources effectively complement one another, contributing valuable information to enhance performance. To further substantiate this hypothesis, we perform an additional experiment involving pairs of knowledge sources, as illustrated in Figure 6. Notably, even when paired with an informative knowledge source such as WIT, incorporating an extra corpus consistently leads to performance improvements.

Analyzing different retrieval training methods. Another core component of REVEAL is the attentive fusion layer, which supports efficient joint training of the retriever and generator. We investigate its performance compared to two existing retrieval training method categories: 1) a frozen retriever based on ALIGN [20] representations ; 2) end-to-end retrieval training methods including **Attention Distill** [17], **EMDR**^{2} [47], and **Perplexity Distill** [18].

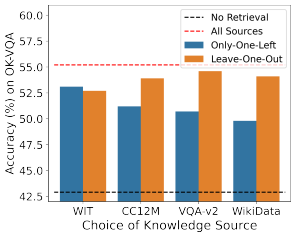

Figure 5. OKVQA Accuracy of REVEAL using 1) **Only-One-Left**: only use a single knowledge source; 2) **Leave-One-Out**: use all without this knowledge source.

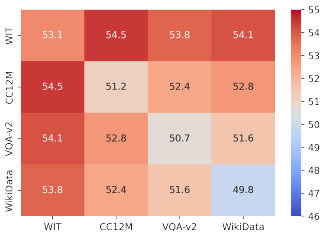

Figure 6. OKVQA Accuracy of REVEAL using all **Pair of Knowledge Sources**. Results show that combining multiple sources could consistently improve performance.

Retrieval Method	Acc@10	Acc@100	OKVQA Acc.	GFLOPs
ALIGN [20] (fixed)	0.638	0.793	44.7	-
Attention Distill [17]	0.674	0.835	45.9	119
EMDR^2 [47]	0.691	0.869	46.5	561
Perplexity Distill [18]	0.704	0.886	46.7	561
Ours (Attentive Fusion)	0.726	0.894	47.3	120

Table 6. **Analysis of Retrieval Training Method**: We train REVEAL-Base (frozen generator, only train randomly initialized retriever) to retrieve from the WIT dataset (only text passage without image), and show the retrieval accuracy at the first 10 or 100 results, as well as fine-tuned OKVQA accuracy.

We use the pre-trained REVEAL-Base model, fix the generator and randomly initialize the retriever (query head and key head). We utilize our modified version of WIT dataset with pseudo ground-truth retrieved labels as the evaluation corpus. We evaluate retrieval performance by checking whether the correct passage appears in top-10/100 results. For the ALIGN model, we directly evaluate the retrieval results from the pre-trained checkpoint, while for other models, we perform retrieval-augmented training on the WIT dataset. To prevent the model from relying on image similarity for accurate results, we only use text passages as knowledge entries and discard images. Subsequently, we finetune the model on OKVQA and report its accuracy. The results are presented in Table 6. We observe that directly using pre-trained encoder does not perform well, even with a strong model like ALIGN. Moreover, among the various end-to-end retrieval training approaches, our attentive fusion method attains better accuracy in both retrieval and OKVQA tasks. Importantly, our technique exhibits a computational cost (quantified by GFLOPs) comparable to that of attention distillation, yet significantly lower than EMDR^2 and Perplexity distillation. This indicates that our proposed method is more efficient and effective for pre-training retrieval-augmented visual-language models.

Analyzing Knowledge Modification. One advantage of utilizing knowledge memory is that we could easily add or update knowledge entries without re-training model's parameters. To validate this, we conducted ablation studies in which we removed a specific percentage of knowledge entries from the corpora and assessed the performance of the REVEAL-Base model on the OKVQA dataset. Subsequently, we add the removed knowledge back into the corpora, allowing the trained model to make predictions using the complete set of corpora. This approach ensured that the removed knowledge was not seen by the model during fine-tuning, enabling us to test its ability to accurately retrieve and utilize that knowledge for problem-solving.

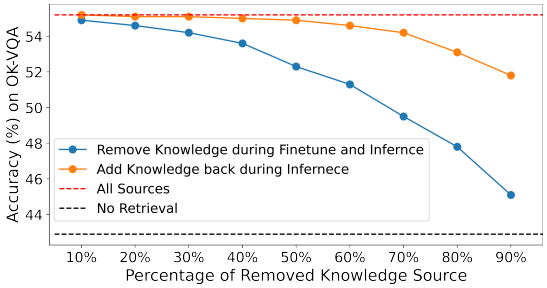

Figure 7. **Study of Knowledge Update**. The blue curve shows result by removing certain percentage of knowledge during both fine-tuning and inference stage. The orange curve shows results by still first removing the knowledge, and then adding the knowledge back during inference, which simulates the knowledge update.

The results are illustrated in Figure 7, with the blue curves representing the inference outcomes without the removed knowledge and the orange curve depicting the results after adding the removed knowledge back. A notable performance improvement was observed upon reintroducing the knowledge (orange curve) compared to the outcomes with the removed knowledge (blue curve). Specifically, for the model fine-tuned with only 10% of the knowledge, the reintroduction of the removed knowledge resulted in an accuracy of 51.8 (+6.7 higher than when removed). This finding demonstrates that the REVEAL model can swiftly adapt to new knowledge by merely updating the memory, obviating the need for re-training model parameters.

6. Conclusion

This paper presents an end-to-end Retrieval-augmented Visual Language model (REVEAL), which contains a knowledge retriever that learns to utilize a diverse set of knowledge sources with different modality. The retriever is trained jointly with the generator to return multiple knowledge entries. We pre-train REVEAL on a massive image-text corpus with four diverse knowledge corpora, and achieves state-of-the-art results on knowledge-intensive visual question answering and image caption tasks. In the future we'd explore the ability of this model to be used for attribution, and applying it to broader class of multimodal tasks.

References

[1] Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE, 2019. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. 1, 7

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015. 6

[4] 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. 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*, 2020. 1, 2

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 5

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325, 2015. 7

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *CoRR*, abs/2209.06794, 2022. 1

[8] Felix Chern, Blake Hechtman, Andy Davis, Ruiqi Guo, David Majnemer, and Sanjiv Kumar. TPU-KNN: K nearest neighbor search at peak flop/s. *CoRR*, abs/2206.14286, 2022. 6

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. 1, 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[11] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online, 2020. Association for Computational Linguistics. 2, 6

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. 1, 5

[13] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 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, Seattle, United States, 2022. Association for Computational Linguistics. 1, 2, 6, 7

[14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. *ArXiv preprint*, abs/2002.08909, 2020. 1, 2, 4

[15] Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. Empowering language models with knowledge graph reasoning for open-domain question answering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates,*

December 7-11, 2022, pages 9562–9581. Association for Computational Linguistics, 2022. 2

[16] Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed H. Chi. Improving multi-task generalization via regularizing spurious correlation. *CoRR*, abs/2205.09797, 2022. 4

[17] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 7, 8

[18] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *CoRR*, abs/2208.03299, 2022. 1, 2, 7, 8

[19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 2021. 4

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 7, 8

[21] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 662–681. Springer, 2022. 6, 7

[22] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 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, 2020. Association for Computational Linguistics. 2

[23] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4

[24] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. REVIVE: regional visual representation matters in knowledge-based visual question answering. *CoRR*, abs/2206.01201, 2022. 2, 6, 7

[25] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6949–6959. IEEE, 2022. 1

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. 6

[27] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2, 6

[28] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14111–14121. Computer Vision Foundation / IEEE, 2021. 2, 6

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 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, 2019. 2, 6

[30] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *ArXiv preprint*, abs/2111.09734, 2021. 6

[31] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 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, 2018. 2

[32] Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. Unified open-domain question answering with structured and unstructured knowledge. *ArXiv preprint*, abs/2012.14610, 2020. 5

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. 1, 3, 5

[34] Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 25968–25981, 2021. 2

[35] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. *CoRR*, abs/2206.01718, 2022. 2, 6

[36] Anshumali Shrivastava and Ping Li. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2321–2329, 2014. 6

[37] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM, 2021. 1, 5

[38] Hao Tan and Mohit Bansal. 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, 2019. Association for Computational Linguistics. 6

[39] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015. 7

[40] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. 1, 5

[41] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296. ijcai.org, 2017. 2

[42] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, 2018. 2

[43] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022. 1

[44] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. VQA-GNN: reasoning with multimodal semantic graph for visual question answering. *CoRR*, abs/2205.11501, 2022. 2

[45] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 5, 7

[46] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2712–2721. AAAI Press, 2022. 2, 6

[47] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 7, 8

[48] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. *ArXiv preprint*, abs/2109.05014, 2021. 2, 6

[49] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022. 1, 7

[50] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1204–1213. IEEE, 2022. 5

[51] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18102–18112. IEEE, 2022. 5

[52] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. 7