Answer Mining from a Pool of Images: Towards Retrieval-Based Visual Question Answering

Abhirama Subramanyam Penamakuri[1] , **Manish Gupta**[2] , **Mithun Das Gupta**[2] , **Anand Mishra**[1]
[1]Indian Institute of Technology Jodhpur
[2]Microsoft
{penamakuri.1, mishra}@iitj.ac.in, {gmanish,migupta}@microsoft.com

Abstract

We study visual question answering in a setting where the answer has to be mined from a pool of relevant and irrelevant images given as a context. For such a setting, a model must first retrieve relevant images from the pool and answer the question from these retrieved images. We refer to this problem as retrieval-based visual question answering (or RETVQA in short). The RETVQA is distinctively different and more challenging than the traditionally-studied Visual Question Answering (VQA), where a given question has to be answered with a single relevant image in context. Towards solving the RETVQA task, we propose a unified Multi Image BART (MI-BART) that takes a question and retrieved images using our relevance encoder for free-form fluent answer generation. Further, we introduce the largest dataset in this space, namely RETVQA, which has the following salient features: multi-image and retrieval requirement for VQA, metadata-independent questions over a pool of heterogeneous images, expecting a mix of classification-oriented and open-ended generative answers. Our proposed framework achieves an accuracy of 76.5% and a fluency of 79.3% on the proposed dataset, namely RETVQA and also outperforms state-of-the-art methods by 4.9% and 11.8% on the image segment of the publicly available WebQA dataset on the accuracy and fluency metrics, respectively.

1 Introduction

Question Answering (QA) over textual as well as visual data has been an active area of research [Hsu *et al.*, 2021; Guo *et al.*, 2023]. In text-based QA, the research focus has recently shifted from highly-explored QA on a single paragraph such as SQuAD [Rajpurkar *et al.*, 2018] to a setting where mining answers from a huge corpus of documents is a requirement [Ahmad *et al.*, 2019; Hsu *et al.*, 2021]. On the contrary, visual question answering (VQA) [Antol *et al.*, 2015] literature has so far largely restricted itself to answering questions about a given relevant visual context (often a single image). However, this does not necessarily suffice to satisfy our information needs since the information may be spread across multiple images and may not be present in some images. For example, consider a natural language question '*Do the rose and sunflower share the same color?*', answering such a question from a pool of images as visual context (refer Figure 1), requires a model to first retrieve relevant images and then perform visio-lingual reasoning on the retrieved images to arrive at a fluent free-form natural language answer. We refer to this problem as RETVQA or retrieval-based visual question answering. The RETVQA setting has potential applications in question answering on web images, e-commerce, environmental monitoring, and health care, among others, e.g., multiple images of a particular area can be analyzed to monitor environmental changes over time; multiple MRI or CT scans of a patient's brain need to be analyzed to detect abnormalities, such as tumors.

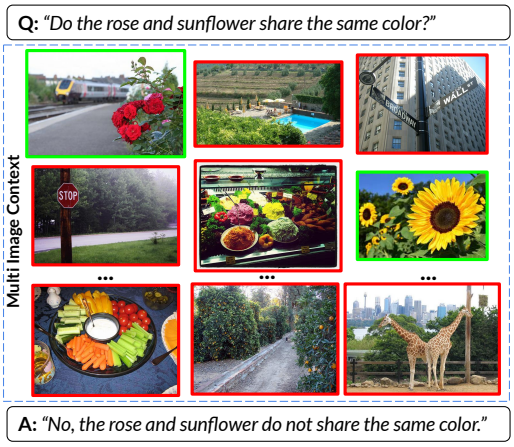

Figure 1: Given a question and a pool of images (multi-image context), RETVQA task involves two stages: (i) retrieve the relevant images from the pool, and (ii) generate a free-form natural language answer by reasoning over the retrieved relevant images as context.

For RETVQA, the input is a pool of images with only a few images being relevant to the question. Close to our setting, there is some exciting progress in the recent literature [Talmor *et al.*, 2021; Bansal *et al.*, 2020; Singh *et al.*, 2021; Chang *et al.*, 2022]. However, these works assume one or more of the following constraints: "without requiring ex-

Measurement	Value
#Distinct questions	418K
#Distinct precise answers	16,205
Train set questions	334K (80%)
Val set questions	41K (10%)
Test set questions	41K (10%)
Avg question length (words)	8.7
Avg answer length (words)	8.5
#Distinct words in questions	10,868
#Distinct words in answers	9,278
#Avg relevant images per question	2
#Avg irrelevant images per question	24.5

Table 1: Key statistics for RETVQA dataset.

Question Category	Binary	Open-ended	Total
Color	50K	50K	100K
Shape	49K	50K	99K
Count	50K	50K	100K
Object-attributes	80K	-	80K
Relation-based	-	38K	38K
Total	229K	188K	418K

Table 2: Distribution of questions by various categories in RETVQA dataset. The answers are of two types: binary-generative and open-ended generative.

plicit retrieval", "having classification-type fixed-vocabulary answers", "assuming the availability of meta-data like *Wiki-Entities, captions*", "having a homogeneous yet limited number of images in the pool", and "having only a small set of questions that need multiple images". Such constraints in the existing datasets point towards a need for a large-scale benchmark to study RETVQA. To this end, we present a *derived* dataset prepared from Visual Genome [Krishna *et al.*, 2017], leveraging its questions and annotations of images. We curate questions under different categories: (i) common attributes such as color, shape, and count, (ii) other object-attributes that include non-common attributes, e.g. length, material, and (iii) subject-object relationships, e.g., 'eats', 'left of'. Further, to facilitate benchmarking capabilities of the VQA models over open-ended answers, we curate questions under binary (yes/no) and open-ended answer categories. Note that the answers are free-form fluent in both the answer categories, e.g. '*No, rose and sunflower do not share the same color*' (a binary answer); '*The color of rose and sunflower is red and yellow, respectively*' (an open-ended generative answer). RETVQA dataset statistics and distribution across the question-answer categories are shown in Table 1 and Table 2, respectively.

Further, to solve the RETVQA task, a model must first retrieve the relevant images for the question and then consume the retrieved images as the context to answer the question. Towards this end, we present a unified Multi Image BART that takes in the question along with the multi-image context retrieved using a relevance encoder to generate the free-form fluent natural-language answer. Our proposed framework, MI-BART, allows joint reasoning over multiple retrieved images along with the question to capture better semantics.

To summarize, our contributions are as follows: (i) We present RETVQA, a 20× larger dataset than the closest dataset [Chang *et al.*, 2022] in this setting. RETVQA dataset is prepared by leveraging questions and image annotations from Visual Genome. It emphasizes on multi-image, metadata-independent questions over a pool of heterogenous collections of images, expecting a mix of classification-oriented and generative answers. We strongly believe that the proposed task, curated dataset and benchmarks presented in this paper will pave the way for further research. (ii) We present Multi Image BART (MI-BART) - a unified method that reasons jointly over the retrieved multi-image context along with the question to generate a free-form fluent answer for the question. (iii) We perform extensive experiments to evaluate the performance of our proposed framework on RETVQA and the image segment of WebQA. Our approach clearly outperforms baseline approaches on RETVQA dataset and achieves state-of-the-art performance on the image segment of WebQA. We make our data and implementation publicly available.[1]

2 Related Work

Visual and Multi-modal QA. Visual Question Answering (VQA) aims at answering a natural language question in the context of a relevant image [Antol *et al.*, 2015]. This area has seen significant progress partly due to the introduction of several challenging datasets [Malinowski *et al.*, 2015; Ren *et al.*, 2015a; Zhu *et al.*, 2016; Antol *et al.*, 2015; Goyal *et al.*, 2017; Johnson *et al.*, 2017; Krishna *et al.*, 2017]. Most methods for VQA either use a multimodal fusion of language and image embeddings [Ren *et al.*, 2015a; Gao *et al.*, 2015; Noh *et al.*, 2016; Kembhavi *et al.*, 2017], attention-based multimodal fusion [Yang *et al.*, 2016; Fukui *et al.*, 2016; Shih *et al.*, 2016; Lu *et al.*, 2016; Xiong *et al.*, 2016] or neural module networks [Andreas *et al.*, 2016; Hu *et al.*, 2017]. More recently, knowledge-based VQA [Shah *et al.*, 2019; Marino *et al.*, 2019] has gained attention where external knowledge is used for answering visual questions. Contrary to these exciting works in VQA literature, our problem setting is distinctively different as we need to mine the answer from a collection of relevant as well as irrelevant images.

Sharing a similar motivation as ours, the following tasks and accompanying datasets have been recently introduced in the literature: (i) MultimodalQA [Talmor *et al.*, 2021], (ii) ISVQA [Bansal *et al.*, 2020], and (iii) WebQA [Chang *et al.*, 2022]. In MultimodalQA [Talmor *et al.*, 2021], only a small part of the dataset (ImageListQ) is relevant to our setting; however, even on this subset, MultimodalQA assumes the availability of extra image metadata, i.e., table or Wiki-Entity linkage. Similarly, in ISVQA [Bansal *et al.*, 2020], every question has a small set of homogeneous images as context. Since images are homogeneous, there is no need for explicit retrieval. Recently, WebQA [Chang *et al.*, 2022] dataset has been proposed to target such practical QA scenarios, where a question has to be answered in the context of multimodal sources; however, the images in the dataset have associated captions. All of these settings have differences from RETVQA in either usage of additional context, constraints on images in the collection, answer schema, or classification instead of generation. In another recent work

[1] https://vl2g.github.io/projects/retvqa/

Dataset	#Questions	Retrieval required	Heterogenous images	Multi-image reasoning	No Meta-data assumption	Answer type	% of questions that need multiple images
MultimodalQA [Talmor *et al.*, 2021]	2K	✗	✗	✓	✗ (WikiEntities)	Classification	6.1%
ISVQA [Bansal *et al.*, 2020]	141K	✗	✗	✓	✓	Classification	33%
WebQA [Chang *et al.*, 2022]	18K	✓	✓	✓	✗ (Captions)	Open-ended	44%
RETVQA (Ours)	**327K**	✓	✓	✓	✓	Open-ended	**100%**

Table 3: Comparison of our curated dataset RETVQA with other relevant QA datasets. For Multimodal QA and WebQA datasets, we have considered their image-only modality questions subset.

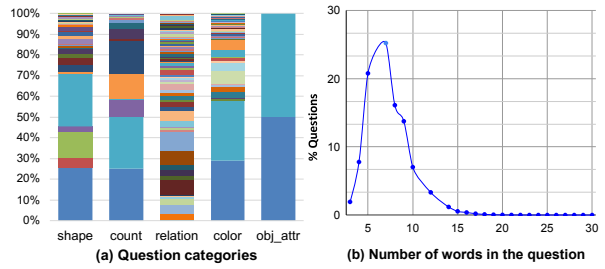

Figure 2: RETVQA questions and answers analysis: (a) Answer distribution over various question categories, (b) Distribution of the number of words across questions.

Figure 3: Word cloud of Top-80 frequent answers.

MIMOQA [Singh *et al.*, 2021], only extractive question answering is performed. In terms of reasoning on more than one image, another related work is NLVR [Suhr *et al.*, 2019]. However, it does not involve any retrieval and open-ended answer generation. Further, in terms of QA over multiple document images or video frames, related works are DocVQA [Tito *et al.*, 2021] and VideoQA [Lei *et al.*, 2018; Lei *et al.*, 2020; Tapaswi *et al.*, 2016]. DocVQA focuses on text-heavy document images, limiting visual reasoning, whereas the VideoQA task does not involve explicit retrieval and open-ended answer generation. Contrary to these works, our newly curated dataset is significantly larger, has no assumption of meta-data availability, and requires retrieval and reasoning over multiple images to arrive at an answer.

Multi-modal modeling. Recently, multi-modal transformers such as VisualBERT [Li *et al.*, 2019], VilBERT [Lu *et al.*, 2020], VILT [Kim *et al.*, 2021], LXMERT [Tan and Bansal, 2019], OSCAR [Li *et al.*, 2020], UNITER [Chen *et al.*, 2020] have shown strong results on the downstream vision and language tasks. However, these encoder-based models are more suited for classification-style VQA settings. Multimodal transformers like VLP [Zhou *et al.*, 2020] and VLBart [Cho *et al.*, 2021] are pre-trained with sequence-to-sequence objectives and hence are more suitable for the current setting that requires the model to generate free-form natural language answers. We follow a similar approach by devising an encoder-decoder framework to jointly reason over multiple images along with the question.

3 RETVQA Dataset

Traditionally, VQA datasets [Antol *et al.*, 2015; Goyal *et al.*, 2017; Singh *et al.*, 2019; Talmor *et al.*, 2021] assume that the context provided is always relevant to the question. Recently, [Chang *et al.*, 2022] proposed a benchmark where given a question and a pool of multimodal sources (containing both image and text snippets as context), only a few of these sources are relevant to the question. Similar to our problem setup, it requires first retrieving the relevant context and then using it to answer the question. However, after carefully observing the WebQA dataset, we found that most questions include rare entities like 'Maracana Stadium', 'Minnetonka Rhododendron', etc. Such questions make the retrieval task of the problem non-trivial without auxiliary information about these images. Methods proposed in [Chang *et al.*, 2022] leverage metadata like image captions, which contain information like the name of the entity in the image. Further, a rule-based retrieval using word overlap of the question with the image caption for the retrieval task has an F1 score of 37, asserting our claim that retrieval is over-dependent on image metadata and not significantly on visual data. Further, the image-based subset of WebQA has a majority of samples (55.6%) with one relevant image per question, thereby, a single image VQA method may perform equally well given the relevant image is retrieved.

To overcome such limitations, we curate a dataset RETVQA from Visual Genome, where we emphasize multi-image, metadata-independent questions over a pool of heterogeneous images, expecting a mix of classification-oriented and open-ended generative answers. We leverage question-answer and object annotations of Visual Genome to curate the dataset. We curate truly multi-image questions spanning over five different categories, namely, color, shape, count, object-attributes, and relation-based. For each question category, we curate binary-generative and open-ended generative answers. Dataset statistics are shown in Tables 1 and 2.

The questions in RETVQA are curated as follows. We start by extracting subjects and relations of the existing question-

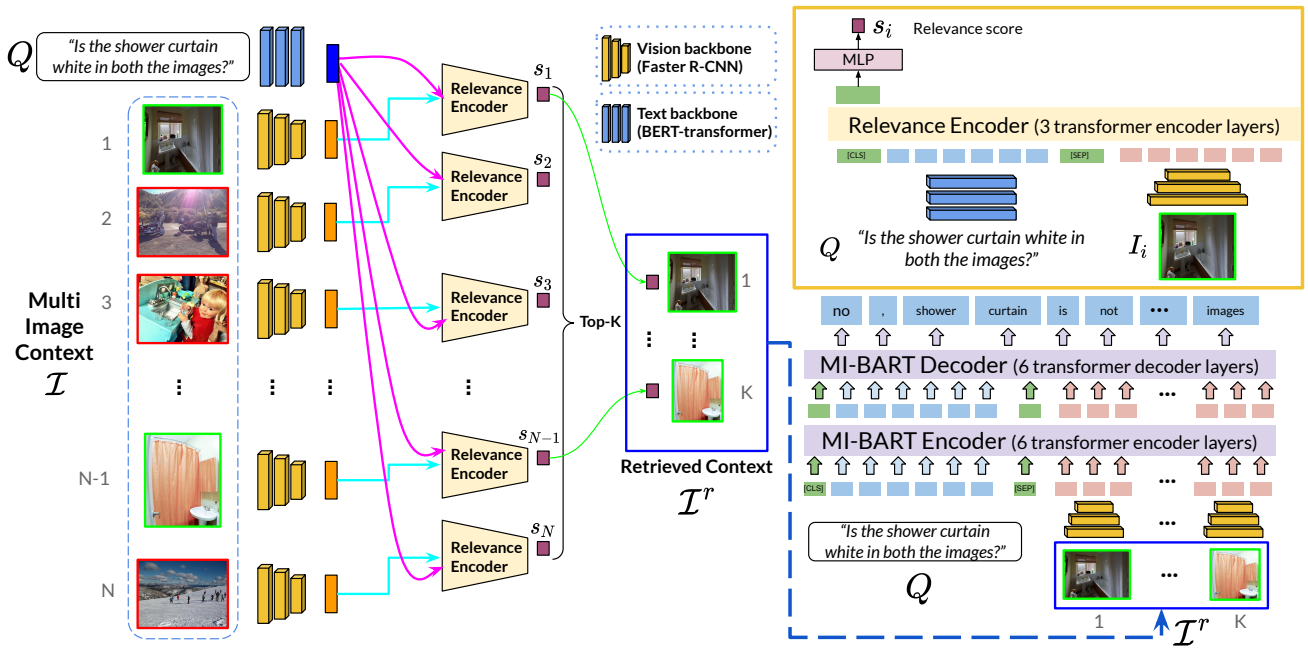

Figure 4: An overview of our proposed framework for retrieval-based VQA. Given a question Q and a pool of images $\mathcal{I}$, we encode the question and each image using a pretrained BERT and a pretrained Faster R-CNN, respectively. Once encoded, our multimodal relevance encoder (shown in the yellow box at the top right) generates relevance scores S for all images in the set with the question. We choose top-K scoring images as the retrieved relevant images $\mathcal{I}^r$. We encode the images in $\mathcal{I}^r$ using Faster R-CNN and feed them to our MI-BART encoder along with Q to facilitate joint reasoning over the multi-image context with respect to the question. Once the MI-BART encoder encodes the question in the context of retrieved images, the MI-BART decoder generates the free-form natural language answer A to the question.

answer pairs from Visual Genome; for example, consider these two question-answer pairs: q_1 (over image I_1): "What is the cow eating in the image?" where the answer is a_1: "grass"; and, q_2 (over image I_2): "what is the sheep eating?" where the answer is a_2: "grass". Given q_1 and q_2, we extract subjects ("cow" and "sheep"), relations ("eating (eat/eats)"), and then we frame combined questions using templates (over images I_1 and I_2) as follows. q_3: "what else eats the same thing as cow does?" with answer a_3: "sheep eats the same thing as cow". Another question could be q_4: "Does cow and sheep eat the same thing?" where the answer is a_4: "Yes, cow and sheep eat the same thing". Thus, we curate binary-generative (like q_4) and open-ended generative (like q_3) types of answers. We further associate negative images for each of the curated questions using their object annotations as follows. A negative image is one where both the subject and object (used to generate the question) do not exist together in the image. This enforces that the answer has to be inferred only when all the relevant images are correctly retrieved and the negatives serve as sufficiently hard negatives.

We use a random 80%-10%-10% train-validation-test split. All the questions in our dataset have at least two relevant images and 24.5 irrelevant images on average. A comparison with the other relevant datasets is shown in Table 3. Further, Figure 2 shows the distribution of unique answers across question types and question length distribution. Figure 3 shows the word cloud of top-80 frequent answers. We observe that most questions are in the 5–10 words range, and there is no noticeable bias towards the majority of answers in the dataset.

4 Retrieval QA Methodology

4.1 RETVQA Problem Formulation

The RETVQA problem is defined as follows. Given a natural language question Q, a set of N heterogeneous images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, the task is to generate an answer (A) for the question Q based on $\mathcal{I}$ where only a few images are relevant for the question. To answer the question Q using $\mathcal{I}$, we need a method that retrieves the relevant images $\mathcal{I}^r \subseteq \mathcal{I}$ and then leverages the retrieved context $\mathcal{I}^r$. Accordingly, our labelled dataset consists of quadruplets $(Q, \mathcal{I}, \mathcal{I}^r, A)$.

4.2 RETVQA Framework

The proposed framework solution: (i) multi-modal relevance encoder for retrieval of relevant sources $\mathcal{I}^r$ from $\mathcal{I}$ for the given question Q and (ii) a unified Multi Image BART (MI-BART) to generate fluent free-form natural language answer for the question Q using the retrieved images $\mathcal{I}^r$ as context.

Image representation. Inspired by recent vision-language pretraining literature [Li *et al.*, 2019; Chen *et al.*, 2020; Li *et al.*, 2020], for every image I_i in $\mathcal{I}$ where $i \in \{1, 2, \ldots, N\}$, we first detect a fixed set of $\mathcal{P}$ objects using Faster R-CNN [Ren *et al.*, 2015b] pretrained on Visual Genome [Krishna *et al.*, 2017]. For every object p, where $p \in \{1, 2, \ldots, \mathcal{P}\}$, we obtain 2048-dimensional regional feature $\mathbf{o}_p^{reg}$ and 4-dimensional bounding box co-ordinates

$\mathbf{o}_p^{bbox}$. Thereby, each image I_i is represented by a set of $\mathcal{P}$ object proposals $\{(\mathbf{o}^{reg}, \mathbf{o}^{bbox})_p\}_i$. Following [Li *et al.*, 2019], for every region p we project both 2048-dimensional regional representation and 4-dimensional bounding box coordinates into the d-dimensional space using a linear projection to obtain $\{\mathbf{o}_p\}_i$ and then concatenate across all regions within the image to obtain image embedding $\mathbf{o}_i$ as follows.

$$\mathbf{o}_i = \{\mathbf{o}_p\}_i, \text{ where } p \in \{1, 2, \ldots, P\}, i \in 1, 2, \ldots, N. \quad (1)$$

Question representation. We encode the textual question Q containing M words using a pretrained BERT [Devlin *et al.*, 2019]. This results into a sequence $\mathbf{q}$ of M d-dimensional vectors, $\mathbf{q} = \{\mathbf{q_m}\}$ where $m \in \{1, 2, \ldots, M\}$. Note that if any additional metadata is available (e.g. captions in WebQA dataset), we augment it to the question.

$$\mathbf{q} = \{\mathbf{q_m}\} = BERT(Q), \text{where } m \in \{1, 2, \ldots, M\}. \quad (2)$$

4.3 Multimodal Relevance Encoder for Image Retrieval

Pretraining. Our multi-modal Relevance Encoder (RE) consists of three transformer encoder layers followed by a multi-layered perceptron (MLP) with a sigmoid unit over the final representation of the $[CLS]$ token. We pretrain our relevance encoder on MS-COCO [Lin *et al.*, 2014] using two unsupervised objectives, Image Text Matching (ITM) and Masked Language Modelling (MLM) similar to [Li *et al.*, 2019].

Question-Image relevance learning. Each sample in our dataset contains a question Q and N images $I_1, I_2, \ldots, I_N$ of which some have been labelled as positive and others negative. Further, for each image, we have P regions. We use each question-image pair (Q, I_i) to learn question-image relevance using our multi-modal Relevance Encoder (RE). Our pretrained multi-modal relevance encoder is fed with question-image pairs, along with two special tokens, $[CLS]$ and $[SEP]$; in short, the input to our relevance encoder is $[[CLS], \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}\}, [SEP], \{\mathbf{o_1}_i, \mathbf{o_2}_i, \ldots, \mathbf{o_p}_i\}]$. Our encoder then allows the input $M + P + 2$ token sequence of question-image features to attend to each other and produces a sequence of contextualized embeddings. The d-dimensional contextualized embedding of $[CLS]$ token is further fed to an MLP with a sigmoid unit to produce a relevance score ($\hat{s}_i$) between 0 and 1 (Eq. 3), indicating whether the given question-image pair (Q, I_i) is relevant or not. We finetune our multi-modal relevance encoder parameters ϕ by minimizing binary cross-entropy loss $\mathcal{L}_{REL}(\phi)$ (Eq. 4).

$$\hat{s}_i = RE_\phi(Q, I_i). \quad (3)$$

$$\mathcal{L}_{REL}(\phi) = -\mathbb{E}_{(Q,I_i)\sim D}[s_i \log(\hat{s}_i) + (1 - s_i)\log(1 - \hat{s}_i)]. \quad (4)$$

Given a question Q and a set of N images $\mathcal{I}$ sampled from our dataset D, we obtain relevance scores $S = \{\hat{s}_i\}_{i=1}^N$ for each question-image pair $(Q, \{I_i\}_{i=1}^N)$ using our fine-tuned relevance encoder (Eq.5). To choose the final set of relevant images $\mathcal{I}^r$ from the pool of images $\mathcal{I}$, we rank all the images in the pool using S and choose top-K images as our relevant context $\mathcal{I}^r$ for the given question Q (Eq. 6).

$$S = \{\hat{s}_i\}, \text{ where } \hat{s}_i = RE(Q, I_i), i \in 1, \ldots, N. \quad (5)$$

$$\mathcal{I}^r = \{I_k\} \text{ where } k \in \text{top-}K(S). \quad (6)$$

4.4 Multi Image BART for Question Answering

Given the question and the retrieved images $\mathcal{I}^r$, the goal of MI-BART is to generate an accurate yet fluent free-form natural language answer for the question. Towards this end, we propose an encoder-decoder architecture similar to SimVLM [Wang *et al.*, 2022]. MI-BART encoder is a stack of six transformer layers [Vaswani *et al.*, 2017], where each transformer layer comprises a self-attention layer, followed by a fully connected linear layer with a residual connection. Similarly, the MI-BART decoder is also a stack of six transformer layers [Vaswani *et al.*, 2017], with an additional cross-attention layer in each transformer layer. We concatenate question embedding $\mathbf{q}$ with image embeddings of each image I_k in $\mathcal{I}^r$ with a special token $[SEP]$ in between. Also, to distinguish the image features belonging to different images in the retrieved image set $\mathcal{I}^r$, we assign image order ids to image features from different images. Note that image order ids are not meant for assigning a sequence number to images in the retrieved set. Their sole purpose is to differentiate image features from different images. In short, the input to our MI-BART encoder is $[[CLS], \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}\}, [SEP], \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_p}\}_1, [SEP], \ldots, \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_p}\}_k]$, where $m \in \{1, 2, \ldots, M\}$, $p \in \{1, 2, \ldots, P\}$ and $k \in \{1, 2 \ldots, K\}$. These inputs attend over each other through various self-attention layers of the MI-BART encoder and produce a sequence of contextualized embeddings $\mathbf{z} = \{\mathbf{z_j}\}$, where $j \in \{1, 2, \ldots, M + (P \times k) + k\}$ (Eq. 7).

$$\mathbf{z} = \{\mathbf{z_j}\} = \text{MI-BARTEncoder}(Q, \mathcal{I}^r). \quad (7)$$

MI-BART decoder auto-regressively predicts the probability of the next token A_t in the answer A by attending to these encoder outputs $\mathbf{z}$ and previously generated answer tokens $a_{<t}$ through cross-attention and self-attention layers, respectively (Eq. 8). We train MI-BART decoder parameters θ by minimizing the generative loss $\mathcal{L}_{GEN}(\theta)$ for generating the target answer token conditioned on the question Q and retrieved image context $\mathcal{I}^r$ (Eq. 9). During training, we leverage the ground truth relevant images as retrieved image context $\mathcal{I}^r$, while during inference, we obtain it from our relevance encoder.

$$P_\theta(A_t | A_{<t}, Q, \mathcal{I}^r) = \text{MI-BARTDecoder}(\mathbf{z}, A_{<t}). \quad (8)$$

$$\mathcal{L}_{GEN}(\theta) = -\mathbb{E}_{(Q,\mathcal{I}^r)\sim D}\left[\sum_{t=1}^{|A|} \log(P_\theta(A_t | A_{<t}, Q, \mathcal{I}^r))\right]. \quad (9)$$

To summarize, our proposed framework works as follows, given a question Q and a pool of N images $\mathcal{I}$, we (i) obtain question-image relevance scores S for each question-image (Q, I_i) pair in $(Q, \mathcal{I})$ using our multimodal relevance encoder, (ii) rank all images in the pool based on S, and choose

	RETVQA						WebQA					
	Oracle Images			Retrieved Images			Oracle Images			Retrieved Images		
Method	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines												
Global popularity	27.4	14.5	7.9	36.2	16.7	9.8	17.7	1.3	0.4	17.7	1.3	0.4
Per-category popularity	27.8	16.0	7.6	27.8	16.0	7.6	25.2	1.3	0.5	25.2	1.3	0.5
Other Baseline Approaches												
Question only	62.4	15.3	10.4	62.4	15.3	10.4	22.2	34.9	13.4	22.2	34.9	22.2
Aggregate VQA	69.2	17.1	13	66.6	16.2	11.9	*	*	*	*	*	*
VLP [Zhou *et al.*, 2020]	65.1	70.2	58.8	65.1	70.2	58.8	45.7	42.2	25.9	44.2	38.9	24.1
MI-BART (Ours)												
Image stitch **MI-BART**	78.2	74.7	70.7	72.1	76.6	66.8	49.6	50.5	27.5	**49.1**	50.3	27.4
MI-BART	**84.2**	**85.6**	**79.8**	**76.5**	**79.3**	**70.9**	**49.8**	**51.1**	**28.1**	48.7	**50.7**	**27.6**

Table 4: Performance comparison of various methods on RETVQA and image segment of WebQA. * WebQA only provides full-sentence answers rather than answer category annotations. Therefore, classification model like AggregateVQA cannot be trained for WebQA.

	Color			Shape			Count			Object-attributes			Relation-based		
Method	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines															
Global popularity	25.4	8.2	0.6	24.5	9.2	1.3	25.6	13.4	6.7	49.5	35.2	30.5	0.0	4.8	0.0
Per-category popularity	25.3	9.1	0.5	24.5	9.2	1.3	24.5	14.4	4.5	49.5	35.2	30.5	6.0	15.6	2.3
Other Baseline Approaches															
Question-only	58.0	12.2	6.1	86.9	13.3	11.8	51.6	15.3	9.6	74.9	21.8	18.3	12.4	14.7	4.1
Aggregate VQA	60.1	12.4	6.7	91.3	14.4	13.5	54.6	15.6	10.3	75.4	21.9	18.6	32.2	20.8	11.9
VLP [Zhou *et al.*, 2020]	62.0	67.3	52.8	84.0	81.0	75.7	50.8	70.0	50.8	76.8	78.2	74.8	36.8	33.8	18.5
MI-BART (Ours)															
Image stitch **MI-BART**	71.8	**76.8**	63.7	**96.2**	**94.4**	**91.1**	62.7	**80.1**	62.6	**81.6**	**87**	**81.4**	52.0	39.5	26.4
MI-BART	**72.1**	75.7	**63.9**	92.4	90.3	87.7	**66.0**	80.0	**66.0**	78.5	83.4	78.4	**69.5**	**50.4**	**43.1**

Table 5: Performance breakdown for various methods by question categories on RETVQA with the retrieved images.

top-K images as retrieved images context $\mathcal{I}^r$, and (iii) question Q along with the retrieved image context $\mathcal{I}^r$ is fed to MI-BART which encodes the provided context and generates the free-form natural language answer A to the question Q.

4.5 Image-stitch MI-BART

Inspired by [Bansal *et al.*, 2020], we consider stitching the retrieved images along the width into a single joint image and use single image VQA on this joint image. We use our proposed MI-BART for this baseline; however, we feed the stitched joint image instead of feeding K images. While the MI-BART combines information across images in the embedding space, the image-stitch MI-BART combines information across images in the input space.

5 Experiments and Results

We conduct our experiments on RETVQA as well as on WebQA. Since our task deals with the image set as a given context, we consider the image-only subset of the WebQA dataset. Following [Chang *et al.*, 2022], we use accuracy (A), fluency (F), and $F \times A$, as metrics to evaluate the generated answers. Accuracy validates whether the correct answer is present in the generated answer, whereas fluency measures the quality of the answer paraphrase. Fluency is computed using a recently proposed natural language generation metric called BARTScore [Yuan *et al.*, 2021]. Further, an F1 score is used for retrieving relevant images from a pool of images.

5.1 Baselines, Ablations and Implementation Details

Baselines. We compare our proposed method (MI-BART) and its variant image-stitch MI-BART with the following baselines: **(i) Popularity-based Baselines**: To check for prior biases associated with frequent answers globally or per question category, we use two popularity-based baselines. (a) Global popularity: the most frequent answer in the training set is always considered as the answer by the model, and (b) Per-category popularity: the most frequent answer for each question category is always considered the answer for questions in the corresponding question category. **(ii) Aggregate VQA**: RETVQA task involves VQA over multiple images. In the Aggregate VQA baseline, we use the traditional single-image VQA method [Antol *et al.*, 2015] for each image and aggregate the results. Given a question Q and its corresponding K retrieved images, $\mathcal{I}^r$ from our relevance encoder, we feed each retrieved image along with the question Q to a single image VQA model to get a joint representation. We concatenate joint representations of all retrieved images into a single representation F and feed to a linear layer (MLP) to predict the final answer, i.e., $A = MLP(F)$. Since traditional VQA methods follow a classification-style answer prediction approach, we use the 1000 most frequent answers as

	Binary			Open-ended		
Method	Acc.	F	F×A	Acc.	F	F×A
Popularity-based Baselines						
Global popularity	49.9	17.3	14.3	0.0	11.1	0.0
Per-category popularity	49.5	17.1	13.3	1.2	14.6	0.5
Other Baseline Approaches						
Question-only	74.2	11.5	9.2	48	20	12
Aggregate VQA	75.6	11.5	9.5	55.5	22	14.9
VLP [Zhou *et al.*, 2020]	73.2	80	72.5	55.1	58.2	42
MI-BART (Ours)						
Image stitch **MI-BART**	**80.4**	**88.3**	**80.3**	69.4	70.2	57
MI-BART	78.7	85.6	78.7	**73.7**	**71.7**	**61.5**

Table 6: Performance breakdown by answer categories for various methods on RETVQA with the retrieved images.

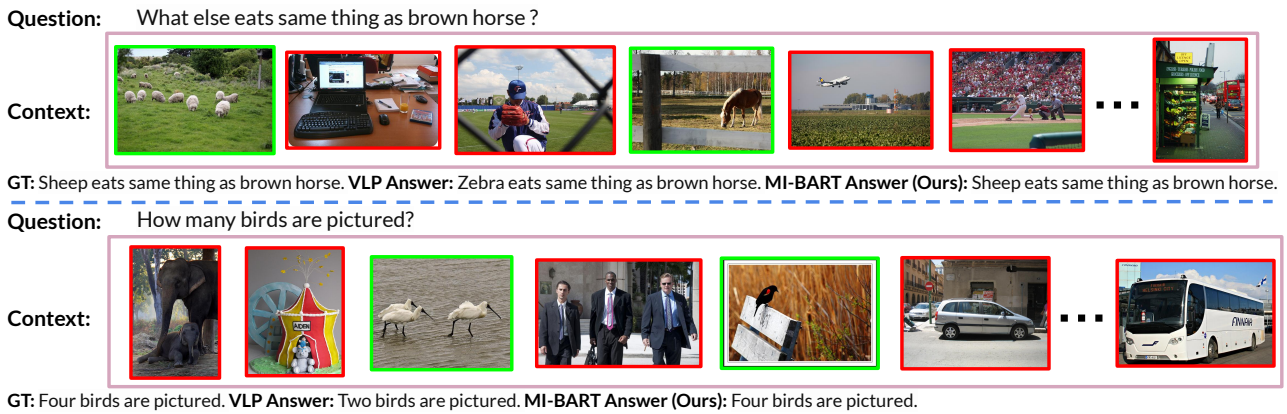

Figure 5: A selection of QA using MI-BART (Ours) and VLP (best baseline) from RETVQA test dataset. The images in the green and red bounding boxes show relevant and irrelevant images, respectively. Our approach consistently performs better than the baseline on different types of questions requiring multiple images to arrive at an accurate answer. (Best viewed in color).

Retrieval	Acc.	F	F×A
Top-1 retrieved image	59.8	61.1	48.8
All retrieved images	**76.5**	**79.3**	**70.9**

Table 7: MI-BART performance on RETVQA using different retrieval strategies.

	Retrieval			QA
Captions	P	R	F1	F×A
w/ captions	32.3	44.7	37.5	19.7
w/o captions	**79.7**	**86.3**	**77.4**	**28.1**

Table 8: Effect of w/ and w/o captions in WebQA: Performance of MI-BART on retrieval and QA (retrieved images setting).

classes in the softmax layer. To generate a fluent answer, we prepend the predicted answer to the question after removing the first word from the question. This baseline is not benchmarked on the WebQA dataset, as the dataset does not provide precise answer annotations for the trainset questions. **(iii) VLP**: As the RETVQA task requires the model to generate the text, encoder-only multimodal transformer models like ViLBERT [Lu *et al.*, 2020], VisualBERT [Li *et al.*, 2019], OSCAR [Li *et al.*, 2020], ViLT [Kim *et al.*, 2021] and UNITER [Chen *et al.*, 2020] are not directly suitable. Hence, we use VLP [Zhou *et al.*, 2020] which is a unified encoder-decoder multimodal transformer as our baseline. We finetune a pretrained VLP on our datasets for evaluation.

Ablations. We perform the following ablations to better understand the various components of our proposed model. (i) *Question-only*: To study the role of the images in generating accurate and fluent answers, we ignore the images and use questions only as input to our model. (ii) *Single-image retrieval*: To study the importance of reasoning over multiple images to generate an answer to the question, we use top-1 retrieved image as our only context instead of a multi-image context. (iii) *Missing captions*: To study the role of image metadata in the relevant source image retrieval and, thereby, the answer generation, we conduct experiments on WebQA without leveraging the image metadata (captions). In this ablation, we augment captions (available in WebQA) as part of the textual input in both the relevance encoder and MI-BART.

Implementation details. We have implemented our framework in PyTorch [Paszke *et al.*, 2019] and Hugging Face's transformers [Wolf *et al.*, 2020] library. Our relevance encoder has three transformer layers, each having eight attention heads. We pretrain our relevance encoder on MS-COCO [Lin *et al.*, 2014] with a constant learning rate of 1e-4 using Adam optimizer [Kingma and Ba, 2015]. Using the same optimiser, we finetune the relevance encoder on both datasets with a constant learning rate of 2e-5. Our MI-BART has six standard transformer encoder layers and six standard transformer decoder layers [Vaswani *et al.*, 2017]; we initialize our MI-BART with VLBart [Cho *et al.*, 2021] pretrained weights to leverage the strong visual-textual learning of VL-Bart. We further finetune MI-BART on a multi-image QA task with a learning rate of 5e-5 using Adam optimizer with a linear warm-up of 10% of the total steps. Our relevance encoder and MI-BART were trained using 3 Nvidia RTX A6000 GPUs with a batch size of 96 and 256 while training and a batch size of 360 and 480 during testing, respectively.

5.2 Results and Analysis

We conduct our experiments in two settings, namely, (i) Oracle images: Here, we use ground-truth relevant images for answer generation, and (ii) Retrieved images: Here, relevant images are retrieved using our relevance encoder. We show the results under both these settings in Table 4. We observe that the popularity-based methods perform poorly. This result is expected as popularity-based methods do not use any question or image context. Methods that involve either questions or images perform better than the popularity-based baselines. However, the question-only baseline has a F×A score of 10.4 on RETVQA, showing that image context is needed to generate accurate yet fluent answers. Transformer-based baseline VLP and image-stitch MI-BART reach a F×A score of 58.8 and 70.7 on our dataset, respectively, in the oracle setting, compared to 79.8 of our proposed MI-BART framework. Image-stitch MI-BART outperforms transformer-based VLP

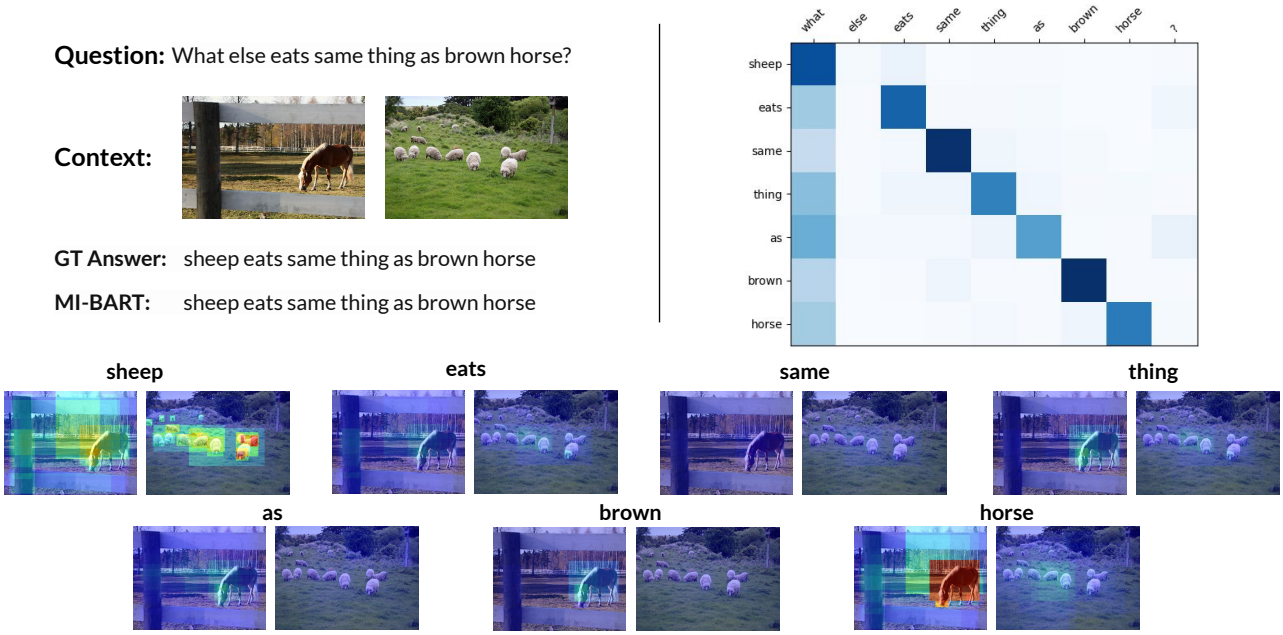

Figure 6: Multimodal attention map over the retrieved images and the question during the answer generation. We observe that our proposed MI-BART attends to the relevant regions of both images while generating the main answer word 'sheep'. Further, we see that it attends to brown horse regions of the first image along with the corresponding question parts while generating 'brown horse'. [Special tokens are removed for visualization.] (Best viewed in color).

by 12% on our dataset and 1.5% on the WebQA dataset, which shows that having a separate decoder in the proposed MI-BART baseline has better reasoning capabilities than a unified encoder-decoder like VLP. We further present the QA results over the retrieved images setting using our relevance encoder, which has an F1 score of 71 at the top-2. All the approaches involving image context outperform question-only baseline, emphasizing that RETVQA has a reasonable utility to develop and benchmark methods capable of jointly reasoning over multi-image context and the question.

Further, in Table 5, we show the QA results over various question categories, and in Table 6, we show the results over answer categories under the retrieved images setting. Our framework outperforms baselines, especially in questions with open-ended generative answers, which constitute nearly half of our dataset. As expected, open-ended generative questions are more challenging than binary ones. However, compared to the baselines, MI-BART provides better improvement for open-ended questions than binary ones by jointly reasoning over multi-image context. Results in Table 7 further emphasize our hypothesis of requiring multiple images to answer the given question. We show the missing caption ablation results on the image-subset of WebQA in Table 8; this result further affirms our claims that the performance of methods on the WebQA dataset depends on the image metadata like captions.

Qualitative analysis. We illustrate a selection of results using our proposed approach and one of the most competitive baselines viz. VLP in Figure 5. In both these results, our approach correctly answers the question in a large heterogeneous visual context. Further, to understand the importance of the multimodal input for question answering, we plot the multimodal attention map over the retrieved images and the question during the answer generation in Figure 6. The figure shows that our proposed MI-BART model attends to the relevant regions of both images while generating the main answer word 'sheep'. Further, we observe that it attends to brown horse regions of the first image along with the corresponding question parts while generating 'brown horse'. Thus, both images are needed and paid attention to when generating the right answer. We further conducted a detailed error analysis on 50 randomly chosen samples where our model failed to generate a correct answer. We categorize the errors into four major categories: (i) partial retrieval: images retrieved by relevance encoder are partially relevant (52%). (ii) Incorrect retrieval: images retrieved by the relevance encoder are entirely irrelevant (26%). (iii) Incorrect reasoning: model generating a partially incorrect answer despite all the retrieved images being relevant (40%).

6 Conclusion and Future Scope

In this paper, we introduced the RETVQA task. We proposed a unified Multi Image BART model to answer the question from the retrieved images using our relevance encoder. Our proposed framework shows promising improvements over the baselines. We have also performed several ablations to further understand the importance of various modules in the proposed framework. In the future, we would like to explore stronger retrieval models and QA on a large pool of images. We firmly believe RETVQA will pave the way for further research avenues in a broader theme of web image QA.

Acknowledgments

Abhirama is supported by the Prime Minister Research Fellowship (PMRF), Government of India. We thank Microsoft for supporting this work through the Microsoft Academic Partnership Grant (MAPG) 2021.

References

[Ahmad *et al.*, 2019] Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. Reqa: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, 2019.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Bansal *et al.*, 2020] Ankan Bansal, Yuting Zhang, and Rama Chellappa. Visual question answering on image sets. In *ECCV*, pages 51–67, 2020.

[Chang *et al.*, 2022] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *CVPR*, pages 16495–16504, 2022.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020.

[Cho *et al.*, 2021] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, pages 1931–1942, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468, 2016.

[Gao *et al.*, 2015] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NeurIPS*, volume 28, pages 2296–2304, 2015.

[Goyal *et al.*, 2017] 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 *CVPR*, pages 6904–6913, 2017.

[Guo *et al.*, 2023] Dalu Guo, Chang Xu, and Dacheng Tao. Bilinear graph networks for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2):1023–1034, 2023.

[Hsu *et al.*, 2021] Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. Answer generation for retrieval-based question answering systems. In *ACL/IJCNLP*, pages 4276–4282, 2021.

[Hu *et al.*, 2017] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813, 2017.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[Kembhavi *et al.*, 2017] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007, 2017.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[Lei *et al.*, 2018] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[Lei *et al.*, 2020] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, pages 8211–8225, 2020.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.

[Li *et al.*, 2020] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, 2020.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, volume 29, pages 289–297, 2016.

[Lu *et al.*, 2020] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, pages 1–9, 2015.

[Marino *et al.*, 2019] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019.

[Noh *et al.*, 2016] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, pages 30–38, 2016.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, pages 8024–8035, 2019.

[Rajpurkar *et al.*, 2018] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL*, pages 784–789, 2018.

[Ren *et al.*, 2015a] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NeurIPS*, volume 28, pages 2953–2961, 2015.

[Ren *et al.*, 2015b] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, volume 28, pages 91–99, 2015.

[Shah *et al.*, 2019] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, pages 8876–8884, 2019.

[Shih *et al.*, 2016] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016.

[Singh *et al.*, 2019] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[Singh *et al.*, 2021] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. Mimoqa: Multimodal input multimodal output question answering. In *NAACL-HLT*, pages 5317–5332, 2021.

[Suhr *et al.*, 2019] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428, 2019.

[Talmor *et al.*, 2021] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: complex question answering over text, tables and images. In *ICLR (Poster)*, 2021.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019.

[Tapaswi *et al.*, 2016] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016.

[Tito *et al.*, 2021] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *ICDAR*, pages 778–792, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, pages 5998–6008, 2017.

[Wang *et al.*, 2022] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR (Poster)*, 2022.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, et al. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, pages 38–45, 2020.

[Xiong *et al.*, 2016] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, pages 2397–2406, 2016.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[Yuan *et al.*, 2021] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, volume 34, pages 27263–27277, 2021.

[Zhou *et al.*, 2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049, 2020.

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016.