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

Abhirama Subramanyam Penamakuri¹, Manish Gupta², Mithun Das Gupta², Anand Mishra¹

¹Indian Institute of Technology Jodhpur
²Microsoft

{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 distinctly 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.

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 [Tal-mor 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-
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 heterogeneous 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 implementa-

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Distinct questions</td>
<td>418K</td>
</tr>
<tr>
<td>#Distinct precise answers</td>
<td>16.205</td>
</tr>
<tr>
<td>Train set questions</td>
<td>50K</td>
</tr>
<tr>
<td>Val set questions</td>
<td>334K (80%)</td>
</tr>
<tr>
<td>Test set questions</td>
<td>41K (10%)</td>
</tr>
<tr>
<td>Avg question length (words)</td>
<td>8.7</td>
</tr>
<tr>
<td>Avg answer length (words)</td>
<td>8.5</td>
</tr>
<tr>
<td>#Distinct words in questions</td>
<td>10.568</td>
</tr>
<tr>
<td>#Distinct words in answers</td>
<td>9.278</td>
</tr>
<tr>
<td>#Avg relevant images per question</td>
<td>2</td>
</tr>
<tr>
<td>#Avg irrelevant images per question</td>
<td>24.5</td>
</tr>
</tbody>
</table>

Table 1: Key statistics for RETVQA dataset.

<table>
<thead>
<tr>
<th>Question Category</th>
<th>Binary</th>
<th>Open-ended</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>30K</td>
<td>50K</td>
<td>160K</td>
</tr>
<tr>
<td>Shape</td>
<td>49K</td>
<td>50K</td>
<td>99K</td>
</tr>
<tr>
<td>Count</td>
<td>50K</td>
<td>100K</td>
<td>150K</td>
</tr>
<tr>
<td>Object-attributes</td>
<td>80K</td>
<td>80K</td>
<td>160K</td>
</tr>
<tr>
<td>Relation-based</td>
<td>-</td>
<td>38K</td>
<td>38K</td>
</tr>
<tr>
<td>Total</td>
<td>229K</td>
<td>188K</td>
<td>417K</td>
</tr>
</tbody>
</table>

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

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

1https://v12g.github.io/projects/retvqa/
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-
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 \( \mathcal{I}' \subseteq \mathcal{I} \) and then leverages the retrieved context \( \mathcal{I}' \). Accordingly, our labelled dataset consists of quadruplets \((Q, I, I', A)\).

4.2 RETVQA Framework

The proposed framework solution: (i) multi-modal relevance encoder for retrieval of relevant sources \( I' \) from \( I \) 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 \( I' \) 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 \( P \) objects using Faster R-CNN [Ren et al., 2015b] pretrained on Visual Genome [Krishna et al., 2017]. For every object \( p \), where \( p \in \{1, 2, \ldots, P\} \), we obtain 2048-dimensional regional feature \( o_p^{reg} \) and 4-dimensional bounding box co-ordinates.
Thereby, each image \( I_i \) is represented by a set of \( P \) object proposals \( \{ (o_{p}^{\text{seg}}, o_{p}^{\text{bbox}}) \}_{p} \). 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 \( \{ o_{p} \} \), and then concatenate across all regions within the image to obtain image embedding \( o \) as follows.

\[
o_i = \{ o_p \}_i, \quad \text{where} \quad p \in \{ 1, 2, \ldots, P \}, \quad i \in \{ 1, 2, \ldots, N \}.
\]

**Question representation.** We encode the textual question \( Q \) containing \( M \) words using a pretrained BERT [Devlin et al., 2019]. This results into a sequence \( q \) of \( M \) \( d \)-dimensional vectors, \( q = \{ q_m \} \) 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.

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

### 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 \([CLS], \{ q_1, q_2, \ldots, q_m \}, [SEP], \{ o_1, o_2, \ldots, o_p \} \). 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 \( \phi \) by minimizing binary cross-entropy loss \( \mathcal{L}_{\text{REL}}(\phi) \) (Eq. 4).

\[
\hat{s}_i = \text{RE}_\phi(Q, I_i).
\]

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

Given a question \( Q \) and a set of \( N \) images \( \mathcal{I} \) sampled from our dataset \( D \), we obtain relevance scores \( S = \{ s_i \}_i \) for each question-image pair \((Q, I_i)\) using our fine-tuned relevance encoder (Eq.5). To choose the final set of relevant images \( \mathcal{T}' \) 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 \( \mathcal{T}' \) for the given question \( Q \) (Eq. 6).

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

\[
\mathcal{T}' = \{ I_k \} \quad \text{where} \quad k \in \text{top-}\!K(S). \quad (6)
\]

### 4.4 Multi Image BART for Question Answering

Given the question and the retrieved images \( \mathcal{T}' \), 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 SimVL [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 \( q \) with image embeddings of each image \( I_k \) in \( \mathcal{T}' \) with a special token \([SEP]\) in between. Also, to distinguish the image features belonging to different images in the retrieved image set \( \mathcal{T}' \), 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 \([CLS], \{ q_1, q_2, \ldots, q_m \}, [SEP], \{ o_1, o_2, \ldots, o_p \}, \{ [SEP], \ldots, [SEP], \{ o_1, o_2, \ldots, o_p \} \_k \), where \( m \in \{ 1, 2, \ldots, M \}, \quad p \in \{ 1, 2, \ldots, P \} \) 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 \( z = \{ z_j \} \), where \( j \in \{ 1, 2, \ldots, M + (P \times k) + k \} \) (Eq. 7).

\[
z = \{ z_j \} = \text{MI-BARTEncoder}(Q, \mathcal{T}'). \quad (7)
\]

MI-BART decoder auto-regressively predicts the probability of the next token \( A_j \) in the answer \( A \) by attending to these encoder outputs \( z \) and previously generated answer tokens \( A_{<t} \) through cross-attention and self-attention layers, respectively (Eq. 8). We train MI-BART decoder parameters \( \theta \) by minimizing the generative loss \( \mathcal{L}_{\text{GEN}}(\theta) \) for generating the target answer token conditioned on the question \( Q \) and retrieved image context \( \mathcal{T}' \) (Eq. 9). During training, we leverage the ground truth relevant images as retrieved image context \( \mathcal{T}' \), while during inference, we obtain it from our relevance encoder.

\[
P_\theta(A_i | A_{<t}, Q, \mathcal{T}') = \text{MI-BARTDecoder}(z, A_{<t}). \quad (8)
\]

\[
\mathcal{L}_{\text{GEN}}(\theta) = -\mathbb{E}_{(Q, \mathcal{T}') \sim D} \left[ \sum_{t=1}^{A} \log(P_\theta(A_i | A_{<t}, Q, \mathcal{T}')) \right]. \quad (9)
\]

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
top-$K$ images as retrieved images context $I'$, and (iii) question $Q$ along with the retrieved image context $I'$ 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 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, $I'$ 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

---

**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.

<table>
<thead>
<tr>
<th>Method</th>
<th>ReTVQA Oracle Images</th>
<th>ReTVQA Retrieved Images</th>
<th>WebQA Oracle Images</th>
<th>WebQA Retrieved Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Popularity-based Baselines</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global popularity</td>
<td>27.4 14.5 7.9 36.2 16.7 9.8</td>
<td>17.7 1.3 0.4 17.7 1.3 0.4</td>
<td>25.2 1.3 0.5 25.2 1.3 0.5</td>
<td></td>
</tr>
<tr>
<td>Per-category popularity</td>
<td>27.8 16.0 7.6 27.8 16.0 7.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Baseline Approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question-only</td>
<td>62.4 15.3 10.4 62.4 15.3 10.4</td>
<td>22.2 3.4 22.2 3.4</td>
<td>45.7 2.2 45.7 2.2</td>
<td></td>
</tr>
<tr>
<td>Aggregate VQA</td>
<td>69.2 17.1 13 66.6 16.2 11.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLP [Zhou et al., 2020]</td>
<td>65.1 70.2 58.8 65.1 70.2 58.8</td>
<td>45.7 2.2 45.7 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI-BART (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image stitch MI-BART</td>
<td>78.2 74.7 70.7 72.1 76.6 66.8</td>
<td>49.6 50.5 27.5 49.1 50.3 27.4</td>
<td>50.7 50.7 27.6</td>
<td></td>
</tr>
<tr>
<td>MI-BART</td>
<td>84.2 85.6 79.8 76.5 79.3 70.9</td>
<td>49.8 51.1 28.1 48.7 50.7 27.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Color</th>
<th>Shape</th>
<th>Count</th>
<th>Object-attributes</th>
<th>Relation-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Popularity-based Baselines</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global popularity</td>
<td>25.4 8.2 0.6</td>
<td>24.5 9.2 1.3</td>
<td>25.6 13.4 6.7</td>
<td>49.5 35.2 30.5</td>
<td>0.0 4.8 0.0</td>
</tr>
<tr>
<td>Per-category popularity</td>
<td>25.3 9.1 0.5</td>
<td>24.5 9.2 1.3</td>
<td>24.5 14.4 4.5</td>
<td>49.5 35.2 30.5</td>
<td>6.0 15.6 2.3</td>
</tr>
<tr>
<td><strong>Other Baseline Approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question-only</td>
<td>58.0 12.2 6.1</td>
<td>86.9 13.3 11.8</td>
<td>51.6 15.3 9.6</td>
<td>74.9 21.8 18.3</td>
<td>12.4 14.7 4.1</td>
</tr>
<tr>
<td>Aggregate VQA</td>
<td>60.1 12.4 6.7</td>
<td>91.3 14.4 13.5</td>
<td>54.6 15.6 10.3</td>
<td>75.4 21.9 18.6</td>
<td>32.2 20.8 11.9</td>
</tr>
<tr>
<td>VLP [Zhou et al., 2020]</td>
<td>62.0 67.5 25.8</td>
<td>84.0 81.0 75.7</td>
<td>50.8 70.0 50.8</td>
<td>76.8 78.2 74.8</td>
<td>36.8 33.8 15.8</td>
</tr>
<tr>
<td>MI-BART (Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image stitch MI-BART</td>
<td>71.8 76.8 63.7</td>
<td>96.2 94.4 91.1</td>
<td>62.7 80.1 62.6</td>
<td>81.6 87 81.4</td>
<td>52.0 39.5 26.4</td>
</tr>
<tr>
<td>MI-BART</td>
<td>72.1 75.7 63.9</td>
<td>92.4 90.3 87.7</td>
<td>66.0 80.0 66.0</td>
<td>78.5 83.4 78.4</td>
<td>69.5 50.4 43.1</td>
</tr>
</tbody>
</table>

**Table 6:** Performance breakdown by answer categories for various methods on ReTVQA with the retrieved images.
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 VL-Bart [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...
What else eats same thing as brown horse?

GT Answer: sheep eats same thing as brown horse

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


