

Multimodal Multihop Question Answering

Aditya Veerubhotla*
Mashrin Srivastava*

Deep Karkhanis*
Soundarya Krishnan*

Harsha Vardhan*
Srijan Bansal*

{adityasv, dkarkhan, vvl, mashrins, soundark, srijanb}@andrew.cmu.edu

Abstract

The problem of multimodal multihop question answering (QA) is not a very well researched field. Most QA datasets are either multimodal or multihop but not both. The WEBQA dataset simulates the way humans search for information on the internet by having multiple modalities of data and requiring aggregating of information across multiple sources. It is the only QA dataset available which is both multimodal and multihop. We propose a novel three-step architecture for such multimodal-multihop datasets. We first use a text-only simple multimodal retriever that uses captioned images and text snippets to perform coarse-grained corpus-level retrieval. We also propose the first ever multimodal and multihop retriever for the fine-grained retrieval of information sources. Moreover, we perform a detailed analysis of visual grounding (X-VLM) and patch-based image encoder techniques (CLIP) on retrieval and reader tasks. We show that simple ROI baselines work well and are also computationally less expensive.

1 Introduction and Problem Definition

People often refer to multiple modalities to answer their questions. This tendency is commonplace when we use search engines for information. The WEBQA ([Chang et al., 2021](#)) dataset is a visual question-answering dataset that resembles the human experience of web search. WEBQA is the first Multimodal - Multihop Question Answering dataset. Performing well on WEBQA requires models to (i) retrieve knowledge from both the visual and textual domain, (ii) reason over and aggregate information contained in multiple sources, and (iii) generate fluent answers in natural language based on this information. This makes WEBQA a challenging dataset that requires multimodal and multihop reasoning.

WEBQA can be posed as a composition of two tasks: retrieval and answering (reader). The baseline retrievals used by the WEBQA authors performed retrieval on a small predefined question-specific subset of the corpus. As is analyzed in [3](#), retrieval on the full corpus level is a much more realistic web search problem. Thus, we have proposed a two-stage retrieval model where we first do a coarse retrieval on the corpus level and then a fine-grained retrieval on the outputs of this coarse retrieval. Based on our analysis, we have found that a text-only SPLADEv2 first stage retrieval works best when images are captioned into text. We also explicitly perform multihop reasoning during the second stage retrieval. We do this by performing sequential retrieval of sources such that retrieval of the next source is conditioned on both the question and all sources retrieved so far. To set up training data for such a retriever, we do additional annotation on the WEBQA data to get a source ordering based on source-question entity overlap. We find that our multihop retriever far outperforms the baseline retriever. To the best of our knowledge, this is the first-ever multimodal-multihop retriever architecture.

In the case of the answering task, our analysis showed that in most cases, the image encoding did not extract areas in the image which were relevant to the question. Thus indicating that the image modality has largely been ignored except for the text caption. We used state of the art Visual Grounding model X-VLM to perform query condition image encoding. As is seen in Section [7.2](#), X-VLM does generate bounding boxes highly relevant to the query. We also compare reader performance when image patch based features are used instead of region of interest (RoI) features. We used CLIP features for our patch based image encoder. We notice, however, that readers trained on both these features perform much poorer than

*Equal contribution, names listed in alphabetical order

the original Detectron-based reader. We notice the same trend for X-VLM and CLIP-based retrievers as well. We attribute this to the fact that perhaps the feature extracted by CLIP (patch-based) and X-VLM (patch-based + RoI) are not aligned with the reader/retriever architecture which worked well for Detectron (RoI).

Our main contributions includes:

1. Perform image captioning followed by text only SPLADEv2 retrieval to get state of the art corpus level retrieval on WEBQA
2. Proposing a novel multi-modal, multi-hop retriever to perform fine-grained second-stage retrieval
3. Analysis of visual grounding (X-VLM) and patch based image encoder (CLIP) on retrieval and reader tasks for multimodal multihop Question answering.

2 Related Work and Background

2.1 Related Datasets

Visual Question Answering is a task that requires understanding of vision, language, and their interactions. Several datasets have been proposed to test this understanding. ManyModalQA (Hannan et al., 2020) focuses on the choice of modality to answer a question, while MultimodalQA (Talmor et al., 2021) attempts to curate questions that require reasoning over multiple modalities with questions that are generated from pre-defined templates. MIMOQA (Singh et al., 2021) explores multimodal outputs and highlights the importance of accompanying a textual response with images. We work on the WEBQA dataset (Chang et al., 2021), which is a new benchmark for multimodal multihop reasoning and question answering. We describe WEBQA dataset in greater detail in 3.

2.1.1 Multimodal Representation

Motivated by the success of transformer-based models like BERT (Devlin et al., 2019) in language representation, numerous attempts have been made to translate this progress to multimodal representations with the proposal of various transformer-like models.

Vision Language Pretraining: Using innovative approaches for pretraining is one of the most popular methods to learn better alignment between modalities. CLIP (Contrastive Language-Image

Pre-Training) consists of a simplified version of ConVIRT (Zhang et al., 2020) trained from scratch, which is an efficient method of image representation learning from natural language supervision. OSCAR (Li et al., 2020) uses object tags (language representation of visual concepts) as anchor points to ease the learning of alignment. VINVL (Zhang et al., 2021b) is a large-scale object attribute detection model that is better designed for Vision + Language tasks and trained on a massive corpus of multiple public object detection datasets. BLIP (Li et al., 2022a) attempts to adopt encoder-decoder models for image-text retrieval tasks. It proposes a multimodal mixture of Encoder-Decoder (MED) model for effective multitask pretraining and flexible transfer learning. OFA (Wang et al., 2022a) is a unified multimodal pre-trained model that unifies modalities and tasks to a simple sequence-to-sequence learning framework, and allows for finetuning with task instructions without extra task-specific layers.

Visual Grounding: Visual grounding aims to locate objects from an image conditioned on a text query. Early attempts like NMTREE (Liu et al., 2019) and LGRAN (Wang et al., 2019) formulated visual grounding as a text-based image retrieval task on a set of candidate regions in a given image. Recent works like Attention-based Regression models (Endo et al., 2017), ZSGNet (Sadhu et al., 2019), and Recursive Sub-query based models (Yang et al., 2020) simplify the visual grounding pipeline by discarding the region proposal generation stage and attempt to locate the referred object directly. X-VLM (Zeng et al., 2021) is the current state of the art on the RefCOCO+ (Yu et al., 2016a), the most popular visual grounding dataset.

2.1.2 Multimodal Information Retrieval

Past works in Multimodal Information Retrieval (mIR) can be divided into two categories: vector-similarity based scoring and query-document interaction based scoring.

Vector-similarity based mIR: Vector-similarity based information retrieval (Dense Retrieval) aims to encode the query and document into vectors in a shared space, and the relevance score is determined by their inner product. Prominent examples of dense retrieval include VSE++ (Faghri et al., 2018) and ACMR (Wang et al., 2017). Training strategies like hard negative mining, in-batch negatives (Karpukhin et al., 2020), and the distillation

of expensive interaction-based models have shown promising results in textual IR (Hofstätter et al., 2021), and can be explored in mIR as well.

Interaction based retrieval Interaction-based IR models compute the relevance of query-document by computing the alignment between query and document representations. With the adoption of pre-trained transformer models, there has been a drastic improvement over previous approaches. These approaches include UNICODER-VL (Li et al., 2019), CLIP4CMR (Zeng and Mao, 2022) and GILBERT (Hong et al., 2021). In Multihop dense retrieval (Xiong et al., 2020), the query is reformulated at each retrieval hop or stage based on previous retrieved results and the retrieval scores are based on maximum inner product search at corpus level. This simple yet effective idea showed state of the art retrieval results on two multihop datasets, HOTPOTQA and multi-evidence FEVER.

2.1.3 Multimodal Question Answering

Question answering involves generating natural language answers from text queries by aggregating information across sources spanning modalities. Recent models use attention mechanisms to learn visual representation and semantic alignment between image and text jointly. In newer approaches, pre-training has become an invaluable precursor to train state-of-the-art QA systems. The purpose is to have model architectures which *a priori* have focussed on encoding and alignment of multiple modalities. As mentioned in 2.1.1, Vision Language Pretraining (VLP) and Visual Grounding are two popular approaches used here.

More recent works in multimodal question answering focus on specific issues of current models to get generalized improvements over multiple datasets. INFACUALITY (Vickers et al., 2021) explicates that simple labeled data may be insufficient to learn complex knowledge representation as required in multihop reasoning. They integrate facts extracted from an external knowledge base like Wikipedia to augment QA knowledge to get 19% improvement on the KVQA dataset. GAR (Mao et al., 2021) performs generation-augmented retrieval. This approach augments queries with heuristically discovered relevant context through text generation without external supervision or costly downstream feedback, and results in SoTA performance on various datasets on extractive QA tasks.

3 Task Setup and Data

People often refer to multiple modalities to answer their questions. This is a commonplace when we use search engines for information search. The WEBQA dataset is a visual question-answering dataset that resembles the human experience of web search. Humans search for information from multiple sources of images and text to arrive at their answers. WEBQA simulates real-world web search by requiring the following QA skills conjunctively:

1. **Multimodality:** Sources and distractors are present as both text and images since human aggregate data across both modalities when parsing web information
2. **Multihop:** The model needs to aggregate information across multiple sources to correctly answer a question.
3. **Caption Necessity:** The caption accompanying the image is essential to associate an image with the text.

A question in WebQA can either have only images as gold sources or only text snippets as gold sources, not a combination of both. However, distractors for questions are usually from both modalities. We thus categorize our questions as visual questions and textual questions based on the modality of their gold sources. Visual questions are further categorized into Color, Shape, Number, Choose, YesNo and Others / Miscellaneous depending on what kind of reasoning is to be done over the images. The number of gold-sources that are present for a question indicates the hop-count. A hop-count of more than 1 indicates that the question is multihop and information across multiple sources needs to be aggregated to answer the question. The statistics related to the number of samples in each set – train, validation, and test, are given in Table. 1. Due to computational constraints, we use a smaller dataset (20% the size of the original) that is sub-sampled within each question category and hop count.

3.1 Need for Retrieval

Most of the previous multimodal question-answering datasets are grounded on a single image that is provided with the query. However, for the WEBQA dataset, grounding image(s) is not

Modality	Train			Dev		
	1	2	>2	1	2	>2
Visual	11968	6986	-	1553	958	-
— Color	1199	452	-	125	54	-
— Shape	421	70	-	66	8	-
— Number	1694	165	-	244	15	-
— Choose	2527	1191	-	330	172	-
— YesNo	2750	3742	-	329	499	-
— Others	3377	1366	-	459	210	-
Textual	75	17092	645	4	2360	91

Table 1: Number of datapoints in our sub-sampled dataset for each question category and number of hops

Modality	Textual		Visual	
	Correct	Distractors	Correct	Distractors
Visual	-	15.5	1.6	15.9
— Color	-	14.9	1.3	15.8
— Shape	-	15	1.1	15.8
— Number	-	15.8	1.1	15.8
— Choose	-	15.1	1.7	15.9
— YesNo	-	15.5	1.6	15.9
— Others	-	15.3	1.3	15.8
Textual	2.0	14.6	-	11.6

Table 2: Statistics for Correct Evidence and Distractors

	Textual	Visual	Overall
Corpus	666225	335456	1001681

Table 3: Number of textual and visual sources at corpus level

Modality	Train	Dev
Visual	49.34	49.35
— Color	46.99	46.25
— Shape	46.15	47.69
— Number	49.59	49.94
— Choose	48.03	48.34
— YesNo	53.37	53.45
— Others	45.89	45.84
Textual	9.81	9.35

Table 4: Jaccard similarity between questions and answers for each question category for train and validation sets

provided with the query. Generating free-form answers directly from the set of input sources is computationally prohibitive and a very difficult learning problem. Thus, a retrieval component is necessary both at the corpus-level and distractor pool level to make sure reader only gets relevant information. Retrieval is done on the entire corpus-level and not only on the predefined set of distractors. This is done to model a more realistic scenario which makes the source-retrieval more challenging.

Table. 2 shows that when operating at a distractor pool level, a retriever needs to only select from a set of around 12-17 sources. However, as Table

3 points out, when retrieval is performed at corpus level, the selection needs to be from among 1 million sources. The latter replicates the scale of a web search task much better and is a much harder but a more realistic problem.

We also show Jaccard similarity between the questions and answers across visual modality categories and textual modality in Table. 4 for train and validation sets. Jaccard similarity would help us investigate if the answers have a lot of word overlap with the questions. This is a proxy to estimate if, and to what extent the answer is a rephrasing of the question.

3.2 Metrics

The WEBQA authors evaluate models based on two metrics:

- F_1 score for retrieval performance
- $Fluency \times Accuracy$ for reader and end-to-end model performance

The *Fluency* is calculated based on normalized *BARTScore* between the prediction and the reference. The metric has been described as follows: given a candidate c and a set of references R , the fluency is computed as follows:

$$Fluency(c, R) = \max\{\min\{1, \frac{BARTScore(r, c)}{BARTScore(r, r)}\}\}_{r \in R}$$

It models the semantic agreement between sentences and tries to be robust to functional-word misplacements. However, it does not account for visual semantics learned in the text. To account for this, an *Accuracy* term is multiplied which focuses on the presence of key entities. The metric can be defined as follows, for a candidate c and a set of keywords K , the accuracy is computed as:

$$Acc(c, K) = \begin{cases} F1(c \cap D_{qc}, K \cap D_{qc}), & \text{if } qc \in \{\text{color, shape, number, Y/N}\} \\ RE(c, K), & \text{otherwise} \end{cases}$$

Based on our baselines analysis, we point out that these metrics are not best suited for the task at hand. Fluency is primarily defined based on a paraphrasing goal and might not be appropriate for a generation task. *Recall@k* is a more commonly used standard metric for evaluating retrieval performance as opposed to F_1 scores. We thus add the following metrics to our experiments and analysis:

- $Recall@k$ for retrieval performance
- ROUGE-L for evaluating text generation performance

ROUGE-L is a much more robust metric which measures sentence sub-sequence overlap and unlike *Fluency*, it isn't sensitive to the original task it was defined on. We measure $Recall@k$ numbers over the entire corpus and not just the gold sources + distractor pool for every question.

4 Baselines

We run baselines that span multiple levels of modality complexities. In addition, we also run component-wise baselines by evaluating the current state-of-the-art retrievers and readers on the WEBQA framework. These models have been evaluated both in pre-trained and fine-tuned settings.

4.1 Unimodal Baselines

We initially run models which only look at a single modality to evaluate the significance of an individual modality. They help establish what part of the performance of final models is a result of cross-modal information. We experiment with different retriever architectures coupled with a fixed reader model. For the reader, we fine-tuned BART (Lewis et al., 2019) on the WEBQA dataset.

1. **Lexical Retrieval:** In our experiments, we use the well-known BM25 (Robertson et al., 1994) retrieval model, which performs an exact match on the words present in the question and the sources in a bag-of-words manner. This method has shown to be a strong baseline across multiple datasets owing to its high retrieval accuracy. We use the standard parameters ($k1 = 1.5$, $b = 0.75$, $\epsilon = 0.25$) in our experiments.
2. **Sparse Retrieval:** Sparse retrieval aims to perform question source matching, by enforcing sparsity constraints on the question and source representations. For sparse retrieval, we perform experiments with the pre-trained SPLADEv2 (Formal et al., 2021) model, which performs question-source matching in the vocabulary space. This technique has shown state-of-the-art performance across different information retrieval benchmarks.

3. **Dense Retrieval:** Dense retrieval-based approaches perform retrieval by computing a single vector representation of the question and the source to an inner product-based matching function. For evaluating dense retrieval, we consider TAS-Balanced (Hofstätter et al., 2021), a state-of-the-art dense retrieval model that performs semantic matching by projecting questions and sources to a common latent space, and compute their inner product.

4.2 Simple Multimodal Baselines

- **Image Captioning:** WEBQA dataset has sources that can be textual snippets or images with captions. We replace images with generated captions using OFA (Wang et al., 2022b) which is a unified multimodal pre-trained model in a sequence-to-sequence learning framework. The generated captions are concatenated with the original captions provided with the images. The evidences for a question are then retrieved using SPLADEv2. The objective of this baseline is to evaluate the relevance of visual features for the question-answering task.

- **BLIP Pretrained (BLIP-pt):** BLIP (Li et al., 2022b) is a flexible VLP framework for vision language understanding and generation tasks. BLIP uses image-grounded text encoder-decoder architecture with parameter sharing. We use BLIP pre-trained on the COCO dataset (Lin et al., 2014) to retrieve on WEBQA dataset which helps us evaluate the effectiveness of image-grounded text encoder-decoder, parameter sharing, and cross-modal attention for retrieval.

4.3 Competitive Baselines

We have analyzed three competitive baselines for this report. We have made changes so that they fit the WEBQA setting.

- 1. **VLP with Detectron2:** Detectron2 is a popular modular object detection library for object detection and related tasks (Girshick et al., 2018). We use Vision Language Pretraining with Detectron2 as an object detection framework. We use this detectron based finetuned model provided by the WEBQA authors as our first competitive baseline.

2. **VLP with VinVL:** We use VinVL (Zhang et al., 2021a) objection detection framework which uses a cross-modal (V and L) fusion model to combine the text input with these visual features. This is followed by the VLP model which is finetuned on the WebQA dataset to achieve competitive results.
3. **BLIP Retrieval fine-tuned (BLIP-ft):** We chose BLIP as our retriever owing to the bi-encoder architecture and image-grounded text encoder sharing the parameters with the text encoder. To perform retrieval, we use a text encoder to encode textual data (questions and textual sources), and an image-grounded text encoder for image sources (including captions). Unlike the BLIP pre-trained model, we made modification to include original captions to retrieve evidences.

5 Proposed Model

Figure 1 describes our approach. The architecture can be broken into four core components: 1. Source-Retriever 2. First Stage Filter 3. Multihop Retriever 4. Reader. The Source-Retriever performs corpus-level retrieval to fetch the top-K sources from the corpus C, where $K \ll |C|$. The retrieved results are then fed into the First Stage Filter, which identifies the modality of the question (into image/text), and filters all sources that do not align with the prediction. These filtered documents are then passed into the Multihop Retriever, which selects the sources that are best aligned with the question by iteratively selecting the most relevant source. These sources are then sent to the reader, which generates the answer conditioned on the question and retrieved sources. In the following sections, we will describe our approach in detail.

5.1 Source Retrieval Module

Our source retriever is based on sparse, lexical matching of words between the words in the question and the source tokens. In our experiments, we use the SPLADEv2 model (Formal et al., 2021) as our first stage retriever, which performs lexical matching of the question and the source. We chose this architecture owing to its good retrieval performance over other text-based baselines and low computational requirements. In this paradigm, there are two encoders: question encoder and source encoder. The encoders project queries and sources into a shared embedding space where the relevance

is computed by the dot product of the embeddings. Mathematically, if $f_{q\theta}(q)$ denotes the query encoder and $f_{d\phi}(d)$ denotes the source encoder, the score $s_{\theta,\phi}(q, d)$ can be defined as:

$$s_{\theta,\phi}(q, d) = f_{q\theta}(q)^\top f_{d\phi}(d) \quad (1)$$

For scoring text sources, we use the text passages, and for image sources, we perform matching over the captions. Ranking is performed by sorting of the matching scores. Our experiments show that the SPLADEv2 model trained on the MSMARCO dataset (Campos et al., 2016) obtains high recall out-of-the-box, and we do not fine-tune the model on this dataset. While we observed an increase in recall by appending the OFA captions to the images, we decided to not go ahead with this approach owing to the high computation cost for inference of the captioning model. We conduct all experiments by setting the number of retrieved sources to 20 (i.e. $K = 20$). The retrieved results are then sent to the First Stage Filter which we describe next.

5.2 First Stage Filter

In WebQA, each question is designed to be answered by a single modality. This scenario is prevalent in web search, where the queries usually require answer in only one modality for retrieval. This motivated us to incorporate a modality selector module over the retrieved results, which takes as input the question and predicts if the question can be answered by a text source or an image source which is formulated as a binary classification problem. The prediction is used to remove the cross-modal distractors and keep the sources which are of the same modality as predicted by the module.

We model the selector by using BERT (Devlin et al., 2019), and applying a classification layer over the [CLS] representation, subsequently training the model using binary cross entropy. The output of the model can be expressed as:

$$\begin{aligned} \hat{y} &= \operatorname{argmax}_i P(y_i|q) \\ P(y_i|q) &= \operatorname{softmax}_i(h_i) \\ h_i &= W_i z_{[CLS]} \\ z_{[CLS]} &= \operatorname{BERT}(q) \end{aligned} \quad (2)$$

Where $W_i \in \mathbb{R}^{1 \times 768}$ is the projection matrix to the label space.

5.3 Multihop Retriever

Owing to the multihop nature of the dataset, the retrieval of evidences independently might not

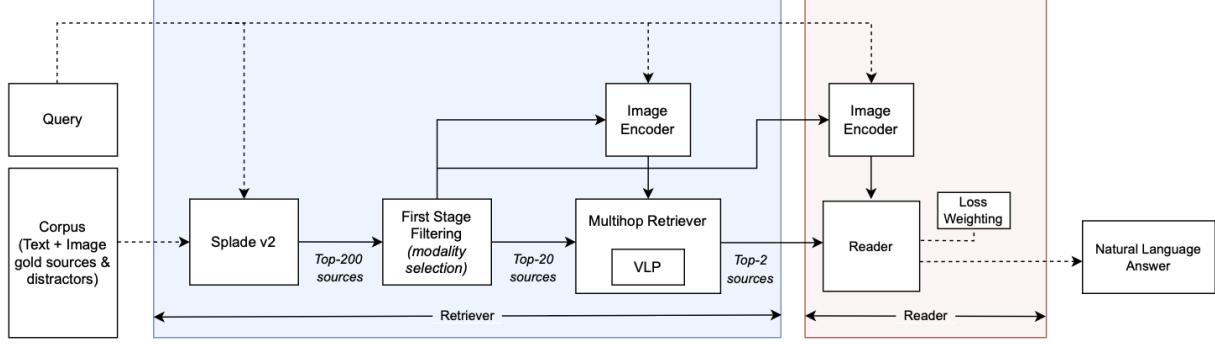


Figure 1: Proposed architecture

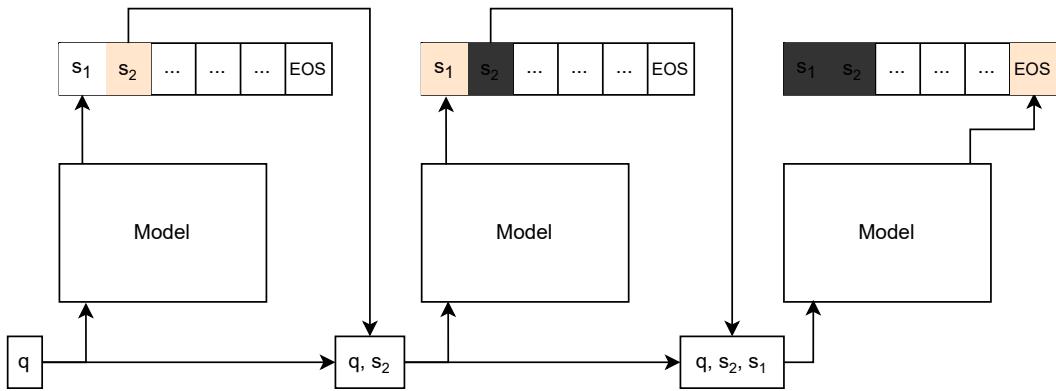


Figure 2: An example of the inference of the multihop retriever. The decoding is initialized with the query, and a source is selected from the pool, which is subsequently appended to the query and another source is retrieved. This process is repeated till the EOS token is selected.

ensure that all the evidences are retrieved. This is especially evident when the entities linking multiple evidences do not occur in the question as shown in Figure 3. In these situations, it is nearly impossible to retrieve the right evidences of the question without knowing the right context.

In an attempt to solve this problem, we take inspiration from the iterative retrieval system described in (Xiong et al., 2020) and adapt their work to this domain. We formulate our reranker as a selection problem where a set of most relevant sources are selected from a filtered pool. The selection begins when model is initialized with the question and outputs a probability distribution of selection over the pool. The source with the highest probability is then picked. The question is reformulated as the question concatenated with the selected source and the process is then repeated. This process continues till a special end-of-sequence [EOS] token is sampled. This allows the model to leverage the previously selected context to pick the right source while

being able to select arbitrarily large sets of contexts. Figure 2 illustrates the discussed approach.

While the inference from the model progresses as described in the previous paragraph, the training of the model is performed by performing teacher forcing, i.e., the reformulated question always contains the ground truth sources from previous hops. As the WEBQA dataset does not provide the order of selecting sources, we develop a heuristic based approach for ordering our sources. We order the sources w.r.t the overlap of the keywords between the question and the sources. Keywords are defined as those words whose part-of-speech-tags are in the following categories: ["NOUN", "PROPN", "ADJ", "VERB"]. We recognize that this heuristic might not work for longer hops of selection, and hence we truncate this approach till two sources are selected.

5.4 Reader

The retrieved sources are then fed into the Reader, which generates the answer conditioned on the question and the retrieved sources. For our reader, we use the VLP model (Zhou et al., 2019) as the

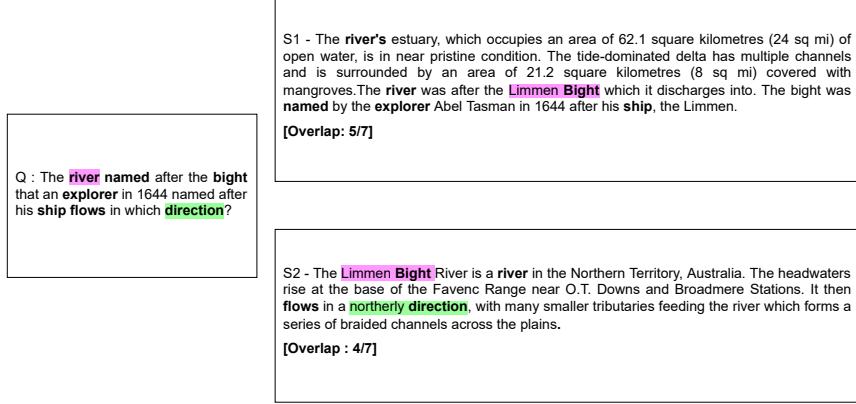


Figure 3: For the given question, we first need to find the river (highlighted in yellow) that was named by the explorer which will help us answer the main question of direction of its flow (highlighted in orange). We find that Limmen Bight river was named by the explorer from S1. This will further help us retrieve S2 which contains Limmen Bight river and its direction of flow. Our ground truth ordering mechanism based on keywords (**bold**) overlap is also highlighted which orders S1 [overlap:5/7] followed by S2 [overlap:4/7].

base for experimentation. The inputs to the model are the concatenated representations from the question and source tokens, while the labels are the answer tokens. In line with the setup described in WebQA, we use the Masked LM loss for training the model, with masking for ensuring autoregressive decoding.

5.5 Input Representations

Our error analysis from the previous report shows that the image representations obtained from state-of-the-art object detectors frequently do not retrieve the objects that are asked in the question. An example from our analysis is shown in figure 4 a) where the question requires reasoning over a bicycle in front of the fence. However, Detectron does not detect the fence as an object, and hence the model fails to answer the question. This motivates us to generate image representations that are grounded in the question. Inspired by the works on visual grounding (Yu et al., 2016b) task that aims to generate bounding boxes for an input caption, we plan to use the representations from models trained on this task. We performed qualitative analysis on XVLM (Zeng et al., 2021), a state-of-the-art visual-grounding system, and showed that this model showed strong performance in generating relevant bounding boxes on the questions in WebQA. As can be seen in 4 b), XVLM correctly identifies the region of interest that encompasses the bicycles and the fence in the image.

To obtain question-grounded image representations, we used the features from the XVLM model, which represents an image as a collection of image

tokens that are obtained from the image patches. The image representations are then refined by performing cross attention with the image tokens. For this model, we take the intermediate representations from XVLM before the bounding box is generated as the question conditioned image representation. We hypothesize that the features obtained from this model should contain sufficient visual information for generating correct outputs. The initial results obtained by directly fine-tuning the model showed poor performance, which led us to believe that the alignment between the modalities is not correct. To combat this, we ran fine-tuning in two stages, 1: where we freeze the VLP head and train the projection layers from the XVLM space to the VLP space, and 2: where we fine-tune the full model on the dataset. Our results show that this approach improves the performance slightly, and we conduct all our analysis/experiments with this checkpoint.

Another potential solution to the low object recall problem is to change the manner in which representations of the image are obtained. To this end, we evaluate the effectiveness of patch-based representations obtained from CLIP (Radford et al., 2021), which has shown state of the art performance in multiple vision alignment tasks.

5.6 Loss functions

5.6.1 Cross-Entropy Loss

We use cross-entropy loss for retriever training where we have two different settings for baseline retriever and multi-hop retriever-

Q: Is the fence in front of The Glass House in Fulham taller or shorter than a bicycle?



a) Detectron output



b) XVLM output

Figure 4: Results of the Image Grounding system XVLM. Detectron is unable to identify the fence mentioned in the query, while XVLM is able to identify it and draw the correct bounding box over the region of interest.

1. The baseline retriever scores each document independently and is trained using binary cross entropy loss.
2. In the multi-hop retriever, the retrieval is formulated as a selection problem and we use negative log likelihood to train the retriever. The retriever loss is computed as follows -

$$NLL(w) = - \sum_{i=1}^M \log(p_w^i)$$

where $p^i = \sum_{s_i \in [P_q]} \text{softmax}_{s_i}(f(q, s_i))$, q is the question, $M = |P_q|$ and P_q are the set of sources. The reader is also trained on negative log likelihood loss where the loss is computed over the set of target answer tokens.

5.6.2 Loss Weighting

An error analysis on the failures showed that the VLP model was unable to generate the right answer words, which decreased the accuracy metric of the model. To combat this, we mined keywords from the training dataset and enforced higher penalty on making a mistake on these examples. We call this technique loss-weighting, and can be expressed as follows: For a set of keywords for the answer of a question $K(q) = \{w_1, w_2, \dots, w_k\}$, the loss function is weighted as follows:

$$L(w) = \sum_{i \in [A]} \lambda_{K(q)}(w_i) l(w)$$

where $\lambda_{K(q)}(w_i) = 1$ when $w_i \notin K(q)$ and $\lambda_{K(q)}(w_i) = \lambda$ otherwise for a fixed hyperparameter λ . The keywords are mined by taking those tokens that are present in the answer but not in the question. This method returns words such as "Yes/No", Colors, and Numbers, that are important in the question. Keyword selection logic for answer weights can be found in Table 5.

Question	Answer	Keywords
What color are both domes on the Imam Husayn Shrine and the Imam Ali Shrine?	Gold is the color used in both domes of the Imam Husayn Shrine and the Imam Ali Shrine.	Gold, dome, used, is, of, in, h
Are both the National Museum of the American Indian in Washington, D.C. and the Xanadu House in Kissimmee, Florida the same color?	Yes, both the National Museum of the American Indian in Washington, D.C. and the Xanadu House in Kissimmee, Florida are beige.	Yes, are, be, ##ige, h
How does a player of Animal Boxing hold their DS game system to play the game?	The DS must be held upside down so that the touch-screen is above.	must, be, held, touch, so, above, that, ##screen, The, is, upside, down
How many dancers participate in the golden flag dance in St.Patricks Festival Dublin?	Six dancers participate in the golden flag dance in St. Patrick's Festival in Dublin.	golden, six, h, 's, gold

Table 5: Keywords for answer weights

5.7 Changes to training data

We performed changes to the training data to facilitate the training of our multihop retriever and

sub-sampled the number of training instances to iterate more quickly. For the multihop retriever, we introduced a mechanism, ordering to the multiple sources present for a question which is based on the number of keywords overlapping between the question and a source. The details of our algorithm is explained in Section 3. For efficient experimentation, we sub-sampled the training dataset to 20% of its original size by stratifying on the Question Category, and the number of positive sources (hops) required to answer the question.

6 Results

In this section we describe the results of our experiments. Table 6 describes the different first stage retriever models based on sparse, lexical retrieval and dense retrieval based approaches. It is evident that lexical approaches outperform dense retrieval based approaches indicating that it is sufficient to perform keyword matching over image captions and those in the passages. In contrast, dense retrieval based on multimodal representations perform poorly, indicating that more work is needed to be done in this field to be able to outperform lexical approaches. The best results are obtained with SPLADEv2 which performs sparse lexical matching on the augmented tokens of the inputs. For improving image recall, the images are then further enhanced by incorporating the captions generated by OFA model. However, performing image captioning is not a scalable solution for large scale information retrieval, and hence we perform subsequent experiments with SPLADEv2.

Model	R@10	R@100
BM25 (original caption)	0.5769	0.7563
TAS-B (original caption)	0.7070	0.9181
SPLADEv2 (original caption)	0.6973	0.9124
SPLADEv2 (original + OFA captions)	0.7329	0.9821
BLIP-pt	0.2265	0.2739
BLIP-ft	0.2736	0.3451

Table 6: Recall at 2, 3, 5, 10 and 100 for text, image and overall modality for various models. Highest value for each recall level for text, image and overall modality is in bold.

For the reader, we run BART and VLP-based models as our readers. We experiment with Detec-

Setting	FL	Acc	FL* Acc	Rouge-L
BART	31.57	54.55	20.54	45.08
VLP (Detectron2)	37.01	45.4	25.3	41.3
VLP (CLIP)	11.25	32.6	8.52	23.76
VLP (X-VLM)	12.41	35.07	9.12	26.17
VLP (X-VLM, 2-stage)	14.44	34.4	10.29	27.05
VLP (Loss weighting)	34.74	43.66	23.09	39.49

Table 7: Results of different Reader configurations. The inputs to the reader were the gold sources

tronV2, CLIP to X-VLM visual feature extraction models for obtaining image features, and provide an analysis on their effectiveness. For our reader, we evaluate the effects of loss weighting in improving the accuracy of the answers. The reader models are evaluated on Fluency (FL), Accuracy (Acc), FL*Acc and Rouge-L metrics as described earlier, while BART gives good performance on FL, Acc and Rouge-L metrics as it learns to just copy the questions in the answers which we further elaborate in the qualitative analysis. VLP with Detectron2 as feature extractor shows very good performance on FL, Acc and Rouge-L metrics. This VLP (Detectron2) acts as a strong baseline for VLP based models. We use CLIP as feature extractor with VLP to test patch based image representation models with VLP as reader. As seen in Table 7, VLP (CLIP) does not perform as well as VLP (Detectron2) which we believe is due to the fact that patch based image representation is not effective for this task.

We generate a question conditioned image representation by taking the intermediate representation from X-VLM as input to our VLP reader. The VLP (X-VLM) performs worse than baseline models as there might be domain shift between the model’s train dataset and WebQA dataset which is not the case with Detectron2. To tackle this, we freeze the VLP reader model and train the X-VLM model to learn projection of these features into VLP space, we call this model VLP (X-VLM, 2-stage). We observe that VLP (X-VLM, 2-stage) performs better than VLP(X-VLM) on FL, FL*Acc and Rouge-L but still underperforms compared to baselines BART and VLP (Detectron2) as limited fine-tuning may not be enough to outperform pre-trained models on those datasets. We also use VLP with weighted loss where we penalize the

model for missing keywords in the answer but the performance decreases compared to baseline models without loss weighting. This can be due to the fact that the retrieved evidences do not contain the required information to answer the questions and forcing the model to generate the desired words pulls the overall performance down.

Model	P@2	R@2	F1
VLP (Detectron)	99.19	68.17	80.8
VLP (CLIP)	99.31	63.99	77.83
VLP (XVLM)	99.26	64.05	77.85
Multihop (Detectron)	84.37	82.57	83.46

Table 8: P@2, R@2 and F1 on VLP retriever models and multihop retriever

Table 8 shows the P@2, R@2, F1 of retriever models like VLP(Detectron), VLP(CLIP), VLP(XVLM), Multihop (Detectron) at corpus level. Though Multihop (Detectron) performance is lower than other VLP models on P@2, it outperforms the other baselines on R@2 and F1 metrics which further assures the better performance of multihop retrieval model.

7 Analysis

7.1 Quantitative

Model	# evidences	
	1	2
VLP (DetectronV2)	80.58	80.85
VLP (CLIP)	81.84	77.01
VLP (X-VLM)	81.39	77.14
Multihop (DetectronV2)	90.88	80.51

Table 9: F1 scores of the Multihop retriever for different number of hops

The performance of our proposed multihop retriever is mentioned in Table 9. The results show that proposed multihop retriever outperforms the other baselines in the first hop, and is second best on the second hop. This can be explained by the fact the model is learning to predict EOS, which helps the model to stop decoding for single evidence tokens. The loss of performance can be attributed to our ordering function which can cause with the model in cases where both evidences are independent of the question.

Table 10 shows the performance of VLP retriever models and multihop (detectron) model across

question categories like text, YesNo, Choose, Number and Color. This gives us the insight about the performance of retriever models across different types of questions. We observe that the retriever models perform uniformly well across different types of question categories. Multihop outperforms other models on all the categories especially by a good margin on YesNo, Others, Choose, Number, Color and Shape categories leading to overall best F1 score on the dataset.

7.2 Qualitative and Examples

We perform qualitative analysis to perform a comparative study of various model components. We first look at specific examples to see why multihop retriever does better than the baseline retriever. We then also try to explicate the reason why X-VLM doesn't perform as well as Detectron based models even though it is the state of the art visual grounding model. We compare X-VLM with detectron in both retrieval and reader settings. In case of retrievers, for each question, we list the gold sources, the ordering of our multihop training data (for multihop retriever), the independence of sources (whether one source's importance to answer the question is only apparent if another is present) and the retrieved sources for each retriever model. For reader, apart from the gold sources, we report the reader output and the bounding box made by their corresponding image encoders.

7.3 Multihop retrieval vs Baseline retrieval

In Figures 5, 6, and 7, we have listed a few interesting examples to compare the multihop detectron model with its baseline counterpart.

The first example shows the typical case of multihop retriever succeeding where baseline retriever couldn't. Only after getting the information from S1 about Keriya Town can the model relate that S2 is relevant to the river asked in the question. This matches the intuition behind using a 2 step multihop retriever. This is also the case in example 5 where the fact that roads are called motorway in Switzerland will only be known after retrieving S2. It is only then that the model can realize the importance of S1 since it talks about motorway speed limits in Switzerland, and not the road's speed limit.

The second example although multihop, involves independent retrieval of sources. Although not obvious from the intent of multihop retriever, it still does better in these examples. Baseline retriever

Model	Q-category							
	Text	YesNo	Others	Choose	Number	Color	Shape	Overall F1
VLP (detectron)	83.49	71.72	77.4	80.8	85.59	78.64	75.75	80.8
VLP (clip)	78.02	73.44	78.34	79.77	86.31	79.27	74.81	77.83
VLP (xvilm)	78.09	73.32	78.42	80.03	87.24	76.72	74.81	77.85
Multihop (detectron)	83.33	77.58	85.84	84.2	93.63	85.92	84.46	83.46

Table 10: Comparison of retrieval results at corpus level across text and visual categories

misses a source due to perhaps lack of confidence. However, multihop, owing to the fact that the first source and question mention about statues, reinforces the information and succeeds in getting the second source as well. The same thing happens in the text setting in example 4 where the retrieval of S2 reinforces Atoll in the Maldives and ensures the retrieval of S1.

Multihop retriever doesn't seem to be having significant gains on color based questions. This is explicated in the fact that in example 3, even multihop retriever failed to retrieve the second source. This can be attributed to the fact that multi-hop retrieval doesn't really add any new information to the color of an entity when independent sources are involved. However, in example 6, the multihop retriever still outperformed baseline retrieval on a color based question. This is perhaps because the sources although independent need to be checked for having the same colour (unlike the previous example where both colors need to be stated). It is much easier to reinforce information when comparing for similarity than extraction.

The last example is one of the rare cases where the dataset annotation for our multihop training is at fault. As is seen, S1 needs to be retrieved first for the need of S2 to be apparent. However, since S2 has more entity overlap, it is chosen first. Choosing S2 first and conditioning next retrieval on it will actually cause hampering of performance since S1 is best retrieved independently.

7.4 X-VLM retrieval vs Baseline retrieval

We compare the X-VLM retrieval with the baseline retrieval using a few interesting examples in Figures 8 and 9. We first compare how does the text retrieval change when the image encoding system is changed to X-VLM. This primarily will happen since the alignment and fusion of text with image will be affected by the image encoder during training. We can see in the first example that the X-VLM retriever performs poorly in text ques-

tions even when simple text-matching needs to be done. S1 is only retrieved when a huge amount of text overlap was present between the source and the question. The performance of the model on text questions seems to degrade when trained on X-VLM based image encoding.

In example 2, it seems that both models fail to retrieve the right sources. This can be attributed to poor caption encoding and lack of grounding on complex concepts like the "hair" on mushrooms.

In case of questions heavily oriented towards visual grounding like the example 3, it seems X-VLM doesn't degrade in performance in spite of possible alignment issues. X-VLM does seem to be able to identify skis and retrieve as well as the baseline.

Example 4 is a typical case where X-VLM is supposed to be used. Since the caption will mention the name of the building, X-VLM can explicitly focus only on the trees in the image as asked by the question. It can thus ground better and outperforms baseline retriever in spite of possible alignment issues and distribution shift.

7.5 X-VLM reader vs Baseline reader

Finally, we analyze specific examples to compare the X-VLM reader with the original baseline reader in figure 10. We explicate why there is a significant drop in performance on the X-VLM reader even though visual grounding seems to be a natural way to improve the baseline reader.

We notice that X-VLM does very good image grounding based on the question. However, the free form answer generation is what lacks substance in the X-VLM reader.

As in 7.4, we first compare the change in performance on text based questions when the image encoding system is changed to X-VLM. We take a simple generative example in the first question. Even in such a simple case, the lack of generative ability of X-VLM is apparent. It fails to copy word from the text even when attempting to give a single

Question	Gold Sources	Our Ordering	Independent or Not	R _B	R _M
What town in Yutian County has a river along it that flows for 519 km from the Kunlun Shan mountain range?	S ₁ : It flows for 519 km (322 mi) from the Kunlun Shan mountain range north into the endorheic Tarim Basin, but is lost in the desert several hundred kilometers south of the Tarim River. The only major settlement along the river is Keriya Town, east of Hotan. S ₂ : Keriya Town or Mugala Town is a town in Yutian (Keriya) County, Hotan Prefecture, Xinjiang, China, on the old Southern Silk Road. As the commercial and administrative centre of Keriya County, it is about 166 km east of Hotan, 80 km east of Qira, and 120 km west of Niya.	S ₁ , S ₂	No	S ₁ , S _{other}	S ₁ , S ₂
Which statue has a taller base: Monument to playwright Carlo Goldoni (in "Campo San Bartolomio" square in Venice) or Statue of Bartolomeo Colleoni by Andrea Verrocchio in Venice?	S ₁ :  S ₂ : 	S ₁ , S ₂	Yes	S ₂ , S _{other}	S ₁ , S ₂

Figure 5: Baseline Retriever vs Multihop Retriever – 1

word answer. Same is the case with example 2. Although baseline reader does correct generation, X-VLM fails to parse the text efficiently. It confuses the name suffix "I" for a proper noun, that is, the name of the kingdom.

In case of example 3, X-VLM identifies the right entity to bound. However, since this doesn't give any gains over the baseline model since the question talks about the most important entity in the images which even detectron can efficiently identify.

We also explicate certain failure cases of the X-VLM bounding box itself. It can be seen in example 4 that when the question is too wordy and when it talks about a lot of entities, the generic X-VLM bounding box is way too loose. It ends up focussing on the largest entity in the question unlike detectron which always detects every object separately. This issue can perhaps be alleviated by

using X-VLM to generate multiple bounding boxes instead of a single generic box. X-VLM also fails to generate a good fluent answer.

Example 5 is a case of X-VLM performing precise grounding and identifies the right entities to bound. Detectron object detection fails severely in this case to label the detected objects correctly. However, the poor free-form generative ability of X-VLM still makes it overall perform much worse than the baseline reader.

Question	Gold Sources	Our Ordering	Independent or Not	R _B	R _M
Are the German Antarctic research base Neumayer Station III and the Halley Research Station painted in the same colors?	<p>S₁: </p> <p>S₂: </p>	S ₂ , S ₁	Yes	S _{other} , S ₂	S ₂ , S ₁
The Spanish vihuela is widely considered to have been the single most important influence in the development of which string instrument with five courses of gut strings and movable gut frets?	<p>S₁: The Spanish vihuela, called in Italian the " viola da mano ", a guitar-like instrument of the 15th and 16th centuries, is widely considered to have been the single most important influence in the development of the baroque guitar.</p> <p>S₂: The Baroque guitar (c. 1600–1750) is a string instrument with five courses of gut strings and moveable gut frets. The first (highest pitched) course sometimes used only a single string. The Baroque guitar replaced the Renaissance lute as the most common instrument found when one was at home.[2][3] The earliest attestation of a five-stringed guitar comes from the mid-sixteenth-century Spanish book Declaracion de Instrumentos Musicales by Juan Bermudo, published in 1555.[4] The first treatise published for the Baroque guitar was Guitarra Española de cinco ordenes (The Five-course Spanish Guitar), c. 1590, by Juan Carlos Amat.</p>	S ₂ , S ₁	No	S ₁ , S ₂	S ₂ , S _{other}

Figure 7: Baseline Retriever vs Multihop Retriever – 3

Question	Gold Sources	R _B	R _X
What are a set S of natural numbers called in the theory whose goal is to determine which problems, or classes of problems, can be solved in each model of computation?	<p>S₁: One goal of computability theory is to determine which problems, or classes of problems, can be solved in each model of computation. A model of computation is a formal description of a particular type of computational process.</p> <p>S₂: In computability theory, a set S of natural numbers is called computably enumerable (c.e.), recursively enumerable (r.e.), semidecidable, partially decidable, listable, provable or Turing-recognizable if: There is an algorithm that enumerates the members of S. That means that its output is simply a list of all the members of S: s₁, s₂, s₃,</p>	S ₁ , S ₂	S ₁
Which fungus appears to be growing more hair like substance atop it; Coprinus comatus or Mycena acicula?	<p>S₁:</p>  <p>S₂:</p> 	S ₃	S ₃

Figure 8: Baseline Detectron Retriever vs XVLM Retriever – 1

Question	Gold Sources	R _B	R _X
How many more skis were used by Anders Södergren during the 2010 Olympics than were used by Martin Rulsch during the 2020 Winter Youth Olympics?	<p>S₁:</p>  <p>S₂:</p> 	S ₂ , S ₃	S ₂
Do the same kind of trees surround the Museo Nacional de Bellas Artes on all sides?	<p>S₁:</p>  <p>S₂:</p> 	S ₁	S ₁ , S ₂

Figure 9: Baseline Detectron Retriever vs XVLM Retriever – 2

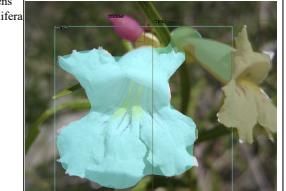
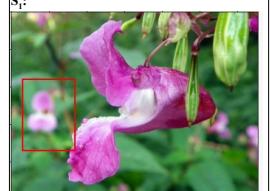
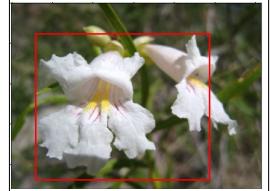
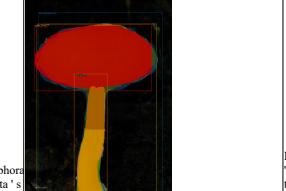
Question	Answer	Gold Sources	D _B	D _{B(BB)}	D _X	X _{BB}
In bacteria, promoter regions may contain a Pribnow box, which serves an analogous purpose to the eukaryotic what, which is a sequence of DNA found in the core promoter region of genes in archaea and eukaryotes?	TATA Box	<p>S₁: In molecular biology, the TATA box (also called the Goldberg-Hogness box) is a sequence of DNA found in the core promoter region of genes in archaea and eukaryotes. The bacterial homolog of the TATA box is called the Pribnow box which has a shorter consensus sequence .</p> <p>S₂: In bacteria, promoter regions may contain a Pribnow box, which serves an analogous purpose to the eukaryotic TATA box. The Pribnow box has a 6 bp region centered around the -10 position and an 8-12 bp sequence around the -35 region that are both conserved.</p>	TATA Box	NA	TAA	NA
From what ancient kingdom did Caranus and Aeropus I rule over?	Macedon.	<p>S₁: A rival tradition is held by Livy, Pausanias, Suidas and Julianus Justinus, with Caranus as the first Macedon king. According to Polyaeus' story, Argaeus founded the Dionysus cult with Maenads (latter depicted on a 480 BC vase fragment).</p> <p>S₂: Aeropus I of Macedon (Greek: Αέροπος Α' ὁ Μακεδών) was the son of Philip I, the great-grandson of Perdiccas I, the first king of the ancient kingdom of Macedon according to Herodotus, and the father of Alcetas .</p>	Macedon	NA	I.	NA
Which flower is lighter in color; Chilopsis linearisssp. arcuata or Impatiens glandulifera?	The Chilopsis linearisssp. arcuata is lighter in color than the Impatiens glandulifera.	<p>S₁: </p> <p>S₂: </p>	<p>S₁: The Chilopsis linearisssp. arcuata is lighter in color than Impatiens glandulifera .</p> <p>S₂: </p>	<p>S₁: </p> <p>S₂: </p>	<p>It is lighter in color .</p>	
Does the Dothan Civic Center or the Vista Civic Center have more flags in front of the building?	The Vista Civic Center has more flags in front than the Dothan Civic Center.	<p>S₁: </p> <p>S₂: </p>	<p>S₁: The Dothan Civic Center has more flags in front of the building</p> <p>S₂: </p>	<p>S₁: </p> <p>S₂: </p>	<p>The Vista Civic Center has more flags in front of the the Dothan Civic than Civic Center</p>	
Which fungus' top appears to be more like a Truffle; Dictyophora indusiata or Amanita albocreata?	Dictyophora indusiata looks more like a truffle than Amanita albocreata.	<p>S₁: </p> <p>S₂: </p>	<p>S₁: Dictyophora indusiata's top appears to be more like a Truffle . Dictyophora indusiata</p> <p>S₂: </p>	<p>S₁: </p> <p>S₂: </p>	<p>Dibobobot s s s truffule appears appears appears to to be be be be be be</p>	

Figure 10: Baseline Detectron Reader vs XVLM reader - 1

8 Future work and Limitations

As is mentioned in 7.2, X-VLM does good grounding but fails at free form answering. Perhaps adding the caption to the query conditioned feature generation might help. This can then be clubbed with the multihop retrieval to get a very good multihop X-VLM based retriever. The proposed model does not perform well on questions related to shape and color.

We have also identified multiple failures with respect to the dataset annotations. Long text answers, for example, are only generated from text snippets, and the image answers are short and can be converted to classification labels. Also, the answer originate from only one modality. An extension could be have answers from both the modalities.

Based on our analysis, F1 and Fluency * Accuracy is not the best suited for the task at hand. Fluency is primarily defined based on a paraphrasing goal and might not be appropriate for a generation task. Also, Recall at k (Recall@k) is a much more standard metric for evaluating retrieval performance than F1 scores.

Most of the models we used, processed either only the images together or only the text together or single image with the text. Processing the text and images separately can lead to poor alignment. There is a lack of multimodal models that takes multiple images and multiple text as an input.

End to end training of retriever and reader models is an important future direction. We believe incorporating reader and retriever in a Retriever Augmented Generation architecture might yield better results due to the end to end training. Also, present models use state of the art object detectors which do not account for the question giving aberrant representations and use either patch-based or ROI based features. Question conditioned image encoding might lead to better representations due to the better context provided. We have shown that multihop retrieval gives very good performance for retrieval task, exploring multihop reasoning based readers is also a good research direction.

9 Ethical Concerns and Considerations

In the WebQA dataset, there are many location-based questions, and a lot of these are American monuments. The dataset is hence biased towards American monuments. Further, Wikipedia is used as a data source. Wikipedia is one of the most popular reference sites on the web, but it is not a

credible source of information as anyone is allowed to be a contributor to the website. This brings in their own biases. Wikipedia Academic explains why it is a bad idea to consider it as a credible academic source ([wik, 2022](#)).

We have made use of pre-trained models like CLIP, XVLM, Detectron, OFA, BART, SpladeV2, BLIP. The biases present in these pre-trained models will propagate to our proposed solution. CLIP is pre-trained on publicly available image-caption data. The dataset was created using website crawling and utilizing commonly-used pre-existing image datasets like YFCC100M ([Thomee et al., 2016](#)). Major chunk of the data is from crawling the internet. This means that the data is more typical of the individuals and society most linked to the internet, which skews towards more developed countries and younger, male consumers. Similarly, other pre-trained models used is also trained on biased dataset. This will lead to some of these biases propagating to our models and hence our proposed solution.

A proposed approach to overcome these ethical issues would be to present a Model Card ([Mitchell et al., 2019](#)), to provide more transparency to the users. Transparency in machine learning models is crucial as it affect people’s lives. The information that downstream users want will differ, as will the details that developers require to determine whether a model is appropriate for their use case. Model Cards will provide a detailed overview of a model’s suggested uses and limitations, which can benefit developers, regulators, and downstream users alike. In our published modal card, we can mention the biases and other ethical concerns of the pre-trained models as well as the data we have used. An example would be to mention that Wikipedia has been used as the data source for WebQA, and it is not a reliable source of information. If the researchers are aware of these biases, they can work towards building models that makes these ethical considerations and attempts to overcome it.

10 Team member contributions

- **Aditya Veerubhotla:** Introduction, Problem definition, Multimodal IR, Related dataset, Task setup, Baselines, Proposed model, Loss function, Result, Analysis, Limitations and Future work.
- **Deep Karkhanis:** Introduction, Problem definition, Multimodal QA, Related dataset, Task setup, Baselines, Proposed model, Loss function, Result, Analysis, Limitations and Future work.
- **Harsha Vardhan:** Introduction, Problem definition, Multimodal IR, Related dataset, Task setup, Baselines, Proposed model, Loss function, Result, Analysis
- **Mashrin Srivastava:** Introduction, Problem definition, Multimodal Representation, Related dataset, Task setup, Baselines, Proposed model, PolyLoss, Result, Analysis, Limitations and Future work, Ethical concerns and considerations.
- **Soundarya Krishnan:** Introduction, Problem definition, Multimodal QA, Related dataset, Task setup, Baselines, Proposed model, Loss function, Result, Analysis, Limitations and Future work.
- **Srijan Bansal:** Introduction, Problem definition, Multimodal Representation, Related dataset, Task setup, Baselines, Proposed model, Loss function, Result, Analysis, Limitations and Future work.

Acknowledgement

We are very grateful to Yingshan for her valuable work in the creation and maintenance of the WEBQA dataset and model, as well as for clarifying our questions over meetings. We are also very grateful to Li-Wei Chen for his comments on our work. We also thank Prof. Yonatan Bisk, and all TAs associated with 11-777: Multimodal Machine Learning (Spring 2022) at Carnegie Mellon University.

References

2022. Wikipedia:academic use.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. [Webqa: Multihop and multimodal qa](#).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Ko Endo, Masaki Aono, Eric Nichols, and Kotaro Fukakoshi. 2017. [An attention-based regression model for grounding textual phrases in images](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3995–4001.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. [Vse++: Improving visual-semantic embeddings with hard negatives](#).

Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv*, abs/2109.10086.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. <https://github.com/facebookresearch/detectron>.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. [Manymodalqa: Modality disambiguation and qa over diverse inputs](#). *ArXiv*, abs/2001.08034.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. [Efficiently teaching an effective dense retriever with balanced topic aware sampling](#). *CoRR*, abs/2104.06967.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. [Efficiently teaching an effective dense retriever with balanced topic aware sampling](#).

Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. 2021. [GilBERT: Generative Vision-Language Pre-Training for Image-Text Retrieval](#), page 1379–1388. Association for Computing Machinery, New York, NY, USA.

Vladimir Karpukhin, Barlas Ouz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). *CoRR*, abs/2004.04906.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *CoRR*, abs/1910.13461.

- Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. [Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training](#). *CoRR*, abs/1908.06066.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. [BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation](#). *CoRR*, abs/2201.12086.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. [Oscar: Object-semantics aligned pre-training for vision-language tasks](#). *CoRR*, abs/2004.06165.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). *CoRR*, abs/1405.0312.
- Daqing Liu, Hanwang Zhang, Zhengjun Zha, and Feng Wu. 2019. Learning to assemble neural module tree networks for visual grounding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4672–4681.
- Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. [Generation-augmented retrieval for open-domain question answering](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). *FAT* ’19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *CoRR*, abs/2103.00020.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*.
- Arka Sadhu, Kan Chen, and Ram Nevatia. 2019. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. [MIMOQA: Multimodal input multimodal output question answering](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. [Multimodal{qa}: complex question answering over text, tables and images](#). In *International Conference on Learning Representations*.
- Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. [YFCC100m](#). *Communications of the ACM*, 59(2):64–73.
- Peter Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. 2021. [In factuality: Efficient integration of relevant facts for visual question answering](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 468–475, Online. Association for Computational Linguistics.
- Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. [Adversarial cross-modal retrieval](#). In *Proceedings of the 25th ACM International Conference on Multimedia*, MM ’17, page 154–162, New York, NY, USA. Association for Computing Machinery.
- Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. [Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. [Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). *CoRR*, abs/2202.03052.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. [Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). *CoRR*, abs/2202.03052.
- Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Ouz. 2020. [Answering complex open-domain questions with multi-hop dense retrieval](#). *CoRR*, abs/2009.12756.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. [Improving one-stage visual grounding by recursive sub-query construction](#). *CoRR*, abs/2008.01059.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016a. [Modeling context in referring expressions](#). *CoRR*, abs/1608.00272.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016b. [Modeling context in referring expressions](#). *CoRR*, abs/1608.00272.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. [Multi-grained vision language pre-training: Aligning texts with visual concepts](#). *CoRR*, abs/2111.08276.

Zhixiong Zeng and Wenji Mao. 2022. [A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval](#). *CoRR*, abs/2201.02772.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021a. [Vinvl: Making visual representations matter in vision-language models](#). *CoRR*, abs/2101.00529.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. [Vinvl: Revisiting visual representations in vision-language models](#). *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.