# WebQA: Multihop and Multimodal QA

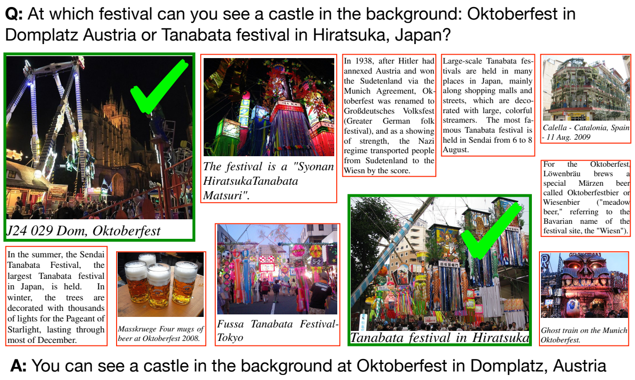
Yingshan Chang[1] Mridu Narang[2] Hisami Suzuki[2]
Guihong Cao[2] Jianfeng Gao[3] Yonatan Bisk[1,3]
[1]Carnegie Mellon University [2]Microsoft, Bing Search [3]Microsoft Research

## Abstract

*Scaling Visual Question Answering (VQA) to the open-domain and multi-hop nature of web searches, requires fundamental advances in visual representation learning, knowledge aggregation, and language generation. In this work, we introduce* WEBQA*, a challenging new benchmark that proves difficult for large-scale state-of-the-art models which lack language groundable visual representations for novel objects and the ability to reason, yet trivial for humans.* WEBQA *mirrors the way humans use the web: 1) Ask a question, 2) Choose sources to aggregate, and 3) Produce a fluent language response. This is the behavior we should be expecting from IoT devices and digital assistants. Existing work prefers to assume that a model can either reason about knowledge in images* ***or*** *in text.* WEBQA *includes a secondary text-only QA task to ensure improved visual performance does not come at the cost of language understanding. Our challenge for the community is to create unified multimodal reasoning models that answer questions regardless of the source modality, moving us closer to digital assistants that not only query language knowledge, but also the richer visual online world.*

## 1. Introduction

Web search is a multimodal experience: Will I find my answer on the image search tab or within text snippets? In contrast, most deployed Question Answering (QA) systems treat the web as a text-only landscape of facts to be extracted, ignoring the knowledge present in images. This has two fundamental limitations: 1. The text-based web is impoverished [3,4], and 2. This form of information extraction is inefficient. For example, when searching to see if a park has picnic tables, surfacing an image of the picnic area answers the question immediately, rather than wading through pages of reviews hoping someone happened to mention this fact. QA engines need to move to treating the Internet as a multimodal trove of information, but this requires multihop reasoning on either images or text.



**Q:** At which festival can you see a castle in the background: Oktoberfest in Domplatz Austria or Tanabata festival in Hiratsuka, Japan?

*J24 029 Dom, Oktoberfest*

*The festival is a "Syonan HiratsukaTanabata Matsuri".*

In 1938, after Hitler had annexed Austria and won the Sudetenland via the Munich Agreement, Oktoberfest was renamed to Großdeutsches Volksfest (Greater German folk festival), and as a showing of strength, the Nazi regime transported people from Sudetenland to the Wiesn by the score.

Large-scale Tanabata festivals are held in many places in Japan, mainly along shopping malls and streets, which are decorated with large, colorful streamers. The most famous Tanabata festival is held in Sendai from 6 to 8 August.

*Calella - Catalonia, Spain - 11 Aug. 2009*

For the Oktoberfest, Löwenbräu brews a special Märzen beer called Oktoberfestbier or Wiesenbier ("meadow beer," referring to the Bavarian name of the festival site, the "Wiesn").

In the summer, the Sendai Tanabata Festival, the largest Tanabata festival in Japan, is held. In winter, the trees are decorated with thousands of lights for the Pageant of Starlight, lasting through most of December.

*Masskruege Four mugs of beer at Oktoberfest 2008.*

*Fussa Tanabata Festival-Tokyo*

*Tanabata festival in Hiratsuka*

*Ghost train on the Munich Oktoberfest.*

**A:** You can see a castle in the background at Oktoberfest in Domplatz, Austria

Figure 1. Example WEBQA dataset pipeline in which the question requires finding and reasoning about two relevant sources and discarding distractors to produce the correct natural language answer.

To this end, datasets are rapidly emerging [10, 24, 28]. But they either use pre-defined templates for the curation of multihop multimodal QA pairs [28], or encourage a "question decomposition + rerouting to uni-modal model" approach to superficially solve the problem [10]. However, when humans absorb knowledge, there is no need to distinguish whether the knowledge was learned from books versus images, or whether a piece of knowledge is a composite of multiple scattered fragments versus being carried by a single one. We argue that genuine progress in reasoning over linguistic notions of meanings and visually grounded meanings under the same representation framework depends on the development of a unified system that indiscriminately treats snippets and images as knowledge carriers. On top of that, the goal includes better extraction, integration and summarization abilities in a heterogeneous information landscape.

To facilitate this research intersection, in this work we propose a novel benchmark, WEBQA, for *multi-hop, multimodal, open-domain question-answering* where all questions are knowledge-seeking and resemble real-world use cases. Success on WEBQA requires a system to a) incorporate both text and images, b) retrieve relevant knowledge in either modality, c) aggregate information from multiple sources via logical or numerical reasoning, and d) generate

| | #Train | #Dev | #Test | #Img | Len Q | Len A |
|---|---|---|---|---|---|---|
| VQA v2 [9] | 443K | 214K | 453K | 200K | 6.1 | 1.2 |
| OKVQA [18] | 9.0K | 0 | 5.0K | 14.0K | 8.1 | 1.3 |
| MultiModalQA [28] | 23.8K | 2.4K | 3.6K | 57.7K | 18.2 | 2.1 |
| ManyModalQA [10] | 2.0K | 3.0K | 5.1K | 2.9K | – | 1.0 |
| MIMOQA [24] | 52.4K | 0.7K | 3.5K | 400.0K | – | – |
| WEBQA (ours) | 34.2K | 5K | 7.5K | 390.0K | 17.5 | 12.5 |

Table 1. Comparison of multimodal knowledge-seeking benchmarks by size and average question/answer lengths.[1]

answers in natural language. We experiment with state-of-the-art multimodal reasoning and text generation models, whose failures indicate promising future directions.

## 2. Related Work

Many datasets and tasks can be broadly considered "question answering." For example, VQA [2, 9, 11, 18] is one of the widely studied tasks at the intersection of language and vision. Nevertheless, it is unclear how VQA models should be adapted to open-domain scenarios. This is largely due to the simplification of VQA tasks into classification over a fixed vocabulary of frequent answers. Recent work on video [15, 30, 32] has also adopted a multiple-choice format. In contrast, OK-VQA [18] broadens the task to knowledge-seeking questions. OK-VQA and our task differ in the role of images. Images in OK-VQA are regarded as part of the query rather than as part of the knowledge source, and can only be processed after retrieval.

Within the natural language community, QA datasets are experiencing a similar transition from multiple-choice and span prediction to the harder free-form answer generation paradigm. Multi-hop question answering has recently taken the spotlight as it aligns with the multi-hop nature of how humans perform reasoning during knowledge acquisition, leading to a proliferation of benchmarks [27, 31, 34].

There have been several recent benchmarks for reasoning over input and contexts in multiple modalities [26]. MultiModalQA [28] made the first foray into complex questions that require reasoning over snippets, tables and images. It focuses on cross-modal heterogeneous knowledge extraction. However, questions are generated from templates. Once a template is detected the task reduces to filling in blanks with modality-specific answering mechanisms.

ManyModalQA [10] also deals with snippets, images and tables. However, the primary challenge their design addresses is the choice of answer modality — rather than knowledge aggregation or extraction. Our focus is more about representing world knowledge in a unified space, than about distinguishing the answer modality, since mastering the former may naturally eliminate the need to classify questions according to the answer modality.

| | Eval Metrics | Answer Schema |
|---|---|---|
| VQA v2<br>OK-VQA | $min\{\frac{\#\text{human agreement}}{3}, 1\}$ | Top training answers |
| MultimodalQA | Exact Match<br>F1 | Txt: span/Y/N<br>Img: Fixed vocab<br>Table: Y/N, cell, or op. |
| ManymodalQA | Classification Accuracy | Context word or vocab |
| MIMOQA | Txt: ROUGE-1/-2/-L or BLEU<br>Img: Precision@1/@2/@3 | Span prediction<br>+ Image retrieval |
| WEBQA (ours) | Fluency: BARTScore<br>Keyword Acc: Recall/F1 | Complete NL sentence |

Table 2. Comparison of knowledge-seeking, multimodal benchmark metrics and answer schema.

Finally, MIMOQA [24] introduces a new concept of "Multimodal Input Multimodal Output" which highlights accompanying a textual answer with an image in order to enhance cognitive understanding. MIMOQA requires selecting a text span and an image from the context as an output pair. Their approach is nicely complementary to ours. Where we differ, is that our task also requires aggregation and summarization before producing the final natural language answer, whereas the outputs required by MIMOQA are not completely digested by the model. Here, we refer "digesting" to the ability to produce a reasonable output which cannot be directly copied from the input.

Tables 1, 2 and Appendix E provide comparisons between WEBQA and related datasets. No existing multimodal or knowledge-seeking benchmark requires the answers to be complete, free-form natural language sentences, as opposed to extractive spans, or elements from a finite set. Additionally, previous work has not supported both natural language generation (NLG) evaluation and accuracy-style evaluation as we do. To this end, we highlight that a) in WEBQA more importance is attached to digesting, aggregating and summarizing information as answers cannot be simply copied from an existing text span or image patch, b) WEBQA requires the source retrieval stage in addition to VQA, which better simulates the full reasoning pipeline during a web search, and c) answers in the form of a natural language sentence better transit to downstream applications such as conversational agents and voice assistants.

## 3. Task Formulation

As in Fig 1, examples consist of a question $Q$, a set of positive sources $s_1, ..., s_m$ (in green), a set of distractor sources $s_{m+1}, ..., s_n$ (in red) and an answer $A$. Each source can be either a snippet or an (image, description) pair. Each image is accompanied by a description to resolve names or geographic information not present in the image itself, but

[1] Note, MultiModalQA and ManyModal QA also contain tables – 3.5K for ManyModal and while 700k were used MultiModalQA's dataset generation, it is unclear how many ended up in the final dataset.

serve as critical links to references in the question. We include both a restricted ($n \approx 40$) and full ($n \approx 900K$) setting.

We decompose the task into two stages. First, given $Q$ and $s_1, s_2, ..., s_n$, the model identifies the sources from which to derive the answer. The second stage is question answering where the model takes $Q$ and the chosen sources as context $C$, to generate an answer $A$. Ideally, a single-stage system would jointly process $Q, s_1, s_2,\ ...,\ s_n$ to produce $A, C$, but we are unaware of any modeling approaches that can consume sufficiently large multimodal contexts to achieve this, so this is left to future work.

## 4. WebQA

Following the paradigm popularized by search engines, we structure our data as having answers that can be found either via image search or general web (text) search. Note, WebQA does not contain questions that need an image and an (independent) snippet as knowledge sources. However, *all* image-based questions already require processing *both* images and text as *image descriptions* provide necessary information. Below we outline how both types of questions are collected, structured, and filtered for quality.

### 4.1. Answers from Images

We collect both multi-image questions that require stitching two images to answer and complex single-image questions. Rich multi-image questions do not naturally exist at scale in user search logs,[2] likely because users do not issue queries they believe search engines cannot handle, thus we turn to crowdsourcing.

We presented annotators with a set of six related images and asked them to produce three QA-pairs by selecting one or two images per pair that are necessary to answer the question. We require that at least one of the three pairs utilizes two distinct images.Additionally, we instructed annotators to avoid questions that: a) are simple facts (e.g. "*How many wheels does a car have*"); b) are easily answered by a text-only search; c) are bound to a specific image; d) ensure every question is meaningful without paired context. This elucidates one of the key differences between the well-known VQA task and ours. In most VQA style tasks, every question is about a paired image, whereas in our task images serve as knowledge sources over which to reason, and do not serve the role of augmenting the question. To assist annotators, each image is accompanied by a description extracted from Wikipedia. This description is only to be used to confirm the name or location of the objects depicted. The answer has to be derived from visual clues.

Images were crawled from Wikimedia Commons via the Bing Visual Search API. Wikimedia's topic list cannot be used directly as most categories are (visually) uninteresting. We seeded with natural scenes and iteratively refined the image pool by removing categories flagged as (visually) uninteresting. This resulted in categories like animals, plants, attractions, and architecture (Fig 3).

**Hard Negative Mining**. We produce a set of both text- and image-based hard negatives for models to sift through for every question. Text sources are extracted from relevant passages on Wikipedia based on noun chunks in the question, while limiting overlap to avoid false negatives. For images, we leverage Bing APIs to find similar images with respect to both the description (via Bing Image Search) and the visual content (via Bing Image Insight). In total, we collect 25K image-based questions, each requiring an average of 1.4 visual sources, and paired with 15.3 text and 15.9 visual distractors. Question prefixes are visualised in Fig 2.

**Categorization**. We categorize questions into open and closed classes. Closed class questions include: color, shape, number (i.e."how many"), yes/no (Y/N), and "multi-choice" (MC). The rest are open class questions.

**Adversarial splits**. We construct our test set to be out-of-distribution when possible to reward models with better generalization and reasoning. For color, shape, and number questions, we partition the answer set and ensure that the majority class during training does not carry over to testing. For the "Y/N" and "MC" classes, we trained models on 10 random train-test splits and consistently difficult samples across splits were placed in the test set. Finally, we randomly split questions from the open-class "other".

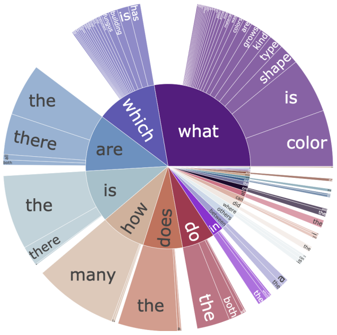


Figure 2. Image question prefixes (see Appendix B).

### 4.2. Answers From Text

We collected multi-hop QA pairs that involve combining knowledge from ≥2 snippets. To generate diverse, yet consistent, topics for mining difficult multi-hop reasoning questions, we construct clusters of similar entities, but where text snippets had low overall n-gram overlap or semantic similarity (yielding 8K clusters). We provide annotators with four snippets to prevent and allow them to contribute facts they researched to help answer the question.

**Hard Negative Mining**. For text distractors we mine passages from Wikipedia that contain noun phrases from the question and choose those with the highest lexical overlap

[2] While details are omitted here, we requested details from a search company that provided us basic statistics about query logs to confirm this.

|  | Question | Answer | Descriptions<br>Correct | Descriptions<br>Distract | Snippets<br>Correct | Snippets<br>Distract |
|---|---|---|---|---|---|---|
| Image | 16.4± 6 | 14.4± 6 | 13.3±11 | 12.6±11 | — | 36.4±10 |
| Text | 18.6± 8 | 10.7±10 | — | 14.1±13 | 45.3±12 | 38.3±10 |

Table 3. Length distribution for different textual components.

but lacking reference to the answer. For image distractors, we use the images and descriptions present on the aforementioned Wikipedia pages, again filtering for those with high lexical overlap. In total, we collected 24K text-based questions, each requiring 2.0 text sources, and paired with 14.6 text and 11.6 visual distractors. Lacking clear criteria for question categorization, we do not construct an adversarial test split, but instead simply sample randomly.

### 4.3. Quality Control

We ensure the data quality via crowdworkers training and expert-feedback-in-the-loop, which are found to be effective ingredients in crowdsourcing [19]. The initial pool of annotators were trained with a tutorial and selected via a qualification task. Additionally, we released the annotation task in batches to spot check quality after every batch, followed by sending constructive feedback to correct any deviation from our expectations. Workers who failed multiple times were de-qualified. Crowdsourcing data is challenging in that crowdworders are usually income-driven and will stick to a fixed answer generation pattern once they find it lucrative. To better align the crowdworkers' incentives with our goal, we generously bonus out-of-the-box thinking. All data was then also run through additional validation HITs to ensure agreement. Annotator pay averaged $13/hr overall (lower on the initial qualification and higher on the annotation/validation). Appendix A contains rubics and interfaces.

### 4.4. Dataset Statistics

In total, WEBQA has over 34K training QA pairs, with an additional 5K and 7.5K held out for development and testing. Overall Statistics are summarized in Table 4 and language distributions are presented in Table 3.

| Modality | Train | Dev | Test |
|---|---|---|---|
| Image | 18,954 | 2,511 | 3,464 |
| Text | 17,812 | 2,455 | 4,076 |

Table 4. Number of samples collected for each modality fold.

**Multi-hop**. 44% of image-based queries and 99% of text-based queries require two or more knowledge sources. This is verified by crowdworkers during validation to ensure that multiple knowledge sources provide non-overlapping information and cannot be replaced by each other. Additionally, as image sources also require understanding the caption, even single-image queries require multi-source reasoning.

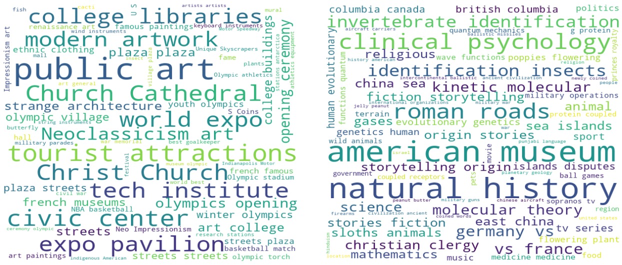


Figure 3. Samples of common topics in the image-based (left) and text-based (right) folds of the data.

**Topics**. Fig 3 provides a qualitative sense of the wide range of topics covered in WEBQA. In contrast to MultiModalQA, the images in WEBQA concentrate on the natural-world, events, and locations rather than digital artifacts (e.g. posters/logos). Snippets also exhibit a wide range of topics from contemporary science to ancient mythology. When comparing the topic clouds, it is clear that image-based queries more often relate to physical entities while text-based queries tend to be more abstract.

## 5. Metrics

WEBQA requires a model to answer open-domain questions and cite its sources. Therefore, we evaluate model performance with respect to both relevant fact prediction and question answering. While fact retrieval is easily evaluated via F1, language fluency and accuracy metrics are nuanced.

### 5.1. Question Answering Metrics

Our task expects fluent and complete sentences as answers, which we believe are appropriate for applications such as voice assistants or conversation agents. Therefore, the quality is measured as both fluency and accuracy. On each testing sample we collected five full-sentence answers written by humans. In addition, we collected one keyword answer by asking human annotators to rephrase the full-sentence answer into a succinct minimal semantic form.

**Fluency**. We measure fluency via BARTScore [35], a newly proposed NLG evaluation metric based on accurate measurement of paraphrase quality. BARTScore($A$, $B$) measures the probability of generating $B$ from $A$. In our setting, this is computed as BARTScore($r$, $c$), which can be interpreted as the probability of generating a candidate given a reference. Since BARTScore is based on the generation likelihoods, it does not distribute neatly across $[0, 1]$. So we normalize BARTScore($r$, $c$) by the identity score BARTScore($r$, $r$). On top of that, we make the normalized score bounded by 1. Finally, we choose the best score for a candidate across all references, as illustrated in Eq. 1.

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

This formulation a) prioritizes semantic agreement and is robust to functional words misplacement, b) does not heavily punish short sentences (i.e. $< 4$ words) as BLEU4 [20] does, c) penalizes word reordering / disfluencies d) and unlike BERTScore [37], which indiscriminately treats all colors or all shapes as nearly identical, BARTScore better captures small but critical differences. However, no language based embedding metrics accurately evaluate visual phenomena, so we also introduce an accuracy metric.

**Accuracy**. To ensure answer accuracy we use the collected keywords. Note, our paradigm differs from both open-domain text QA which focuses on lexical F1 and visual QA which uses a multiple choice evaluation. F1 rewards copying the question even if the key information is missing (e.g. the wrong color or count is chosen). Conversely, multiple-choice paradigms are not applicable to evaluate generated sentences. The goals of measuring accuracy on WebQA are: 1. Detect the presence of key entities. 2. Penalize the use of any incorrect entities. 3. Avoid penalizing semantically relevant but superfluous words. We are unaware of any solution to all of these criteria in the naturally mixed setting of our data (open-domain entities with a nearly closed-domain set of properties), so we propose an appropriate metric to tackle the different styles of answers.

Given the aforementioned question categorization for visual queries, questions having closed answer domains should be evaluated via F1 that tests for precision (avoiding a model producing both Yes and No to game the metric). We define the answer domains $D_{qc}$ of those question categories ($qc$) in Table 5. For the remaining visual queries and all textual queries, they have diverse and unrestricted answer domains. So, there are good reasons to believe that the probability of cheating by guessing a long list of keywords is small and would be penalized by BARTScore, so we evaluate accuracy via recall (RE). With $c$ as a candidate output, $K$ for correct answer keywords, and $qc$ for question category, Equation 2 sketches our **Acc** score.

| $qc$ | ”Answer Domain” $D_{qc}$ |
|---|---|
| | Union of keywords |
| color | ... across color queries |
| shape | ... across shape queries |
| number | ... and #s in references |
| Y/N | {’yes’, ’no’} |

Table 5. “closed” classes

$$\mathbf{Acc}(c, K) = \begin{cases} \text{if } qc \in [\text{color}, \text{shape}, \text{number}, \text{Y/N}] : \\ \qquad F1\big(c \cap D_{qc}, K \cap D_{qc}\big) \\ \text{otherwise :} \\ \qquad RE\big(c, K\big) \end{cases} \tag{2}$$

Finally, we report the average combined fluency and accuracy score **FL*Acc** across all test samples as a single evaluation result for a system.[3]

## 6. Baseline Models

We test existing models on WebQA in both fune-tuned and few-shot settings. The former fine-tunes a pre-trained vision-and-language transformer [38] on our source retrieval and QA tasks, while the latter (PICa [33]) prompts GPT-3 [5] with engineered prefixes. Note, since the answer space in WebQA is inappropriate for the classification approach (3K answers) considered by most VQA models, these models [6, 17, 25, 29], cannot be applied in our generative task.[4] At present, VLP [38] and Oscar [16] are the top generative multimodal transformers. Oscar is built on VLP so we chose VLP as more canonical but include the state-of-the-art visual features of VinVL implemented in Oscar+ [36]. Other recent models [7] may also have complementary strengths. To test the largest possible language model, we also run PICa [33] which leverages VinVL based captioning to augment GPT-3 with oracle source knowledge. Finally, to simulate the full retrieval setting, we ran zero-shot sparse and dense retrieval models over the entire collection of sources.

### 6.1. Fine-tuning Approach

We train two separate models for source retrieval and question answering on from released VLP [38] weights.

**Input Representation**. Text segments, including the questions, answers, textual sources and image captions, are tokenized by the `Bert-base-cased` [8] tokenizer. Each image is represented by 100 regions predicted by an object detection model, which is a variant of Faster RCNN with an ResNeXt-101 FPN backbone, pretrained on Visual Genome [14]. We take the output of fc1 layer from the object detection network an 2048-dim feature and finetune the fc2 layer. We also experiment with the latest state-of-the-art representations from VinVL [36]. Comparing to ResNeXt-101 FPN, the major advances of VinVL include a larger backbone (ResNeXt-152), replacement of FPN by C4[5] and better pretraining enriched by attribute information.

**Source Retrieval**. Candidate sources $s_1, s_2, ..., s_n$ are fed to the model one by one. Each pass takes the concatenation of $<$`[CLS]`, $s_i$, `[SEP]`, $Q$, `[SEP]`$>$ and estimates probability of a particular source being selected. Let $\mathcal{G}$ and $\mathcal{D}$ denote the set of gold sources and distractors for a sample. The loss function is as follows.

[3] Our metric does not solve NLG evaluation. Specifically, the “MC” question type often takes the form: ”which one in set $S$ has property *xyz?*”. Unlike the categories in Table 5 where it is wrong to output incorrect elements, including more elements in additional to the correct element in an answer may be correct if asked to compare the elements. We leave this to future NLG evaluation research as outside the scope of this work.

[4] see Appendix C for limitations of classification in WebQA

[5] Prior work [12, 36] has shown that C4 features are more effective for VL tasks due to its ImageNet weight initialization and inductive bias of the convolutional head. Both factors are not present in the MLP head of FPN.

$$Loss_{retrieval} = \sum_{s_i \in \mathcal{G}} logp_{s_i} + \sum_{s_i \in \mathcal{D}} log(1 - p_{s_i}) \quad (3)$$

**Question Answering**. We feed $<$`[CLS]`, $S$, `[SEP]`, $Q$, $A$, `[SEP]`$>$ to the Transformer, where attention masks are applied to tokens in $A$ to satisfy the auto-regressive property. We use standard Masked-Language-Modeling [8] loss during fine-tuning. We decode by iteratively appending a `[MASK]` to the end of the input, replacing it with a predicted token and appending a new `[MASK]` for the next timestep. Generation stops upon seeing `[SEP]`, `[PAD]`, or reaching a maximum length. We use beam search ($n = 5$) and choose the most confident output for evaluation.

**Model Variants**. In addition to the standard VLP trained on full data, we also include two modality-specific variants VLP$^{\mathcal{I}}$ and VLP$^{\mathcal{T}}$, which are trained on image- or text-based queries only as opposed to the full data, in order to reveal gains and losses resulted from the complexity of presenting models with data from both modalities,

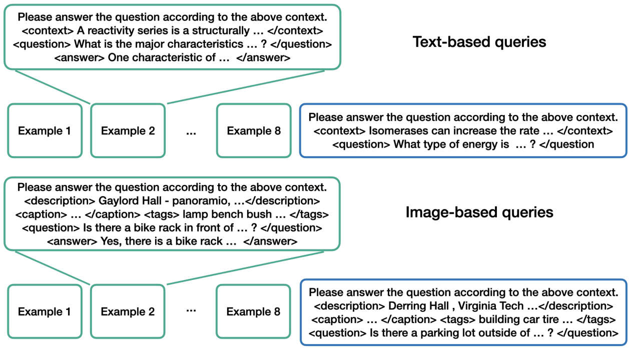
## 6.2. Zero-shot Full-scale Retrieval Approach

For end-to-end performance in an open-domain setting, we consider the entire collection of sources as our retrieval space (390k images and 540k text sources). Since running VLP-based retrieval of the test set over the entire source collection is prohibitively expensive (~3 years), we consider both sparse retrieval (BM25 [23]) and dense retrieval for a coarse filtering. Dense retrieval was achieved via CLIP [21] encoding all image and text sources, as well as all questions.[6] Next, using the modality knowledge, we rank all image/text sources based on the question-source similarity.

## 6.3. Few-shot Question Answering Approach

PICa [33] is the strongest model on OK-VQA [18], where GPT-3 is prompted to generate answers given a few training samples as prefix. We adapt PICa to our QA task using oracle sources to provide an upper-bound for the best possible performance of the strongest known models. PICa (and GPT-3) exhibit unstable behaviors on source prediction when presented with >4 choices (as it is most familiar with 4-way multiple choice tasks). Due to the inability to fine-tune, we cannot construct a truly fair comparison of PICa with our other baselines on our full pipeline.

We construct an input prefix by concatenating the pre-selected training examples, the context and the question of a testing sample. Since PICa's transformer backbone does not accept visual input, each image is described by three text segments, namely 1) a wikipedia description, b) a caption generated by Oscar+ [36] and c) a list of tags predicted by Oscar+. Limited by the maximum input length, we experimented with an 8-shot setting. If the input length exceeds the maximum length, we decrease the number of shots until it fits in the length budget.



Please answer the question according to the above context.
<context> A reactivity series is a structurally ... </context>
<question> What is the major characteristics ... ? </question>
<answer> One characteristic of ... </answer>

Please answer the question according to the above context.
<context> Isomerases can increase the rate ... </context>
<question> What type of energy is ... ? </question

Please answer the question according to the above context.
<description> Gaylord Hall - panoramio, ...</description>
<caption> ... </caption> <tags> lamp bench bush ... </tags>
<question> Is there a bike rack in front of ... ? </question>
<answer> Yes, there is a bike rack ... </answer>

Please answer the question according to the above context.
<description> Derring Hall , Virginia Tech ...</description>
<caption> ... </caption> <tags> building car tire ... </tags>
<question> Is there a parking lot outside of ... ? </question>

Figure 4. Few-Shot GPT-3 prompts.

**Training Example Selection**. Training examples to be included in the prefix for each testing sample are selected according to both question and source similarities. We use CLIP [22] to extract text or image encodings for questions, oracle snippets and oracle images. When multiple sources exist, we take the average of pairwise similarities between sources in one sample and sources in the other.

**Prompt Design**. We use XML-style brackets [13] to denote different text segments. See Fig 4 for what constitutes a prompt for a text- or image-based query.

# 7. Results & Analysis

Below we present results and analysis of our baselines' performance on WebQA. We include question-only baselines for both VLP and PICa to investigate how effectively models use the sources. VLP scores 22.6 on the proposed metric when evaluated end-to-end (Table 6). Modest improvement can be achieved by knowing the gold sources, showing room for growth on retrieval correctness. We observe that the latest best-performing visual encoder, VinVL, does not lead to significant gains. This may support the argument that the missing aspects from the status quo are more reflected in cross-modal information sharing than in the imperfection of uni-modal representations. PICa achieves a large gain over VLP. Promising as it is, we later show that, while pursuing the benefits of scaling up is one thing, there is still a lot remaining to be done to combat the diminishing returns involved with scale [1]. We show that humans can perform our task with ease (i.e. achieving >94 **Acc** and >55 **FL**) computed via cross-evaluation on multiple (3-6) references provided by different annotators to

[6] CLIP never assigns an image as more similar to a question than any text snippet, so we assume knowledge of what modality to retrieve. A BERT modality classifier can also achieve near perfect accuracy, but future unified approaches will hopefully not require this simplification.

| | | Source | QA Pred. source | | | QA Oracle source | | |
|---|---|---|---|---|---|---|---|---|
| | | **F1** | **FL** | ×**Acc** | = | **FL** | ×**Acc** | = |
| Restricted | VLP (Q-only) | —— | 34.9 | 22.2 | 13.4 | 34.9 | 22.2 | 13.4 |
| | VLP | 68.9 | 42.6 | 36.7 | 22.6 | 44.6 | 40.4 | 24.5 |
| | + VinVL | **70.9** | **44.2** | **38.9** | **24.1** | 45.7 | 42.2 | 25.9 |
| | PICa (Q-only) | — | — | — | — | 47.6 | 43.4 | 28.8 |
| | PICa | — | — | — | — | 57.1 | 61.6 | 40.1 |
| Full | CLIP$_{(2)}$ +VLP | 12.0 | 34.2 | 24.1 | 14.6 | — | — | — |
| | CLIP$_{(20)}$+VLP | **24.0** | **36.1** | **27.2** | **16.1** | — | — | — |
| | Human | **90.5** | — | — | — | **55.1** | **94.3** | **52.4** |

Table 6. We present both a "restricted" setting with relevant sources to pick between and a "full" setting in which retrieval includes all sources. Both VLP [38] and PICa [33] leverage VinVL [36] features. CLIP$_{(20)}$ uses VLP to further filter to two sources for QA (Table 7) and is 8pts weaker than the restricted setting.

prove robustness and consensus. While models' **FL** scores are high, reaching human-level accuracy is not within sight.

### 7.1. Source Retrieval

Crucial to a complete system design is multi-modal source retrieval. We investigate the effect of retrieval scale (Table 7) and dense versus sparse retrieval approaches. For the VLP-based model, sources are selected if its binary classification confidence is above a specified threshold. While the optimal thresholds for different models may vary, for fair comparisons we use 0.2, which is optimal for VLP on the development set.

| | Query Type | Image | Text |
|---|---|---|---|
| Restr. | BM25 | 25.61 | 43.75 |
| | VLP | **68.13** | **69.48** |
| Full | BM25 | 20.43 | **28.15** |
| | CLIP$_{(2)}$ | 9.71 | 13.96 |
| | CLIP$_{(20)}$+VLP | **21.68** | 26.01 |

Table 7. Source Retrieval (F1↑) over ∼40 sources (Restr.) or the full corpus. In CLIP$_{(20)}$+VLP, VLP reranks the top 20 sources retrieved.

VLP achieves >68% F1 given a restricted set of candidates. Indicating that it can model semantic relevance, despite its lack of scalibility. In comparison, we use simpler and less expensive approaches when scaling up to the full collection which causes our overall performance to degrade substantially (likely due both to the ambiguity and the weaker underlying document representations). The dense retrieval method suffers from a greater performance drop compared to sparse retrieval. Having VLP rerank the top 20 sources predicted by CLIP doubles F1, which holds promise for a future of large-scale coarse-to-fine retrieval that strikes a better accuracy-efficiency balance. See Appendix D for additional retrieval results.

| | | Image | | | | | | Text |
|---|---|---|---|---|---|---|---|---|
| | | Y/N | MC | Color | Shape | Number | Other | |
| | # samples | 935 | 981 | 228 | 62 | 200 | 1058 | 4067 |
| Fine-Tune | VLP (Q-only) | 16.1 | 49.0 | 3.9 | 0.8 | 0.5 | 27.9 | 18.1 |
| | VLP | **17.2** | 52.9 | 2.8 | 0.0 | 0.5 | **28.6** | **50.4** |
| | VLP$^{\mathcal{I}}$ | 11.6 | **55.3** | **3.9** | **2.4** | **0.5** | 26.3 | —— |
| | VLP$^{\mathcal{T}}$ | —— | —— | —— | —— | —— | —— | 48.6 |
| Few | PICa (Q-only) | 26.7 | 70.1 | 30.8 | 19.0 | 14.2 | 45.3 | 42.8 |
| | PICa | 27.4 | 70.1 | 42.5 | 17.3 | 13.8 | 48.7 | 74.8 |
| | Human | **100** | **96.8** | **95.8** | **94.8** | **95.0** | **87.8** | **94.0** |

Table 8. QA performance breakdown by question categories when presented with oracle sources: **Acc** ↑

### 7.2. Question Answering

Table 8 provides an accuracy breakdown with respect to question categories. A noticeable pattern is that models are more capable of solving text-based queries than image-based queries. Both VLP and PICa greatly surpasses the question-only baseline and VLP performs favorably against VLP$^{\mathcal{T}}$, demonstrating reasonable use of sources and the effectiveness of combined training.

On the other hand, image-based queries pose a much harder challenge. VLP and VLP$^{\mathcal{I}}$ are no better than the question-only baseline on image-based queries. While this may be an issue of the sources being ignored, we also attribute this to the fact that the image-based testing samples are intentionally constructed to prevent the success of any superficial correlations that can be drawn from the training set (e.g. the majority answers in each category). We observe a similar issue with PICa. Although PICa consistently outperforms VLP, it does not demonstrate an appropriate utilization of the provided sources, which is especially true on "Y/N", "MC", "Shape", "Number" and "Other" question categories. PICa has a surprising amount of knowledge embedded in its parameters, but unlike with text, on images it shows very little improvement from the inclusion of visual sources, as such it is still lacking the ability to explicitly and effectively use the retrieved sources, which might be crucial for further progress towards human accuracy.

We argue that performance is bottlenecked by the lossy textual representation of images consumed by PICa, thereby calling for concerted effort from both language and vision sides to build a unified representation rather than simply relying on one modality being translated to the other. For future research, we expect to explore whether symbolic or compositional representations in a structured problem space could equip a generative model with skills to perform aggregation beyond simple extraction.

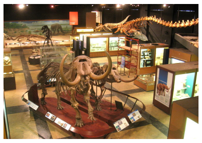
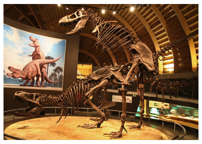
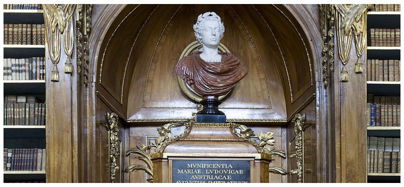
| Source(s) | Question (**Q**), Predicted Answer (**Pred**), & Correct Keywords (**KW**) |
|---|---|
|   | **Q:** Are the land dinosaurs guarded by rail in both the Display Museum of Natural History in University of Michigan and the Museo Jurassic de Asturias?<br>**Pred:** No, the land dinosaurs are not guarded by rail. **KW:** Yes<br>**Notes:** The prediction is wrong but the output sentence is consistent in terms of negation. |
|  | **Q:** What is the sculpted bust at the Baroque library, Prague wearing on its head?<br>**Pred:** The sculpted bust at the Baroque library, Prague is wearing a helmet on its head .<br>**KW:** A flower wreath<br>**Notes:** The model does not seriously consider the visual information. Comparing to a wreath, helmet is more likely to appear on a head and thereby being a safer choice. |
| 1. After the heavy 707 quad-jet was introduced in 1958, Boeing addressed the demand for shorter flight lengths from smaller airports. On December 5, 1960, the 727 was launched with 40 orders each from United Airlines and Eastern Air Lines .<br>2. The first airliner with jet power only was the Nene-powered Vickers VC.1 Viking G-AJPH, which first flew on 6 April 1948. | **Q:** How many years after the flight of the first jet airliner was the Boeing 727 released ?<br>**Pred:** 727<br>**KW:** 12 years |
| **Notes:** The model realizes a number is required, but being unable to perform arithmetic, simply copies a number from the snippets. | |
| 1. Posterior reversible encephalopathy syndrome (PRES), also known as reversible posterior leukoencephalopathy syndrome (RPLS), is a rare condition in which parts of the brain are affected by swelling, usually as a result of an underlying cause.<br>2. The diagnosis is usually made by brain scan (MRI) on which areas of swelling can be identified. The treatment for PRES is supportive: removal of the cause or causes and treatment of any of the complications, such as anticonvulsants for seizures. | **Q:** How is the condition also known as reversible posterior leukoencephalopathy syndrome (RPLS) diagnosed?<br>**Pred:** It is diagnosed by swelling , usually as a result of an underlying cause.<br>**KW:** by brain scan (MRI) |
| **Notes:** The selected span from the first source is relevant but does not inform the diagnostic method | |

Table 9. Common failures (see supplementary for additional predictions) include attempts at extraction or language model hallucinations.

### 7.3. Qualitative Analysis

Finally, we perform a qualitative analysis of the model's failures for both image- and text-based questions. Table 9 includes two image-based and two text-based examples with commentary (additional analysis in Appendix F). Both image questions are clean examples of producing logically consistent and fluent sentences which are incorrect. The first matches the negation but the answer should have been yes, while in the second, the model runs away with a very logical hallucination (heads wear helmets).

In the text examples, we see a different pattern. Here the model is more easily able to copy facts from the source texts, but still demonstrates a lack of understanding or reasoning. In the first example, the model appears to know it is looking for a number, but choosing one via direct copying rather than performing the arithmetic necessary to combine both facts. In the second case, the model finds a relevant span selection (as is commonly the only thing necessary for text QA tasks), but does not understand that the question is asking about a method of diagnoses versus a symptom.

None of the questions presented here require complex problem-solving skills. They follow rather simple implication, addition, or visual extraction patterns which are out of reach for current models (uni- or multi-modal).

## 8. Conclusion

WEBQA is a new multi-hop, multi-modal question answering challenge for our community. Designed to simulate the heterogeneous information landscape one might expect during a web search, WEBQA covers a series of open-domain general visual queries while also forcing models to still reason about text. Our task requires a system to determine relevant sources, perform aggregation and reasoning. We also propose a novel general recipe for evaluation on WEBQA which measures both fluency and accuracy.

Neither the versatile V&L transformer nor the large-scale text generator present a nearly-there solution. We provide both a restricted and full retrieval setup, to bridge multimodal QA and IR research. This dataset not only mirrors our everyday experience on the web, but provides a playground for the community to explore important sub-challenges, targeting the creation of a single model for multimodal reasoning, knowledge aggregation, and open-domain visual understanding.

WEBQA aims to facilitate research into constructing a single model which can 1) retrieve relevant documents, and 2) integrate information across a large context window including multiple paragraphs and images, in order to 3) generate fluent natural language answers.

# References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. *arXiv preprint arXiv:2110.02095*, 2021. 6

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[3] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020. 1

[4] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 1

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, 05 2020. 5

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 5

[7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5, 6

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[10] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886, 2020. 1, 2

[11] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. 2

[12] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 5

[13] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021. 6

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018. 2

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 5

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32:13–23, 2019. 5

[18] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. 2, 6

[19] Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R Bowman. What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. 4

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[22] 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. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 02 2021. 6

[23] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. 6

[24] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 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, 2021. 1, 2

[25] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 5

[26] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[27] Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, 2018. 2

[28] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. 1, 2

[29] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 5

[30] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[31] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. 2

[32] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, 2018. 2

[33] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021. 5, 6, 7

[34] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018. 2

[35] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*, 2021. 4

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 5, 6, 7

[37] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5

[38] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press, 2020. 5, 7