A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge

Dustin Schwenk[1], Apoorv Khandelwal[1], Christopher Clark[1], Kenneth Marino[2], and Roozbeh Mottaghi[1]

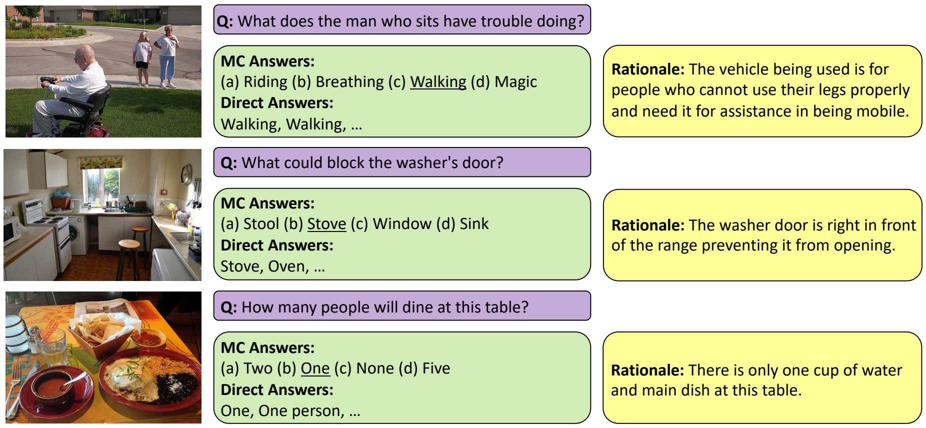
[1] PRIOR @ Allen Institute for AI
[2] Carnegie Mellon University
http://a-okvqa.allenai.org

Abstract. The Visual Question Answering (VQA) task aspires to provide a meaningful testbed for the development of AI models that can jointly reason over visual and natural language inputs. Despite a proliferation of VQA datasets, this goal is hindered by a set of common limitations. These include a reliance on relatively simplistic questions that are repetitive in both concepts and linguistic structure, little world knowledge needed outside of the paired image, and limited reasoning required to arrive at the correct answer. We introduce A-OKVQA, a crowdsourced dataset composed of a diverse set of about 25K questions requiring a broad base of commonsense and world knowledge to answer. In contrast to existing knowledge-based VQA datasets, the questions generally cannot be answered by simply querying a knowledge base, and instead require some form of commonsense reasoning about the scene depicted in the image. We demonstrate the potential of this new dataset through a detailed analysis of its contents and baseline performance measurements over a variety of state-of-the-art vision–language models.

1 Introduction

The original conception of the Visual Question Answering (VQA) problem was as a Visual Turing Test [11]: can a computer answer questions about an image well enough to fool us into thinking it's human? To truly solve this Turing Test, the computer would need to mimic several human capacities including: visual recognition in the wild, language understanding, basic reasoning abilities, and a background knowledge about the world. In the years after VQA was formulated, many of these aspects have been studied. Early datasets mostly studied the perception and language understanding problem on natural image datasets [2,12,30]. Other datasets studied complex chains of reasoning about procedurally generated images [21]. More recent datasets include questions requiring factual [32, 47, 48] or commonsense knowledge [53] to answer.

But, VQA has largely been a victim of its own success. With the advent of large-scale pre-training of vision–language models [5, 8, 28, 29, 38, 50, 54] and other breakthroughs in multi-modal architectures, much of the low-hanging fruit in the field has been plucked and many of the benchmarks have seen saturated



Q: What does the man who sits have trouble doing?

MC Answers:
(a) Riding (b) Breathing (c) Walking (d) Magic
Direct Answers:
Walking, Walking, ...

Rationale: The vehicle being used is for people who cannot use their legs properly and need it for assistance in being mobile.

Q: What could block the washer's door?

MC Answers:
(a) Stool (b) Stove (c) Window (d) Sink
Direct Answers:
Stove, Oven, ...

Rationale: The washer door is right in front of the range preventing it from opening.

Q: How many people will dine at this table?

MC Answers:
(a) Two (b) One (c) None (d) Five
Direct Answers:
One, One person, ...

Rationale: There is only one cup of water and main dish at this table.

Fig. 1: **A-OKVQA dataset** includes questions that require reasoning via a variety of knowledge types such as commonsense, world knowledge and visual knowledge. We provide Multiple-Choice (MC) as well as Direct Answer evaluation settings. There are 3 rationales (one shown) associated to each question in the train set providing the explanation/knowledge for answering the question.

performance. Even performance on newer knowledge-based datasets has been improved by such models [54]. So how can we continue developing yet more challenging datasets? A good start is to ask which human capabilities are not yet expressed by current models.

We propose the following options. First, continuing the direction of past work in knowledge-requiring VQA, we further expand the areas of knowledge required. Our dataset requires diverse forms of outside knowledge including explicit fact-based knowledge that is likely to be contained in knowledge bases, commonsense knowledge about human social behavior, intuitive understanding of physics, and visual knowledge. Second, we increase the complexity of reasoning needed to answer questions. Our questions require models to recognize the image, understand the question, recall relevant knowledge, and use reasoning to arrive at an answer. For instance, in the first question shown in Figure 1, the model should reason that people use that type of cart to avoid walking, and therefore the old man likely has trouble with this activity. In general, our dataset requires additional types of world knowledge compared to our previous work OK-VQA [32]. Hence, we call it Augmented OK-VQA (A-OKVQA).

A-OKVQA is composed of about 25K questions paired with both multiple choice (MC) answer options and ten free-form answers to allow for direct answer (DA) evaluation. The MC component of the dataset bypasses many difficulties inherent in (DA) evaluation and allows for a simple, clean accuracy score. This is particularly helpful given the greater variety in answers in A-OKVQA questions. At the same time, we believe direct answer evaluation is important to encourage models with more real-world applicability. In addition to the questions and answers, we provide *rationales* for each question. These brief explanatory statements provide extra information that can be used for training reasoning or knowledge retrieval methods, or to build more explainable VQA models.

In this work, our contributions are: (i) A new benchmark VQA dataset requiring diverse sources of outside knowledge and reasoning; (ii) A detailed analysis of the dataset that highlights its diversity and difficulty; (iii) An evaluation of a variety of recent baseline approaches in the context of the challenging questions in A-OKVQA; (iv) An extensive analysis of the results leading to interesting findings (e.g., how well models perform when answers are in the tail of the distribution, and the complementarity of the studied models).

2 Related Work

Visual Question Answering. Visual Question Answering (VQA) has been a common and popular form of vision–language reasoning. Many datasets for this task have been proposed [2,9,23,30,40,46,52,56] but most of these do not require much outside knowledge or reasoning, often focusing on recognition tasks such as classification, attribute detection, and counting.

Knowledge-based VQA datasets. Several previous works have studied the problem of knowledge-based VQA. The earliest explicitly knowledge-based VQA datasets were KB-VQA [47] and FVQA [48]. While these benchmarks did specifically require knowledge for questions, the knowledge required for these benchmarks is completely "closed". FVQA [48] is annotated by selecting a triplet from a fixed knowledge graph. KVQA [41] is based on images in Wikipedia articles. Because of the source of the images, these questions tend to mostly test recognizing specific named entities (e.g., Barrack Obama) and then retrieving Wikipedia knowledge about that entity rather than commonsense knowledge.

Most similar to our work is OK-VQA [32]. This dataset was an improvement over prior work in terms of scale, and the quality of questions and images. It also has the property that the required knowledge was not "closed" or explicitly drawn from a particular source, and could be called "open"-domain knowledge. While this is an improvement over the previous works, it still suffers from problems which we address in this work. The knowledge required, while "open" is still biased towards simple lookup knowledge (e.g., what is the capital of this country?) and most questions do not require much reasoning. In contrast, our dataset is explicitly drawn to rely on more common-sense knowledge and to require more reasoning to solve. In addition, our dataset includes "rationale" annotations, which allow knowledge-based VQA systems to more densely annotate their knowledge acquisition and reasoning capabilities. S3VQA [19] analyzes OK-VQA and creates a new dataset which includes questions that require detecting an object in the image, replacing the question with the word for that object and then querying the web to find the answer. Like OK-VQA, its questions have the shortcoming of generally requiring a single structured knowledge-retrieval, rather than commonsense knowledge and reasoning.

Another related line of work is Visual Commonsense Reasoning (VCR) [53]. VCR is also a VQA dataset, but is collected from movie scenes and is quite focused on humans and their intentions (e.g. "why is [PERSON2] doing this"), whereas our dataset considers questions and knowledge about a variety of objects. Additionally, the Ads Dataset [18] is a dataset requiring knowledge about the topic and sentiments of the ads. Other datasets have considered knowledge-based question answering for a sitcom [10] and by using web queries [6].

Explanation / Reasoning VQA. Visual reasoning on its own has been studied in several VQA datasets. In CLEVR [21], the image and question are automatically generated from templates and explicitly require models to go through multiple steps of reasoning to correctly answer. This dataset and similar datasets which rely on simulated images suffer from lack of visual realism and lack of richness in the images and questions and are thus prone to be overfit to with methods achieving nearly 100% accuracy [51]. Our dataset requires reasoning on real images and free-form language. Other works [24, 34] have collected or extracted justifications on the VQAv2 [12] dataset. However, VQAv2 mostly focuses on questions about object attributes, counting and activities, which do not require reasoning on outside knowledge.

3 A-OKVQA Collection

Image source. The most important attribute of an image source for this knowledge-based VQA task is that it contains an abundance of visually rich and interesting images. Images containing a small number of objects are typically quite challenging to write interesting questions requiring outside knowledge to answer. We used images from the 2017 partitioning of the COCO dataset [25] in the creation of A-OKVQA, because it (1) has many images cluttered with multiple objects and entity types, and (2) is an established dataset with many associated models already in existence. To ensure suitable images for annotation, we do some additional filtering to remove uninteresting images: for the training and validation sets, we define images with more than three objects as "interesting" and select those for question writing. For the test set, which lacks object annotations, we train a ResNet-50 classifier to distinguish such "interesting" images, achieving an accuracy of **78%** on our validation set. After multiple rounds of filtering (described below), we obtain 23.7K unique images.

Question collection & filtering. The questions in A-OKVQA were written and refined over several rounds of annotation by 437 crowd-workers on the Amazon Mechanical Turk platform and refined through several manual and automated filtering steps to increase overall quality. As a first quality assurance measure, workers completed a qualification task to demonstrate their ability to write questions that met our criteria, namely that questions: (1) require looking at the image to answer, (2) need some commonsense or specialized knowledge, (3) involve some thinking beyond merely recognizing an object, and (4) not be too similar to previous questions.

To help ensure the last point, we clustered images by CLIP [37] visual features and batched similar images together such that the same worker wrote questions sequentially for related images (e.g., a worker might write questions for several images showing baseball games in one task) to cut down on repetitive questions. As an added measure to encourage question diversity, we maintained a database of questions written and required users to check a new question against these by displaying the five previous questions most similar in terms of their RoBERTa [27] embeddings. We used a simple VQA model (Pythia [44] pre-trained on VQAv2) to automatically find and remove questions we considered trivial (which the model answered correctly). Questions were then screened by three other workers and only included if the majority agreed that it met our criteria for inclusion. In all, 37,687 questions, or **60%** of post-qualification questions were excluded from the dataset by this process.

Answers. We asked workers to provide the correct answer along with three distractors for the questions they wrote. After all questions and multiple-choice options were gathered, we collected nine additional free-form answers per question from a separate pool of workers.

Rationales. After questions and answers were collected and validated, we performed a separate task to collect three rationales per question. Workers were given a question and its multiple choice options and asked to explain why a particular answer was correct (in one to two simple sentences, including any necessary facts or prior world knowledge not shown by the images).

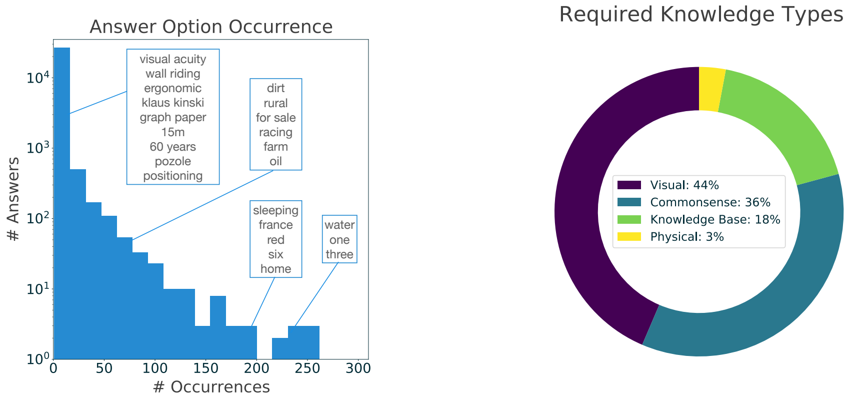
4 Dataset Statistics

The A-OKVQA dataset contains **24,903** (Question + Answer + Rationale) triplets in 17.1K (train) / 1.1K (val) / 6.7K (test) splits. These preserve the train/val/test splits in COCO 2017. The average lengths of and numbers of unique words in the questions, answers, and rationales are shown in Table 1.

In Figure 2a we show the distribution of answer options in our dataset. What we see is a fairly typical long-tail distribution of labels, as is seen in many open-labeled image tasks [55]. A few answers occur quite often in the dataset, but most fall into the long tail of the distribution.

We are also interested in the amount of answer set overlap between the training set and the validation / testing sets. We find that 87.6% of val set and 82.7% of test set questions have correct answers that appear as options in the train set. While this demonstrates a reasonable similarity between our splits, there remains a significant portions the test set that requires an answer not seen during training. Thus, models must be able to generate answers that are out-of-distribution or based on some knowledge outside of the dataset.

Comparison with other datasets. In Table 1, we show dataset properties and statistics for A-OKVQA compared to related datasets. We have 2-10x more questions than the more knowledge-focused natural image datasets, such as OK-VQA, while VCR (focused on images of people in movies) has 10x more than ours. This is unsurprising because we intentionally filter out similar questions,



(a) Answer occurrence distribution of A-OKVQA.

(b) Knowledge type distribution for a random subset of the dataset.

Fig. 2: **Dataset statistics.**

Table 1: **Comparison of knowledge-based VQA datasets.** Data based on reported numbers and available annotations. Thus, some statistics may exclude test sets. Answer statistics for A-OKVQA is based on the direct answer set. Rationales are not available (or not in sentence form) in all datasets. Q: question, I: image, A: answer, R: rationale, DA: Direct Answer, MC: Multiple Choice.

	Q	I	Rationale	Knowledge type	Ans type	Avg. length (Q/A/R)	unique words (Q/A/R)
KB-VQA [47]	2,402	700	✘	fixed KB	DA	6.8/2.0/—	530/1,296/—
FVQA [48]	5,826	2,190	✓	fixed KB	DA	9.5/1.2/—	3,010/1,287/—
OK-VQA [32]	14,055	14,031	✘	factoid	DA	8.1/1.3/—	5,703/11,125/—
S3VQA [19]	7,515	7,515	✘	factoid	DA	12.7/2.8/—	7,515/8,301/—
VCR [53]	290k	99,904	✓	people actions	MC	8.7/7.7/16.8	11,254/18,861/28,751
A-OKVQA	24,903	23,692	✓	common/world	DA/MC	8.8/1.3/11.0	7,248/17,683/20,629

making our questions more diverse (see Table 2). However, these are also difficult to collect at scale. Unlike these other datasets, ours has both multiple choice and direct answer annotations. Our dataset also has rationales, unlike OK-VQA, S3VQA and KB-VQA. Rationales in FVQA are in the form of knowledge tuples, rather than full sentences. VCR has the most similar rationales to our own. Since our rationales are more knowledge-based and have more possible variations per question, we collect three, unlike both FVQA and VCR which collect just one. Our questions are longer on average than in all datasets besides S3VQA and FVQA. Ours also contain the most unique words besides S3VQA (which has a similar number) and VCR (which has many more questions).

Knowledge types. The most significant factor differentiating our dataset is the kind of knowledge required. Datasets such as FVQA have fixed knowledge bases that are used to write the questions, and so the knowledge required can be found in e.g. ConceptNet [26] directly. OK-VQA and S3VQA focus on more factoid knowledge (e.g., years of invention or countries of origin). Researchers have found that these datasets take the form of finding an entity in the image and/or question and searching and retrieving knowledge about that particular entity [19]. VCR requires images to have people in them and overwhelmingly depicts people interacting in television shows and movies. Thus, the required knowledge is very focused on commonsense about human behavior and intentions. In our dataset, we require broader areas of knowledge, including the factoid knowledge likely to be contained in knowledge bases (as in FVQA, KB-VQA, OKVQA and S3VQA) and commonsense knowledge (similar to VCR, but broader in scope).

To analyze the knowledge required in A-OKVQA more quantitatively, we annotated a randomly sampled subset of 1,000 questions in the test set. In this experiment, we asked the annotators to label the knowledge type required to answer each question: (1) **Commonsense** knowledge about human social behavior (e.g. that many donuts being made in a cart implies they are for sale rather than for personal consumption), (2) **Visual knowledge** (e.g. muted color pallets are associated with the 1950s), (3) **Knowledge bases** (e.g. hot dogs were invented in Austria), and (4) **Physical knowledge** about the world that humans learn from their everyday experiences (e.g., shaded areas have a lower temperature than other areas). The distribution is shown in Figure 2b. Most of our questions cluster around commonsense and visual knowledge. It should be noted that sometimes there is no clear distinction between these two categories, and a question can belong to either category.

Question diversity. To compare the diversity of A-OKVQA to other datasets, we use the average pairwise cosine distance between questions for every dataset. We embed our questions with a sentence transformer[3]. We see from Table 2 that our dataset has the most diversity on this metric. In particular, we see a large difference compared to VCR which has many similar questions such as "What is going to happen next?" and questions relating to what specific people in the scene are doing and why. We also compare the diversity of rationales to VCR and VQAv2 (using rationales from VQA-X [34] rationales), and we find that our rationales are much more diverse than in these datasets. Qualitatively, we also find that our dataset tends to have much more varied questions because it is taken from the more visually diverse COCO dataset (a quality shared by OK-VQA and VQAv2 which do almost as well on this metric) and requires more diverse kinds of knowledge.

Finally, we use the same mean pairwise distance to look in particular at how different our questions are from OK-VQA which is the most similar prior work to ours. To do this we compare the minimum pairwise distance between every question in the OK-VQA training set to every question in the OK-VQA test set and our test set. We find that the average minimum distance from OK-VQA train to test is **0.256** compared to **0.311** between OK-VQA train and our test

[3] Specifically multi-qa-MiniLM-L6-cos-v1 [15] to avoid overlap with RoBERTa.

Table 2: **Question and Rationale Diversity.** Mean pairwise cosine distances in a sentence transformer space. ✗ indicates lack of rationale. Rationales for VQAv2 come from the VQA-X. dataset [34].

Dataset	Mean Q distance	Mean rationale distance
FVQA [48]	0.6199	✗
VCR [53]	0.7095	0.8017
KB-VQA [47]	0.7192	✗
S3VQA [19]	0.8050	✗
VQAv2 [12]	0.8405	0.8228
OK-VQA [32]	0.8428	✗
A-OKVQA	0.8564	0.8779

set[4]. This shows that there is in fact a significant difference between our question set and OK-VQA in this feature space.

5 Experiments

Next, we benchmark the A-OKVQA dataset and compare the performance of different models. We consider three classes of methods: (1) **large-scale pre-trained models** such as CLIP [37] and GPT-3 [4], (2) **models that generate and use rationales**, and (3) **specialized models** that are designed for knowledge-based VQA (KRISP [31]) or tested for VQA (e.g., VilBERT [28]).

5.1 Evaluation

In the *multiple choice (MC)* setting, a model chooses its answer from one of four options and we compute accuracy as the evaluation metric. In the *direct answer (DA)* setting, a model can generate any text as its answer and we use the standard VQA evaluation from [2].

5.2 Large-scale Pre-trained Models

We compare three types of large-scale pre-trained models (discriminative, contrastive, and generative) in Table 3. We also test these models in different input settings (where questions, images, or both are provided).

We compute BERT [8,16] and CLIP ViT-B/32 text encoder representations for questions. We also compute ResNet-50 [13] and CLIP ViT-B/32 features for images. These are provided as inputs to the appropriate discriminative and contrastive models. We provide questions as tokens and CLIP RN50x4 image

[4] To make this comparison even, we chose a random subset of our test set to be the same size as OK-VQA test set so that the minimum is over the same number of possible choices in both cases.

representations as inputs to the generative models. We generate a vocabulary from a subset of training set answers and choices to use across all appropriate models. We describe this vocabulary further in Appx. B.

Discriminative models. We train a multi-label linear classifier (i.e. MLP with one hidden layer and sigmoid activation function) on top of BERT (row *d*), ResNet (row *i*), and CLIP (rows *e/j/m*) representations to score answers from the vocabulary. When questions and images are both provided, we first concatenate their representations. For the DA setting, we predict the top scoring vocabulary answer. For the MC setting, we instead predict the nearest neighbor[5] choice to the top scoring vocabulary answer.

Contrastive models. We also evaluate models which match input questions and/or images with answers using their CLIP encodings. First, we evaluate the zero-shot setting (rows *f/k/n*). If both questions and images are provided as inputs, we first add their representations. We select the answer whose encoding has the greatest cosine similarity to our input representation. We select from vocabulary answers in DA and the given choices in MC.

We also train a single-layer MLP on top of our input representations (rows *g/l/o*). If both questions and images are provided, we first concatenate their representations. Our MLP produces a 512-d embedding and we train this with a CLIP-style contrastive loss between embeddings and their corresponding answers. We describe this loss further in Appx. B. We repeat the evaluation from the zero-shot setting, using these learned embeddings.

Generative models. We also evaluate models (GPT-3 [5] and ClipCap [33]) that generate answers directly as text. For both models, we predict the generated text for DA and the generated text's nearest neighbor choice for MC.

We prompt GPT-3[6] (row *h*) with 10 random questions and answers from the training set, followed by a new question, and let GPT-3 generate an answer to that question, in a manner similar to [49]. We provide GPT-3 with the prompt template "Question: ... Answer: [...]", expecting it to complete the answer for each evaluation question.

ClipCap [33] (row *p*) is an image captioning method that passes CLIP image features through a trained network to GPT-2 (as input tokens). We adapt this model by adding question tokens (and answer choices if applicable) to the prompt of GPT-2, generate answers instead of captions, and fine-tune on our data. We provide additional details, diagrams, and variations in Appx. B.

Results. Table 3 shows the results of our evaluation of these models. Rows *a-c* show the biases in our dataset, but that the direct answer setting is appropriately challenging. Question-only baselines (rows *d-h*) show poor performance in both MC and DA settings. However, it is interesting that GPT-3 performs similarly to the fine-tuned CLIP models (whichever is better per setting). The zero-shot CLIP model (row *f*) is least effective, indicating that training is necessary to repurpose CLIP text encodings for language-only tasks. Unsurprisingly, CLIP image features are very strong for zero-shot multiple choice matching (row

[5] Cosine similarity between mean GloVe [17,35] word embeddings.

[6] We use the second largest available GPT-3 model, Curie, as in [49].

Table 3: **Large-scale pre-trained models.** We also compare with no input heuristics (rows *a-c*). *Random* is a uniform sampling from choices (for MC) or answers in the training set (for DA). *Random (weighted)* uses weighted sampling proportional to correct answer frequencies. *Most Common* selects the most frequent answer in train.

	Multiple Choice		Direct Answer	
Method	**Val**	**Test**	**Val**	**Test**
(a) Random	26.70	25.36	0.03	0.06
(b) Random (weighted)	29.49	**30.87**	0.15	0.10
(c) Most Common	**30.70**	30.33	**1.75**	**1.26**
Question				
(d) BERT [8] (classifier)	32.93	33.54	9.52	8.41
(e) CLIP [37] (classifier)	32.74	33.54	**13.10**	10.24
(f) CLIP [37] (zero-shot)	30.42	30.58	0.44	0.57
(g) CLIP [37] (contrastive)	**37.40**	**38.58**	5.56	3.83
(h) GPT-3 [4]	35.07	35.21	12.98	**11.49**
Image				
(i) ResNet [13] (classifier)	28.19	28.81	2.68	2.30
(j) CLIP [37] (classifier)	33.21	32.56	**5.15**	**4.38**
(k) CLIP [37] (zero-shot)	**56.28**	**53.94**	2.24	2.29
(l) CLIP [37] (contrastive)	52.56	50.09	2.33	2.45
Question & Image				
(m) CLIP (classifier)	40.84	38.30	18.95	14.27
(n) CLIP (zero-shot)	48.19	45.72	1.08	0.71
(o) CLIP (contrastive)	53.77	51.01	10.36	7.10
(p) ClipCap [33]	**56.93**	**51.43**	**30.89**	**25.90**

k). However, they are not as strong as for the fine-tuned classifier (row *j*) in DA. ClipCap (row *p*) outperforms all other baselines in DA, because we use powerful image features and also fine-tune a strong language model for our task.

5.3 Rationale Generation

We are interested in whether we can improve GPT-3 prompting results by providing additional image- and question- specific context and report results for the following methods in Table 4. So, we fine-tune ClipCap (given images and questions, but not choices) as above, but for the task of generating rationales instead of answers. Our model scores **10.2** (val) / **9.58** (test) on SacreBLEU [36] and **0.271** (val) / **0.256** (test) on METEOR [3]. We can then prompt GPT-3 (as above) but also provide these generated rationales as "Context: ...". This model is denoted by 'ClipCap → Ratl. → GPT'. We provide additional details, diagrams, and examples of generated rationales in Appx. C. We repeat this experiment using captions (generated from only images) from the original ClipCap model: 'ClipCap → Cap. → GPT'.

Results. We show results from these experiments in Table 4. Interestingly, prompting GPT-3 with ground-truth rationales (row *d*) is competitive with the best model in Sec. 5.2 (Table 3, row *p*) in MC and significantly outperforms the question-only GPT-3 method (Table 3, row *h*). When we prompt GPT-3 with

Table 4: **Models using generated and GT rationales** as described in Sec. 5.3. We are unable to evaluate the GT Caption → GPT setting on the test set, as captions are not available in the COCO [7] test set.

	Multiple Choice		Direct Answer	
Method	**Val**	**Test**	**Val**	**Test**
(a) ClipCap → Cap. → GPT	42.51	43.61	16.59	15.79
(b) ClipCap → Ratl. → GPT	**44.00**	**43.84**	**18.11**	**15.81**
Oracles				
(c) GT Caption → GPT	45.40	—	16.39	—
(d) GT Rationale → GPT	**56.74**	56.75	**24.02**	20.75

Table 5: **Specialized models results.** Baselines trained for VQA or knowledge-based VQA, and fine-tuned on A-OKVQA. The bottom two rows are not comparable with the others since they use ground-truth rationales at test time.

	Multiple-Choice		Direct Answer	
Method	**Val**	**Test**	**Val**	**Test**
(a) Pythia [20]	49.0	40.1	25.2	21.9
(b) ViLBERT [28] - OK-VQA	32.8	34.1	9.1	9.2
(c) ViLBERT [28] - VQA	47.7	42.1	17.7	12.0
(d) ViLBERT [28]	49.1	41.5	30.6	25.9
(e) LXMERT [45]	51.4	41.6	30.7	25.9
(f) KRISP [31]	51.9	42.2	33.7	27.1
(g) GPV-2 [22]	**60.3**	**53.7**	**48.6**	**40.7**
Oracles				
(h) GPV-2 [22] + Masked Ans.	65.1	58.3	52.7	43.9
(i) GPV-2 [22] + GT Ratl.	73.4	67.2	58.9	51.7

ground-truth rationales (row d), we see higher performance than when we provide ground-truth captions (row c). This affirms that rationales contain useful information (i.e. specific to our questions and answers) in addition to captions. However, the additional performance of prompting GPT-3 using generated rationales (row b) over generated captions (row a) is not as significant. This indicates potential room for improvement in our approach for generating rationales.

5.4 Specialized Models

In this section, we evaluate some recent high-performing, open-source models trained on knowledge-based VQA or the traditional VQA. The models we consider are Pythia [20], VilBERT [28], LXMERT [45], KRISP [31], and GPV-2 [22]. As the first four models are part of MMF [43], it is easier to compare them fairly. KRISP is a high-performing model on OK-VQA [32]. It provides a suitable baseline as it was designed to perform well on knowledge-based VQA. GPV-2 performs multiple vision and vision–language tasks and has learned a large number of concepts, so it can be a strong baseline for A-OKVQA. All of these models are fine-tuned on A-OKVQA. We adapt them to MC using the nearest choice method described above. See Appx. D for the details of each model.

Results. Unsurprisingly, these models, which are specialized for DA and some of which are specialized for knowledge-based VQA perform very well on the DA evaluation and quite well on MC. Of the models trained only on A-OKVQA KRISP does the best, likely because it is trained to directly use outside knowledge graphs. GPV-2, however, performs best of all, beating all other models (that do not use ground-truth rationales) in all settings, possibly because of the large number of concepts it has learned.

Transfer results. We train ViLBERT on VQAv2 and OK-VQA datasets (denoted by 'ViLBERT-VQA' and 'ViLBERT-OK-VQA' in Table 5) to evaluate whether the knowledge from those datasets is sufficient for A-OKVQA. The low performance shows that significant differences exist between these datasets and A-OKVQA.

Ground-truth Rationales. To evaluate how well the model performs if it is provided with high-quality rationales, we use ground-truth rationales at test. We show these results with GPV-2 (our best model). Ground-truth rationales are appended to questions as additional input text ('GPV-2 + GT Ratl.'). For this experiment, we used only one of the rationales. Comparing rows *g* and *i* of Table 5 shows rationales are helpful. To evaluate how much of this improvement can be attributed to rationales and not the fact that sometimes rationales contain the answer, we replaced answers in the rationales with [answer] token. The performance drops (row *i* vs row *h*), however, it is still higher than the case that we do not use rationales (row *h* vs row *g*).

6 Analysis of Models

Next, we analyze the predictions that our baseline models make to see if we can learn more about A-OKVQA: what kinds of questions do different types of approaches do better / worse on? For these experiments, we choose some of the best performing models on Direct Answer: VilBERT [28], LXMERT [45], KRISP [31], ClipCap [33] and GPV-2 [22]. We also use the ClipCap → Rationale → GPT model from Table 4, which will be referred to as 'GR-GPT' for Generated Rationales GPT.

Answer Frequency. First, we look at how answer frequency affects performance in Table 6. We first count the number of times any answer appears in the direct answers in the training set. We then divide these into bins and look at the direct DA test accuracy of our baselines for each of these frequency bins. We find that GPV-2, and to a lesser extent ClipCap and GR-GPT perform better on questions whose answers do not appear often in the training set (1-5 and 6-10 columns of Table 6). GPV-2 in particular (which is fine-tuned on several vision and language tasks) is able to predict these tail answers much better than other methods, especially the discriminative methods such as LXMERT.

Table 6: **Results across different answer frequencies.** The questions are categorized based on the frequency of the GT answer in the training set. Columns show accuracy for answers that appear 1-5 times, 6-10 times, etc. If multiple direct choices, we default to most common one.

Model	1-5	6-10	11-20	21-50	51-100	101-200	201+
VilBERT [28]	0.00	0.00	3.68	10.97	19.95	26.53	35.91
LXMERT [45]	0.00	0.00	4.29	13.73	20.18	26.69	34.31
KRISP [31]	0.00	0.61	6.34	13.99	21.78	28.55	35.22
ClipCap [33]	4.71	4.24	9.10	17.90	25.93	29.44	33.99
GR-GPT	8.18	9.29	9.41	17.39	18.31	21.98	24.65
GPV-2 [22]	**10.16**	**12.12**	**22.60**	**31.04**	**38.40**	**41.60**	**44.69**

Prediction overlap/difference. Finally, we look at some statistics on a question by question level in the A-OKVQA test set. Specifically we look at the overlap in which methods answered which questions correctly[7].

First, we find that only **5.85%** of questions in test were answered correctly by all models and **30.96%** of questions had no model predict a correct answer for. Considering the worst performing model of these gets **15.81%** DA accuracy and the best gets **40.7%**, it implies that there is actually a large variation between these models beyond some just being generally better than others and thus getting "hard" questions right and keeping performance on "easy" questions.

In Table 7, we show the difference between the questions each model gets right on A-OKVQA test. Each row shows the percentage of that method's correctly answered questions that were not correctly answered by the comparison model in each column. If we look at the row for the lowest performing model (GR-GPT) for the column for the best performing model (GPV-2), we still see that **29.2%** of GR-GPT's correctly answered questions are answered wrongly by GPV-2!

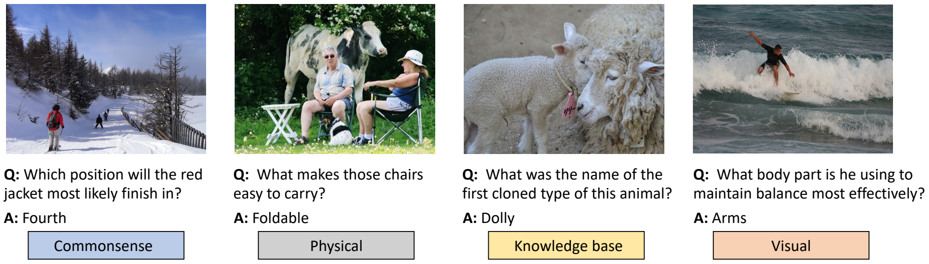
Finally, to further illustrate the point that different models have very different mistake patterns, we take the prediction of all of these models except for GPV-2 for each question and take the majority vote between these. This majority vote combination gets an accuracy of **29.5** compared to the best of these models which gets **27.1**. This does not work when GPV-2 is added (this majority model gets **35.60** which is lower than GPV-2's **40.7**). We can also look at the Oracle combination accuracy. That is, from our six models, choose the answer with the highest ground-truth value and take that as the oracle combination answer. This DA accuracy is **56.87** versus the single best performance of **40.7**, again showing that even worse performing models get many questions right that the best model gets wrong.

Qualitative Analysis. We extracted questions that all of the discussed models fail at. Figure 3 shows an example from each knowledge type. This shows what type of reasoning is missing in top performing models.

[7] For ease of analysis we count a binary yes/no of whether a model answered correctly if it answered any possible answer in the direct answer set.

Table 7: **Pairwise difference between correctly answered questions.** For row i and column j of this table the value is percentage of questions answered correctly by model i that j did not answer correctly.

Model	VilBERT	LXMERT	KRISP	ClipCap	GR-GPT	GPV-2
VilBERT [28]	0.00	29.00	27.19	43.72	59.72	26.33
LXMERT [45]	28.07	0.00	26.57	44.39	59.73	27.44
KRISP [31]	30.44	30.76	0.00	44.18	60.29	27.43
ClipCap [33]	48.72	49.98	46.76	0.00	55.94	26.64
GR-GPT	50.27	50.91	48.67	40.30	0.00	29.20
GPV-2 [22]	51.09	52.46	49.57	46.56	61.94	0.00


Q: Which position will the red jacket most likely finish in?
A: Fourth

Q: What makes those chairs easy to carry?
A: Foldable

Q: What was the name of the first cloned type of this animal?
A: Dolly

Q: What body part is he using to maintain balance most effectively?
A: Arms

Fig. 3: **Example questions that all discussed models fail at.**

Collectively, these analyses reveal several interesting findings. First, aside from being generally difficult, the A-OKVQA dataset shows a surprising lack of overlap in the specific questions different models answer correctly. Second, we see that different methods handle rare answers very differently. Thirdly, different methods perform differently based on the type of knowledge required to answer questions. Together, these features suggests that A-OKVQA contains a wide variety of challenging questions which are able to reveal and contrast the strengths and weaknesses of VQA methods.

7 Conclusion

Vision–language models have become progressively more powerful, however, evaluation of the reasoning capabilities of these models have not received adequate attention. To take a step in this direction, we propose a new knowledge-based VQA benchmark called A-OKVQA, which primarily includes questions that require reasoning using commonsense and world knowledge. We provide *rationales* for each question so models can learn the line of reasoning that leads to the answer. We evaluate a large set of recent, high performance baselines. While they show impressive performance on the proposed task, it is evident that they lack the reasoning capability and/or the knowledge required to answer the questions, and there is a large room for improvement.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 23
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: ICCV (2015) 1, 3, 8
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (2005) 10
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS (2020) 8, 10
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020) 1, 9
6. Chang, Y., Narang, M.B., Suzuki, H., Cao, G., Gao, J., Bisk, Y.: WebQA: Multihop and multimodal qa. arXiv (2021) 4
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO Captions: Data collection and evaluation server. arXiv (2015) 11, 23
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019) 1, 8, 10
9. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. In: NeurIPS (2015) 3
10. García, N., Otani, M., Chu, C., Nakashima, Y.: KnowIT VQA: Answering knowledge-based questions about videos. In: AAAI (2020) 4
11. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences (2015) 1
12. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017) 1, 4, 8, 23
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 8, 10
14. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019) 23
15. HuggingFace: `https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1` 7
16. HuggingFace: `https://huggingface.co/sentence-transformers/nli-bert-base` 8
17. HuggingFace: `https://huggingface.co/sentence-transformers/average_word_embeddings_glove.6B.300d` 9
18. Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A.: Automatic understanding of image and video advertisements. In: CVPR (2017) 4

19. Jain, A., Kothyari, M., Kumar, V., Jyothi, P., Ramakrishnan, G., Chakrabarti, S.: Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In: SIGIR (2021) 3, 6, 7, 8
20. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0.1: the winning entry to the VQA challenge 2018. arXiv (2018) 11, 23
21. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017) 1, 4
22. Kamath, A., Clark, C., Gupta, T., Kolve, E., Hoiem, D., Kembhavi, A.: Webly supervised concept expansion for general purpose vision models. arXiv (2022) 11, 12, 13, 14, 24
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 3, 23
24. Li, Q., Fu, J., Yu, D., Mei, T., Luo, J.: Tell-and-answer: Towards explainable visual question answering using attributes and captions. In: EMNLP (2018) 4
25. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 4
26. Liu, H., Singh, P.: ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal (2004) 7, 24
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv (2019) 5
28. Lu, J., Batra, D., Parikh, D., Lee, S.: VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019) 1, 8, 11, 12, 13, 14, 23, 24
29. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: CVPR (2020) 1
30. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NeurIPS (2014) 1, 3
31. Marino, K., Chen, X., Parikh, D., Gupta, A.K., Rohrbach, M.: KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In: CVPR (2021) 8, 11, 12, 13, 14, 24
32. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: CVPR (2019) 1, 2, 3, 6, 8, 11
33. Mokady, R., Hertz, A., Bermano, A.H.: ClipCap: CLIP prefix for image captioning. arXiv (2021) 9, 10, 12, 13, 14, 21, 24
34. Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: CVPR (2018) 4, 7, 8
35. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: EMNLP (2014) 9
36. Post, M.: A call for clarity in reporting BLEU scores. In: Conference on Machine Translation (2018) 10
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 5, 8, 10

38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog (2019) 1
39. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020) 24
40. Ren, M., Kiros, J., Zemel, R.S.: Exploring models and data for image question answering. In: NeurIPS (2015) 3
41. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: KVQA: Knowledge-aware visual question answering. In: AAAI (2019) 3
42. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 23, 24
43. Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D.: MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf (2020) 11
44. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: CVPR (2019) 5
45. Tan, H.H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: EMNLP (2019) 11, 12, 13, 14, 23, 24
46. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: understanding stories in movies through question-answering. In: CVPR (2016) 3
47. Wang, P., Wu, Q., Shen, C., Dick, A.R., van den Hengel, A.: Explicit knowledge-based reasoning for visual question answering. In: IJCAI (2017) 1, 3, 6, 8
48. Wang, P., Wu, Q., Shen, C., van den Hengel, A., Dick, A.R.: FVQA: fact-based visual question answering. TPAMI (2017) 1, 3, 6, 8
49. West, P., Bhagavatula, C., Hessel, J., Hwang, J.D., Jiang, L., Bras, R.L., Lu, X., Welleck, S., Choi, Y.: Symbolic knowledge distillation: from general language models to commonsense models. arXiv preprint arXiv:2110.07178 (2021) 9
50. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of GPT-3 for few-shot knowledge-based VQA. arXiv (2021) 1
51. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In: NeurIPS (2018) 4
52. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: ICCV (2015) 3
53. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019) 1, 4, 6, 8
54. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting visual representations in vision-language models. In: CVPR (2021) 1, 2, 24
55. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: CVPR (2014) 5
56. Zhu, Y., Groth, O., Bernstein, M.S., Fei-Fei, L.: Visual7W: grounded question answering in images. In: CVPR (2016) 3, 23