

CommVQA: Situating Visual Question Answering in Communicative Contexts

Nandita Shankar Naik
Stanford University
nanditan@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Elisa Kreiss
University of California, Los Angeles
ekreiss@ucla.edu

Abstract

Current visual question answering (VQA) models tend to be trained and evaluated on image-question pairs in isolation. However, the questions people ask are dependent on their informational needs and prior knowledge about the image content. To evaluate how situating images within naturalistic contexts shapes visual questions, we introduce CommVQA, a VQA dataset consisting of images, image descriptions, real-world communicative scenarios where the image might appear (e.g., a travel website), and *follow-up* questions and answers conditioned on the scenario and description. CommVQA, which contains 1000 images and 8,949 question–answer pairs, poses a challenge for current models. Error analyses and a human-subjects study suggest that generated answers still contain high rates of hallucinations, fail to fittingly address unanswerable questions, and don’t suitably reflect contextual information. Overall, we show that access to contextual information is essential for solving CommVQA, leading to the highest performing VQA model and highlighting the relevance of situating systems within communicative scenarios.

1 Introduction

Visual question answering (VQA), the task of providing an answer to a question given an image, measures a model’s ability to synthesize visual and textual modalities, and has many promising real-world applications. For example, when images online can’t be seen and accessed, it severely affects people’s abilities to educate themselves, socially engage, and stay informed (Morris et al., 2016; MacLeod et al., 2017; Voykinska et al., 2016; Gleason et al., 2019), and VQA models are an opportunity for providing interactive accessibility to such visual content at scale (Gurari et al., 2018; Baker et al., 2021). While most prior VQA datasets focus on investigating image-text alignment as a decontextualized task (Antol et al., 2015; Goyal et al.,

2017; Hudson and Manning, 2019; Marino et al., 2019), we aim to reframe it as a human-centric communicative problem, moving it closer to a real-world interactive setting.

According to a pragmatic Bayesian view of communicative actions, people tend to ask questions that maximize the contextually relevant information gain based on their existing prior beliefs about the world (Frank and Goodman, 2012), following fundamental pragmatic principles (Grice, 1975). Based on these linguistic insights, we argue that prior VQA datasets largely do not consider two central communicative drives that limit their utility. The information people aim to obtain (and consequently the types of questions they ask) varies with (1) the person’s *information needs* based on their goals when encountering the image, and (2) the person’s *prior knowledge* of the image content. Thus we introduce CommVQA, a benchmark that treats VQA as an inherently communicative task.¹

To investigate people’s *information needs*, CommVQA consists not only of images, questions, and answers, but also of image descriptions and plausible communicative scenarios for each image. Prior image accessibility research with blind and low-vision (BLV) participants shows that the information people want from an image is dependent upon the scenario in which the image appears (Stangl et al., 2021, 2020; Kreiss et al., 2022a; Muehlbradt and Kane, 2022). For example, if a person encounters an image when they are shopping online, they are likely to ask questions about the brands within the image, while if they’re browsing the news, the perceived purpose of the image shifts, and they are likely to ask questions about the event occurring within the image (Stangl et al., 2021, 2020). We therefore define a scenario as the type of website (e.g., a shopping website) combined with a goal for viewing it (e.g., to buy a gift) and expect it to shape

¹All data and code are available at: <https://github.com/nnaik39/commvqa>.

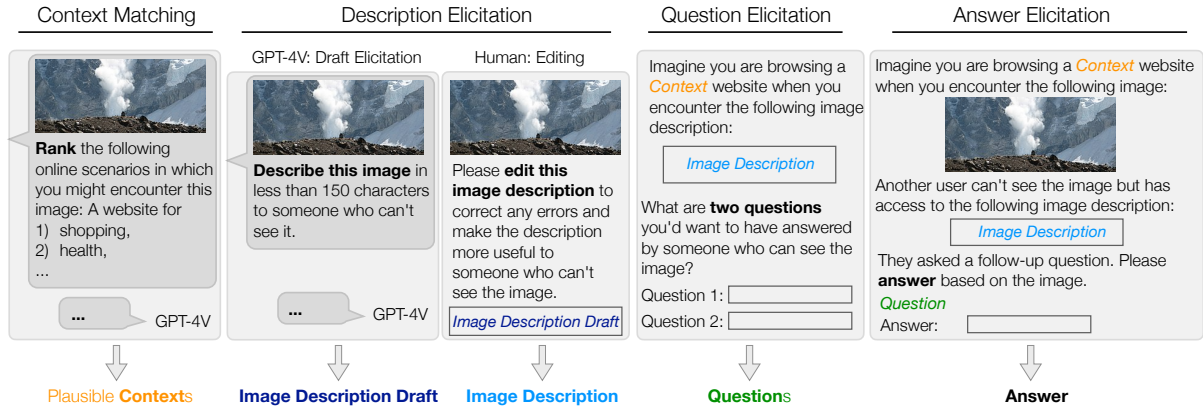


Figure 1: **Overview of the CommVQA Dataset Construction.** Images were sourced from Wikipedia and paired with relevant scenarios. The description were first generated by GPT-4V, then edited by humans. Other participants then provided questions and answers based on the scenario and description, resulting in at least three answers for each of the 2,983 unique visual questions. Simplified instructions are shown here; full details are in Appendix A.

people’s information needs in a VQA task.

In addition, the relevant questions a person might ask are predicted to be guided by their *prior knowledge* about the image. For most images we encounter online, we can rely on rich cues that allow sophisticated inferences about what an image contains. An image on a shopping website, for instance, would likely be accompanied by an article label, such as “Colorful Summer Skirt”, or it could have an informative alt text description. In the standard VQA task, annotators are asked to write questions in isolation to “fool a smart robot”, an adversarial task where the goal of the questioner is to trick an AI model (Goyal et al., 2017). However, in the naturalistic setting, these visual questions are better conceptualized as *follow-up* questions, since they are conditioned on already available information. In CommVQA, we situate the task by providing people with quality-controlled image descriptions instead of the image itself when collecting visual questions. Together with the contextual grounding of the images, this pipeline, presented in Figure 1, generates a challenging dataset with context-sensitive, highly diverse questions, as well as longer answers compared to prior VQA datasets.

We benchmark four state-of-the-art VQA models on CommVQA, and show that the situated nature of CommVQA poses a significant challenge for current state-of-the-art vision-language models and allows for insights into model generation behavior due to a highly controlled data generation pipeline. We find that the most successful model needs to integrate contextual information, suggesting that

context shapes VQA in communicative settings.

In summary, our main contributions are: (a) we introduce CommVQA, a benchmark consisting of images, contexts, descriptions, questions, and answers, where models must infer the details most relevant to the questioner’s goals, (b) we benchmark this dataset on current VQA approaches and explore whether models can be instructed to integrate context, and (c) we show via error analyses and a human-subjects experiment that even the best models generate false information at high rates.

2 Related Work

With CommVQA, we aim to situate the abstract VQA task within communicative settings. This builds on prior VQA dataset work (Section 2.1), and communicative insights from linguistics and human-computer interaction (Section 2.2)

2.1 VQA Datasets

Most VQA datasets focus on image-question-answer triplets, which are constructed in isolation from the real-world contexts in which the images might appear (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019; Marino et al., 2019). The VizWiz dataset (Gurari et al., 2018) stands out among VQA datasets as it focuses on real-world accessibility. The images in VizWiz are elicited from BLV users, who uploaded a picture to a phone app with a question about their real-world environment. The goals behind the questions and images is specific to BLV users exploring their physical environment. CommVQA is complementary to

VizWiz, since it investigates how contextually situating images online affects a model’s VQA task performance.

Some VQA datasets have explored integrating supplementary information. Visual Dialog (Das et al., 2017) aims to create visual chatbots, which can answer a question based on an image and the prior dialogic context. Similarly to CommVQA, these questions can be conceptualized as follow-up questions based on an image description. However, the images, questions, and answers are still decontextualized from a specific information goal beyond the questioner wanting to understand the image. CommVQA instead extends the focus on varying the questioner’s broader goals.

ScienceQA (Lu et al., 2022) consists of multi-modal science understanding questions, where each multiple choice question and answer is associated with a lecture and explanation, functioning as the context. VQAOnline (Chen et al., 2023) consists of images, questions, and context sourced from StackExchange, where the context is the body of the post. The notion of context in these datasets is limited to texts that provide supplementary information, whereas we specifically focus on the effect of changing scenarios on the VQA task.

PromptCap (Hu et al., 2022) similarly considers the notion of “context” for models generating answers in the VQA task. First, the model generates a visual description relevant to the question and then generates an answer solely relying on the description, conceptualized as “context.” In contrast to PromptCap, the contextual condition in CommVQA is strictly complementary to the image (see Section 3.3), therefore fundamentally changing the task and modeling demands.

2.2 VQA as Communication

Within pragmatics, there is a general consensus that questions are grounded in contexts and sensitive to the goals of the interlocutors (Groenendijk and Stokhof, 1984; Ginzburg, 1996; Roberts, 1996; van Rooy, 2003). In line with this prediction, prior human-subject studies with BLV participants show that context influences the information that people want about an image (Stangl et al., 2021, 2020; Muehlbradt and Kane, 2022; Kreiss et al., 2022a). Stangl et al. (2021) find that on social media, they wanted to know more about the people and the activity of the person who posted the image, while if people were on a shopping website to purchase

a gift for a friend, they expressed a desire to learn more about the objects within the image. Inspired by Stangl et al. (2021), we utilize a similar type of scenario. With CommVQA, we constitute a large-scale dataset where context-sensitivity emerges with sighted participants in question and answer behavior and can be studied at scale.

3 The CommVQA Dataset

The CommVQA dataset was constructed in five main steps, as visualized in Figure 1. First, we sourced images from Wikipedia, and elicited both plausible scenarios and descriptions for the images from GPT-4V (OpenAI et al., 2024). To ensure description quality, we conducted a human-subject study to edit the descriptions. We then elicited questions and answers for the dataset from US-based crowdworkers on Prolific. All human-subject studies were conducted with IRB approval.

3.1 Image-Scenario Matching

We extracted 1000 images from Wikipedia pages of topics related to at least one of the scenario conditions. We placed each image in two scenarios, which allows us to investigate how the scenario can induce variations in questions and answers across the same visual stimuli. Each image-scenario pair is annotated with on average 1.54 questions and each question with three answers, which results in 8,949 unique question-answer pairs. We prioritized high coverage of individual datapoints over breadth since the focus of this work is on uncovering the diverse contextual effects on the VQA task.

We chose six potential scenarios: Shopping, Travel, Science Magazines, News, Health, and Social Media. These scenarios were informed by prior work that showed people’s information needs varied across these scenarios (Stangl et al., 2021), and allowed for an overlap between the images that appear in each scenario—for instance, images in travel blogs (e.g., a picture of a waterfall) could plausibly appear in an online science magazine.

To assign plausible scenarios to the image, we instructed GPT-4V (OpenAI et al., 2024) to rank them in order of descending plausibility, guided by the task in Stangl et al. (2021). We validated the assignment on a subset of data with a human-subject study before using GPT-4V to scale (see Appendix A.2 for the prompt). From the top-three scenarios for each image, two scenarios were selected in order to balance out the co-occurrence of contexts.

3.2 Image Description Elicitation

We elicited descriptions for all images in our dataset. They form the basis for the follow-up questions participants asked, simulating the effect of someone encountering an alt text associated with the image online. Importantly, descriptions were generated out-of-context so we could analyze context effects on the questions and answers without the description being a confounding factor.

Automatic Description Draft Elicitation We elicited initial description drafts by prompting GPT-4V with the phrase: Describe this image in less than 150 characters to someone who cannot see it. The length constraint follows a commonly-issued guide on best-practices for accessibility alt text writing (OSU, 2024). We chose Wikipedia for sourcing images due to the copyright permissions and image variety. All selected images were in the public domain.

Description Editing To ensure description quality, we conducted a human-subject study where participants edited the descriptions generated by GPT-4V (OpenAI et al., 2024). These edits were intended to help balance for potential inaccuracies, possible misalignments with human description preferences, and the current design choices of GPT-4V. For instance, as of May 2024, the model refrains from explicitly identifying the people within an image (OpenAI, 2023), even though proper names can be an important detail for a useful image description (Bennett et al., 2021; MacLeod et al., 2017).

Each participant was shown six randomized trials with an image and description. They were instructed to edit the image description to correct any errors and make it more useful to someone who cannot see it. Importantly, participants were not shown the context that each image was placed in to control for question variation across contexts. Participants could also choose to skip if no edits were needed. We recruited 369 participants and compensated them at the rate of \$13.50/hr.

We collected three description edits for each image description draft, and selected a random edited description for the final description.

3.3 Question Elicitation

To elicit visual questions, we recruited 619 participants, who were paid \$13.50/hr with an average completion time of seven minutes. In each trial, participants were given two pieces of information: an image description (e.g., “A group of people of

various ages walking along a grassy path, with trees on one side.”), and a scenario for the image (e.g., Imagine you are browsing a Health website, with the goal of learning how to live a healthier lifestyle). Crucially, participants didn’t see the image to avoid priming for specific questions and simulate the visual inaccessibility of the image in the real-world scenario. They also rated how likely the image would appear within the provided scenario, and were prompted to ask two questions they would like answered by someone who can see the image.

In total, we elicited 2,983 questions, with 1.54 questions on average for each image-scenario pair. Based on a separate human-subject study (Section 5.2), we find that 80% of the questions require the image to answer, emphasizing the difficulty and inherent multimodal nature of CommVQA.

3.4 Answer Elicitation

We elicited answers from 870 participants, who were paid \$13.50/hr, with an average completion time of seven and a half minutes. Participants were shown the image, question, context, and description. They were told that another user asked the question based on this image description, and asked to write an answer that would help the other person visualize the image. Each question was answered by at least three separate participants.

We also included a checkbox for participants to indicate if a question was unanswerable. If a question was voted unanswerable by two or more annotators, we labeled the question as “unanswerable” within the dataset. In total, we collected 283 questions that were labeled unanswerable (9.5%). Within the naturalistic setting, where people will ask questions about an image they cannot see, it’s expected that people will ask unanswerable questions, and that models should have the capability to decline answering. The unanswerable portion of CommVQA can help assess whether a model can abstain from answering instead of providing incorrect information (MacLeod et al., 2017).

4 Dataset Analysis

The final CommVQA dataset consists of 8,949 unique question-answer pairs, spanning across 2000 unique image-scenario pairs. To better understand the challenges posed by this dataset, we now provide an analysis of CommVQA. Examples from the dataset, sampled to cover all scenarios, are given in Appendix Figure 5. We additionally

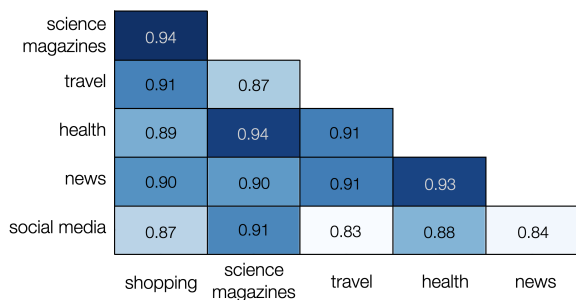


Figure 2: **Heatmap of BERT Classification Accuracy Across Scenario Pairs.** When fine-tuned on different scenario pairs, BERT exhibits varying performance in its ability to distinguish between these scenarios. For instance, BERT achieved 94% accuracy when distinguishing between science magazines and shopping, but only 83% accuracy for travel and social media.

conducted an analysis of a subset of the data and found similar patterns in model responses on this smaller dataset, suggesting our data is sufficiently large to obtain generalizable patterns. For more details, see Appendix C.

4.1 Analysis of Descriptions

In CommVQA, the descriptions are the basis for the VQA task and were collected in a two-stage process: first, automatic description generation by GPT-4V and then a human editing phase. The automatically generated descriptions had an average length of 63.663 characters and in 42% of trials, people didn’t make any edits to the descriptions. When editing, participants added extra information, increasing the average length to 97.29 characters.

4.2 Analysis of Questions

The main goal of CommVQA is to situate the VQA task in a communicative context, assuming that the context shapes what becomes relevant and therefore has implications for downstream model performance. In this section, we investigate the context-sensitivity of the questions in CommVQA.

If questions are context-sensitive, then a trained classifier should be able to predict the correct scenario from the question and achieve a performance reliably higher than random. To investigate whether this is true for CommVQA, we fine-tuned BERT (Devlin et al., 2019) on the task of predicting whether each question appears within a certain context or not. We expect this task to be difficult even for a human, given that questions such as “What time of day is it?” might appear in any scenario. We split all questions in the dataset into an 80-10-

10 train/test/val split, and fine-tuned for 200 epochs with LoRA (Hu et al., 2021). Fine-tuned BERT achieved an accuracy of 56% on this task, a significant improvement over random choice (16%), showing that questions elicited within different contexts are inherently distinct.

While we find overall evidence that questions vary between all scenarios, a classifier analysis further allows us to investigate which individual scenarios have the least and most overlap in the questions asked. Figure 2 shows the BERT classifier’s performance when fine-tuned on distinguishing pairs of scenarios (e.g., shopping from science magazines), and indicates which scenarios are more easily distinguishable. Intuitively, certain scenarios are more related than others—for instance, shopping and social media content is more likely to contain images of famous models, while shopping and science magazines are less likely to have overlap. Our analysis confirms this intuition and highlights where models may fail to translate across scenarios.

We further inspect the question interrogatives for a potential source of between-scenario variation. The results of a Welch’s t-test demonstrated that, for example, compared to all other scenarios, “Who” questions are significantly more likely to appear in the social media scenario ($t(589.6) = 3.85, p < 0.001$), and “Where” questions are significantly more likely to appear in the travel scenario ($t(627.3) = 3.58, p < 0.0001$) compared to any other scenario. These results indicate that different question types are more likely to be asked in certain scenarios, which carries implications that model evaluation should be contextual. A model that has poor performance on “Who” questions might seem competent in non-social media scenarios, but fail to generalize to social media due to the distinct nature of the user’s information needs. However, question type and scenario are still distinct conceptually.

Taken together, we find converging evidence that the questions in CommVQA fundamentally vary based on the scenario the images were presented in, highlighting the diverse requirements for building robust communicatively situated VQA models.

4.3 Analysis of Answers

We now turn to a general analysis of the *answers* in CommVQA and investigate the extent to which they are contextually situated.

Firstly, with an average length of 10.98 words, the answers in CommVQA are surprisingly long

compared to prior datasets that used a similar on-line answer elicitation setup (e.g., 1.1 words for VQA-v2 (Goyal et al., 2017) and 2.1 words for Visual Dialog (Das et al., 2017)). Contrasting prior work, the instructions of CommVQA are framed as a communicative task, where answerers were asked to help someone else who cannot see the image. It’s plausible that the longer responses are partially due to the participant’s wish to faithfully communicate with the questioner (Grice, 1975).

While we focus our analysis on the effects context directly has on the questions being asked (see Section 4.2), we also find that access to context becomes strictly necessary to answer those contextualized questions. Most strikingly, questions of the form “What else is in the image?” directly require the answerer to know what information users already have (see Figure 3 and Figure 6 for examples).

5 Model Benchmarking

In this section, we investigate the performance of four state-of-the-art vision-language models on CommVQA and to what extent providing context to the models improves their performance.

We selected four models to benchmark on this dataset: LLaVA (Liu et al., 2024), BLIP-2 (Li et al., 2023), mPLUG-OWL (Ye et al., 2023), and IDEFICS (Laurençon et al., 2023). To maximally enable the reproducibility of this work and support the public development of models, we focus on open-source models. LLaVA is trained on a multimodal instruction-following dataset, and exhibits strong performance across several VQA datasets (Goyal et al., 2017; Hudson and Manning, 2019). mPLUG-Owl (Ye et al., 2023) is also instruction-tuned, and displays competitive performance on various VQA benchmarks, including VQA-v2 (Goyal et al., 2017). IDEFICS (Laurençon et al., 2023) is an open-access reproduction of Flamingo (Alayrac et al., 2022), and was trained on a naturalistic web-scale dataset of interleaved image-text documents, including 141 million web pages. We selected LLaVA, mPLUG-Owl, and IDEFICS since their instruction-tuned nature allows us to straightforwardly integrate context. We also included a non-instruction-tuned model, BLIP-2, a general-purpose vision-language model that is additionally fine-tuned on the VQA task.

Evaluation Metrics We used three types of evaluation methods for complementary insights on the

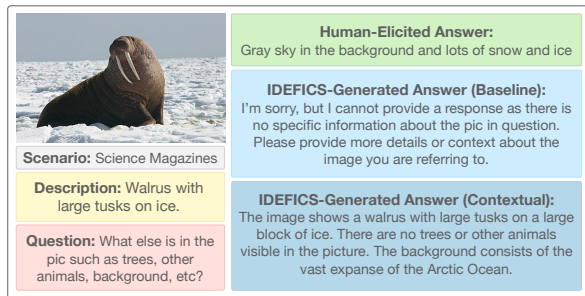


Figure 3: **Example of Context Dependency in Answer Generation.** In this example, the question explicitly asks for content that is not in the description. While the human-elicited answers do not repeat information in the description, IDEFICS (contextual) provides an answer, but does repeat content that is in the description.

quality of the model-generated answers. To obtain insights on the overall string similarity of the model-generated and the human-written answers, we use the well-established reference-based NLG metrics BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004). However, these metrics do not adequately capture object hallucinations (Rohrbach et al., 2018), which is why we supplement this analysis with a human-subject study specifically aimed at capturing hallucinated content. Finally, to measure how closely an answer’s visual details matches with the image and associated visual description, we also include an analysis using CLIPScore (Hessel et al., 2021).

Experimental Setup For all models, we conducted two main experiments. The *contextual* condition simulated the situation human participants were in, where the models had complete access to the image, description and situational scenario while answering the question.

We include our prompt for the contextual condition in Appendix A.2. In addition, we tested a *baseline* condition where the models only had access to the image and question, in order to assess the model capabilities for incorporating contextual information. All evaluations were conducted with the greedy decoding method to ensure reproducibility.

Overall Results In Table 1, we show the performance of all models when provided with the image and question (baseline) and the full context that was available to human answerers (contextual). The IDEFICS model with the contextual condition has the highest performance for all models across

Model	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
IDEFICS	0.273	0.084	0.179	0.378	0.758
IDEFICS (Contextual)	0.285	0.092	0.195	0.392	0.839
LLaVA	0.195	0.052	0.172	0.317	0.453
LLaVA (Contextual)	0.152	0.037	0.161	0.253	0.219
mPLUG-Owl	0.199	0.048	0.169	0.320	0.462
mPLUG-Owl (Contextual)	0.187	0.046	0.172	0.295	0.355
BLIP-2	0.267	0.059	0.098	0.282	0.434
BLIP-2 (Contextual)	0.015	0.001	0.028	0.043	0.014

Table 1: **Comparison of Baseline and Contextual Conditions Across Models.** This table presents results for both baseline and contextual conditions across all models. IDEFICS (contextual) achieved the highest scores across all metrics. Results are averaged over three random data splits and model-generated answers.

all metrics. Evidently, integrating contextual information made the content itself more aligned with the ground-truth answer references, as evaluated by the reference-based metrics. While IDEFICS significantly improves when prompted with the contextual condition, this pattern is reversed for all other models. We hypothesize that less well-performing models closely reiterate the visual information they receive from the image description rather than adding new information to answer the question, which we turn to next.

5.1 Repetition of Visual Information Within Generated Answers

In order to investigate how much visual information models re-iterate within their generated answers, we conducted a similarity analysis of the model-generated answers to the image based on CLIPScore (Hessel et al., 2021) and a similarity analysis of the generated answers to the human-written image descriptions using Sentence-BERT embeddings (Reimers and Gurevych, 2019). The analyses show convergent results.

While the reference-based metrics tend to decrease in the contextual condition, CLIPScore ratings (i.e., the similarity of the generated answer to the image) largely increase, as seen in Table 2. The only exception is BLIP-2, which significantly deteriorates in performance across metrics when contextual information is provided, largely rendering it uninterpretable.

Similarly, Figure 4 shows that the similarity between the description and the generated answers significantly increases with the contextual condition across models, according to a two-sample t-test analysis. However, the increase in description

similarity between the baseline and contextual condition is the lowest for IDEFICS. We conclude that in their answers, the contextual versions of LLaVA, mPLUG-Owl, and BLIP-2 mimic the description more than IDEFICS does, which explains the CLIP-Score increase across the baseline and contextual conditions for these three models. These findings suggest that those models might overly emphasize the descriptive details when available, leading to less alignment with the ground-truth answers.

5.2 Quantifying Hallucinations

While IDEFICS (contextual) best approximates the human answers, it’s still far from achieving human-like performance. When a model is intended to make information available that can’t be verified by a user, it is especially important that this model only generates truthful content (MacLeod et al., 2017). While this is straightforward to evaluate for accuracy-based VQA datasets, this isn’t easily captured by the similarity-based metrics that are used for long-form evaluation, as measured by BLEU and CIDEr. In this section, we aim to give a brief intuition about the rate of hallucinated or incorrect pieces of information contained in the answers of the best-performing model, IDEFICS (contextual).

To estimate how many answers contain wrong, hallucinated or unverifiable content generated by the model, we conducted a human-subject study where participants were asked to rate model-generated answers. We randomly selected 100 answers from IDEFICS (contextual) to provide an assessment of the best-performing model. 70 participants were recruited, and compensated at a rate of \$13.80/hr. For each image, participants were asked to rate whether each answer contained infor-

Model	CLIPScore	CLIPScore (C.)
IDEFICS	0.679	0.691
LLaVA	0.731	0.786
mPLUG-Owl	0.720	0.758
BLIP-2	0.613	0.586

Table 2: **CLIPScore Improvement Across Baseline and Contextual Conditions.** When each model is prompted with the contextual condition, CLIPScore increases, indicating that the model is repeating visual details. “C.” represents the contextual condition.

mation that was clearly not within the image, and to evaluate whether the image was strictly necessary for providing an answer to the question.

Overall, participants indicated that 23% of the model-generated answers contained clearly erroneous information (with a Fleiss kappa inter-annotator agreement of 0.47) and for another 22%, the truthfulness of the answers couldn’t be clearly established. These results indicate that even the best-performing model (IDEFICS (contextual)) generates a high degree of erroneous information, making it unreliable for downstream use.

5.3 Evaluating Unanswerability

Since the questions in CommVQA are asked by people who cannot see the image, these questions are not always answerable. When a person asks an unanswerable question, ideally, models should abstain from answering rather than provide incorrect, hallucinated information (Whitehead et al., 2022).

We evaluated all models on the 238 unanswerable questions in CommVQA. For each question, we assessed whether the model was able to successfully abstain from answering or if the model provided a hallucinated answer. To classify responses as “Abstention” or “Non-Abstention,” we conducted a direct string matching analysis to search for language that models commonly use to abstain (e.g., “I cannot answer.”) A full list of strings in this analysis is provided in Appendix B.

Overall, IDEFICS (contextual) had the highest rate of successful abstentions at 21%, and BLIP-2 had the lowest rate, at 0% (in both contextual and baseline conditions). Without access to contextual information, IDEFICS’s abstention rate drops to 14%, suggesting that context is helpful for accurately identifying unanswerable questions. Appending the string “If you don’t know, say ‘unanswerable’” to the model prompt improved the rate

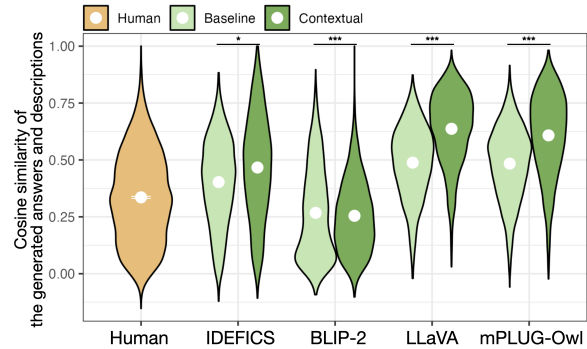


Figure 4: **SBert Cosine Similarity Analysis in Human and Model Responses.** Significance levels are marked with asterisks based on a two-sample t-test analysis.

of IDEFICS’ (contextual) abstention performances on the unanswerable questions to 87%, suggesting that implicit instructions could help models abstain when necessary.

Questions most human participants judged as *unanswerable* tended to be answered by the models as well. Without the contextual condition, LLaVA and IDEFICS declined to answer for answerable questions at a rate of 8.3% and 8.6%, respectively. For both models, the rate of false positives decreased in the contextual condition (LLaVA: 7.6%; IDEFICS: 7.1%).

Taken together, these results highlight that CommVQA poses a challenging problem for state-of-the-art models. Our analyses suggest that models might not be able to leverage contextual information effectively and the high degree of hallucinations makes them unreliable, highlighting two important areas for future research.

6 Conclusion

Visual question answering models are a promising tool for making visual content accessible to all. With CommVQA, we move the contextually isolated VQA task into a communicative setting that starts reflecting the diversity of downstream use, while keeping close control over the nature of the dataset to ease interpretability. We find strong evidence from dataset analysis that the types of relevant questions and answers change with the contextual domains where images appear. We also find evidence that integrating contextually relevant information improves model performance. Our results suggest that the path towards building viable VQA systems requires a focus on the wider communicative context where images appear.

Limitations

In this work, we show that the scenarios images are presented in fundamentally affect the VQA task. To investigate this, we varied the broad type of website where we embedded the image (Social Media, Shopping, etc). However, context effects are likely much more diverse than the effects studied in this work. For example, recent work suggests that even topic changes within a website domain (a Wikipedia article on Mountains vs. Body of Water) change the information needs that sighted and BLV users have for image descriptions (Kreiss et al., 2022b). This result likely translates to the VQA task and needs further investigation.

To allow for easy manipulations of the context domains and a highly controlled recruitment, participants were put in simulated scenarios where they were told about the website domain where the image appears. While even in these induced contextual setups, we find significant contextual variations, future work needs to explore the way this extends to real-world user experience outside of simulated scenarios.

In CommVQA, we elicited questions and answers from sighted participants. Evidence from prior work (Stangl et al., 2021) indicates that these results would likely transfer more directly to the accessibility scenario, but future work needs to analyze how the sighted user behavior translates to the BLV population more directly.

During the image description generation phase, initial descriptions are generated by GPT-4 and then edited by humans. This raises the potential issue that humans may be biased towards the GPT-4 responses, including keeping in details that may not be factually accurate. There is in fact evidence that starting out from an alt text shapes what type of descriptions people write and the quality of them (Mack et al., 2021). Work on alt text captioning has primarily focused on old systems, which were far less competent, and more research is now needed to investigate how that interacts with newer systems. In our dataset, the descriptions themselves are not proposed as potentially ideal descriptions for accessibility. They simply provide the contextual framing for the questions after.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel

Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. *IEEE International Conference on Computer Vision*, abs/1505.00468.

Katie Baker, Amit Parekh, Adrien Fabre, Angus Addlessee, Ruben Kruiper, and Oliver Lemon. 2021. The spoon is in the sink: Assisting visually impaired people in the kitchen. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 32–39.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. “It’s complicated”: Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Chongyan Chen, Mengchen Liu, Noel Codella, Yunsheng Li, Lu Yuan, and Danna Gurari. 2023. Fully authentic visual question answering dataset from online communities. *arXiv preprint arXiv:2311.15562*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman, editor, *Language, Logic, and Computation*, volume 1, pages 221–237. CSLI, Stanford, CA.

- Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*, pages 549–559.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition*.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Jeroen Groenendijk and Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.
- Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.
- Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022a. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697.
- Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022b. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697.
- Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing tools for high-quality alt text authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.
- Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.
- Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 2016. "With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516.
- Annika Muehlbradt and Shaun K Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Transactions on Accessible Computing (TACCESS)*, 15(1):1–32.
- OpenAI. 2023. GPT4-V System Card. [Online; accessed 10-Feb-2024].

- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#).
- OSU. 2024. Alternative (Alt) Text Guide. [Online; accessed 10-Feb-2024].
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-Bert: Sentence embeddings using siamese bertnetworks. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Craige Roberts. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Jae Hak Yoon and Andreas Kathol, editors, *OSU Working Papers in Linguistics*, volume 49: Papers in Semantics, pages 91–136. The Ohio State University Department of Linguistics, Columbus, OH. Revised 1998.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Abigale Stangl, Nitin Verma, Kenneth Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021.

Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*.

Robert van Rooy. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26(6):727–763.

Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1584–1595.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality.

A Appendix

A.1 Statistical Comparisons with Other VQA Datasets

In Table 3, we provide a statistical comparison of CommVQA with other VQA datasets.

A.2 Dataset Collection Overview

In this section, we provide a full list of the tasks and prompts provided in each step of the dataset collection process.

For the scenario matching stage, we prompted GPT-4V with this prompt:

Imagine you are a person browsing the Internet. Please rank the following scenarios in which you might encounter this image:

- 1) You are browsing a shopping website, with the goal of purchasing an item or experience.
- 2) You are browsing science magazines (such as National Geographic), with the goal of learning more about recent science developments.
- 3) You are browsing news websites (such as New York Times), with the goal of learning more about recent news developments.
- 4) You are browsing a health website, with the goal of learning how to live a healthier lifestyle.
- 5) You are browsing social media, with the goal of learning more about your connections.
- 6) You are browsing a travel website, with the goal of traveling to a new location.

For the description elicitation, we asked participants to: “Please edit this description of the image to correct any errors and make the description more useful to someone who cannot see it.” This phrase was intended to incentivize people to both fix errors and include any communicative details that they felt was missing.

For the question elicitation study, we asked participants: “Imagine you are browsing a {scenario} website when you encounter the following image: {description}. If you encounter this image on a {scenario} website, what are two questions you’d want to have answered by someone who can see the image?”

For the answer elicitation study, we asked participants to: “Imagine you are browsing a {scenario} website when you encounter the following image.

Dataset	Who Asked?	\bar{Q}	\bar{A}	\bar{C}	# Imgs	# Sce- narios	# QA Pairs
VQA-v2	Crowdworkers	6.1	1.2	\times	204,700	1	658,111
VizWiz- VQA	Blind people	6.7	1.7	\times	20,500	1	31,000
OK-VQA	Crowdworkers	8.1	1.3	\times	14,031	1	14,055
DocVQA	Remote workers	9.5	2.4	\times		1	
Visual Dialog*	Crowdworkers	5.1	2.9	\times	120,000	1	1.2M
VQAOnline	Stack Overflow users	9.3	173.2	127	64,700	1	64,700
ScienceQA	Elementary and high school cur- ricula	12.1	4.4	41	6,500	1	21,208
CommVQA (ours)	Crowdworkers who can only see the descrip- tion	7.3	10.98	17.2	1000	6	8949

Table 3: **Statistical Comparison: CommVQA and Other VQA Datasets.** This table compares CommVQA with seven other VQA datasets. The first four rows cover image-question-answer inputs, while the bottom three rows cover image-question-answer-context inputs. \bar{Q} , \bar{A} , and \bar{C} denote the average answer, question, and context lengths (as measured by the number of words), respectively. *Since Visual Dialog contains multi-turn conversations, we only included the average question length for the first round in the conversation.

{image} Another user cannot see the image directly but has access to the following image description: {description}. Based on the description, they asked a follow-up question. Please answer based on the image. {question}.”

We recruited annotators through Prolific [3], an online annotation platform similar to Amazon’s Mechanical Turk. We recruited only US-based participants, and the only prescreener we used was to exclude participants who had taken a previous stage of the description generation process. For instance, participants who edited descriptions were excluded from providing questions or answers.

A.3 Context Integration Prompts for Models

We integrated context by prompting models with the format:

Assume someone is browsing a {scenario} website when they encounter this image. They cannot see the image directly, but they can access this image description: {description}. Based on this description, they asked this follow-up question. Please answer based on the image. In your answer, prioritize details relevant to this person. Question: {question}

A.4 Additional Dataset Composition Analysis

A.5 Models Chosen

We evaluated the following model versions, which are freely available on HuggingFace:

LLaVA llava-v1.5-14b-3GB

IDEFICS idefics-9b-instruct

mPLUG-Owl mPLUG-owl-llama-7b

BLIP-2 blip2-opt-2.7b

B Unanswerability Analysis

In this section, we include a list of the strings used to evaluate whether a model abstained from answering.

- I cannot answer
- I cannot determine
- I cannot provide
- unanswerable
- I don’t have enough information
- AI language model
- I don’t have the context
- I don’t have enough context
- I’m sorry
- I don’t have the capability
- I don’t have enough information





Image	Description	Context	Question	Human Answer
	A group of people of various ages walking in line along a grassy path, with trees on one side.	Travel Health	Do the people look happy? How accessible is the path?	The group of people has various expressions. Some people in the back are shown to be giggling and look happy while people in the front are shown to have neutral expressions. The path isn't very accessible. It is a dirt path with lots of small rocks, in what looks to be a wooded area.
	A massive eruption of lava is spewing from a volcano into the sky, with a fiery red and orange glow.	Science News	How far away is the photo being taken? Are there any people or human-made structures in the image?	It's difficult to say given it is probably zoomed in, but at least 50 feet away from the eruption No humans or man made structures in the image
	A diver, equipped with scuba gear, explores an underwater wreck surrounded by vibrant coral formations. The deep blue ocean provides a mysterious backdrop.	Shopping Science	What color are the swimming flippers? What does the wreck look like? (Structure, wear and tear, etc.)	The swimming flippers are blue It looks rusted and full of barnacles. It's also very mossy.
	A group of people are working out on indoor rowing machines at a gym.	Health Social Media	How old are they? What kind of clothes are they wearing?	The people in the image are likely to be in their low to mid 30s The people are mostly wearing workout clothes, such as shorts, t-shirts, tank tops, and tennis shoes.

Figure 5: **CommVQA Dataset Examples.** Four example entries from the CommVQA dataset, each paired with a randomly selected answer.




	Social Media	News	Travel	Health
Scenario	A burrito bowl with a drink on a wooden table.	Lush green tea fields roll over hills under a sunlight-dappled tree in Boseong, Korea.	In front of you is a platter full of delicious Indian food. The largest portion on this platter is a bowl of fluffy white rice. Surrounding the rice are 6 smaller ramekins full of various curries in a multitude of colors.	A person hiking on a winding trail through a lush green forest with tall pine trees. Majestic snow-capped mountains rise in the distance
Question	What is in it?	What else can you see in the image?	What does the rest of the image look like?	Is there any fencing?
Image				
Human Answer	The burrito bowl looks to have lettuce, pico de gallo, rice and meat in it as well as a spoon. The drink appears to be a tea cup with a clear liquid in it - either tea or water.	The green tea fields are set up in perfect rows of landscaped bushes on the hills.	The platter is made of metal and is on a flat surface, which is most likely a table. There is a clear bottle of water behind the platter with no label on it and a checkered table cloth with small blue and white checkers and some doilies as well.	No, there does not appear to be any fencing. It is just an open trail.
Model Answer (IDEFICS) baseline	The bowl contains guacamole and salsa.	In addition to the tea plantation, I can see a large number of trees, some of which are quite tall. The trees are scattered throughout the image, with some located in the foreground and others in the background. There are also some bushes visible in the image.	The rest of the image shows a meal served on a metal plate. The meal consists of rice, curry, and a variety of other foods. There are also some condiments and a bowl of water. The plate is placed on a table, and the table is covered with a tablecloth.	No, there is no fencing visible in the image.
Model Answer (IDEFICS) contextual	The burrito bowl has a variety of ingredients including rice, beans, lettuce, tomatoes, and avocado.	In the image, you can see a beautiful landscape of lush green tea fields that are rolling over hills under a sunlight-dappled tree in Boseong, Korea.	The rest of the image is a close-up of the platter of food. The rice and curries are the main focus of the image.	No, there is no fencing visible in the image.

Figure 6: **Examples of Model-Generated Answers.** This figure shows examples of dataset entries and model responses from IDEFICS, the leading model. Sample responses from IDEFICS, the leading model, indicate that context aids in producing answers that better match human ground-truths, despite some limitations in context integration.

C Effects of Dataset Size

We conducted an analysis on a subset of our dataset and uncovered similar effects of model behavior. In Table 4, we present model benchmarking results on a subset of 150 images and 1,215 unique visual questions within our dataset. Largely, we surface similar patterns as found within the full dataset. Namely, IDEFICS (contextual) exhibits the highest performance, and LLaVA and mPLUG-Owl both decrease their performance with their contextual condition, but this is explained by an increase in CLIPScore, as shown in Table 5.

D Potential Use Cases of the CommVQA Dataset

One potential use case for the CommVQA dataset is to assess whether a model performs better within a certain scenario than others. For instance, given a general-purpose VQA model, the CommVQA dataset provides a way to analyze whether this model excels at questions asked within the Shopping scenario, but struggles on questions asked within the Science Magazines scenario. These insights could prove useful for assessing when to deploy general-purpose VQA models versus specialized, domain-specific models.

Another use case is to evaluate a model’s ability to integrate contextual information along with the image. Prior work shows that people value visual explanations that incorporate contextual information (Stangl et al., 2021, 2020; Muehlbradt and Kane, 2022). But how well do models integrate this contextual information in practice? In Figure 4, we find that LLaVA and mPLUG-Owl both tend to repeat information from the description, while IDEFICS is more successful at integrating the contextual condition. CommVQA can help assess the ability of VQA models to integrate this contextual information, which as prior work has shown, is crucial for ensuring that model-generated outputs align with human preferences.

E Clarifying Context

This section serves to clarify the role of context and how it is distinct from prior work. In our case, “context” means not only the scenario (i.e., the website where the image was encountered), but the scenario and the description. Prior work suggests that for blind and low vision users, this broader context should shape people’s informational needs

about an image (Stangl et al., 2021, 2020). Although other types of context have been shown to result in an improvement in the VQA and image captioning literature, this pragmatic framing of context is unique to our dataset. Our work bridges the gap between the studies showing that this version of context matters, and puts this observation into practice in generating a VQA dataset with a more realistic distribution of questions and answers.

Model	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
IDEFICS	0.326	0.106	0.189	0.401	0.825
IDEFICS (Contextual)	0.356	0.122	0.198	0.413	0.904
mPLUG-Owl	0.191	0.045	0.166	0.311	0.414
mPLUG-Owl (Contextual)	0.178	0.043	0.169	0.284	0.335
LLaVA	0.192	0.050	0.171	0.311	0.431
LLaVA (Contextual)	0.149	0.037	0.161	0.247	0.200
BLIP-2	0.289	0.067	0.100	0.286	0.435
BLIP-2 (Contextual)	0.059	0.010	0.078	0.224	0.343

Table 4: **Model Benchmarking Results on a Subset of 150 Images from CommVQA.** Reflecting similar results from the full dataset benchmark, the highest-scoring model is IDEFICS (contextual), and LLaVA and mPLUG-Owl both exhibit decreased performance within the contextual condition.

Model	CLIPScore	CLIPScore (C.)
IDEFICS	0.667	0.666
LLaVA	0.729	0.784
mPLUG-Owl	0.720	0.756
BLIP-2	0.604	0.662

Table 5: **CLIPScore Improvement Across Baseline and Contextual Conditions for Subset of 150 images.** This table displays the increase in CLIPScore for each model when comparing baseline and contextual conditions for the subset of 150 images, finding a similar trend as when this result is run on the full dataset. In particular, LLaVA (C.) has the highest CLIPScore condition, and the CLIPScore stays stable across IDEFICS and IDEFICS (contextual).