

Fully Authentic Visual Question Answering Dataset from Online Communities

Chongyan Chen¹, Mengchen Liu², Noel Codella², Yunsheng Li², Lu Yuan², and Danna Gurari^{1,3}

¹ University of Texas at Austin, ² Microsoft, ³ University of Colorado Boulder

Abstract. Visual Question Answering (VQA) entails answering questions about images. We introduce the first VQA dataset in which all contents originate from an authentic use case. Sourced from online question answering community forums, we call it VQAonline. We characterize this dataset and how it relates to eight mainstream VQA datasets. Observing that answers in our dataset tend to be much longer (i.e., a mean of 173 words) and so incompatible with standard VQA evaluation metrics, we instead utilize popular metrics for longer text evaluation for evaluating six state-of-the-art VQA models on VQAonline and report where they struggle most. Finally, we analyze which evaluation metrics align best with human judgments. We publicly-share the dataset at: <https://vqaonline.github.io/>.

Keywords: Dataset Creation · Visual Question Answering

1 Introduction

We are witnessing a renewed excitement in visual question answering (VQA), the task of answering questions about images, with recent successes of large foundation models in zero-shot and few-shot settings. Moreover, many papers are reporting that such models can achieve or even exceed human performance [24, 32, 40].

While modern VQA datasets enabled this exciting progress in VQA model development, they have two key limitations. First, most publicly-available VQA datasets [10, 33, 35] are artificially-constructed to probe model performance rather stemming from authentic use cases. Yet, repeatedly, prior work [12, 13, 42] has revealed that models developed for artificially-constructed settings generalize poorly to authentic use cases due to significant domain shifts. The second key limitation of existing datasets lies in a limited diversity of content representing authentic use cases. The only VQA dataset reflecting an authentic use case is VizWiz-VQA [11]. However, VizWiz-VQA is constrained to one population, specifically blind individuals seeking assistance in learning about their visual surroundings. A further limitation of the dataset is that the answers were collected years later to support model benchmarking, rather than having been validated as acceptable by those who asked the questions.

We address these limitations by introducing a new dataset called *VQAonline*. Sourced from online question answering communities, it is the the first VQA dataset with all contents originating from an authentic use case. We conduct analysis to highlight how this dataset differs from prior datasets; examples include that it contains: (1) authentic

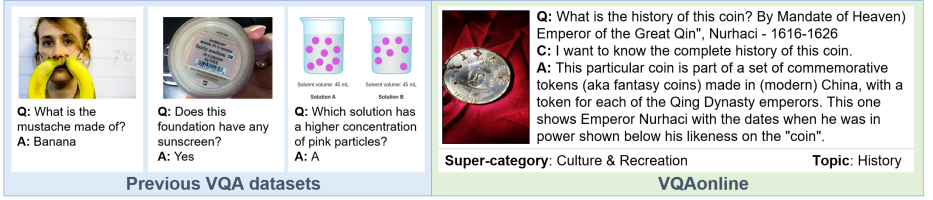


Fig. 1: VQA examples from our VQAonline dataset and three mainstream VQA datasets [10,33,35]. VQAonline is the first VQA dataset to originate from an *authentic* use case end to end, including with authentic context, answers, and topics/categories labels. It also is the first VQA dataset from an online question answering platform. A critical distinction of VQAonline, which necessitates a new evaluation methodology, is its lengthy answers. (Q=question, C=context, A=answer).

context that clarifies the question, (2) an answer the individual asking the question validated as acceptable, (3) answers that are considerably longer (*i.e.*, a mean of 173 words versus typically 11 words or fewer in prior work), and (4) user-chosen topics for each visual question from 105 diverse topics. Such differences are exemplified in Figure 1. Next, observing that the mainstream VQA evaluation metric is ill-suited our dataset’s lengthy textual answers, we benchmarked the performance of six modern foundation models with respect to popular evaluation metrics for long-form text. Our findings reveal that modern models perform poorly overall, underscoring of this dataset to facilitate the research community’s continued improvement of VQA models. We conclude with assessing which related evaluation metrics for long-form text best align with human judgement for our dataset to also facilitate future work in improving evaluation metrics for lengthy textual answers.

We expect improved models designed for VQAonline can also benefit many other real-world applications including for other community question answering platforms (*e.g.*, Yahoo! Answers, Wiki Answers, Quora), learning assistance VQA tools for children [15], virtual assistants, visual dialog and query, as well as visual search and browsing. To encourage community-wide progress, we publicly-share the dataset at <https://vqaonline.github.io/>.

2 Related Work

Visual Question Answering Datasets. A commonality of all VQA datasets (*e.g.*, [1,4,10,12,17,19,35]) is the inclusion of artificially-generated content. The closest dataset to reflect an authentic use case is VizWiz-VQA [12], as the visual questions were all submitted by individuals seeking assistance to learn about their visual surroundings. However, the answers were crowdsourced and never shared with or validated by those asking the questions. To our knowledge, our work is the first to provide an authentic dataset end-to-end, with each example including a natural language question, image, context, and answer validated by the individual asking the question. We conduct extensive analysis to reveal unique aspects of our dataset (Section 3.2).

Our work also aligns with VQA datasets providing context that supplements each visual question: Context-VQA [29] and ScienceQA [33]. However, *context* in Context-VQA [29] is defined as the kind of website the image is sourced from rather than our definition of supplementary data that supports answering the visual question. More similar to our work is ScienceQA [33], however the context is both artificially-generated and has distinct compositions and content, as discussed in Section 3.2.

Community Question Answering (CQA) Datasets. Numerous datasets have been developed around CQA. However, all focus on text-based question answering [36] or natural language questions with multimedia answers [30]. To our knowledge, our work is the first to introduce a community multimodal question answering dataset specifically for visual question answering.

VQA Models. Traditional VQA models generate short answers [38] and do not consider extra context to help answer the questions. Yet, our dataset features answers that are lengthy, typically containing multiple sentences, and provides additional context that can help answer visual questions. Accordingly, we benchmark large Vision and Language Models (VLMs) [5, 21, 23, 32, 41, 44] which, by their generative design, are better-suited to generate long-form answers using their strong natural language understanding capabilities. Overall, we find these models perform poorly and provide fine-grained analysis to reveal opportunities for improving such models.

3 VQAonline Dataset

We now introduce our VQAonline dataset originating from an authentic use case where people used a web-based platform to post and answer visual questions.

3.1 Dataset Creation

Source. We created our dataset by scraping data from the online question-answering platform Stack Exchange. We used the publicly-available data that spanned from August 2010 to November 2022, and adhered to all data license restrictions (cc-by-sa 4.0) set forth by Stack Exchange.

From the Stack Exchange archive, we collected all examples containing four components: a natural language question, a natural language context, an image, and at least one answer. Originally, this information was collected from Stack Exchange users via four user entry interactions. The natural language question was collected as the post "title", using a text entry field paired with instructions "be specific and imagine you're asking a question to another person" when writing the question. The image was collected as a URL link. The natural language context was provided via a "body" text entry field supporting inline images (via a URL link) with users instructed to supply "all the information someone would need to answer your question". These three pieces of information were then shared publicly, after which individuals on the site could publicly provide

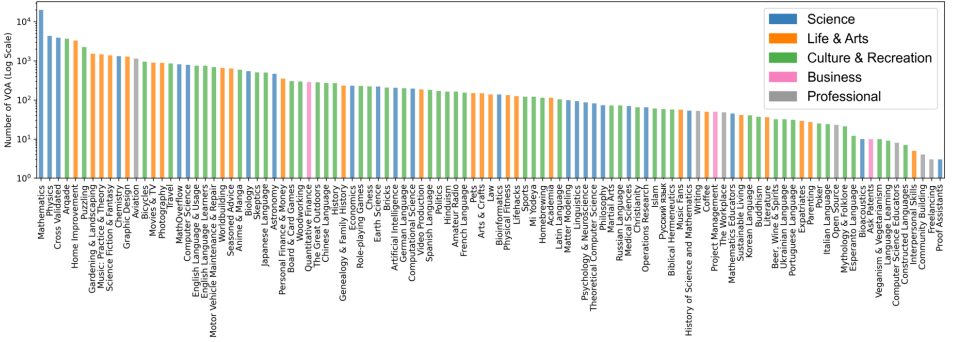


Fig. 2: Number of visual questions per topic (in log scale). The colors, as defined in the legend, indicate the super-category for each topic.

answers.¹ Finally, individuals who posted the visual questions could mark which answer was the “accepted answer”; we only use examples for which individuals did this.

Of note, the Stack Exchange platform has two key mechanisms that contribute to high quality VQAs. First, the platform enforces a stringent reputation system with moderation policies.² Second, it features topic-specific sites for users to engage with a relevant community; *e.g.*, music.stackexchange.com and gardening.stackexchange.com. We used the data from 105 topical sites that belong to five super-categories, with 46 about culture & recreation (*e.g.*, travel, sports, Christianity, French language), 25 about life & arts (*e.g.*, physical fitness, parenting, pets), 24 about science (*e.g.*, physics, chemistry, biology), 7 about being professional, and 3 about business.³ All topics are shown in Figure 2.

Filtering. We conducted two filtering steps to obtain our final dataset of **64,696 VQAs**. First, we removed low-quality VQAs based on Stack Exchange’s feature that enables users to up-vote or down-vote each post (either question or answer). Each up-vote added 1 to the score and each down-vote subtracted 1 from the score. We removed visual questions with scores 0 or less for either the question or answer. Second, we excluded visual questions with multiple images (*e.g.*, GIF) or with visual answers.

Dataset Splits. VQAonline supports few-shot learning by containing training, validation, and test splits of 665, 285, and 63,746 examples respectively. We created training and validation splits by randomly selecting from each of the 105 topics in our dataset 7 examples for training and 3 examples for validation for topics with at least 20 examples.⁴

¹ Individuals could either propose edits to existing answers, which had to then be approved by high-reputation users, or could propose new answers. Further community deliberation to arrive at high-quality answers could occur when users posted comments under the question or answers. The mean number of answers per visual question is 1.45.

² The moderation policies are documented at <https://meta.stackexchange.com/help>.

³ We excluded 75 technology topics (*e.g.*, Stack Overflow, LaTeX), because they typically are language problems. In particular, the questions typically ask about code shown in screenshots.

⁴ No topics are excluded from the test dataset.

Table 1: Characterization of existing VQA datasets and our new VQAonline in terms of mean question length (i.e., $\bar{Q}$), mean answer length (i.e., $\bar{A}$), number of images (i.e., N_{img}), number of authentic topics (i.e., # Auth. Topics), inclusion of context, inclusion of authentic visual questions (i.e., Auth. VQ), inclusion of authentic context (i.e., Auth. C), and inclusion of answers validated by those asking the questions (i.e., Vld. A).

VQA Dataset	$\bar{Q}$	$\bar{A}$	N_{img}	# Auth. Topics	Context	Auth. VQ	Auth. C	Vld. A
VQAonline (Ours)	9.3	173.2	64.7K	105	✓	✓	✓	✓
Context-VQA [29]	6.5	11.0	42	✗	✗	✗	✗	✗
VQAv2 [10]	6.1	1.2	204.7K	✗	✗	✗	✗	✗
VizWiz-VQA [12]	6.7	1.7	20.5K	✗	✗	✓	✗	✗
OK-VQA [25]	8.1	1.3	14K	✗	✗	✗	✗	✗
DocVQA [27]	9.5	2.4	12.8K	✗	✗	✗	✗	✗
ScienceQA [33]	12.11	4.4	6.5K	26	✓	✗	✗	✗
InfographicVQA [26]	11.5	1.6	5.5K	✗	✗	✗	✗	✗
INFOSEEK [4]	8.9	1.5	1.36M	✗	✗	✗	✗	✗

3.2 Dataset Analysis

Comparison to Existing Datasets. To contextualize how our dataset compares to the current focus of the research community, we compare VQAs in our dataset to those in eight existing popular VQA datasets. We chose VQAv2 [10] because it is the most popular (e.g., highly cited) VQA dataset, Context-VQA [29] and ScienceQA [33]⁵ because they similarly claim to provide context, VizWiz-VQA [12] because it similarly originates from an authentic use case, INFOSEEK [4] because it similarly centers on information seeking rather than probing an AI model, OK-VQA [25] because it similarly requires domain-specific knowledge, as well as DocVQA [27] and Infographic VQA [26] because they similarly contain VQAs with documents and infographics.

For each dataset, we report the mean number of words in each question, mean number of words in each answer, number of images, number of topics which were authentically used to label clusters of VQAs by theme, whether context is included, whether the visual questions are authentic, whether the context is authentic, and whether the answers are validated by those asking the questions. Results are reported in Table 1.

Our findings reveal commonalities of our VQAonline dataset with existing datasets. For example, we observe that the visual questions themselves are similar in terms of the typical question length and number of images. Questions have a mean length of 9 words in our VQAonline compared to 6 to 11 words in other datasets and the nearly 65,000 images in our dataset are comparable in size to existing datasets.

Our analysis also demonstrates significant differences from existing datasets across two key dimensions. The first dimension is the *dataset authenticity*. Our VQAonline dataset is the only fully authentic dataset for the entire VQA pipeline, meaning all contents originate from an authentic use case. The second key difference of our dataset is the typical *answer length*. Answers in our VQAonline dataset are orders of magnitude longer than in existing datasets, with a mean length of 173 words in VQAonline versus 11 or fewer words in existing datasets. This larger length is not anomalous, as the median

⁵ We focus on the subset of ScienceQA with both image and context.

answer length in VQAonline is 120 words. This longer length in part reflects responses are provided in paragraph form, with most commonly 3 sentences per answer. This distinction renders the mainstream evaluation metric for VQA models largely ineffective, as it assumes brief responses [1]. This issue will be explored in Section 4.2.

VQAonline Topics. The authentic user-based posting of each VQA within a topic community reveals the diversity of content in our dataset. We report the number of VQAs in each of the 105 topics in Figure 2. As shown, there can be as few as 10 examples in a topic to over 20,000.

VQAonline Questions. Our VQAonline dataset is the first with authentic visual questions asked by users of an online question-answering platform. This results in a domain shift from the only other VQA dataset with authentically asked visual questions, VizWiz-VQA [12], which were asked by blind users of a mobile phone application. For instance, blind people typically asked about their physical surroundings (e.g., "What is this?", "What does this say?") while Stack Exchange users often asked for more fine-grained or specialized information about the images (e.g., "What's a better term to describe these 'stripes' in blouse (Screen shot attached)", "What is the third character on this sign seen in Taipei?"). Another distinction stems from Stack Exchange's feature for avoiding redundant questions by letting users remove duplicates of other already submitted questions. The consequence is that only 0.2% of natural language questions in our VQAonline dataset are duplicated compared to 37.9% of questions in VizWiz-VQA.

VQAonline Context. Our VQAonline dataset is the first to include authentic context. The only other VQA dataset containing context, as defined by supplementary data that supports answering the visual question, is ScienceQA [33]. However, the context was contrived with heuristic rules for extracting data from elementary and high school science curricula. Consequently, context in ScienceQA is typically narrated in third-person, while context in our VQAonline often are narrated in first-person. Another difference is ScienceQA has a shorter average length of the context compared to that in our dataset; i.e., a mean of 41 words, median of 28 words, and most commonly 3 sentences versus a mean of 127 words, median of 94 words, and most commonly 4 sentences. A further difference is that the context in ScienceQA only relates to 26 science topics while the context in our dataset spans 105 topics that extend beyond science.

VQAonline Image Contents. We next quantify the tendency for images to include screenshots and infographics to reflect the tendency for visual questions to be related to the digital realm and data analysis. We identified an image as a "screenshot" when contents were copied digitally from a computing device and an "infographic" when the image shows a chart, table, graph, attention map, map, or diagram. We manually labeled 105 images sampled *from each of the 105 topics* as well as an independent random selection of 100 images from the *entire dataset*. We found for the topic-sampled set that 79% are screenshots and 17% show infographics and for the samples from the entire dataset that 84% are screenshots and 33% show infographics. The similarly high prevalence of screenshots across both samples underscores Stack Exchange users' tendency to ask questions about the digital realm. We hypothesize the different prevalence

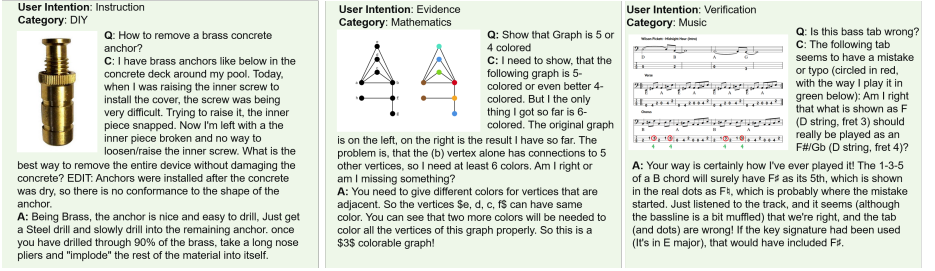


Fig. 3: Examples of three visual questions with three different user intents.

of infographics across both samples is due to topical differences, as 52% of all VQAs (i.e., 33,384 out of 64,696) belong to only 23% of the topics (i.e., 24 science topics). Supporting this hypothesis, we found from inspecting 100 random VQAs from each subset of the science topics and non-science topics that infographics were in 58% and 6% of images respectively.

VQAonline Importance of Images. We next examine the value of images for answering the questions. To do so, we flagged each image as *question-aligned* if elements described in the question are found in the image and *context-aligned* if elements described in the context are found in the image. We additionally flagged whether an image is *necessary*, based on whether the answer can be obtained without the image. We manually labeled 105 examples from each of the 105 topics. We found that 91.4% of the images are *question-aligned* and 99% of the images are *context-aligned*. This finding contrasts that of the only other dataset with authentic visual questions, VizWiz-VQA [12], where only 71% are deemed answerable because blind photographers struggled to capture contents of interest in their images. We also found that 65.7% images are *necessary* for answering the question in our dataset. Such examples are shown in Figures 3 and 5. Often, the purpose of unnecessary images was to motivate or show scenarios in which the asker encounters the question. Altogether, these findings underscore that VQA models often must analyze the images to arrive at the answers.

VQAonline User Intentions. We finally characterized intents behind the visual questions, complementing two prior works that identified user intentions for three community question answering platforms (Answerbag, Metafilter, and Yahoo! Answer) [14] and the music StackExchange community [8]. Our work is the first to investigate user intentions for VQAs. Through a multi-round analysis of the data, described in the Supplementary Materials, we defined eight potential intents: *instruction*, *evidence*, *verification*, *advice*, *identification*, *opinion*, *reason*, and *other*. Examples for three of these categories are shown in Figure 3. We then recruited crowdworkers to annotate the intent for 105 examples, with one from each of the 105 topics. The resulting distribution of assigned categories is 19% are instruction, 18% evidence, 17% verification, 16% advice, 10% identification, 10% opinion, 8% reason, and 1% other. Examples from the three most common categories are shown in Figure 3.

4 Experiments

4.1 Model Benchmarking

We next analyze the performance of modern models on our VQAonline dataset to reveal what makes it challenging for the research community.

Models. We evaluate the following six Vision and Language Models (VLMs) that achieve strong performance for various VQA datasets and subtasks in a zero-shot setting:

- **BLIP-2** [21]: state-of-the-art (SOTA) for zero-shot setting on four VQA datasets, including ScienceQA and VizWiz-VQA according to LVLM-eHub, a comprehensive evaluation benchmark for VLMs [39].
- **MiniGPT-4** [44]: ranked first in open-world VQA according to LVLM-eHub [39].
- **LLaVA** [23]: top-two performance in the zero-shot setting for three VQA datasets, including for DocVQA [2, 27] according to LVLM-eHub [39].
- **mPLUG-Owl** [41]: SOTA for VQAv2 [10] in early 2023 and ranked second overall in open-world VQA scenario according to LVLM-eHub [39].
- **InstructBLIP** [5]: SOTA for zero-shot setting on four VQA datasets, including OK-VQA [3, 17, 25, 28] according to LVLM-eHub [39].
- **GPT-4V** [32]: SOTA for numerous VQA datasets, such as Infographic VQA [27] and TVQA [20]. While we aim to assess its performance, note that its reproducibility is a major concern due to its proprietary nature, which means the model can evolve without transparency or even cease to exist. Given that concern paired with its costliness (i.e., access is possible only through a fee-based API [7]), we evaluate GPT-4V only on a subset of data. Due to space constraints, we report key findings here in the main paper and all results in the Supplementary Materials.

Of note, most of these models are believed to be trained on Stack Exchange data and so potentially saw our VQAonline data⁶, with three (MiniGPT-4, LLaVA, and mPLUG-Owl) confirmed to have language encoders trained using Stack Exchange data⁷ and three others (BLIP2, GPT-4V, InstructBLIP) suspected (the original source of parts of these models’ training data are not publicly shared). While models exposed to such data during training bring an unfair advantage, evaluation with such models is still valuable. Their poor performance *strengthens* the argument that our dataset offers a difficult problem for the research community since they perform poorly despite their advantage.⁸

Evaluation Metrics. Given the deficiency of the existing VQA evaluation metric for our dataset, in that it is designed for short responses [1], we instead leverage five evaluation metrics for evaluating longer text from related fields (*e.g.* long-form question answering and image captioning). Like prior work, we evaluate with respect to multiple metrics to address a well-known issue that each brings deficiencies.

⁶ This issue also applies for all Flickr-based datasets including COCO (the source for many VQA datasets [12]), CommonCrawl, and L400 [34].

⁷ 2% of their pre-training data are from the 28 largest topics on Stack Exchange.

⁸ A possible step for future work is retraining open-sourced VLMs with publicly-disclosed training datasets by removing all data available in VQAonline.

Three metrics assess **similarity** between model-generated answers and human-generated answers, with their values ranging from 0 (bad) to 1 (perfect). *ROUGE-L* [22] identifies the longest sequence of words shared between the reference and the predicted answer, and consequently neglects semantic similarity (*e.g.* synonyms), context (*e.g.* different meanings of a word based on context, such as for "the bank of America", "the bank of a river" would receive a higher score than "bank account"), and negation (*e.g.* for "Exercise is good for human health", "Exercise is NOT good for human health" would receive a higher score than "Humans can benefit from exercise"). *METEOR* [6] offers a more sophisticated utilizes more advanced techniques to assess word similarity between the reference and predicted answers, such as stemming and synonyms, but it also neglects context and negation. *BERTscore* [43], measures the similarity of BERT’s contextual embeddings between the reference and predicted answers to incorporate consideration of context. However, context can reward semantically related but factually incorrect content as well as grammatically correct, well-written responses that are inaccurate [9].

Two metrics assess **image compatibility** between the image and model-generated answer, with their values ranging from 0 (bad) to 1 (perfect). *CLIP-S* [16], designed to assess the compatibility between an image and a caption by measuring the similarity between image and text embeddings, has been shown to have a positive correlation with human evaluations on answer faithfulness (*i.e.* if the answer can be grounded in the image versus is hallucinated) [18]. However, CLIP-S can reflect the biases of the pretraining data [16] and ignores the reference answers. *RefCLIP-S* [16] incorporates consideration of the reference answer by calculating a harmonic mean of the CLIP-S score and the cosine similarity between the reference answer and the predicted answer.

Overall Results. Results are shown in Table 2.⁹

⁹ As noted earlier, results for GPT-4V on the dataset subset are provided in the Supplementary Materials. They show GPT-4V achieves the best performance followed closely by mPLUG-Owl.

Table 2: Performance of VLMs on our VQAonline dataset with respect to five metrics both for the zero-shot setting as well as for the top-performing zero-shot model in a one-shot setting. As shown, mPLUG-Owl is the best-performing model in the zero-shot setting and has a small drop in performance regardless of which of four types of in-context exemplars that it receives. All metrics range from 0 to 1 with larger values signifying better performance.

Models	ROUGE-L	METEOR	BERTscore	CLIP-S	RefCLIP-S
mPLUG-Owl [41]	0.16	0.12	0.76	0.72	0.75
LLaVA [23]	0.15	0.08	0.75	0.73	0.75
MiniGPT-4 [44]	0.15	0.09	0.75	0.70	0.74
InstructBLIP [5]	0.09	0.05	0.67	0.68	0.71
BLIP2 [21]	0.09	0.04	0.69	0.66	0.71
mPLUG-1shot-Random-noImg	0.138	0.106	0.750	0.715	0.747
mPLUG-1shot-MatchedTopic-noImg	0.139	0.108	0.756	0.716	0.749
mPLUG-1shot-Random	0.135	0.104	0.745	0.717	0.748
mPLUG-1shot-MatchedTopic	0.137	0.105	0.752	0.717	0.748

We first examine the ranking of all models. The best-performing model is mPLUG-Owl, based on four of the five evaluation metrics. The next best models are LLaVA and MiniGPT-4, which consistently outperform the remaining two models BLIP2 and InstructBLIP across all evaluation metrics. This ranking of models was also found in prior work [39]. We hypothesize this observed ranking arises in part due to mPLUG-Owl, LLaVA, and MiniGPT-4 bringing disproportionately greater advantages in having previously seen Stack Exchange data and mPLUG-Owl and LLaVA having been fine-tuned with instruction-following data for VQA tasks. The one exception to this ranking is that LLaVA is the top-performing model with respect to the CLIP-S score followed mPLUG-owl. This suggests LLaVA generates answers that are more compatible with the image, rather than answers that are necessarily factually accurate.

While the evaluation scores support ranking the models’ performance, such scores for each evaluation metric are compressed in a small range. We attribute this to a known limitation that such metrics perform best at distinguishing poor versus excellent generated text rather than between mediocre and good generated text [31]. We suspect all benchmarked models fall into this latter range, as we will show that evaluation scores can vary much more than in this compressed range for specific topics, signifying poor and excellent performance for subportions of the dataset (Figure 4).

Best Performing Model in One-shot Setting. We explore for the best-performing model to what extent in-context learning examples can provide a further performance boost. We choose two types of training exemplars: (1) randomly selected and (2) matching topic tag. We explore two variants of both types, where we include versus exclude the images in the exemplars. We perform experiments in a one-shot setting and design prompts based on best practices established by prior work.

Results are shown in Table 2. As shown, none of the one-shot examples enhance performance. This could be due to the model overfitting to or becoming distracted by the exemplar. This also may relate to the inherent diversity of VQAs within our datasets, where even VQAs within the same topic can require very different skills and knowledge. This concern is further supported by the fact that, overall, models benefit from a simpler exemplar that lacks images. Moreover, exemplars with matching topic tags consistently outperform randomly chosen ones, underscoring that greater exemplar similarity to the test case has the potential to boost performance.

Analysis With Respect to Input Types. We next analyze the predictive power of each input of the top-performing model, mPLUG-Owl, by removing each independently: i.e., Q+C, Q+I, and C+I. Results are shown in Table 3.

Overall, the best-performing model all the information (i.e., Q+C+I), and this is also observed for GPT-4V (see Supplementary Materials). This underscores a complementary benefit provided by each of the three input types.

The findings suggest that context (which is typically language-based) is the most valuable source for arriving at the answer. That is because the worst-performing variant across both models, according to four of the five evaluation metrics, lacks *context* as input (i.e., Q+I). The one exception is for CLIP-S, which ranks the model lacking *context* (i.e., Q+I) as the *best-performing* model. We attribute this conflicting finding from different evaluation metrics to the specifics of CLIP-S; specifically, it measures how compatible

Table 3: Analysis of the top-performing VQA model, mPLUG-Owl, when fed different input types (Q+C+I, C+I, Q+C, Q+I), including for each of the five VQA super-category types.

Models	ROUGE-L	METEOR	BERTscore	CLIP-S	RefCLIP-S
mPLUG(Q+C+I)	0.16	0.12	0.76	0.72	0.75
mPLUG(C+I)	0.14	0.12	0.76	0.72	0.75
mPLUG(Q+C)	0.14	0.11	0.76	0.71	0.75
mPLUG(Q+I)	0.13	0.07	0.74	0.73	0.73
Science	0.15	0.14	0.78	0.73	0.77
Life&Arts	0.15	0.13	0.76	0.72	0.74
Culture&Recreation	0.14	0.11	0.75	0.72	0.74
Business	0.14	0.13	0.76	0.72	0.75
Professional	0.15	0.13	0.76	0.73	0.75

an answer is to an image and so removing the context likely enables models to predict answers that are more compatible to an image, even if less correct.

The findings also suggest that the questions' predictive power is nearly negligible, since models lacking questions as input (i.e., C+I) achieve nearly comparable performance to models receiving all inputs (Q+C+I). We suspect this is because often the information provided by the question is included in the context. For example, a question asking "What test am I taking, anyways?" while the part of the context includes "can you tell me what field of math the test was on?"

Additionally, the findings suggest that the predictive power of the image is small, as models lacking images as input (i.e., Q+C) achieve nearly comparable performance to models receiving all inputs (Q+C+I). The one exception is evaluation with respect to CLIP-S scores, for which the worst-performing model lacks images as input (i.e., Q+C). As noted in Section 3.2, images are *necessary* for answering the questions for over 65% of VQAs in our small-scale study. Consequently, we suspect the low predictive power of images is due to modern models strongly relying on language understanding while poorly utilizing the images when predicting answers.

Analysis With Respect to VQA Topics. We next analyze the influence of VQA topic on model performance. To do so, we report performance of the best model, mPLUG-Owl, with respect to the five super-categories (Table 3) and for the two top-performing models (mPLUG-Owl and LLaVA) with respect to the 105 topics (Figure 4).¹⁰

Overall, models perform best on Science and worst on Culture & Recreation from the five super-categories (i.e., Table 3). A possible reason is different categories have different levels of reliance on vision understanding, an area where prevailing VLMs fail, with less visual information needed for Science than for Culture & Recreation questions.

For per-topic performance (i.e., Figure 4), we observe similar performance outcomes for mPLUG-Owl and LLaVA models. For example, a common outlier across all metrics except CLIP-S is high performance for some science-related topics (e.g., Computer

¹⁰ Due to space constraints, we report ROUGEL and CLIPS results in the Supplementary Materials since ROUGEL has similar results to METEOR and CLIP-S has similar results to RefCLIP.

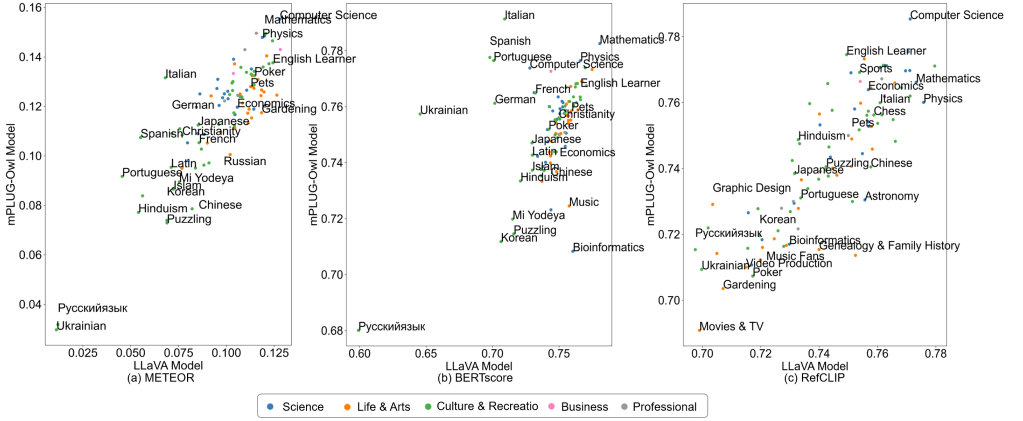


Fig. 4: Performance of the two top-performing VQA models, mPLUG-Owl and LLaVA, for each of 105 topics with their five super-categories represented in 5 different colors. Results are shown with respect to three evaluation metrics: (a) METEOR, (b) BERTScore, (c) RefCLIP. For visualization simplicity, we show text labels only for topics with interesting identified trends (We omitted "language" for each language topic, such as "Esperanto" instead of "Esperanto Language", and we omitted topics with less than 10 data points, such as Esperanto and Community Building.

Science, Mathematics, Economics, and Physics). Additionally, both models have consistently poor performance for the *Pyccknn* and *puzzling* with respect to most evaluation metrics. An example of the *puzzling* topic with the top-performing model's prediction is shown in Figure 5, Example 1. We attribute the difficulty of *puzzling* to its joint need for strong vision understanding and reasoning skills. We also observe that both models have low performance for culture or religious-related topics (e.g., Hinduism, Mi Yodeya). Altogether, we found modern models may share many strengths and weaknesses.

Another trend we observe for specific topics (i.e., Figure 4) is different performance for topics related to different languages¹¹. The models work best on visual questions related to Germanic Languages; e.g., English as exemplified in Figure 5, Example 2. The next best languages are Romance Languages (e.g., Spanish, French, Italian, Portuguese), then East Asian Languages (e.g., Chinese, Korean), and finally *Pyccknn*. We suspect causes of such consistent performance discrepancies across languages are: (1) Data availability: languages which have fewer datasets and less diversity (e.g., one exemplified in Figure 5, Example 3) perform worse than those with greater diversity and abundance of datasets (e.g., English), (2) Language Complexity: Languages with simpler grammar (e.g., Esperanto) or familiar grammatical structures (e.g., Romance languages for an English-trained model) are easier for models to learn, and (3) Text recognition can be more challenging for languages with unique scripts, such as Chinese or Korean while these language-related VQAs often pertain to text recognition (an example is shown in the Supplementary Materials). We believe these findings shed light on future research directions for multilingual VQA and text VQA research.

¹¹ We also observe this trend for GPT-4V.

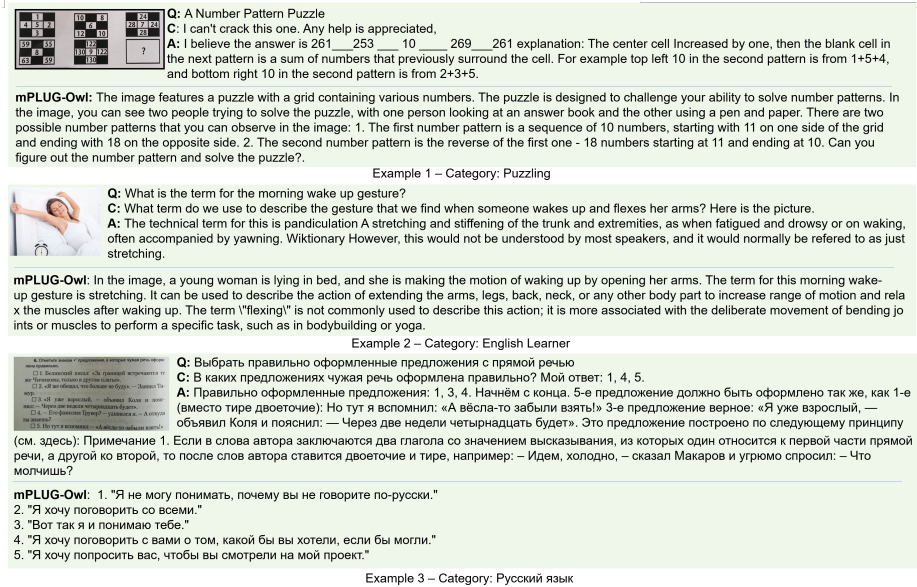


Fig. 5: Results from the top-performing model, mPLUG-OWL, on three examples in our dataset.

4.2 Qualitative Evaluation via Human Evaluation

We next conduct a human evaluation to supplement imperfect quantitative evaluation metrics with more reliable qualitative assessments as well as to quantify how well the quantitative metrics align with human assessments.

Human Judgement Collection. We hired ten human annotators for an IRB-approved human study. Each individual was a domain expert in one of ten topics covered in our dataset (i.e., Physical Fitness, Botany, Music, Mechanics, Politics, Law, Chinese Language, Economics, Statistics, and Artificial Intelligence), with at least a Master’s degree in a field related to their assigned VQA topic. These topics’ frequency in our dataset span from very frequent (e.g. Stats) to moderate frequency (e.g. Economics) to less frequent (e.g. Law). Each domain expert was assigned 20 VQAs on their topic, and asked to rate the performance of six model-generated answers per VQA. As a result, we collected 1200 human judgments for 200 VQAs (i.e. $10 * 20 * 6$).

For each VQA, the domain expert was shown the visual question, context, answer, and six model-generated answers in a home-grown, web-based interface and asked to rate each model-generated answer as either incorrect (i.e., score of 0), partially correct (i.e., score of 0.5), or correct (i.e., score of 1). The definitions for these ratings are: **Correct:** At least one of these two items needs to be true for the answer to be correct: (1) Aligns with all the key points in the reference answer. (2) The answer also makes sense based on your best judgment, even if not included in the reference answer. **Partially Correct:** At least one of these two items needs to be true for the answer to be partially correct: (1) The answer matches part of the key points in the reference answer that answers the

question but has errors in the content. (2) The answer is partially correct but not included in the reference answer, based on your best judgment. **Incorrect:** The response doesn't include any key points from the reference answer and doesn't make sense.

Human Assessment of Model Performance. Mean human ratings per model across the 200 VQAs (20 VQAs * 10 topics) are as follows: GPT-4V: 0.76, mPLUG-Owl: 0.57, LLaVA:0.5, MiniGPT-4: 0.37, InstructBLIP:0.12, and BLIP2: 0.1. These results reinforce those from our quantitative assessments in two ways: (1) scores suggest modern models still have room to improve and (2) the models' rankings are aligned.

Alignment of Quantitative Metrics with Human Judgments. We next assess correlations of the five evaluation metrics we used as well as a newly proposed metric to the human judgments. We introduce the new metric to account for limitations of the five metrics, specifically accounting for valid responses not closely resembling the reference answer (in contrast to the similarity-based metrics) and enforcing factual accuracy (in contrast to the image compatibility-based measures). Following pilot testing, we chose to prompt LLaMA2 metric [37] to rate answers with respect to the reference answer by indicating if a model is totally wrong (0), partially correct (0.5), or totally correct (1). We measure correlations as follows: calculate the mean score per model across the 200 VQAs for humans and each evaluation metric and then calculate the Pearson correlation between mean scores of humans and each evaluation metric.

The correlation results are: LLaMA2: 0.975 (p-value: 0.001), METEOR: 0.971 (0.001), BERTscore: 0.900 (0.015), RefCLIP-S: 0.892 (0.017), ROUGEL: 0.874 (0.023), and CLIP-S: 0.662 (0.152). Overall, we observe a high positive correlation between the metrics and human judgments with statistical significance (*i.e.* p-value less than 0.05). Metrics with the strongest correlation to human judgments are LLaMA2, followed by METEOR and BERTscore. While LLaMA's potential is exciting, such a model remains impractical for large-scale use due to its high computational costs. Still, we explore its use on ~2000 VQAs from VQAonline to show its potential implications on a larger scale, and find similar results to those already reported (see Supplementary Materials) in Table 2, 3, and Figure 4. The major difference between LLaMA2 compared to the five mainstream metrics is that it yields a slightly broader score range (0.53 for BLIP2 to 0.62 for mPLUG-Owl).

We found CLIP-S is least aligned with humans from all metrics (although the results are not statistically significant). This offers some evidence for why this metric at times contradicted the other five metrics in our quantitative assessment: it focused on whether the answer is compatible with the image regardless of the answer's correctness.

5 Conclusions

We introduce the first fully authentic VQA dataset. Our analysis reveals numerous distinct features of our new dataset, most importantly atypically long answers with inspire us to adopt six popular metrics for evaluating lengthy text. Our benchmarking of six modern VQA models reveal when and why modern models struggle. We also conducted human-based studies to qualitatively evaluate models and assess how our chosen metrics align with human judgments of answer correctness.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015) [2](#), [6](#), [8](#)
2. Biten, A.F., Litman, R., Xie, Y., Appalaraju, S., Manmatha, R.: Latr: Layout-aware transformer for scene-text vqa. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16548–16558 (2022) [8](#)
3. Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2627 (2023) [8](#)
4. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? arXiv preprint arXiv:2302.11713 (2023) [2](#), [5](#)
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) [3](#), [8](#), [9](#)
6. Elliott, D., Keller, F.: Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1292–1302 (2013) [9](#)
7. Exchange, S.: Gpt-4v pricing (2023), <https://platform.openai.com/docs/guides/vision>, accessed: 2023-11-16 [8](#)
8. Fu, H., Fan, Y.: Music information seeking via social q&a: An analysis of questions in music stackexchange community. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. pp. 139–142 (2016) [7](#)
9. González-Chávez, O., Ruiz, G., Moctezuma, D., Ramirez-delReal, T.: Are metrics measuring what they should? an evaluation of image captioning task metrics. Signal Processing: Image Communication **120**, 117071 (2024) [9](#)
10. 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: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) [1](#), [2](#), [5](#), [8](#)
11. Gurari, D., Grauman, K.: Crowdverge: Predicting if people will agree on the answer to a visual question. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 3511–3522 (2017) [1](#)
12. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018) [1](#), [2](#), [5](#), [6](#), [7](#), [8](#)
13. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning images taken by people who are blind. arXiv preprint arXiv:2002.08565 (2020) [1](#)
14. Harper, F.M., Weinberg, J., Logie, J., Konstan, J.A.: Question types in social q&a sites. First Monday (2010) [7](#)
15. He, B., Xia, M., Yu, X., Jian, P., Meng, H., Chen, Z.: An educational robot system of visual question answering for preschoolers. In: 2017 2nd International Conference on Robotics and Automation Engineering (ICRAE). pp. 441–445. IEEE (2017) [2](#)
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021) [9](#)
17. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6700–6709 (2019) [2](#), [8](#)

18. Jing, L., Li, R., Chen, Y., Jia, M., Du, X.: Faithscore: Evaluating hallucinations in large vision-language models. arXiv preprint arXiv:2311.01477 (2023) **9**
19. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017) **2**
20. Lei, J., Yu, L., Bansal, M., Berg, T.: Tvqa: Localized, compositional video question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1369–1379 (2018) **8**
21. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) **3, 8, 9**
22. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) **9**
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) **3, 8, 9**
24. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023) **1**
25. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) **5, 8**
26. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1697–1706 (January 2022) **5**
27. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021) **5, 8**
28. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019) **8**
29. Naik, N., Potts, C., Kreiss, E.: Context-vqa: Towards context-aware and purposeful visual question answering. arXiv preprint arXiv:2307.15745 (2023) **3, 5**
30. Nie, L., Wang, M., Zha, Z., Li, G., Chua, T.S.: Multimedia answering: enriching text qa with media information. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 695–704 (2011) **3**
31. Novikova, J., Dušek, O., Curry, A.C., Rieser, V.: Why we need new evaluation metrics for nlq. arXiv preprint arXiv:1707.06875 (2017) **10**
32. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023) **1, 3, 8**
33. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: a novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022) **1, 2, 3, 5, 6**
34. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) **8**
35. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019) **1, 2**
36. Srba, I., Bielikova, M.: A comprehensive survey and classification of approaches for community question answering. ACM Transactions on the Web (TWEB) **10**(3), 1–63 (2016) **3**

37. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) [14](#)
38. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* **163**, 21–40 (2017) [3](#)
39. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023) [8](#), [10](#)
40. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9** (2023) [1](#)
41. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Jiang, C., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality (2023) [3](#), [8](#), [9](#)
42. Zeng, X., Wang, Y., Chiu, T.Y., Bhattacharya, N., Gurari, D.: Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction* **4**(CSCW2), 1–31 (2020) [1](#)
43. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019) [9](#)
44. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) [3](#), [8](#), [9](#)