Fully Authentic Visual Question Answering Dataset from Online Communities: Supplementary Material

Chongyan Chen[1], Mengchen Liu[2], Noel Codella[2], Yunsheng Li[2], Lu Yuan[2], and Danna Gurari[1,3]

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

1 Supplementary Materials

This document supplements the main paper with more information about:

1. Dataset Collection (Supplements Section 3.1)
2. Dataset Analysis (Supplements Section 3.2)
3. Algorithm Benchmarking (Supplements Section 4.2)

I Dataset Collection

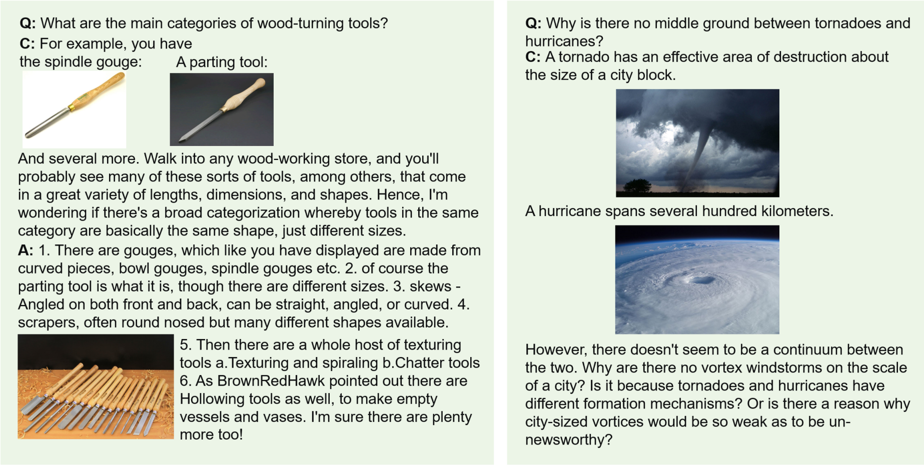
I.1 Dataset Source and Filtration

The Stack Exchange data is hosted at *https://archive.org/details/stackexchange* in XML format. We started with 330,705 candidate visual questions. After removing visual questions without an accepted answer, it resulted in 165,766 visual questions. As mentioned in the main paper, we then conducted two filtering steps. After removing visual questions with scores of 0 or less for either the question or answer, we had 119,177 visual questions. Next, after removing visual questions with multiple images, we had 85,573 visual questions. Subsequently, removing visual questions with visual answers, left 65,849 visual questions. Examples of filtered visual questions with visual answers and multiple

Table 1: Comparison of visual questions from eight existing VQA datasets and our new VQAonline dataset regarding image and question sources. Like OVEN [25], INFOSEEK sources images from nine image classification and retrieval datasets.

VQA Dataset	Which Images?	Who Asked?	From User?
Our dataset	StackExchange Users	StackExchange Users	✓
Context-VQA [42]	Six types of websites	Annotators	✗
VQAv2 [22]	MSCOCO	Crowd workers (AMT)	✗
VizWiz-VQA [23]	Captured by Blind people	Blind people	✓
OKVQA [39]	MSCOCO	Crowd workers (AMT)	✗
DocVQA [41]	UCSF Industry Documents Library	Remote workers	✗
ScienceQA [48]	Online K-12 learning platform	Online K-12 learning platform	✗
InfographicVQA [40]	Bing and Google Image Search	Annotators	✗
INFOSEEK-Wikidata [12]	9 datasets following OVEN	Template-based auto-generation	✗
INFOSEEK-Human [12]	9 datasets following OVEN	Annotators	✗

images are shown in Figures 1 (a), and (b), respectively. Finally, after removing examples for which the image could not be downloaded from the provided link, we got to our final dataset of 64,696 visual questions. For data format consistency, we converted all images to png format.



(a) (b)

Fig. 1: (a) An example of a visual question with a "visual answer". (b) An example of a visual question with multiple images. The answer is omitted to save space.

I.2 User Intention Taxonomy and Definitions

User Intention Taxonomy. We first brainstormed 11 free-form user intention categories. To do so, we solicited help from GPT-3.5 by having it indicate via a zero-shot prompt the user intentions when given a question and context. We identified an initial 11 free-form user intention categories, which are shown in the second column of Table 2. We then refined these categories based on the frequency of their occurrence in the literature on user intention (shown in the third to the last column in Table 2) for related fields such as question answering and search queries [3, 7, 10, 24, 26, 28, 38, 53] to the following 7 categories: advice, evidence, identification, instruction, opinion, reason, and verification. We then drafted the definitions for each taxonomy, mainly by adapting definitions from [3, 7, 24].

We then finalized the taxonomy with definitions via four rounds of annotator analysis on a total of 105 visual questions. Specifically, two annotators (authors) tagged the category for each visual question. Any disagreements that arose were discussed and resolved, and the definitions associated with each taxonomy category were adjusted accordingly. We observed slight agreement accuracy [1] improvement and Cohen's kappa agreement

[1] The agreement accuracy is calculated by dividing the number of agreements between two annotators by the total number of examples annotated.

improvement after each round. Specifically, from round 1 to round 4, agreement accuracy improved from 53.3% to 87.5% and Cohen's kappa agreement improved from 45.24% to 84.69%. This process culminated in our final taxonomy and their definition, which includes the following eight categories mentioned in the main paper: *advice, evidence, identification, instruction, opinion, reason, verification*, and *other*.

As shown in Table 2, our final taxonomy is different from [3,7,10,24,26,28,38,53] as they focus solely on text-based question or search queries whereas our work centers on visual questions. [4] also explores user intentions for authentic visual questions for blind individuals. However, blind people's user intents of visual questions significantly differ from that of online users' visual questions. In fact, only one of the user intents identified for blind individuals (i.e., identification) overlaps with the user intents we have identified for our dataset.

Final user intention taxonomy and definitions. We show the final user intention taxonomy and definitions as follows:

- **Verification**: These fact-checking questions aim to confirm or refute a hypothesis through affirmative answers (yes/no or multiple choice options). Since this question is based on fact, it cannot include subjective questions which should be classified as Opinion.
- **Identification**: The expected answer for these questions is a named entity or object identification. The answer is objective.
- **Reason**: These questions require answers that explain causes underlying a particular action or event. Most "why" questions fall into this category.
- **Evidence-based**: This category primarily includes questions that ask for the definition, features, description, or process of a concept, idea, object, event, or one of its attributes. It also covers translation questions of seeking the meaning of a sentence or word. If a question can be classified as Verification, it should not be classified as Evidence-based.
- **Instruction**: These questions typically involve "how to do" inquiries, seeking instructions, guidelines, or procedures for a specific action in real life.
- **Advice**: This category includes questions where the user seeks personalized guidance on a specific topic. Advice questions differ from Instruction questions in that they expect subjective recommendations, while Instruction questions seek objective, step-by-step processes for performing a specific action. Advice questions may also involve finding resources or seeking better translations of sentences/words. Additionally, this category can include questions where users are looking for ideas or comments on how to improve an existing solution. If a question can be classified as Instruction, it should not be categorized as Advice.
- **Opinion**: These questions aim to elicit subjective opinions on a topic of interest (e.g., "what do you think about" or "is X good/bad"). It might include religious questions. This category excludes Advice questions, where the focus is on the user asking the question.
- **Other**: Other.

I.3 User Intention Annotations

Hiring Crowdworkers. We hired three crowdworkers from Amazon Mechanical Turk to perform our task. For quality control, we only accepted workers located in the United States who had completed more than 500 Human Intelligence Tasks (HITs) with over a 95% approval rating.

Table 2: Eight user intent taxonomy categories: one from our dataset, one for free-form intent (derived through GPT-3.5 bootstrapping and manual labeling), and six from existing research. Our taxonomy specifically addresses user intentions for visual question answering, which is in contrast to the taxonomy from [7] which focuses on Bing query search and others [3, 19, 24, 26, 53] that are dedicated to community question answering.

Ours	free-form intent	Ignatova (2009) [26]	Harper (2010) [24]	Toba (2014) [53]	Cambazoglu(2021) [7]	Bolotova (2022) [3]	Fu (2016) [19]
Instruction	How	Procedural	Prescriptive	Proceure	Process	Instruction	Instruction
Evidence-based	Comprehend fact	General Info need	Factual	Factoid	Description	Evidence-based	Factual
Reason	Why	Causal	-	Reason	Reason	Reason	-
Verification	Validate	Verification	-	Yes/No	Verification	Debate	-
Opinion	-	Disjunctive	Disapproval/Quality	Opinion	Opinion	-	Opinion
Identification	Recognize	Concept completion	Identification	-	Entity	-	Identifying resources
Advice	Advise	-	Advice	-	Advice	Experience	Recommend/Solution
-	Prove	Quantification	-	-	Quantity	-	
-	Compare	Comparison	-	-	-	Comparison	
-	-	Definition	-	Definition	-	-	
-	Find	-	-	-	Resource	-	
-	Language	-	-	-	Language	-	
-	-	-	-	-	Temporal	-	
-	-	-	-	-	Calculation	-	
-		-	-	-	Attribute	-	
-	-	-	-	-	List	-	
-	-	-	-	-	Weather	-	
-		-	-	f -	Location	-	
	Explain	-	-	-	-	-	
-	-	-	-	-	-	-	Request research
Other	-	-	-	-	-	-	

User Intention Annotations Task Design. Crowdworkers labeled one primary intent per visual question when shown the question, image, context, and answer. For the final label, we used the majority vote label per VQA from 3 crowdsourced labels per visual question.

Annotation Task Design and Collection. We provided instructions, taxonomy definitions, and two examples per user intention category. The annotation task interface showed for each visual question the question, context, and answer.

To facilitate collecting high-quality results, we provided each crowdworker with one-on-one Zoom training. We also provided a qualifying annotation test that all workers passed to verify they understood the instructions.

To enable the assessment of high-quality results, we collected two user intention annotations from two independent workers (worker A and B) per VQA instance (i.e. image-question-context-answer) for 105 randomly selected VQAs, one from each of 105 topics. We then detected if their annotations matched. We found that 25 out of 105 didn't match. Thus, we hired a third worker (worker C) to break the tie by instructing that individual to select one of the two provided user intentions. We paid \$40 Amazon Gift Card to workers A and B and \$5 Amazon Gift Card to worker C.

Instructions for User Intention Annotation. The following are the instructions provided for workers A and B:
- Read definitions and two examples for each category.
- We will present to you 105 questions. Each question contains a question, context, and a reference answer. Please read them and categorize the primary user intention according to step 1.
- If there are any non-English questions, or if the questions contain many proper nouns that are difficult to interpret, you can use a translation tool to translate them into your native language.
- Fill in the user intent category for each question in the provided spreadsheet.

I.4 Potential Benefits of User Intention in VQA

Though we only provide an initial exploration of user intention in the main paper, we discuss the potential usage of user intention here to encourage future extensions. Overall, understanding and identifying user intent can potentially help create human-centered VQA models for better user experience. First, it facilitates the creation of datasets with better answers tailored to user needs. Second, analysis of intention prevalence in real-world scenarios—examining alongside model performance for each type can help developers in prioritizing their efforts. Third, a model with the ability to understanding user intent also can potentially provide answers more directly meeting users needs, rather than related, true responses not meeting users' needs. For instance, when asked a visual question such as "why doesn't the bulb work," users may prefer practical solutions to fix the bulb rather than just reasons and explanations. Without recognizing user intentions, models might only offer reasons.

I.5 Potential Limitations and Societal Impact of VQAonline Dataset.

Despite the VQAonline dataset's inclusion of multilingual VQAs and coverage of users' generated visual questions from various countries, we highlight that this dataset may poorly represent people in poverty-stricken areas since VQAs are posted by Stack Exchange users with access to mobile devices and Internet. Additionally, the geographic representation of Stack Exchange is unbalanced, with more from countries like US, India, Europe and less from countries like Egypt.

II Dataset Analysis

We supplement the main paper by comparing the sources of the visual questions for eight datasets in Table 1. Only our dataset and VizWiz-VQA dataset are sourced from authentic use cases with both the images and visual questions coming from real users.

III Algorithm Benchmarking

Architectures. Details about each of the six benchmarked models are provided in Table 3 and Table 4. Specifically, we report each model's image encoder, language encoder, adapter, and their training data [15, 33, 37, 44, 60, 62, 64]. Dataset sources that have data contamination are marked in red and sources suspected with data contamination are marked in blue [2].

Model Implementations. For GPT-4V, we used the gptv-2023-07-01-preview version.

For one-shot settings, we created the prompts by adjusting prompts from mPLUG's official repository and Azure's few-shot prompting examples for groundedness evaluation. The prompts we used is exemplified in Figure 2.

[2] All models have data coming from LAION400 (L400), LAION2B(L2B), or C4, which are in turn derived from Common Crawl that is known to contain random web pages (and so could have come from Stack Exchange). GPT-4V training details are not publicly-available but are believed to involve internet data.

Table 3: Details about the six benchmarked models' model configuration and training data. CC* comprises datasets from COCO [11], CC3M [52], and CC12M [9]. CC stands for Conceptual Caption [9, 52]; VG stands for Visual Genome [29]; CY stands for COYO-700M [5]; L400 stands for LAION 400M [51]; SBU [45] contains 1 million images with captions. LLaVA-I stands for 158K multimodal instruction-following data in LLaVA [37]. QA* stands for 13 question-answering datasets in InstructBLIP [15].

	Model Configuration			Img-text Data		Visual Instruction Data	
Models	Img Encoder	Lang Encoder	Adapter	Source	Size	Source	Size
BLIP2 [33]	ViT-g/14	FlanT5XL	Q-Former	CC*, VG, SBU, L400	129M	-	-
MiniGPT-4 [64]	BLIP2-VE	Vicuna 13B	FC layer	CC, SBU, L400	5M	CC+ChatGPT	3.5K
LLaVA [37]	ViT-L/14	LLaMA 13B	FC layer	CC3M	595K	LLaVA-I	158K
mPLUG-Owl [62]	ViT-L/14	LLaMA 7B	LoRA	CC*, CY, L400	204M	LLaVA-I	158K
InstructBLIP [15]	ViT-g/14	FlanT5XL	Q-Former	-	-	QA*	16M
GPT-4V [44]	Unknown	Unknown	-	Unknown	-	-	-

Table 4: Details about the six benchmarked models' language encoders and image encoders. L400, L2B, and C4 are all derived from Common Crawl. (L400=LAION-400; L2B=LAION-2B; C4=Colossal Clean Crawled Corpus).

Encoder	Source
ViT-L-14 [46]	L400 [51]
ViT-g/14 [18]	L2B [50]
FlanT5XL [13]	Finetuned based on T5 [47] while T5 was trained with C4 [47]. Finetuned with 69 Multi-task finetuning with instructions (Muffin) datasets [2, 6, 16, 34, 58, 61], 55 datasets from T0-SF [49], and 375 natural instructions v2 (NIV2) datasets [57] and 9 Chain-of-Thought (CoT) datasets derived from [1, 8, 14, 21, 27, 30, 35, 43, 56]
LLaMA [54]	CommonCrawl, C4 [47], Github, Wikipedia, Gutenberg and Books3 [20], ArXiv [31], and QAs from StackExchange's 28 largest websites.
Vicuna	Finetuned based on LLaMA [54] with user-shared conversations, so also with CommonCrawl, C4 [47], Github, Wikipedia, Gutenberg and Books3 [20], ArXiv [31], and QAs from StackExchange's 28 largest websites.

"The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

Example Task
Human: <image_example>
Human: {question_example} + {context_example}
AI: {reference_answer}"

Actual Task:
Human: <image>
Human: {question} + {context}
AI: "

Fig. 2: The prompt for one-shot setting.

As mentioned in the main paper, to complement five popular evaluation metrics, we also introduce a new evaluation metric based on LLaMA2 [55]. We tested four different prompts for it to assess the correctness of model-generated answers: prompting to output continuous scores from 0 to 1, discrete scores (0, 0.5, 1), following [63], or Azure groundness evaluation examples. We selected the best one from our preliminary analysis, which is shown in Figure 3.

System: "You are a helpful AI assistant. You will be presented with a REFERENCE ANSWER and a PREDICTED ANSWER. Your task is to rate the correctness of the PREDICTED ANSWER. Chose one of the following rating: 0 (Totally Wrong), 0.5 (Partially Correct), or 1 (Totally Correct).
Just complete the last space of the correctness score."

User: "REFERENCE ANSWER: {Reference}
PREDICTED ANSWER: {Prediction}
Score: "

Fig. 3: The prompt for LLaMA2 metric.

Subset Creation. As mentioned in the main paper, due to GPT-4V model requiring a fee for their usage [17], and due to the computational and financial cost of running LLaMA2 [3], we only evaluate GPT-4V and LLaMA2 on a subset of data.
We created this subset by randomly selecting 20 VQAs from each of the 105 topics in the test set, except for the 17 topics containing less than 20 examples where we used all

[3] As mentioned in the main paper, LLaMA2 is impractical at scale due to computing time. For example, it takes 4 days for 4 Quadro RTX 8000 GPUs to compute LLama2 score for one model on the entire VQAonline dataset. Considering we have 13 models to evaluate (six baseline models, mPLUG-Owl models with three different input types, and mPLUG-Owl models with four different one shot-settings), such evaluation would take 52 days.

Table 5: Performance of VLMs on the **subset** with 1,903 random samples from VQAonline dataset with respect to six (ROUGE-L, METEOR, BERTscore, CLIP-S, RefCLIP-S, and LLaMA2) for the zero-shot setting. As shown, GPT-4V is the best-performing model in the zero-shot setting.

Models	ROUGE-L	METEOR	BERTscore	CLIP-S	RefCLIP-S	LLaMA2
GPT-4V [44]	**0.16**	**0.12**	**0.76**	0.70	**0.75**	**0.70**
mPLUG-Owl [62]	0.14	0.10	0.75	0.71	0.74	0.62
LLaVA [37]	0.14	0.08	0.75	**0.72**	0.74	0.62
MiniGPT-4 [64]	0.13	0.08	0.74	0.69	0.73	0.59
InstructBLIP [15]	0.09	0.06	0.69	0.68	0.71	0.59
BLIP2 [33]	0.07	0.04	0.69	0.65	0.70	0.53

available VQAs. The resulting subset contains 1903 VQAs. Another reason for creating the subset with random sampling is due to the original dataset displaying a long-tail distribution across different topics.

III.1 Results on Subset.

Overall Results on Subset. Results are shown in Table 5. As mentioned in the main paper, GPT-4V achieves the best performance on the subset, based on five of the six evaluation metrics. Other models have the same rankings as that shown in the main paper.

mPLUG-Owl Model in One-shot Setting, Evaluated with LLaMA2 Metric. We explore the mPLUG-Owl model in one-shot-setting on subset, evaluated with LLaMA2 metric. Results are shown in Table 6. It strengthens our finding in the main paper that none of the one-shot examples enhance performance for mPLUG-Owl and that exemplars with matching topic tags consistently outperform randomly chosen ones.

Table 6: Performance of mPLUG-Owl on the **subset** of VQAonline dataset with respect to the LLaMA2 metric with four different one shot settings.

Models	LLaMA2
mPLUG-1shot-Random-noImg	0.609
mPLUG-1shot-MatchedTopic-noImg	0.617
mPLUG-1shot-Random	0.608
mPLUG-1shot-MatchedTopic	0.617

Anaysis With Respect to Input Types, Evaluated with LLaMA2 Metric. We then analyze the predictive power of each input of the top-performing models, mPLUG-Owl and GPT-4V on subset evaluated with LLaMA2 metric. Results are shown in Table 7. The findings from LLaMA2 metric strengthens the findings we discussed in the main paper

Table 7: Fine-grained analysis of the top-performing VQA models on the **subset** of the VQAonline dataset, GPT-4V and mPLUG-Owl, when fed different input types (Q+C+I, C+I, Q+C, Q+I), evaluated with LLaMA2 metric.

Models	LLaMA2
mPLUG (Q+C+I)	**0.62**
mPLUG (C+I)	0.61
mPLUG (Q+C)	0.60
mPLUG (Q+I)	0.58
GPT-4V (Q+C+I)	**0.70**
GPT-4V (C+I)	0.69
GPT-4V (Q+C)	0.69
GPT-4V (Q+I)	0.62

Table 8: Fine-grained analysis of the top-performing VQA models on the **subset** of the VQAonline dataset, GPT-4V and mPLUG-Owl, with respect to each of the five VQA super-category types, evaluated with LLaMA2 metric.

Models	LLaMA2
mPLUG (Science)	0.69
mPLUG (Life&Arts)	0.59
mPLUG (Culture&Recreation)	0.60
mPLUG (Business)	**0.72**
mPLUG (Professional)	0.66
GPT-4V (Science)	0.75
GPT-4V (Life&Arts)	0.67
GPT-4V (Culture&Recreation)	0.67
GPT-4V (Business)	**0.76**
GPT-4V (Professional)	0.74

that (1) best-performing model is with all information (Q+C+I), (2) the worst-performing model is the model lacking context (i.e., Q+I) for both GPT-4V and mPLUG-Owl models, and (3) the context is the most valuable information source for arriving at the target answer, while the predictive powers of questions and images are nearly negligible.

Analysis With Respect to VQA Topics. We next analyze the influence of VQA topic on the top-performing models, mPLUG-Owl and GPT-4V on the subset, with respect to the five super-categories in Table 8 and with respect to 105 topics, evaluated with LLaMA2 metric in Figure 4.
Overall, the models perform best in the Business category and worst in the Life & Arts category. Note that this does not contradict our findings in the main paper, as models still perform well in Science and poorly in Culture & Recreation. We suspect that the differences between this and the main paper's findings are due to the distribution differences (different proportion of each topic in each super-topics) of the subset compared to the entire dataset. To support our hypothesize, we also calculated METEOR, and

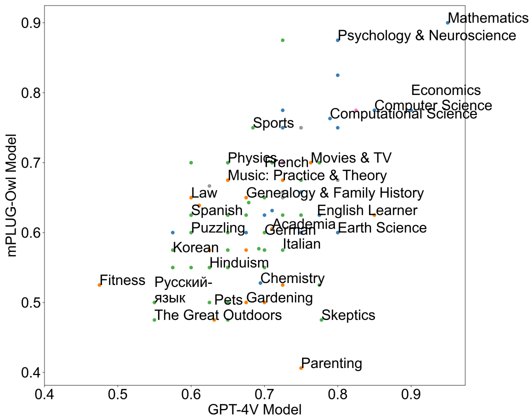


Fig. 4: Performance of mPLUG-Owl and GPT-4V on the subset of VQAonline, for each of 105 topics with their five super-categories represented in 5 different colors. Results are shown with respect to LLaMA2 metric.

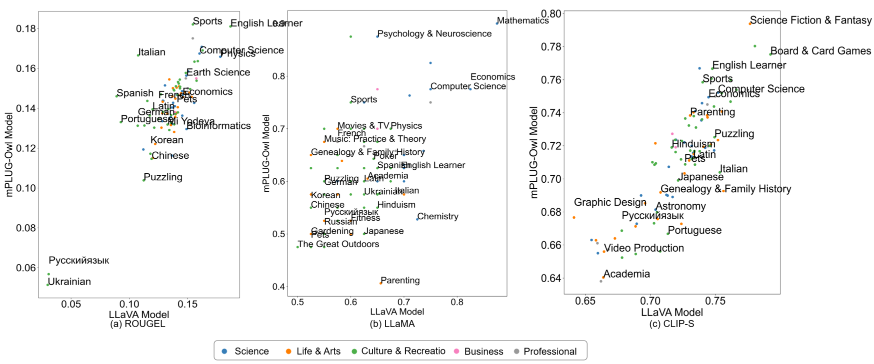


Fig. 5: Performance of mPLUG-Owl and LLaVA on the VQAonline, for each of 105 topics with their five super-categories represented in 5 different colors. Results are shown with respect to three evaluation metrics: (a) ROUGEL, (b) LLaMA2, and (c) CLIP-S. Note that we only report LLaMA2 metric on the subset. For visualization simplicity, we show text labels only for topics with identified interesting trends. We omitted "language" for each language topic and omitted topics with less than 10 data points. We also shortened some topics' names, such as "Gardening" instead of "Gardening & Landscaping", and "Fitness" instead of "Physical Fitness".

BERTscore, which are highly correlate with human judgments and observed the same trend as LLaMA2 metric on subset.
To complement the main paper, we also reported mPLUG-Owl and LLaVA across 105 topics with three other metrics in Figure 5. Overall, it strengthens our findings in the main paper regarding what topics are relatively easy (*e.g*. Mathematics, Economics, and Computer Science) and hard (*e.g*. Pyccknn, Hinduism, Puzzling). Besides, LLaMA2

indicates that some Culture & Recreation topics (e.g., "The Great Outdoors") and some Life & Arts topics (e.g., "Pets", "Gardening & Landscaping", "Physical Fitness") are relatively hard for GPT-4V, mPLUG-Owl, and LLaVA models.

Analysis With Respect to Image Necessity. We conducted a small scale model performance analysis for the Image Necessity subset, created in section “VQAonline Importance of Images". The top-performing model's results are shown in Table 9.

Analysis With Respect to Text Presence. We conducted text spotting model to determine the presence of text within the images. We utilized the state-of-the-art text spotter model DeepSolo to detect text for all images in our dataset. We found 76% (i.e., 48,161 images) contain text. The metadata is shared at: https://vqaonline.github.io/. The top-performing model, mPLUG-Owl, performs better overall for images with text, as shown in Table 9. While interesting, we want to highlight that the presence of text does not necessarily mean that questions ask about that text.

III.2 Experiments with Domain Experts

Topic Selection. We chose the 10 topics based on initial manual selection and expert availability. Initially, we manually identified 20 candidate topics for human evaluation. This selection considered diversity including using three of the seven topics from [59] as well as topics on math (i.e., stats), language (i.e., Chinese), and daily life (e.g., Gardening and Fitness). The topics' frequency span from very frequent (Stats, Gardening, and Music) to moderate frequency (Economics, AI) to less frequent (Law). Subsequently, we selected ten topics based on the availability of domain experts to represent each area. We acknowledge that these 10 topics might not fully present the entire dataset, as it is a chicken-and-egg problem: we don't know the difficulty level of the 105 topics beforehand and thus have to first conduct a human evaluation on a few topics to decide which metrics are most human-aligned for quantitatively evaluating the entire dataset.

Domain Expert Hiring. We hired ten domain experts to represent each of the ten fields (topics). To guarantee English proficiency, as most of our visual questions are written in English, we only accepted experts located in the United States.

Data	ROUGE-L	METEOR	BERTscore	CLIP-S	RefCLIP-S
Necessary	0.146	0.096	0.748	0.713	0.738
Not Necessary	0.139	0.102	0.753	0.692	0.734
With Text	0.144	0.12	0.767	0.725	0.757
Without Text	0.143	0.111	0.753	0.723	0.740

Table 9: Fine-grained analysis of the top-performing model, mPLUG-Owl with respect to five evaluation metrics when analyzing with respect to whether the image is necessary to answer the question and whether text is in the image.

Annotation Task Interface. We show a screenshot of the annotation instructions in Figure 6 and task interface in Figure 7. The link to the code for our web-based annotation tool is available at https://github.com/VQAonline/VQAonlineVisualization.

Data Collection Quality Control Mechanism. We provided a one-on-one training session with each domain expert to introduce the task and instructions. Afterwards, we gave each expert our contact information so that they could contact us with any questions about their tasks and receive feedback quickly. We also inspected the time each expert spent on each visual question as a proxy for assessing whether the hired domain experts were finishing the task by quickly selecting random options. The median and mean times are 1.4 minutes and 1.29 minutes respectively for assessing each model answer per visual question. We compensated each domain expert with a $75 Amazon Gift Card, resulting in an average hourly wage of $26/hour.

Spearman Correlation. To supplement the main paper, we also report the Spearman correlation scores for each evaluation metric: GPT-4 (1 - correlation, 0.000 - statistical significance), ROUGEL (1, 0.000), METEOR (0.943,0.005), BERTScore (0.829, 0.042), RefCLIP-S (0.771, 0.072), and CLIP-S (0.600,0.208). Spearman correlation rankings for the models follow the order from human judgments, reinforcing the main paper's findings that reference-based metrics and human judgments are highly correlated.

Correct, Partially Correct, and Incorrect Examples From Expert Evaluation. We show examples of expert annotations for the six benchmarked models in Tables 10, 11, 12, 13, 14, 15, spanning those that are correct (10), partially correct (Table 11 and Table 12), and incorrect (13, 14, 15).
One mechanics's example shows where most of the models can answer correctly for a closed-ended visual question related to recognition (Table 10).
Two economics examples show where the models partially fail for analyzing an *infographic* and requiring *specific domain knowledge* (Table 11 and Table 12). These highlight that models labeled "partially correct" can occur because of correctly answered closed-ended questions with incorrect explanations or insufficient explanations matching the reference answer. For example, GPT-4V, mPLUG-OWl, MiniGPT-4, and LLaVA answer "yes" but don't match any key points of the reason in the reference answer and have factual errors in their given reasons [4]. InstructBLIP answered "yes" without providing any reason, but simply rephrased the context. For each example, key points in the reference answer were either unsatisfactorily conveyed in the model's answer or were altogether absent.
We also show examples where models gave incorrect answers from a failure to handle the conflict in the question, context, and image. These are shown in Tables 13, 14, and 15. In Table 13 and 14, while the context and question provided incorrect information by indicating the plant has mold and asking how to remove the mold, the ground-truth answer corrects the questioner saying it's not mold but moss/algae so there is no need to remove it. All the models trusted the natural language context and question more than

[4] Of note, the errors in all models' answers to the economic VQ examples are highlighted in red by a Ph.D. student in economics.

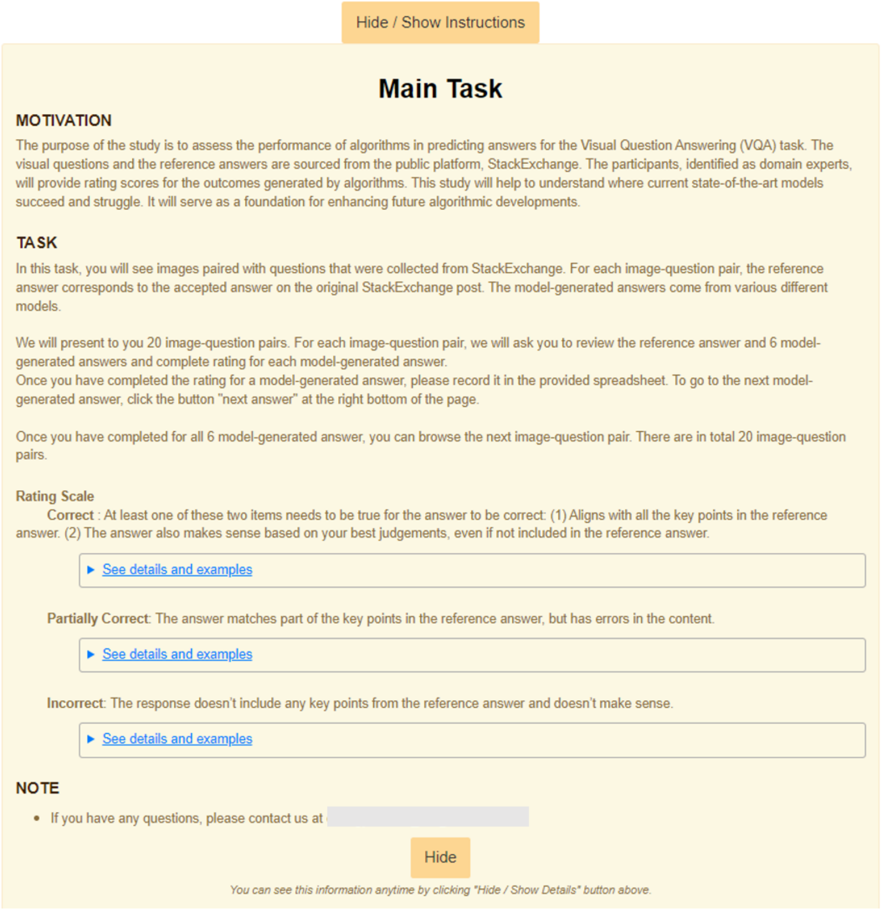
the image and then thus were misled to answer the question (partially) incorrectly[5] by offering methods to remove mold[6]. The failure to handle conflict in question, context, and image is also common in other topics (e.g., Chinese Language), as shown in Table 15. While the context provided incorrect information by indicating the characters are 心兴, the ground-truth answer corrects the questioner saying it's indeed 心头. All the models trusted the context more than the image and then thus were misled to answer the question incorrectly. This type of failure, also known as the hallucination issue, is also discussed in [32,36], where VLMs may generate descriptions or answers conflicting with the given image to align with the natural language instructions or questions, influenced by the strong language priors of large language models. We will release our dataset to facilitate further exploration of the hallucination issue in the authentic use cases.

III.3 Qualitative Examples from Different Models

We additionally provide a qualitative example from different models in Table 16, where the visual questions require significant reasoning. We think the visual question in Table 16 is interesting as algorithms like AlphaGo or MuZero can surpass humans in chess games, but the vision and language models fail in this domain. For example, MiniGPT4, InstructBlip, and BLIP2 didn't even understand the visual question. LLaVA directly copied the example provided in the context. mPLUG-Owl also directly copied the example provided in the context except omitted two moves, leading to an illegal move at round 8 (half move #15 Rh4). In contrast, GPT-4V gave an answer that on the surface-level seemed to make sense. Yet, GPT-4V violated the requirement of not capturing any pawn in the third round (half move #5 Bxf7+) as well as made an illegal move in round 11th (halfmove #21 gxf6). Of note, we verified all models' answers with https://www.apronus.com/chess/pgnviewer/.

[5] Among them, GPT-4V and LLaVA are partially correct as they each matched one key point with the reference answers: The method 2 mentioned by GPT-4V can also remove moss/algae; LLaVA mentioned the moss (though LLaVA identified as mold) might not be harmful.

[6] Of note, models' answers to the Gardening & Landscaping VQ examples are verified with a Botany Ph.D. student.


Hide / Show Instructions

Main Task

MOTIVATION

The purpose of the study is to assess the performance of algorithms in predicting answers for the Visual Question Answering (VQA) task. The visual questions and the reference answers are sourced from the public platform, StackExchange. The participants, identified as domain experts, will provide rating scores for the outcomes generated by algorithms. This study will help to understand where current state-of-the-art models succeed and struggle. It will serve as a foundation for enhancing future algorithmic developments.

TASK

In this task, you will see images paired with questions that were collected from StackExchange. For each image-question pair, the reference answer corresponds to the accepted answer on the original StackExchange post. The model-generated answers come from various different models.

We will present to you 20 image-question pairs. For each image-question pair, we will ask you to review the reference answer and 6 model-generated answers and complete rating for each model-generated answer.
Once you have completed the rating for a model-generated answer, please record it in the provided spreadsheet. To go to the next model-generated answer, click the button "next answer" at the right bottom of the page.

Once you have completed for all 6 model-generated answer, you can browse the next image-question pair. There are in total 20 image-question pairs.

Rating Scale

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 judgements, even if not included in the reference answer.

▸ See details and examples

Partially Correct: The answer matches part of the key points in the reference answer, but has errors in the content.

▸ See details and examples

Incorrect: The response doesn't include any key points from the reference answer and doesn't make sense.

▸ See details and examples

NOTE

- If you have any questions, please contact us at

Hide

You can see this information anytime by clicking "Hide / Show Details" button above.

Fig. 6: Instructions for our domain expert annotation task.

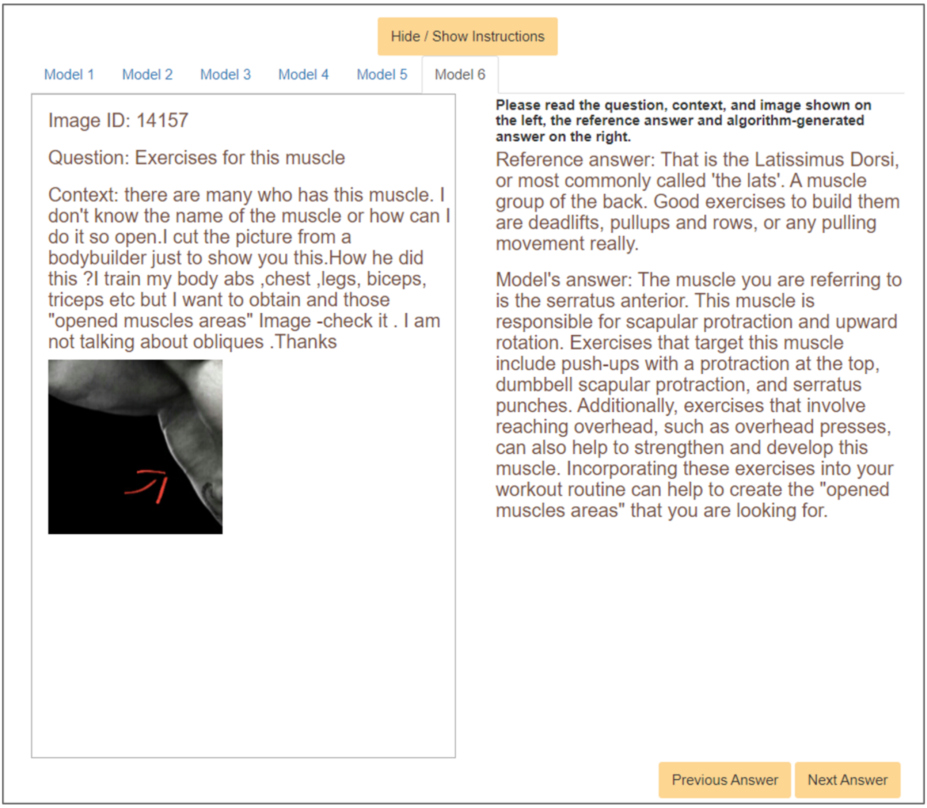


Fig. 7: Screen shot of the annotation task interface. Experts recorded their results in spreadsheets.

A visual question where most of the models can answer correctly:

Category	Mechanics 
Question	Tire sidewall damage, replace immediately?
Context	Somehow I was lucky enough to hit something when going downhill. And now my tire looks like in the picture. So should I still drive with this tire or get it replace immediately? And to replace it, is it safe to drive to the tire shop, which is a few miles away?
Reference Answer	Sidewall damage is a serious issue. Sidewall is the structural part of the tire, and damage to it can result in dangerous consequences. Replace your tire immediately. If you know how to put the spare tire on (and there is a spare tire in your car), please do it now, and drive only to the tire shop. Spare tires are not designed for extended drives, so don't start driving around in the spare tire. If you have driven more than a few miles already with the tire before looking at the damage, you may have enough data to indicate that it may survive to the tire shop. I would however prefer the spare tire approach in your case. If there's no spare tire, then that would be a difficult decision: to have your car towed with the broken tire off the ground, or to drive the car. You know more about the details in your case than I do, because you didn't indicate how many miles exactly you drove with the tire after the incident.
GPT-4V (Correct)	This looks like a significant sidewall damage and driving on this tire can be dangerous. It is recommended to replace the tire immediately. If you have a spare tire, it would be safer to use the spare and drive to the tire shop to get a new tire. If you don't have a spare, drive slowly and cautiously to the tire shop, and avoid highways and high speeds.
mPLUG-Owl (Partially correct)	Yes, it is advisable to replace the tire immediately if it has suffered significant damage. The sidewall of the tire is crucial for maintaining its structural integrity while driving. Damage to the sidewall can lead to air loss, which can increase the risk of a blowout, especially on high speeds or when making sharp turns. In such cases, driving with a damaged tire can be extremely dangerous and could result in a catastrophic accident. It is safer to replace the tire as soon as possible to ensure your safety on the road. To drive to the tire shop, it is usually safe to do so as long as the tire is not leaking air. However, it's important to check for any visible signs of air leakage while driving and to ensure that the tire pressure is maintained at the recommended level. If you notice any signs of air leakage, pull over to a safe location and stop as soon as possible to check for any further issues.
LLaVA (Correct)	Given the severe sidewall damage to the tire, it is recommended to replace it immediately. While it is technically possible to drive with the damaged tire, there are risks associated with doing so. The damage could worsen, leading to a potential blowout, which could result in loss of control, accidents, or further tire damage. Additionally, driving with a damaged tire can impact the vehicle's handling and braking capabilities. It is safer to replace the damaged tire with a new one, ensuring optimal road safety and vehicle performance. As for driving to the tire shop, it is advisable to do so if the tire shop is nearby and if there are no other more severe damages to the vehicle that require immediate attention. If in doubt, consult an expert or seek assistance from roadside services if available.
MiniGPT-4 (Correct)	It is not recommended to drive with a damaged tire, as it could result in a blowout or loss of control of your vehicle. If the tire is severely damaged, it is important to replace it immediately, as it is not safe to continue driving on it. If the tire is only slightly damaged, it may still be safe to drive on it for a short distance, but it is still recommended to have it inspected and potentially replaced. It is safe to drive to the tire shop a few miles away, but it would be wise to make sure that the tire shop is open and able to serve you before making the trip. If you cannot replace the tire immediately, you should drive as little as possible and have the tire inspected as soon as possible.
InstructBLIP (Correct)	replace immediately
BLIP2 (Incorrect)	yes, it is safe to drive to the tire shop

Table 10: An example where most of the models can answer correctly. Models' answers were verified and the errors were highlighted in red by a Mechanics Ph.D.

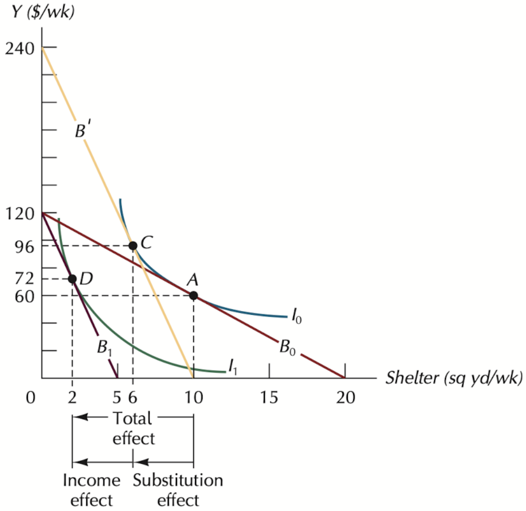
A visual question needing specific domain knowledge:	
Category	Economics
	
Question	Measuring income effect before knowing substitution effect
Context	Suppose on the x axis we have shelter and on the y axis we have composite goods. Now, if the price of shelter increases, the optimal bundle changes from point A to point D. Standard textbook tells me to draw a line parallel to the new budget line B1, which is tangent to the indifference curve I0. In this way we can get the substitution effect. The income effect follows. My question is, can we reverse the two procedures and measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference line? Thanks!
Reference Answer	You can definitely do so! There are two versions of the income effect. The one that you are referring to is called the Equivalent variation. The two ways of looking at the problem were introduced by Sir John Richard Hicks. To see the impact of a price change in monetary terms, we ask how much money should have been taken before the price change to leave a consumer at the same utility level he attains after the price change. This is the same thing as you had asked: can we reverse the two procedures and measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference line? I hope this has helped but if you wish to explore further, see the 'Consumer Surplus' chapter in Hal Varian's 'Intermediate Microeconomics'.
GPT-4V (Partially correct)	Yes, you can reverse the two procedures and measure the income effect first, but it is less intuitive and less accurate compared to measuring the substitution effect first. The reason for this is because the income effect is the difference between the old and new optimal bundles (A and D) after removing the substitution effect. Since the substitution effect is a result of changes in relative prices, in order to accurately measure the income effect, you would first need to find the new optimal bundle that would have been chosen if the relative prices had remained the same. In practice, this is difficult to do because you would have to make assumptions about how the individual's preferences would have changed, and you would need to find the new optimal bundle on the new indifference curve. It would be much easier and more accurate to simply measure the substitution effect first, and then subtract that from the total effect to find the income effect.

Table 11: A visual question that requires specific domain knowledge in Economics and infographic understanding. Models' answers are from GPT-4V. Errors are marked in red by an economics PhD student for reader convenience.

A visual question needing specific domain knowledge:	
mPLUG-Owl (Partially correct)	Yes, you can measure the income effect before knowing the substitution effect by drawing a line parallel to the original budget line and tangent to a new indifference curve. The income effect is the effect on consumption of a given good when the income of a consumer increases or decreases. To measure the income effect, we need to find the new indifference curve that represents consumption choices of consumers after their income has changed. The original budget line represents consumption choices of consumers before the change in income, while the new indifference curve represents consumption choices after the change in income. To find the new indifference curve, we can use the budget line equation: y = C(x, y) = w(y - t(x, y)) where x and y represent consumption of shelter and composite goods, respectively, and t(x, y) represents the marginal utility of composite goods. C(x, y) is the budget line, and w represents the consumer's budget. The budget line is the line of points where the total utility is equal to a given consumer's budget. To measure the income effect, we need to find the point of intersection between the original budget line and the new indifference curve, which represents consumption choices after the change in income. The new indifference curve can be found by solving the budget equation using the new income level and the new marginal utility of composite goods. To find the point of intersection between the original budget line and the new indifference curve, we need to solve the budget equation using the new income level and the new marginal utility of composite goods: w(y – t(x, y)) = y – t(x, y) Substituting the new marginal utility of composite goods into the budget equation, we get: w(y – t(x, y)) = y – t(x, y) = y – t((x1, y1), (x2, y2)) where (x1, y1) and (x2, y2) are the consumption points on the new indifference curve. We can then solve for the point of intersection between the original budget line and the new indifference curve by substituting x1 and y1 into the budget equation: w(y – t(x1, y1)) = y – t(x1, y1) = y.
LLaVA (Partially correct)	Yes, it is possible to measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference curve. This can be done by following the steps you described in your question. The income effect represents the change in consumption due to a change in income, and it can be calculated by comparing the consumer's initial position (point A) to the new optimal consumption point (point D) after the change in the price of shelter. The substitution effect, on the other hand, is the effect of changing the relative price of shelter and composite goods on the consumer's choice of consumption bundle. It can be derived from the difference between the tangent line parallel to the new budget line and the indifference curve, and the income effect.
MiniGPT-4(Partially correct)	Yes, you can certainly measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference curve. This approach is commonly used in economics to measure the income effect of a change in the price of a good or service. The income effect is the change in the quantity demanded of a good or service due to a change in income. When income increases, the income effect causes the consumer to shift the demand curve to the right, leading to an increase in the quantity demanded of the good or service. Similarly, when income decreases, the income effect causes the demand curve to shift left, resulting in a decrease in the quantity demanded. To measure the income effect using a substitution effect approach, you would draw a line parallel to the original budget line (B1) that is tangent to the new indifference curve (I0') that represents the consumer's maximum utility point after the change in the price of the good or service. The point of tangency is denoted by the letter 'P' on the new budget constraint. The ratio of the areas of the triangles PBC and PC defines the magnitude of the income effect, which can be measured in terms of the quantity demanded of the good or service. Note that this approach is not the most intuitive way to measure income effect. It's more common to draw a new budget constraint to measure the income effect. The new budget constraint after
InstructBLIP (Partially correct)	Yes, it is possible to reverse the two procedures and measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference line.
BLIP2 (Incorrect)	No, you can't measure the income effect first by drawing a line parallel to the original budget line and tangent to the new indifference line.

Table 12: A visual question that requires specific domain knowledge in Economics and infographic understanding. Models' answers are from mPLUG-owl, LLaVA, MiniGPT-4, InstructBLIP, and BLIP2. Errors are marked in red by an economics Ph.D. student for reader convenience.

A visual question when the question and the context are misleading while the image contains true information: :	
Category	Gardening & Landscaping 
Question	How do I safely eradicate mold without hurting cactus seedlings?
Context	Here are some of my seedlings, sown a little bit less than a month ago. As you can see, there's mold starting to form. I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect from contamination. What would be an easy way to get rid of the mold without hurting the seedlings? Someone suggested to use hydrogen peroxide, but I'm not sure it's safe for the seedlings. Ideas?
Reference Answer	That isn't mold, it's moss/algae. You can tell (by the green) that it's growing from energy obtained by photosynthesis, so it's not a big threat competition-wise. It's a sign that the mix is constantly moist on the surface, and also grows faster in high humidity. It will not damage the cactus seedlings, but here's a tip. When cacti germinate, they grow very slowly/remain the same size for a long time. That's because they're developing a strong root system. A sign they are established is that they will begin to grow at a faster rate. This means the root system is growing deep, and you can allow the top 1/2 of mix to dry between waterings. This will keep moss from establishing. For now, you can use a toothpick and stir up the top layer if you want, to slow the moss, as described here: Is there anything wrong with allowing moss to grow around herbs? Or you can use a pair of tweezers and carefully remove the moss. In the long run, it really doesn't matter. The moss won't be a problem after the seedlings mature a little, and you don't have to water so frequently. And by the way, those seedlings look great! I hope you have success with them.
GPT-4V (Partially correct)	To get rid of mold without hurting your cactus seedlings, try the following methods: 1. Cinnamon: Cinnamon has antifungal properties and can be used as a natural fungicide. Simply sprinkle some ground cinnamon over the affected areas, and it should help control the mold growth. 2. Baking soda: Baking soda is also effective at controlling mold growth. Mix one tablespoon of baking soda with a quart of water, and spray the solution over the affected areas. 3. Vinegar: Vinegar also has antifungal properties and can be used as a natural fungicide. Mix one part vinegar with three parts water, and spray over affected areas. Some other tips to help prevent mold growth in the future include: - Limiting the amount of moisture in the soil - Increasing air circulation around the seedlings - Avoiding overwatering - Keeping soil temperature between 70-80 degrees Fahrenheit - Keeping relative humidity between 40-60%. Don't use hydrogen peroxide as it can be harmful to the seedlings.
mPLUG-Owl (Incorrect)	It is essential to take steps to safely eradicate any mold growth on your cactus seedlings without harming them. One option is to use hydrogen peroxide. However, it is important to be cautious when using it on seedlings. Hydrogen peroxide is a strong oxidizing agent that can cause burns or damage to the cactus seedlings if used improperly. To safely use hydrogen peroxide to treat mold on the seedlings, follow these steps: 1. Use a dilute solution of hydrogen peroxide. The best dilution ratio is around 1 tablespoon of 3% hydrogen peroxide per gallon of water. 2. Pour the hydrogen peroxide solution into a shallow dish or container. 3. Soak the affected cactus seedlings in the solution for a few minutes. 4. Remove the seedlings from the solution and place them back in the pot with the Ziploc bag. 5. Seal the Ziploc bag, making sure to leave a small opening at the top to allow ventilation. By following these steps, you can ensure that the hydrogen peroxide solution will not come in direct contact with the cactus seedlings and will effectively eradicate any mold without harming the plants. It is essential to monitor the seedlings carefully and remove them from the hydrogen peroxide solution if any signs of injury or damage appear. If you are not confident in using hydrogen peroxide or have any concerns about its effectiveness, you may consider using a more gentle method, such as washing the seedlings with a gentle soap and water solution and allowing it to dry thoroughly.

Table 13: A visual question when the question and context are misleading while the image contains true information. Model answers are from GPT-4V and mPLUG-Owl, all misled by the question and context. Models' answers verified by a Ph.D. in Botany major.

A visual question when the question and context are misleading while the image contains true information:	
LLaVA (Partially correct)	To safely eradicate mold without hurting the cactus seedlings, you can follow these steps: 1. Identify the mold: Look closely at the mold's appearance and determine if it's harmful to cacti. Some molds can be beneficial, such as those that create the beautiful Cordyceps fungus, while others can be detrimental to the plant's health. 2. Remove the mold: If the mold is detrimental, gently pick it off using a clean, dry cotton swab or a soft brush. Avoid using water or wetting the cotton swab, as this could spread the mold spores and affect the seedlings. 3. Improve air circulation and light: Ensure that your seedlings are receiving adequate light and air circulation. Mold thrives in dark, damp environments, so providing proper light and ventilation can help prevent mold growth. 4. Keep the seedlings well-spaced: Overcrowding can lead to mold, as it allows for insufficient air circulation. Keep the seedlings well-spaced and avoid overwatering to maintain a suitable environment for healthy growth. 5. Use hydrogen peroxide cautiously: If you must use hydrogen peroxide, dilute it to a low concentration (1-3%), and apply it only to the affected area, avoiding the seedlings. Remember that hydrogen peroxide can be harmful to some plants, so use it with caution and dilute it properly to minimize risk. Always prioritize the health and safety of your seedlings, and avoid using harsh chemicals or treatments that could harm them. Focus on maintaining a clean and well-lit environment to encourage healthy growth and prevent mold.
MiniGPT-4 (Incorrect)	It's important to take mold in your cactus seedlings seriously and take measures to prevent its spread. While hydrogen peroxide (H2O2) is known for its antimicrobial properties, it can be harmful to plants, especially at high concentrations. I would not recommend using H2O2 to get rid of mold in your seedlings. Instead, you can use the following methods: 1. Water and sunlight: Water the affected area thoroughly and place it in direct sunlight. The sun's UV rays can help to kill the mold spores. 2. Vinegar: Mix a solution of equal parts water and white vinegar in a spray bottle, and spray the mold off the seedlings. Vinegar is a natural fungicide. 3. Bleach: Mix a solution of 1 part bleach and 9 parts water and spray it onto the seedlings. Be careful not to overdose the bleach, as it can be harmful to plants and the environment. Remember, after treating the mold, to thoroughly clean and disinfect the affected area and surrounding areas to prevent the mold from re-appearing. Also keep the humidity and temperature of your grow area in check, to prevent mold from appearing again. Please let me know if you
InstructBLIP (Incorrect)	The cactus seedlings in the image were sown a little bit less than a month ago. As you can see, there's mold starting to form. I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination. As you can see, there's mold starting to form. I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination. What would be an easy way to get rid of the mold without hurting the seedlings? Someone suggested to use hydrogen peroxide, but I'm not sure it's safe for the seedlings. Ideas?
BLIP2 (Incorrect)	I'm using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination I am using the baggie method whereby the pot with the seeds is enclosed in a Ziploc bag, to maintain high humidity and protect against contamination.

Table 14: A visual question when the question and context are misleading while the image contains true information. Model answers are from LLaVA, MiniGPT-4, InstructBLIP, and BLIP2, all misled by question and context. Models' answers verified by a Ph.D. in Botany major.

A visual question requires text recognition skills:

Category	Chinese Language

Question	What is the meaning of 心兴?
Context	What is the meaning of 心兴 in the picture below (towards bottom left in image)? I am guessing it means to think, but doesn't 想 mean to think?
Reference Answer	That character is not「兴」(to rise > excited, happy), it is「头」(head). 「心头」(literally heart + head) means mind - the picture is saying that you use your mind to think/reason.
GPT-4V (Incorrect)	心兴 is a colloquial expression in Chinese. It roughly means "to have a happy mood or personally interested in something". It's often used in the context of "心兴来了" (to suddenly become interested or happy) or "心兴去了" (to suddenly lose interest or become unhappy). In the picture, the character is saying "我心兴来了" which means "I am suddenly interested" or "I am suddenly in a good mood". As for your question about 想, 想 does mean "to think", but it can also mean "to want" or "to wish". Chinese is a contextual language, so the meaning of a word or phrase can change depending on the context it is used in.
mPLUG-Owl (Incorrect)	Yes, you are correct! The meaning of "心兴" is to be enthusiastic or have enthusiasm. It is often used to describe someone who is excited or passionate about something, such as a new project or idea. The meaning of the phrase "心兴" in the picture is to have enthusiasm for learning. It is a part of a Chinese character that is often used in children's books and learning materials, such as the colorful picture you have provided. The image shows a young girl holding a book with this character on it, indicating that she is interested in learning and has an enthusiasm for reading and education.
LLaVA (Incorrect)	In this context, "心兴" likely means "thinking" or "intelligence" in Chinese. The image, which includes pictures of children with various expressions and a book with Chinese characters, suggests that the scene involves a learning activity or a discussion about the concepts covered in the book. The presence of "心兴" in the picture implies that it is a theme or an important idea related to the content being presented.
MiniGPT-4 (Incorrect)	"心兴" (xīn xìng) in the picture is the Chinese word for "heartfelt" or "heartfelt" (a feeling of contentment or happiness). The character for "心" (xīn) means "heart" or "mind," while the character for "兴" (xìng) means "to feel happy or content." Therefore, "心兴" (xīn xìng) can be translated as "feeling contented or happy." The character 想(cāng) in Chinese means to think or to ponder, so it's not the same as "心兴" (xīn xìng) .
InstructBLIP (Incorrect)	yes
BLIP2 (Incorrect)	,,

Table 15: A visual question related to the Chinese Language and requires text recognition skills. Models' answers verified by a Chinese native speaker.

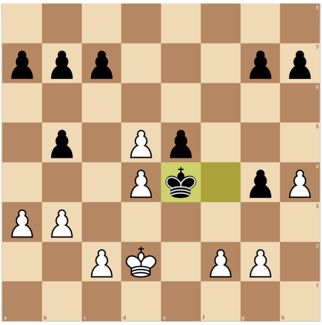
A visual question requires multi-hop reasoning process:	
Category	Chess 
Question	Devouring Nobility
Context	Provide, with the minimum number of moves, a legal chess game where all pieces are captured except the sixteen pawns and the two kings. Example of final configuration: Which is done in 21 moves: 1. e4 Nc6 2. Bb5 e5 3. Bxc6 dxc6 4. Nf3 Bg4 5. h4 Bxf3 6. Rh3 Bxd1 7. Kxd1 f5 8. Rg3 Nf6 9. Rg4 fxg4 10. d4 Ba3 11. Bg5 Qd5 12. Bxf6 Qb3 13. bxa3 Rd8 14. Bxd8 Kf7 15. axb3 Rxd8 16. Nc3 Rd5 17. exd5 Kf6 18. Nb5 cxb5 19. Kd2 Kf5 20. Re1 Kf4 21. Re4+ Kxe4 Can you beat 21 moves?
Reference Answer	13 moves: 1. e3 e6 2. Na3 Bxa3 3. bxa3 Na6 4. Bxa6 bxa6 5. Rb1 Qf6 6. Rb8 Qa1 7. Rxc8+ Ke7 8. Ne2 Qxc1 9. O-O Qxd1 10. Rxg8 Qxe2 11. Rxh8 Qxf1+ 12. Kxf1 Re8 13. Rxe8+ Kxe8 $12\frac{1}{2}$: 1. e3 e6 2. Na3 Bxa3 3. bxa3 Na6 4. Bxa6 bxa6 5. Qf3 Qf6 6. Qxa8 Qxa1 7. Qxc8+ Ke7 8. Qxg8 Qxc1+ 9. Ke2 Qxg1 10. Qxh8 Qxh1 11. Qe8+ Kxe8 12. a4 Qe1+ 13. Kxe1 Try it online!
GPT-4V (Incorrect)	Yes, it can be done in 19 moves: 1. e4 Nc6 2. Bc4 e5 3. Bxf7+ Kxf7 4. Qh5+ g6 5. Qxe5 Nxe5 6. d4 Ng4 7. h4 Nh2 8. Rxh2 d5 9. Bg5 Qxg5 10. hxg5 Bf5 11. gxf6 Bc5 12. exf5 Bxd4 13. Rxh7+ Rxh7 14. fxg6+ Kxg6 15. Nf3 Rf8 16. Nxd4 Rxf6 17. Nd2 Rh1+ 18. Nxh1 Rf4 19. Ng3 Kg5 (Note: There may be even more efficient solutions, but this is the best I was able to come up with.
mPLUG-Owl (Incorrect)	Yes, it is possible to beat the given game of chess with 19 moves. The game is in a drawish position, and it's possible to force a win by sacrificing a piece and exploiting some of the tactics present in the game. Here are the moves to beat the game: 1. e4 Nc6 2. Bb5 e5 3. Bxc6 dxc6 4. Nf3 Bg4 5. h4 Bxf3 6. Rh3 Bxd1 7. Kxd1 f5 8. Rh4 fxg4 9. d4 Ba3 10. Bg5 Qd5 11. Bxf6 Qb3 12. bxa3 Rd8 13. Bxd8 Kf7 14. axb3 Rxd8 15. Nc3 Rd5 16. exd5 Kf6 17. Nb5 cxb5 18. Kd2 Kf5 19. Re1 Kf4 (19. Re4+ Kxe4 would lead to a stalemate as white can capture the king with checkmate on the next turn. The move 19. Re1 Kf4 is therefore a better option to avoid stalemate).
LLaVA (Incorrect)	I cannot play a game of chess directly with you, but I can provide you with a solution to the problem you have presented in the image. The image shows a chess board with a total of 16 pawns and two kings, with a total of 18 pieces on the board. In order to capture all the pieces except for the two kings and the 16 pawns, the following sequence of moves can be followed: 1. e4 Nc6 2. Bb5 e5 3. Bxc6 dxc6 4. Nf3 Bg4 5. h4 Bxf3 6. Rh3 Bxd1 7. Kxd1 f5 8. Rg3 Nf6 9. Rg4 fxg4 10. d4 Ba3 11. Bg5 Qd5 12. Bxf6 Qb3 13. bxa3 Rd8 14. Bxd8 Kf7 15. axb3 Rxd8 16. Nc3 Rd5 17. exd5 Kf6 18. Nb5 cxb5 19. Kd2 Kf5 20. Re1 Kf4 21. Re4+ Kxe4 This sequence of moves will capture all the pieces except for the two kings and the 16 pawns. The solution has been found using the minimax algorithm, which is an optimal algorithm for solving two-player games such as chess.
MiniGPT-4 (Incorrect)	As an AI language model, I can certainly analyze the position you provided and try to find a continuation that may lead to a win for White, but it's important to note that I am not a chess engine and my analysis will not be as deep or as accurate as a specialized chess engine like Stockfish. Given the position you provided, I can see that Black's king is exposed and vulnerable to a quick checkmate. One move that may be try to force a win for White is 21. Rh8+. This move aims to put the king in check and force black king to move, and then exploit the exposed king with a series of moves like Rh6, Rh5, and Rh4, putting pressure on the king and threatening a checkmate. It is also possible to try to exploit the open position with a pawn storm with moves like e4 and e5, to put pressure on the black king and try to win the game. It's important to note that this is just one possible move and that a skilled chess player player could find other other other moves to try to to try to win the the game. Also, Stockfish or other chess engines will likely provide a more detailed analysis of the position and a better best continuation.
InstructBLIP (Incorrect)	Chess
BLIP2 (Incorrect)	bxa3 b .

Table 16: A visual question requires a significant reasoning process.

References

1. Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., Garg, D.: Explanations for commonsenseqa: New dataset and models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3050–3065 (2021) 6
2. Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via question rewriting. arXiv preprint arXiv:2010.04898 (2020) 6
3. Bolotova, V., Blinov, V., Scholer, F., Croft, W.B., Sanderson, M.: A. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1196–1207 (2022) 2, 3, 4
4. Brady, E., Morris, M.R., Zhong, Y., White, S., Bigham, J.P.: Visual challenges in the everyday lives of blind people. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 2117–2126 (2013) 3
5. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset (2022) 6
6. Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Cedilnik, A., Kim, K.Y.: Taskmaster-1: Toward a realistic and diverse dialog dataset. arXiv preprint arXiv:1909.05358 (2019) 6
7. Cambazoglu, B.B., Tavakoli, L., Scholer, F., Sanderson, M., Croft, B.: An intent taxonomy for questions asked in web search. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. pp. 85–94 (2021) 2, 3, 4
8. Camburu, O.M., Shillingford, B., Minervini, P., Lukasiewicz, T., Blunsom, P.: Make up your mind! adversarial generation of inconsistent natural language explanations. arXiv preprint arXiv:1910.03065 (2019) 6
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021) 6
10. Chen, L., Zhang, D., Mark, L.: Understanding user intent in community question answering. In: Proceedings of the 21st international conference on world wide web. pp. 823–828 (2012) 2, 3
11. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 6
12. 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) 1
13. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 6
14. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021) 6
15. 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) 6, 8
16. Dai, Z., Chaganty, A.T., Zhao, V.Y., Amini, A., Rashid, Q.M., Green, M., Guu, K.: Dialog inpainting: Turning documents into dialogs. In: International conference on machine learning. pp. 4558–4586. PMLR (2022) 6

17. Exchange, S.: Gpt-4v pricing (2023), https://platform.openai.com/docs/guides/vision, accessed: 2023-11-16 7
18. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023) 6
19. 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) 4
20. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020) 6
21. Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., Berant, J.: Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics **9**, 346–361 (2021) 6
22. 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
23. 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
24. Harper, F.M., Weinberg, J., Logie, J., Konstan, J.A.: Question types in social q&a sites. First Monday (2010) 2, 3, 4
25. Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. arXiv preprint arXiv:2302.11154 (2023) 1
26. Ignatova, K., Toprak, C., Bernhard, D., Gurevych, I.: Annotating question types in social q&a sites. In: Tagungsband des GSCL Symposiums 'Sprachtechnologie und eHumanities. pp. 44–49 (2009) 2, 3, 4
27. Khot, T., Clark, P., Guerquin, M., Jansen, P., Sabharwal, A.: Qasc: A dataset for question answering via sentence composition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 8082–8090 (2020) 6
28. Kofler, C., Larson, M., Hanjalic, A.: User intent in multimedia search: a survey of the state of the art and future challenges. ACM Computing Surveys (CSUR) **49**(2), 1–37 (2016) 2, 3
29. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017) 6
30. Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E., Soares, L.B., Collins, M.: Qed: A framework and dataset for explanations in question answering. Transactions of the Association for computational Linguistics **9**, 790–806 (2021) 6
31. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al.: Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems **35**, 3843–3857 (2022) 6
32. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 13
33. 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) 6, 8

34. Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al.: Competition-level code generation with alphacode. Science **378**(6624), 1092–1097 (2022) 6
35. Ling, W., Yogatama, D., Dyer, C., Blunsom, P.: Program induction by rationale generation: Learning to solve and explain algebraic word problems. arXiv preprint arXiv:1705.04146 (2017) 6
36. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023) 13
37. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 6, 8
38. Liu, Z.: Understanding and modeling user behavior in social question and answering (2015) 2, 3
39. 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) 1
40. 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) 1
41. 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) 1
42. Naik, N., Potts, C., Kreiss, E.: Context-vqa: Towards context-aware and purposeful visual question answering. arXiv preprint arXiv:2307.15745 (2023) 1
43. Onoe, Y., Zhang, M.J., Choi, E., Durrett, G.: Creak: A dataset for commonsense reasoning over entity knowledge. arXiv preprint arXiv:2109.01653 (2021) 6
44. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023) 6, 8
45. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011) 6
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 6
47. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020) 6
48. 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
49. Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A., et al.: Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207 (2021) 6
50. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 6
51. 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) 6
52. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 6
53. Toba, H., Ming, Z.Y., Adriani, M., Chua, T.S.: Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Information Sciences **261**, 101–115 (2014) 2, 3, 4
54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 6
55. 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) 7
56. Wang, C., Liang, S., Zhang, Y., Li, X., Gao, T.: Does it make sense? and why? a pilot study for sense making and explanation. arXiv preprint arXiv:1906.00363 (2019) 6
57. Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A.S., Naik, A., Stap, D., et al.: Benchmarking generalization via in-context instructions on 1,600+ language tasks. arXiv preprint arXiv:2204.07705 **2** (2022) 6
58. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021) 6
59. Xu, F., Song, Y., Iyyer, M., Choi, E.: A critical evaluation of evaluations for long-form question answering. arXiv preprint arXiv:2305.18201 (2023) 11
60. 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) 6
61. Yasunaga, M., Liang, P.: Graph-based, self-supervised program repair from diagnostic feedback. In: International Conference on Machine Learning. pp. 10799–10808. PMLR (2020) 6
62. 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) 6, 8
63. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36** (2024) 7
64. 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) 6, 8