VQA Therapy: Exploring Answer Differences by Visually Grounding Answers

Anonymous ICCV submission

Paper ID 11568

Abstract

Visual question answering is a task of predicting the answer to a question about an image. Given that different people can provide different answers to a visual question, we aim to better understand why different answers are given by leveraging answer grounding. Towards this goal, we introduce the VQA-AnswerTherapy dataset, the first dataset that visually grounds each unique answer to each visual question. We then propose two novel problems of predicting if a visual question has a single answer grounding and localizing all valid answer groundings for a visual question. We benchmark several algorithms for these novel problems to show where models succeed and struggle. The dataset, evaluation server, and leaderboard all can be found publicly at the following link: https://anonymous.com.

1. Introduction

Visual question answering (VQA) is the task of predicting the answer to a question about an image. A fundamental challenge for VQA is how to account for the fact that a visual question can have multiple natural language answers, a scenario that has been shown to be common [12]. Prior work [3] has revealed reasons for these differences, such as due to subjective or ambiguous visual questions. However, it remains unclear to what extent answer differences arise because *different visual content* in an image is being described versus because the *same visual content* is being described differently (e.g., using different language).

Our work is designed to disentangle the vision problem from other possible reasons that could lead to answer differences. To do so, we introduce the first dataset where all valid answers to each visual question are grounded, meaning where answers can be found are segmented in the image. This new dataset, which we call VQA-AnswerTherapy, consists of 5,825 visual questions from the popular VQAv2 [11] and VizWiz [13] datasets. Our analysis of this dataset helps uncover why different answers are given to a visual question. For example, we found that only 15.7% of the time, answer differences arise due to people looking at different visual evidence (answer groundings).

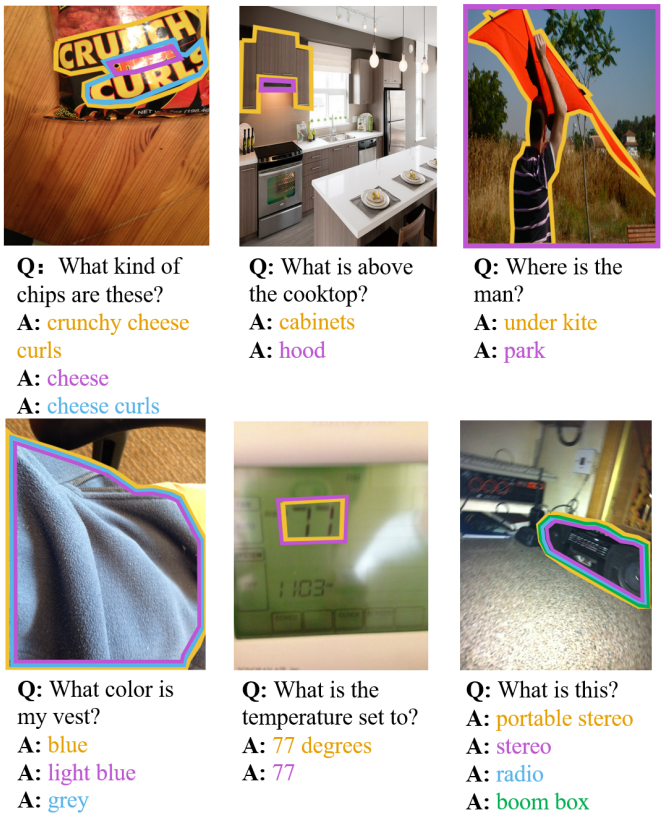

Figure 1: Examples from our new VQA-AnswerTherapy dataset. The first row shows that visual questions can have different groundings for each unique answer, which is a long-standing fact overlooked in VQA answer grounding research. The second row shows visual questions with the same grounding for all valid answers, similar to prior works [10, 16, 8, 37, 19, 15, 4]

We also introduce two novel algorithmic challenges, which are illustrated in Figure 1. First, we introduce the **Single Answer Grounding Challenge**, which entails predicting for a visual question whether all valid answers will describe the same grounding or not. Next, the **Grounding**

Answer(s) Challenge entails localizing the answer groundings for all valid answers to a visual question. We benchmarked models for these novel tasks to demonstrate the baseline performance for modern architectures and to highlight where they are succeeding and struggling.

We offer this work as a valuable foundation for improving our understanding and handling of **answer differences**. It can benefit the many communities interested in VQA applications, including individuals with visual impairments [13], students in educational environments [14], and patients visiting medical doctors [1]. We also expect this work to generalize in informing how to account for **annotator differences** for other related tasks like image captioning and visual dialog and improve efficiency in crowdsourcing. To facilitate future extensions of this work and foster a sense of community, we publicly-share our dataset and a public evaluation server with leaderboard for our dataset challenges at the following link: https://anonymous.com.

2. Related Work

Answer Differences in VQA Datasets. While many datasets have been created to support the development of VQA algorithms [26, 11, 13], a long-standing challenge has been how to account for the common situation that, for many visual questions, different answers are observed from different people [12]. Prior work has offered initial steps. For example, prior work characterized when [12], to what extent [31], and why answers differ in mainstream VQA datasets (e.g., for visual questions that are difficult, ambiguous, or subjective as well as answers that are synonymous) [3]. Other work introduced ways to evaluate VQA models that acknowledge there can be multiple valid answers, whether provided explicitly from different people [2, 20] or augmented automatically from NLP tools to capture plausible, semantically related answers [20]. Another work focused on rewriting visual questions to remove ambiguity regarding what are valid answer(s) [27]. Complementing prior work, we explore answer differences in the VQA task from the perspective of grounding, specifically exploring whether different answers arise because *different visual content* in an image is being described.

Answer Grounding Datasets. Numerous datasets have been proposed to support developing models that locate the visual evidence humans rely on to answer visual questions. This has been motivated by observations that answer groundings can serve as a valuable foundation for debugging VQA models, providing explanations for VQA model predictions, protecting user privacy by enabling obfuscation of irrelevant content in images, and facilitating search behaviors by identifying any relevant visual content in images. A commonality of prior work [10, 22, 16, 8, 5, 37, 19, 15, 16, 22, 4, 6, 30] is that only one answer grounding for one selected answer is provided for each visual question. Our work, in contrast, acknowledges that a visual question can have multiple valid answers and so multiple valid answer groundings. We introduce the first dataset where all valid answers to each visual question are grounded. This new dataset, which we call VQA-AnswerTherapy, enables us to introduce two novel tasks of predicting for a given visual question whether all answers will be based on the same visual evidence and predicting for a visual question the groundings for all valid answers.

Automated VQA Methods. Modern automated VQA models typically only return a single answer; e.g., the predicted answer with the highest probability from a softmax output layer of a neural network. Yet, people often ask visual questions that lead to multiple valid answers [12]. To account for this practical reality, we propose novel tasks and introduce the first models for sharing richer information with end users by (1) indicating when there are multiple plausible answer groundings to a visual question and (2) locating those grounded regions in images.

3. VQA-AnswerTherapy Dataset

We now introduce our new dataset that includes groundings for all valid answers to each visual question. We call this dataset VQA-AnswerTherapy.

3.1. Dataset Creation

VQA Source. Our work builds upon two popular VQA datasets that reflect two distinct scenarios: VizWiz-VQA [13] and VQAv2 [11]. The images and questions of the VizWiz-VQA dataset come from visually impaired people who shared them in authentic use cases where they were trying to obtain assistance in understanding their visual surroundings. In contrast, the images and questions of the VQAv2 dataset come from different sources: while the images come from the MS COCO dataset [7] (and so were collected from the Internet), the questions were generated by crowd workers. Despite these differences, these datasets have in common that they both include for each image-question pair 10 crowdsourced answers, each of which was curated based on the same crowdsourcing interface.

VQA Filtering. Our goal is to unambiguously ground each answer for *visual questions that have more than one valid answer*. To focus on these visual questions of interest, we applied filters to the original VQA sources, which consist of 32,842 image-question pairs for VizWiz-VQA and 443,757 for VQAv2 training dataset. First, we removed answers indicating a visual question is unanswerable (i.e., "unsuitable" or "unanswerable"). Then, we only focused on the remaining visual questions that have two or more valid natural language answers, where we define valid answers

as those for which at least two out of the ten crowdworkers gave the exact same answer (i.e., using string matching). [1]

Similar to prior work [4], we also filter visual questions that embed multiple sub-questions. An example is "How big is my TV and what is on the screen, and what is the model number, and what brand is it?" Following [4], we first removed visual questions that contain more than five words that contain the word "and". We also trimmed visual questions which contained a repetition of a question down to a single question, e.g., from "what is this? what is this?" to "what is this?". We finally filtered visual questions that were flagged as "containing more than one question" in metadata provided by [4].

We selected 27,741 visual questions with 60,526 unique answers as candidates for our new dataset. Included are all visual questions from VizWiz-VQA that met the criteria we selected for; i.e., 9,528 visual questions (with 20,930 unique answers). We also sampled a similar amount from VQAv2's training set; i.e., 18,213 visual questions (with 39,596 unique answers). Our samples included all overlapping visual questions with prior work [3], which provides reasons why answers to each visual question differ, to support downstream analysis. The remaining visual questions were sampled randomly. Statistics for each step of filtration are summarized in the supplementary material.

Answer Grounding Task Design. We designed a user interface to ground the different answers for each visual question. It presents the image-question pair alongside one of its associated answers at a time.

For each answer, two questions were asked to ensure the answer could be unambiguously grounded to one region. First, a worker had to indicate if a given answer is correct or not. If correct, then the worker had to specify how many polygons must be drawn to ground the answer from the following options: zero, one, or more than one. Only when exactly one polygon was needed to ground answer was the worker instructed to ground the answer. This ensured answers could be grounded while avoiding the situation when there are multiple polygons (e.g., "How much money is there?" for an image showing multiple coins)

To ground an answer, a worker was instructed to click a series of points on the image to create a connected polygon. After one answer grounding was generated for a visual question, the annotator could then choose for a new answer to select a previously drawn polygon or draw a new polygon. Guidance was provided for completing the task, including for many challenging annotation scenarios (e.g., objects with holes or complex boundaries).

[1] We follow the status quo established by prior work [4, 12] to obtain valid answers with ESM to provide an upper bound for expected differences). Around 36% of visual questions in VizWiz and VQAv2 datasets have more than one valid answer with ESM. More sophisticated NLP methods could be interesting to explore.

Figure 2: High-quality annotations from our dataset. Visual questions related to *text recognition* often have multiple answer groundings while *recognizing color* is often associated with a *single grounding*. Different visual evidence to answers is more prevalent in the real-world (22% for VizWiz) than the internet-gathered dataset (7% for VQAv2).

Answer Grounding Annotation Collection. We hired crowd workers from Amazon Mechanical Turk to generate answer groundings, given their on-demand availability. Like prior work [4], we only accepted workers from the United States who had completed at least 500 Human Intelligence Tasks (HITs) with over a 95% acceptance rate. For each candidate worker, we provided a one-on-one zoom training on our task. We then provided a qualifying annotation test to verify workers understood the instructions, and only accepted workers who passed this test.

For annotation of our VQAs, we collected two answer groundings per image-question-answer triplet to enable examination of whether the annotations match and so are likely unambiguous, high-quality results. To support the ongoing collection of high-quality results, we also conducted both manual and automated quality control mechanisms. We summarize these in the Supplementary Materials.

Ground Truth Generation. We next analyzed the two sets of annotations collected for each of the 27,741 visual questions in order to establish ground truths for our dataset.

First, we filtered answers and visual questions from our dataset based on our examination of each worker's annotations. Second, we removed any answers from our dataset if at least one person flagged an answer in the preliminary two questions as "incorrect". Additionally, we filtered any visual questions that had answers referring to no polygon since that means there is no answer grounding. Next, we follow the precedence from [4] to filter any visual questions that had answers referring to multiple polygons, it is driven

by the observation that visual questions requiring multiple polygons are rare in the VizWiz-VQA dataset. Also, multiple polygons prevented us from identifying whether all valid answers to a visual question have the same grounding and complicated the analysis of the relationship between different groundings. This left us with 12,290 visual questions and 26,682 unique image-question-answer triplets.

Our next phase of filtering focused on examining the similarity of groundings collected from two different annotators. For each answer, we calculated the intersection-over-union (IoU) between the two answer groundings. If the IoU was large (i.e., equal to or larger than 75%), we used the larger of the two groundings as ground truth since often the smaller one is contained in the larger one. Otherwise, we deemed that answer has an ambiguous grounding and so removed the answer from our dataset.[2]

We only kept visual questions with two or more unique answers in our final dataset. In total, our final dataset includes 5,825 visual questions with 12,511 unique visual-question-answer triplets. This includes 7,426 answer groundings for 3,442 visual questions from VizWiz-VQA dataset and 5,085 answer groundings for 2,383 visual questions from VQAv2 dataset. Examples of high-quality answer grounding results are shown in Figure 2, where answers to a visual question can either have different groundings (e.g., "What is over the elephant" and "What does this logo say") or a shared grounding region (e.g., "shirt's color").

3.2. Dataset Analysis

We now characterize our dataset of 5,825 visual questions with 12,511 visual-question-answer-grounding sets.

Prevalence of different visual evidence versus one visual region . We first explore how often visual questions have different answers that all describe the same visual evidence versus different visual evidence. To do so, we flag a visual question as having different answers describing *one visual region* if an answer grounding pair has an IoU score larger than 0.9. This stringent IoU threshold is motivated by the fact that during the annotation process, workers could indicate that a grounding for an answer is identical to a previously drawn answer grounding. Details are discussed in the supplementary material.

We found 15.7% (i.e., 916/5,825) of visual questions describe *different visual evidence*. Consequently, for the majority of visual questions, people used different natural language to describe a *single visual region*. This imbalance has resulted in the long-standing neglect of the possibility that different answers might refer to different visual evidence in VQA answer grounding research [10, 22, 16, 8, 5, 37, 19, 15, 10, 16, 22, 4, 6, 30] where in these datasets, only one visual region for one selected answer is provided for each visual question. Thus, existing models might struggle with these 15.7% questions, which is confirmed by our subsequent algorithm experiments. We propose that future work should prioritize the development of new evaluation metrics as well as more robust VQA algorithms capable of attending to different visual evidence and providing different answers.

We found that this trend of answers describing different visual content is more prevalent for visual questions coming from VizWiz than VQAv2, accounting for 22% (i.e., 761/3,442) versus 7% (i.e., 155/2,383) of visual questions respectively. This indicates a noteworthy difference between real-world and internet-gathered visual questions regarding visual evidence diversity. This difference highlights the necessity of developing models that cater to the requirements of real-world users, such as facilitating follow-up clarification to specify the referenced visual evidence, to enable effective human-computer communication. Our subsequent analysis sheds light on the reason behind this disparity: VizWiz-VQA has a higher proportion of answers that refer to text, which often leads to different regions, as demonstrated in Figure 2 and Figure 5.

We next identify the most common questions for visual questions that have *multiple* as well as a *single* answer grounding. To do so, we tally how often each question leads to different answer groundings as well as to a single answer grounding respectively. Results are shown in Table 1. We observe *recognizing objects* is common for both scenarios whereas *recognizing text* is uniquely more prevalent for visual questions with *multiple answer groundings* while *recognizing color* is uniquely more prevalent for visual questions with a *single grounding*. Another observation is that questions related to location often lead to different answer groundings, as shown in Table 1 (Top-3 "Where is the pizza") and exemplified in Figure 1 ("where is the man") and Figure 2 ("What is over the elephant"). These findings suggest model can consider what different vision skills are needed for visual questions when predicting whether a visual question has a single region for all valid answers.

Reasons for Visual Questions Having Different Answer Groundings. We next analyze the 916 visual questions that have more than one answer grounding. For each visual question, we flag which relationship types arise between every possible answer grounding pair from the following options: disjoint, equal, contained, and intersected. We categorize an answer pair as *disjoint* when IoU equals 0, *equal* when the value is larger than 0.9, *contained* when one region is part of the other region, such that the size of their intersection is equal to the minimum of their sizes and the size of their union is equal to the maximum of their sizes, and *intersected* when $0.9 \geq$IoU >0 and they do not have a contained relationship. We provide figures to exemplify

[2]We provide fine-grained analysis as to why annotator differences were observed in the Supplementary Materials.

		All	VQAv2	VizWiz-VQA
Different	Top-1	What is this?	What is the man wearing?	What is this?
	Top-2	What is in this box?	What is on the table?	What is in this box?
	Top-3	What does this say?	Where is the pizza?	What does this say?
	Top-4	What is it?	What does the street sign say?	What is it?
	Top-5	What kind of coffee is this?	What does the sign say?	What kind of coffee is this?
Same	Top-1	What is this?	What color is the train?	What is this?
	Top-2	What color is this?	What color is the cat?	What color is this?
	Top-3	What is it?	What is the man holding?	What is it?
	Top-4	What's this?	What room is this?	What color is this shirt?
	Top-5	What color is this shirt?	What color is the bus?	What color is this shirt?

Table 1: The five most common questions that lead to *different* answer groundings and *same* answer groundings for different answers for all visual questions in VQAv2, VizWiz-VQA, respectively. [3]

	All	VQAv2	VizWiz-VQA
1	89% (812)	86% (133)	89% (679)
2	11% (103)	14% (22)	11% (81)
3	0% (1)	0% (0)	0% (1)

Table 2: Number of different kinds of relationships that one visual has with respect to our dataset and each image source.

	All	VQAv2	VizWiz-VQA
Disjoint	10% (99)	16% (28)	8% (71)
Intersected	67% (685)	60% (107)	68% (578)
Contained	15% (151)	12% (21)	15% (130)
Equal	8% (86)	12% (21)	8% (65)

Table 3: Percentage of visual questions with multiple answer groundings having each relationship type between its answer groundings, overall and for each dataset source.

these different relationships in supplementary materials.

We first tally how many relationship types each visual question exhibits between its different answer grounding pairs, overall as well as with respect to each VQA source. Results are shown in Table 2. We find that most visual questions (i.e., 89%) have just one relationship type between their answer groundings. We suspect it is because most of the visual questions only have two valid answers, two answer groundings, and thus one kind of relationship. When comparing results from the two VQA sources, we observe VQAv2 has slightly more relationships than VizWiz-VQA dataset. We suspect this is due to a more even percentage distribution across the four types of relationships we analyzed, as shown in Table 3.

We next tally how many visual questions have each type of relationship, overall as well as with respect to each VQA source. Results are shown in Table 3. The most common relationship between answer groundings for a visual question is intersection, with this occurring for over half of the visual questions. This finding has important implications for both human visual perception and model development. Specifically, it suggests that when multiple individuals provide different answers based on distinct visual evidence, they may be focusing on *salient objects* while paying varying levels of attention to other *details*, resulting in an intersection of visual evidence. This insight could be valuable in designing models that can consider these distinct attention focuses (i.e., *salient object detection* and *attributions grounding*) when predicting answer grounding for different answers.

Overall, we find VizWiz-VQA and VQAv2 have a similar distribution of answer grounding relationships. For future work, it would be interesting to explore whether this pattern may hold true for other VQA datasets as well.

Relationship Between Why Answers Differ and Number of Answer Groundings. We next analyze the tendency visual questions that lead to a single versus multiple answer groundings to be associated with various reasons why natural language answers can differ. For each visual question, we obtain the reasons why answers can differ using the following seven labels provided in the VQA-Answer-Difference dataset [3]: low-quality image (LQI), insufficient visual evidence (INV), difficult questions (DFF), ambiguous questions (AMB), subjective questions (SUB), synonymous answers (SYN), and varying levels of answer granularity (GRN).[4,5] Results are shown in a bar chart in Figure 3, with the left part showing percentages for visual questions that have multiple answer groundings and the right part showing percentages for visual questions with a single answer grounding.

Overall, the visual questions with different answer groundings commonly are associated with different levels of detail or specialization granularity (GRN), ambiguous

[4]We exclude the reasons "Spam answer" and "Invalid question" from the original 9 reasons because they rarely occur in our VQA-AnswerTherapy dataset due to our filtration steps; i.e., spam is observed for 23 visual questions and invalid questions for 47 visual questions.

[5]As done in [3], we assign labels using a 2-person threshold.

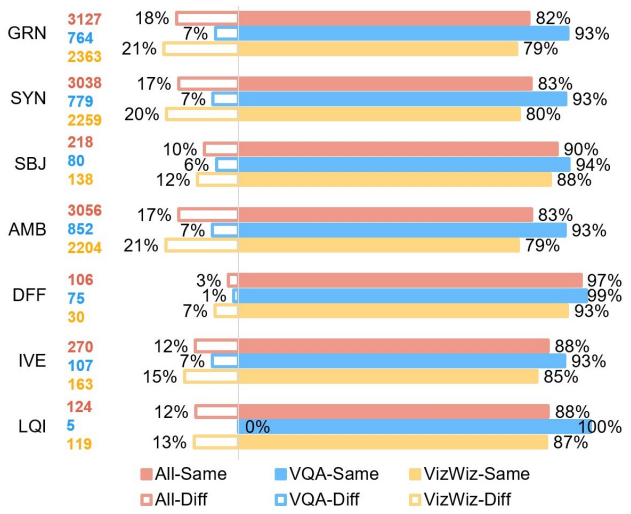

Figure 3: Relationship of whether a visual question has the same grounding for all answers and reasons for different answers for the VQAv2 and VizWiz dataset sources.

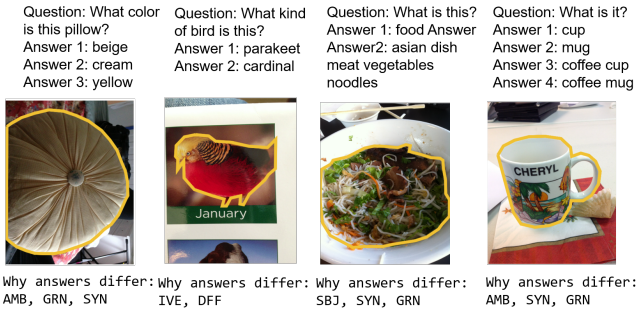

Figure 4: Visual questions that lead to one visual region.

questions (AMB), and synonymous answers (SYN). Note that in the VQA-Answer-Difference dataset, if a VQA is labeled as SYN, it occurs with AMB in over 85% of cases in both VizWiz and VQAv2 datasets, which leads to nearly identical results between AMB and SYN.

Visual questions labeled with difficult (DFF) tend to have the same grounding. Intuitively, this makes sense as there is consensus around what the question is asking about but people simply struggle to know what is the correct answer. An example of this scenario is shown in Figure 4, with the question "what kind of bird is this?"

When comparing results from the two VQA sources, we observe VQAv2 dataset and VizWiz-VQA dataset have large differences (larger than 10%) for reasons between GRN, SYN, AMB, and LQI images. We suspect that visual questions in VQAv2 tend to have a much larger same-to-difference ratio for visual questions labeled as GRN, SYN, or AMB because (1) GRN, SYN or AMB questions could lead to same grounding region if the visual questions are related to the object recognition, shown in Figure 4 (col 3 and

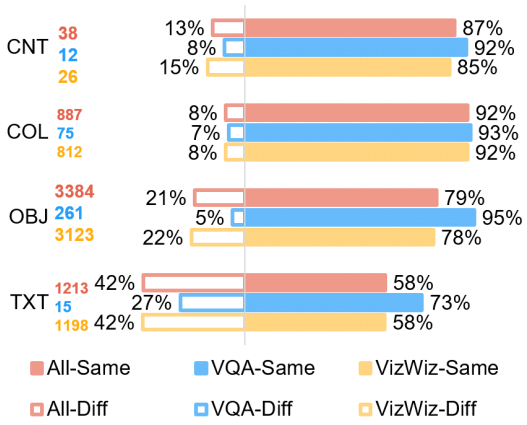

Figure 5: Amount of different answer groundings per visual question for four vision skills, overall and per dataset source (VQAv2 and VizWiz-VQA).

Img Sources	Skills	Relationship per skill Percentage% (actual number)			
		Disjoint	Intersected	Contained	Same
Overall	TXT	9% (50)	71% (403)	12% (69)	7% (42)
	OBJ	8% (62)	70% (538)	15% (117)	7% (54)
	COL	5% (4)	71% (54)	14% (11)	9% (7)
	CNT	0% (0)	83% (5)	0% (0)	17% (1)
VQAv2	TXT	0% (0)	80% (4)	0% (0)	20% (1)
	OBJ	8% (1)	92% (12)	0% (0)	0% (0)
	COL	17% (1)	67% (4)	0% (0)	17% (1)
	CNT	0% (0)	100% (1)	0% (0)	0% (0)
VizWiz-VQA	TXT	9% (50)	71% (399)	12% (69)	7% (41)
	OBJ	8% (61)	69% (526)	15% (117)	7% (5)
	COL	4% (3)	71% (50)	16% (11)	9% (6)
	CNT	0% (0)	80% (4)	0% (0)	20% (1)

Figure 6: The heatmap table shows the percentage and the number of relationships between answer groundings with respect to four vision skills for our dataset (overall) and for each image source (VQAv2 and VizWiz-VQA).

4) (2) For VQAv2, when the visual question is related to object recognition, it has a much larger same-to-difference grounding ratio, shown in Figure 5.

Relationships Between Vision Skills Needed to Answer a Visual Question and Number of Answer Groundings. We next evaluate whether a visual question has the same grounding with respect to the vision skills needed to answer a visual question. We determine if a visual question has the same grounding following the same process of the previous analysis. Labels for the four vision skills (i.e., object recognition (OBJ), text recognition (TXT), color recognition (COL), and counting (CNT)) are provided in the VizWiz-VQA-Skills dataset [34]. Following [34] we use a threshold of larger than 2 out of 5 people agreements to

determine the vision skills' labels. The chart is generated based on the percentage for each VQA source. We perform our analysis over all visual questions in our dataset as well as with respect to each VQA source independently. Results are shown in Figure 5.

Overall, we found that visual questions trying to read *text* tend to have different groundings. One common example stems from visual questions about products; e.g., as exemplified in Figure 1 (e.g., chips product). More examples are provided in supplementary materials. In contrast, questions related to recognizing *color* tend to have the same groundings. We suspect people might express 'color' in different ways because of individual or cultural differences, despite often looking at the same region. For example, a question asking "What is the color of this cloth?" might get different of answers "khaki", "tan", and "brown" despite all referring to the same region (i.e., the cloth).

We also evaluate relationships between visual questions that result in different answer groundings with respect to four vision skills for our dataset overall as well as with respect to each VQA source respectively. We determine if a visual question has the same grounding and what skills are needed following the same process of the previous analysis. The results for VQs from VQAv2 might not be representative because only a few of VQs from our dataset have labels overlapping with [34]. Results are shown in Figure 6. Overall, we observe that visual questions related to "text recognition" and "object recognition" are more likely to have a "disjoint" relationship compared to "color recognition" and "counting" skills. Examples are shown in Figure 1 ("cabinets" and "hood" are disjoint), Figure 2 ("blanket" and "umbrella" are disjoint; "rsb" and "royal society for blind" are disjoint), and in supplementary materials.

4. Algorithm Benchmarking

Using the VQA-AnswerTherapy dataset, we now quantify how well modern architectures support two novel tasks of (1) predicting if a visual question shares the same groundings for all unique answers and (2) localizing all answer groundings for all answers to a visual question.

Dataset Splits. Our VQA-AnswerTherapy dataset contains 3,794, 646, and 1,385 for train/val/test sets, respectively. The visual questions from the VizWiz-VQA dataset are split to match the train/val/test splits of the original VizWiz-VQA dataset [13]. Our dataset also has visual questions originating from the training set of the VQAv2 dataset [11], which is split into train/val/test splits using 70%, 10%, and 20% of the data respectively.

4.1. Single Answer Grounding Challenge

The task is to predict if a visual question will result in answers that all share the same grounding. This task can be used to enhance the efficiency of crowdsourcing system and improve human-computer communication (details are discussed in the Supplementary Materials). We evaluate methods using precision, recall, and f1-score metrics.

GD	Model	Precison	Recall	F1-Score
Same	ViLT	0.86	0.87	0.87
	Naïve (S)	0.74	1.0	0.85
Different	ViLT	0.62	0.6	0.61
	Naïve (D)	0.26	1.0	0.41

Table 4: Performance comparison of baseline VQA methods when evaluated on the VizWiz-Grounding-Difference (VizWiz-GD) test set for whether a VQ will result in answers that all share 'same' or 'different' grounding. (GD = grounding difference, S = same, D = different)

Baseline Models. We benchmark two models. We finetune the recently published ViLT [18] model for the VQA task on the training set on VizWiz-GD and VQAv2-GD datasets. To do so, we modified the output layer of the architecture with a two-class softmax activation to support binary classification. We also benchmark a naïve baseline that always predicts one label, i.e., all samples share the same grounding or all samples share different grounding.

Results. Results for VizWiz-GD are reported in Table 4 and results for VQA-GD are in the supplementary materials due to space constraints. Overall, we find that it is more difficult to predict if the answers share different groundings than the same grounding.[6] We suspect part for what makes this task challenging to learn is the large imbalance in labels.

4.2. Answer(s) Grounding Challenge

Baseline Models. Inspired by the winner of the VizWiz-VQA-Grounding Challenge [4], we use SeqTR [36], a segmentation model, as our baseline.

Evaluation Metric. We employ IoU to measure the similarity of each binary segmentation to the ground truth. We report two IoU scores, mIoU and IoU-PQ, where IoU-PQ calculates the IoU scores for all answer groundings within one visual question and averages them over the number of answers to that question. The mIoU score is used for overall model performance and the IoU-PQ score is used for analyzing algorithms with respect to worker alignment, vision skills, and same/different groundings.

Baseline Models. Many algorithms are proposed to locate the region that the answer is referring to, or the visual

[6] It is worth noting that we observe poor performance for the ViLT model despite that during training it 'cheated' by observing the COCO images that are used in the VQAv2 dataset. This further reinforces the challenge of learning features that are effective for our novel task.

Models	All	VQAv2	VizWiz-VQA
SeqTR (I+Q+A)	50.93	45.71	53.83
SeqTR (I+Q)	44.68	42.09	46.11
SeqTR (I+A)	50.29	46.47	52.41

Table 5: IoU performance of four models on our dataset.

Models	TXT	OBJ	COL	CNT
All	34.19	51.21	67.21	46.47
VQAv2	39.16	45.64	59.36	55.31
VizWiz-VQA	33.97	52.95	69.87	28.76

Table 6: IoU scores with respect to four vision skills overall and dataset source.

Models	Same	Diff
All	56.73	24.46
VQAv2	46.67	32.77
VizWiz-VQA	63.72	23.09

Table 7: IoU scores for visual questions with respect to same or different groundings. Current state-of-the-art model lacks the ability to predict grounding for different answers, especially when answers are referring to different regions.

evidence for a visual question. Some predict the answer grounding by showing where the model is attending to or in bounding box format [29, 35, 28]. However, the state-of-the-art visual grounding algorithms do not work well on the VizWiz-Answer-Grounding Challenge [4]. Referring segmentation [36, 9, 25] is a well-developed challenge that naturally aligns with the answer grounding task: we can treat the answer as the phrase/referring expression to predict the region that the answer is referring to. Thus, we benchmark the state-of-the-art referring segmentation model, SeqTR [36] on our dataset to examine to what extent they succeed in correctly grounding for each answer. Given an image and a phrase (question, answer, or a question-answer pair), SeqTR predicts the region that the answer is referring to.

Overall Results. We evaluated three models based on SeqTR model pretrained on a large corpus of datasets (i.e., [19, 32, 21, 17, 24, 23]). If the input is image-question-answer pair (SeqTR(I+Q+A)), the grounding region is served as visual evidence that supports the answer. For example, it can be used to accelerate browsing and searching behavior. If the input is image-question pairs (SeqTR(I+Q)), the grounding region is served as the visual evidence that leads to the answer, which can be used to explain why the model predicts a certain answer. If the input is the answer and image (SeqTR(I+A), it represents the performance of the phrase grounding model on our dataset. The results are shown in Table 5. Not surprised, model taking more information as input (both question and answer) can predict best overall. Since the model we evaluate is a model designed for phrase grounding, it also performs well when taking a phrase (answer) as input, especially for visual questions from VQAv2 dataset.

Analysis With Respect to Vision Skills. We next evaluate the SeqTR (I+Q+A) model's performance with respect to four vision skills [34]. Table 6 presents the IoU-PQ scores for visual questions with respect to four different vision skills for VizWiz-Therapy and VQA-Therapy datasets. Overall, this result shows that visual questions require recognizing *color* are relatively easier to ground. Visual questions requiring recognizing *text* and *counting* are relatively harder to ground. This observation can be explained by 5 as *color* tends to share the same grounding among different answers while *text* and *counting* tend to have different groundings. Moreover, visual questions require recognizing *text* and *counting* from VizWiz-VQA dataset are harder than VQAv2 dataset for answers grounding task as they tend to lead to different groundings compared to VQAv2 dataset, shown in Figure 5. In contrast, visual questions related to recognizing *color* are easier for VizWiz-VQA dataset compared to VQAv2 dataset.

Analysis With Respect to Same vs Different Groundings. Table 7 presents the IoU-PQ scores for visual questions with respect to visual questions with same grounding and different groundings. We use the SeqTR(I+Q+A) model as it has the best performance overall. Table 7 shows that the current state-of-the-art model lacks the ability to predict grounding for different answers (when answers are referring to different regions). We believe it is because the state-of-art models are designed based on an incorrect assumption that only one answer grounding is needed for a visual question. Thus, it performs better on the visual question having different answers that all describe the same grounding. Qualitative results can be found in supplementary materials.

5. Conclusions

Acknowledging the fact that a visual question can have different answers, we introduce a new dataset VQA-AnswerTherapy which provides visual evidence for all unique answers to a visual question and propose two novel challenges. Our algorithm benchmarking results highlight challenges and opportunities in this new problem space. For future versions, we will expand this work to reasoning datasets (e.g., VCR [33], GQA [15]) to compare it with perception-based datasets (VizWiz-VQA, VQAv2). We will share all our crowdsourcing source code and annotated metadata to aid this direction of future work.

References

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (Working Notes)*, 2(6), 2019. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280, 2019. 1, 2, 3, 5

[4] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 1, 2, 3, 4, 7, 8

[5] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. *arXiv preprint arXiv:2007.14419*, 2020. 2, 4

[6] Shi Chen and Qi Zhao. Rex: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15586–15595, June 2022. 2, 4

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[8] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 1, 2, 4

[9] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 8

[10] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820, 2017. 1, 2, 4

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 7

[12] Danna Gurari and Kristen Grauman. 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*, pages 3511–3522, 2017. 1, 2, 3

[13] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 1, 2, 7

[14] Bin He, Meng Xia, Xinguo Yu, Pengpeng Jian, Hao Meng, and Zhanwen Chen. An educational robot system of visual question answering for preschoolers. In *2017 2nd International Conference on Robotics and Automation Engineering (ICRAE)*, pages 441–445. IEEE, 2017. 2

[15] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 1, 2, 4, 8

[16] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 2, 4

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 8

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 7

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1, 2, 4, 8

[20] Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in the development and evaluation of open-ended visual question answering (vqa) tasks. *arXiv preprint arXiv:2103.15022*, 2021. 2

[21] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 8

[22] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable TextVQA models via visual and textual explanations. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 19–29, Mexico City, Mexico, June 2021. Association for Computational Linguistics. 2, 4

[23] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 8

[24] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazeb-

nik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 8

[25] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. *arXiv preprint arXiv:2207.13325*, 2022. 8

[26] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[27] Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in vqa. *arXiv preprint arXiv:2211.07516*, 2022. 2

[28] Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8465–8474, 2021. 8

[29] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 8

[30] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 2, 4

[31] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018. 2

[32] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 8

[33] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 8

[34] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31, 2020. 6, 7, 8

[35] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE, 2019. 8

[36] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. *arXiv preprint arXiv:2203.16265*, 2022. 7, 8

[37] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 4