VQA Therapy: Exploring Answer Differences by Visually Grounding Answers

Chongyan Chen[1], Samreen Anjum[2], Danna Gurari[1,2]
[1] University of Texas at Austin [2] University of Colorado Boulder

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 with answer groundings. We introduce the first dataset that visually grounds each unique answer to each visual question, which we call VQA-AnswerTherapy. We then propose two novel problems of predicting whether a visual question has a single answer grounding and localizing all answer groundings. We benchmark modern algorithms for these novel problems to show where they succeed and struggle. The dataset and evaluation server can be found publicly at https://vizwiz.org/tasks-and-datasets/vqa-answer-therapy/.

1. Introduction

Visual question answering (VQA) is the task of predicting the answer to a question about an image. A fundamental challenge is how to account for when a visual question has multiple natural language answers, a scenario shown to be common [10]. Prior work [2] revealed reasons for these differences, including 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 described versus because the *same visual content* is 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 we segment for each answer the visual evidence in the image needed to arrive at that answer. This new dataset, which we call VQA-AnswerTherapy, consists of 5,825 visual questions from the popular VQAv2 [9] and VizWiz [11] datasets. We find that 16% of the visual questions have multiple answer groundings, and provide fine-grained analysis to better elucidate when and why this arises.

We also introduce two novel algorithmic challenges, which are exemplified 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 annotator differences. Success on our **Single Answer Grounding Chal-**

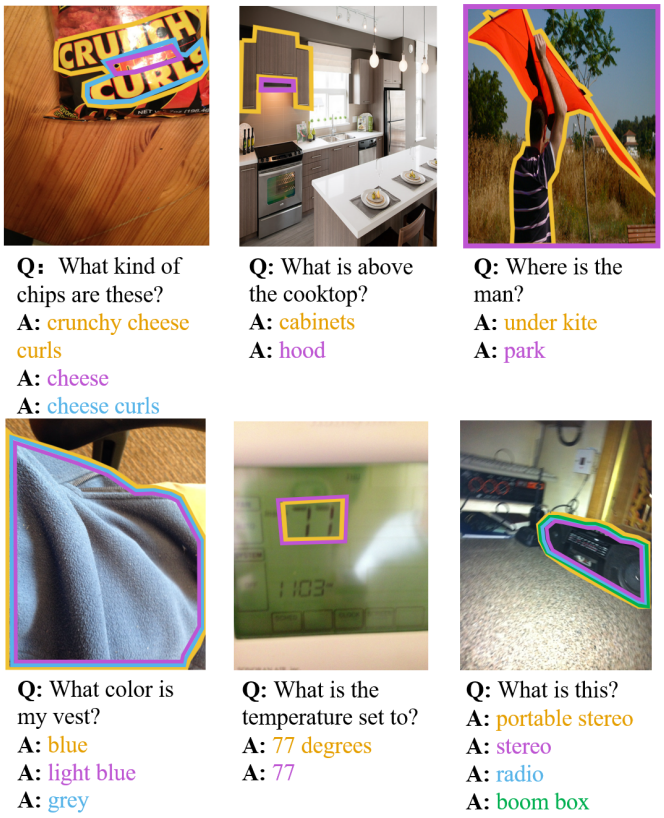

Figure 1: Examples from our VQA-AnswerTherapy dataset showing that visual questions with different natural language answers can have multiple answer groundings (first row) or all share the same answer grounding (second row).

lenge will enable solutions to notify users when there is uncertainty which visual evidence to consider, enabling users to clarify a visually ambiguous visual question. This can immediately benefit visually impaired individuals since a portion of our dataset comes from this population (i.e., VizWiz [11]). Success on the **Answer(s) Grounding Challenge** can enable users of VQA solutions to better understand the varied reasoning processes that can lead to different natural language answers, while also contributing to enhanced model explainability and trustworthiness. More generally, this work can inform how to account for **annotator differences** for other related tasks such as image captioning, visual dialog, and open-domain VQA (e.g., VQAs found on Yahoo!Answers and Stack Exchange). This work also contributes to ethical AI by enabling revisiting how VQA models are developed and evaluated to consider the diversity of plausible answer groundings rather than a single (typically majority) one. To facilitate future extensions, we publicly-share our dataset and a public evaluation server with leaderboard for our dataset challenges at the following link: https://vizwiz.org/tasks-and-datasets/vqa-answer-therapy/.

2. Related Work

Answer Differences in VQA Datasets. While many datasets have been created to support the development of VQA algorithms [22, 9, 11], 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 [10]. Prior work has offered initial steps. For example, prior work characterized when [10], to what extent [28], 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) [2]. Other work introduced ways to evaluate VQA models that acknowledge there can be multiple valid answers, whether provided explicitly from different people [1, 17] or augmented automatically from NLP tools to capture plausible, semantically related answers [17]. Another work focused on rewriting visual questions to remove ambiguity regarding what are valid answers [23]. 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 highlighting relevant visual content in images. A commonality of prior work [8, 19, 13, 7, 4, 34, 16, 12, 13, 19, 3, 5, 26] 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 [10]. 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

3.1. Dataset Creation

VQA Source. Our work builds upon two popular VQA datasets that reflect two distinct scenarios: VizWiz-VQA [11] and VQAv2 [9]. 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 [6] (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 [3], 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 [3], we removed visual questions with more than five words and the word "and", trimmed visual questions containing a repeated question down to a single question (e.g., from "what is this? what is this?" to "what is this?"), and filtered visual questions flagged as "containing more than one question" in metadata provided by [3].

We then selected 27,741 visual questions with 60,526 unique answers as candidates for our dataset. Included are all visual questions from VizWiz-VQA that met the aforementioned criteria (i.e., 9,528 visual questions with 20,930 unique answers) and a similar amount sampled from VQAv2's training set (i.e., 18,213 visual questions with 39,596 unique answers). We included all visual questions used in [2], which indicates why answers to each visual question differ, to support downstream analysis.

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. To simplify the task, we only instructed the worker to ground the answer when exactly one polygon was needed to ground answer. We leave future work to explore 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. Instructions were provided for how to complete the task, including for many challenging annotation scenarios (e.g., objects with holes or complex boundaries).

Answer Grounding Annotation Collection. We hired crowd workers from Amazon Mechanical Turk to generate answer groundings, given their on-demand availability. Like prior work [3], 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.

Ground Truth Generation. We analyzed the two sets of annotations collected for each of the 27,741 visual questions to establish ground truths. We did this after removing answers that at least one person flagged as "incorrect" and visual questions with answers referring to no polygon or multiple polygons. This left 12,290 visual questions and 26,682 unique image-question-answer triplets. 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. Examples of high-quality answer grounding results are shown in Figure 2, where answers to a visual question can either have multiple groundings (e.g., "What is over the elephant" and "What does this logo say") or a single grounding (e.g., "shirt's color").

Figure 2: High-quality annotations from our dataset. These also illustrate a trend that visual questions related to *text recognition* often have multiple answer groundings while *recognizing color* often have a *single grounding*.

[1] We follow the status quo established by prior work [3, 10] to obtain valid answers by using exact string matching (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.

		All	VQAv2	VizWiz-VQA
Multiple	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?
Single	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 *multiple* answer groundings and a *single* answer grounding for all visual questions as well as for VQAv2 and VizWiz-VQA independently. Of note, the overall frequency is dominated by VizWiz-VQA's frequency since the most common questions is far larger for this dataset than observed for VQAv2.

3.2. Dataset Analysis

We now analyze our final dataset, which includes 5,825 visual questions with 12,511 unique visual-question-answer-grounding sets. 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. This final dataset excludes all visual questions with less than two unique answers.

Prevalence of Single Versus Multiple Groundings. We first explore how often visual questions have different answers describing the same visual evidence (a single grounding) versus different visual evidence (multiple answer groundings). We flag a visual question as having different answers describing the *same grounding* if an answer grounding pair has an IoU score larger than 0.9.

We found 15.7% (i.e., 916/5,825) of visual questions with answers leading to *mulitiple answer groundings*. Yet, the status quo for VQA research neglects this reality that different answers can refer to different visual evidence [8, 19, 13, 7, 4, 34, 16, 12, 8, 13, 19, 3, 5, 26]. We suspect existing models would struggle with these 15.7% questions, both for VQA and answer grounding, due to their visual ambiguity.

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 questions about *recognizing objects* is common for both scenarios. In contrast, questions about *recognizing text* is more prevalent when there are *multiple answer groundings* while questions about *recognizing color* is more prevalent for visual questions with a *single answer grounding*. We also observe that questions related to a location often leads to multiple 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 finding suggests that a valuable predictive cue for AI models to predict whether there is a single grounding or multiple groundings for all answers are identifying the vision skills needed to answer a visual question.

When comparing the trend for visual questions to have multiple answer groundings across both data sources, we observe it is more prevalent for visual questions coming from VizWiz-VQA than VQAv2; i.e., it accounts for 22% (i.e., 761/3,442) versus 7% (i.e., 155/2,383) of visual questions respectively. Consequently, multiple answer groundings are more common for an authentic VQA use case than is captured in the most popular, yet contrived VQA dataset.

Reasons Visual Questions Have Multiple 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 $0.9 \geq \text{IoU} > 0$ and they do not have a contained relationship.

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

	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 a visual question's answers have, overall and per data 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 data source.

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. Overall, we find VizWiz-VQA and VQAv2 have a similar distribution of answer grounding relationships. 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. We suspect that when multiple individuals provide different answers based on distinct visual evidence, they may be focusing on the *same object* while paying attention to distinct *details*, resulting in an intersection of visual evidence.

Relationship Between Why Answers Differ and Number of Answer Groundings. We next analyze the tendency for 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 [2]: low-quality image (LQI), insufficient visual evidence (INV), difficult questions (DFF), ambiguous questions (AMB), subjective questions (SBJ), synonymous answers (SYN), and varying levels of answer granularity (GRN).[2,3] 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.

[2]We exclude the reasons "Spam answer" and "Invalid question" because, by definition, they cannot have grounded answers.

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

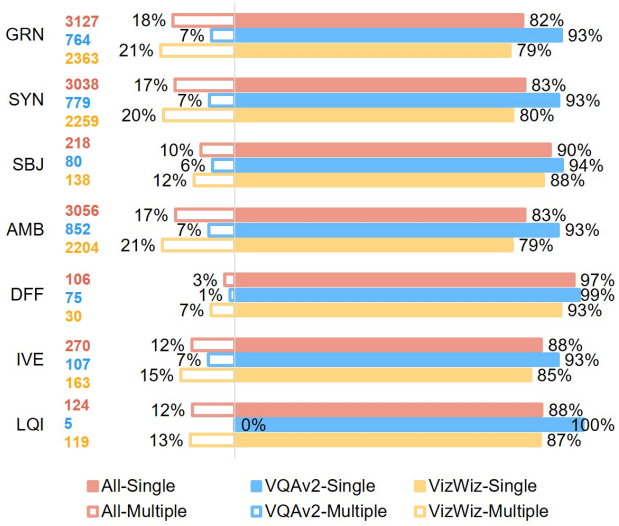

Figure 3: Relationship of whether a visual question has a single grounding for all answers and reasons for different answers for the VQAv2 and VizWiz dataset sources.

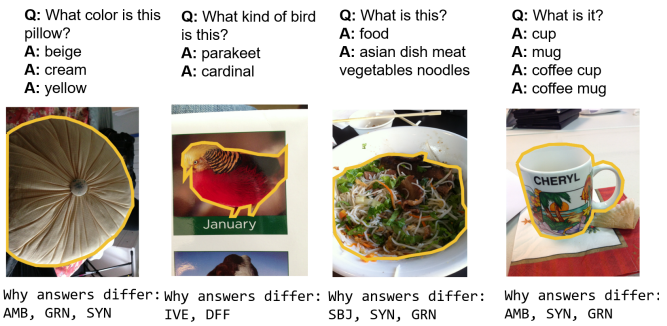

Figure 4: Visual questions with one answer grounding alongside annotations indicating why answers differ [2].

Overall, visual questions with multiple answer groundings commonly are associated with varying levels of answer granularity (GRN), ambiguous questions (AMB), and synonymous answers (SYN). The nearly identical results for AMB and SYN are not surprising since over 85% of VQAs labeled as SYN also occur with AMB in both the VizWiz-VQA and VQAv2 datasets.

Visual questions labeled with difficult (DFF) tend to share a single 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,

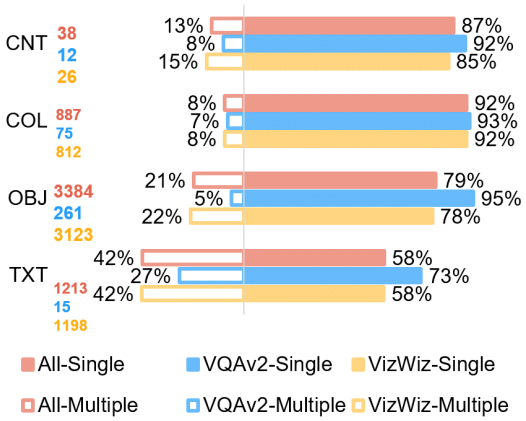

Figure 5: Amount of multiple answer groundings per visual question for four vision skills, overall and per dataset source (VQAv2 and VizWiz-VQA).

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 each of the four vision skills for our dataset (overall) and for each image source (VQAv2 and VizWiz-VQA).

we observe that VQAv2 and VizWiz-VQA have large differences (larger than 10%) for four reasons: GRN, SYN, AMB, and LQI. Examples of visual questions that are labeled as AMB, SYN, GRN are exemplified in Figure 4 (col 1, 3, and 4).

Relationships Between Vision Skills Needed to Answer a Visual Question and Number of Answer Groundings. We next evaluate how the vision skills needed to answer a visual question relate to whether a visual question has a single grounding. The following labels for the four vision skills are provided in the VizWiz-VQA-Skills dataset [31]: object recognition (OBJ), text recognition (TXT), color recognition (COL), and counting (CNT). Following [31] we use majority vote from the 5 annotations to determine the vision skill labels. We perform our analysis over all visual questions as well as with respect to each VQA source independently. Results are shown in Figure 5, based on observed percentages for each VQA source.

Overall, we found that visual questions trying to read *text* tend to have multiple answer groundings. One common example is visual questions about products, as exemplified in Figure 1 (e.g., chips product). In contrast, questions related to recognizing *color* tend to have a single answer grounding. 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 multiple answer groundings with respect to each of the four vision skills overall as well as with respect to each VQA source independently. We determine if a visual question has a single grounding and what skills are needed following the same process of the previous analysis. Results are shown in Figure 6.[4] 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) and Figure 2 ("blanket" and "umbrella" are disjoint; "rsb" and "royal society for blind" are disjoint).

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 grounding for all answers and (2) localizing all 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 [11]. Our dataset also has visual questions originating from the training set of the VQAv2 dataset [9], 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. For completeness, we also explore the predicting if a visual question will result in answers that multiple groundings. We evaluate methods

[4]Results for VQAv2 may not be representative since only a small amount of our dataset's visual questions have vision skill labels.

Model Type:	Precision				Recall			
	ViLT	*mPLUG-Owl*	*Naïve (M)*	*Naïve (S)*	*ViLT*	*mPLUG-Owl*	*Naïve (M)*	*Naïve (S)*
All:S	**0.86**	0.80	-	0.80	0.94	0.82	0.0	**1.0**
VQAv2:S	**0.93**	**0.93**	-	0.92	0.97	0.81	0.0	**1.0**
VizWiz-VQA:S	**0.82**	0.74	-	0.74	0.92	0.83	0.0	**1.0**
All:M	**0.59**	0.20	0.20	-	0.37	0.20	**1.0**	0.0
VQAv2:M	**0.24**	0.11	0.08	-	0.11	0.26	**1.0**	0.0
VizWiz-VQA:M	**0.63**	0.25	0.26	-	0.42	0.17	**1.0**	0.0

Table 4: Performance of methods at predicting whether a visual question will result in answers that all share single (i.e., 'S') or multiple (i.e., 'M') groundings respectively, overall as well as with respect to each data source. Of note, no values ('-') are entered for some models because they do not yield valid scores. This includes the Naïve (M) for task 'S' with respect to precision and Naïve (S) for task 'M' with respect to precision, because no positives are predicted, making the denominator zero (i.e., precision = True Positive/(True Positive+False Positive). This also includes the Naïve (M) model for task 'S' with respect to recall and Naïve (S) model for task 'M' with respect to recall, because there are no positives and so the numerator is again zero (i.e., recall = True Positive / (True Positive + False Negative).

using two standard metrics for binary classification tasks: precision and recall.

Models. We benchmark four models. We fine-tune a top performing algorithm for the VQA task, ViLT [15], on the training set on of our entire dataset. To do so, we modified the output layer of the architecture with a two-class softmax activation to support binary classification. We also benchmark the state-of-the-art vision and language foundation model which was the first to achieve human parity on the VQA Challenge, mPLUG-Owl [29], in a zero-shot setting with the prompt "Do all plausible answers to the question indicate the same visual content in this image?" This zero-shot setting is a useful baseline because of the imbalanced and relatively small size of our dataset to support training. We finally benchmark two naïve baselines that each only predict one label, i.e., all samples share the same answer grounding *or* all samples have multiple answer groundings.

Results. Results are reported in Table 4 for predicting whether a visual question has answers that all share single grounding ('S') and predicting whether a visual question has answers with multiple groundings ('M'). Testing results are reported for the entire dataset, as well as on each VQA source (VQAv2 and VizWiz-VQA) independently.

We observe that it is much more difficult to predict when answers have multiple groundings compared to a single grounding, i.e., both ViLT and mPLUG-Owl receive much lower precision and recall when predicting whether there are multiple answers grounding compared to predicting if there is a single answer grounding. This is true both for the fine-tuned ViLT model, which is susceptible to failing from the data imbalance (i.e., only 15.7% of visual questions have multiple answers grounding), as well as the the zero-shot mPLUG-Owl solution.

Figure 7: Qualitative results for the fine-tuned ViLT model on the Single Answer Grounding Challenge, alongside the question-answer pair and ground truth answer groundings.

We next analyze overall performance for the models. While ViLT is an inferior VQA model compared to mPLUG-Owl, it achieves better performance once fine-tuned on our dataset for our novel task. This enhancement is striking, considering the limited size of our dataset. This observation underscores the value of our dataset, illustrating how even a modest number of samples can bolster the robustness of current models. In contrast, the foundation model, mPLUG-Owl, achieves inferior or comparable performance to a naïve baseline. We manually inspected all examples where the top-performing fine-tuned ViLT struggles, and show examples in Figure 7. For instances where there is a single answer grounding and ViLT predicts multiple answer groundings, across both VizWiz-VQA and VQAv2 sources, often images show text or multi-

ple objects while the question typically references the entire object or a particular area, as illustrated in Figure 7 column 1. Conversely, in situations where there are multiple answer groundings but ViLT predicts only one (143 examples for VizWiz-VQA and 34 for VQAv2), we observe distinct patterns between the VizWiz-VQA and VQAv2 datasets. Specifically, in the VizWiz-VQA dataset, 98 out of the 143 instances occur because the image contains only one significant object, with the answer primarily focusing on text recognition (Figure 7 column 2). In contrast, for the VQAv2 dataset, this discrepancy arises in 27 out of the 34 cases mainly because the question is ambiguous about which object/area it is asking about and the image contains multiple objects (Figure 7 column 3).

When comparing the performance across datasets, despite that we permitted both models to have an unfair advantage that they could observe during training the COCO images that are used in the VQAv2 dataset [5] and it's cheating to test it on the training set of the VQAv2 dataset, we only observe higher performance when predicting "VQAv2:S" compared to "VizWiz-VQA:S" and didn't observe higher performance when predicting "VQAv2:M" compared to "VizWiz-VQA:M". We suspect the reason is that the VQA-Single Answer Grounding dataset is highly imbalanced with 93% of visual questions having different answers that all describe the same visual evidence.

4.2. Answer(s) Grounding Challenge

Given an image and a question, the task is to predict the image region to which the answer is referring.

Evaluation Metric. We measure the similarity of each binary segmentation to the ground truth with IoU. We report two IoU scores, IoU and IoU-PQ (IoU-Per Question). The IoU-PQ uses as the score for each visual question the average of the IoU scores for all answer groundings to that visual question. We utilize IoU-PQ in place of IoU because the metadata (e.g., single/multiple annotations, vision skills annotations) for fine-grained analysis pertains to each visual question rather than each answer grounding.

Models. We evaluate three variants for each of the following three models: SeqTR, UNINEXT [27], and SEEM [35].[6] For the three variants per model, we feed the model the image-question pair (i.e., Model(I+Q)), the image-question-answer triplet (i.e., Model(I+Q+A)) and the image-answer pair (i.e., Model(I+A). We fine-tuned a top-performing referring segmentation algorithm, SeqTR, on our entire dataset. SeqTR [33] is pretrained on a large corpus of datasets (i.e., [16, 30, 18, 14, 21, 20]). We also evaluated zero-shot performance for both the UNINEXT [27] and SEEM [35] models. We selected UNINEXT because of its state-of-the-art performance for the Referring Expression Segmentation task and SEEM since it claims to "segment everything everywhere".

Overall Results. Results are shown in Table 5.[7] Performance is reported for the entire dataset (column 2) as well as with respect to each VQA source independently (column 3 and column 4).

As shown, all analyzed models perform poorly. For example, the top-performing SeqTR(I+Q+A) overall only achieves an IoU of 66.68%. This arises despite that all three models were exposed to COCO images in the pretraining phases; performance on VQAv2 dataset is still similar to that for VizWiz-VQA. We suspect the referring segmentation pretraining may result in models taking shortcuts by remembering images while ignoring the language (the language inputs when pretraining are referring expressions, which can differ considerably from our inputs).

We also analyze the results for each model. While part of the poor performance of SeqTR could be attributed to the relatively small amount of training examples available for fine-tuning, our results in Table 6 offer strong evidence that the challenge of grounding different answers is also an important factor. That is because SeqTR(I+Q+A) scores 72% on visual questions with a single answer grounding versus 43.66% on visual questions with multiple answer groundings, underscoring a greater difficulty for the latter task. Our results on UNINEXT and SEEM also underscore how current large segmentation models lack sufficient zero-shot generalization capabilities, a necessary prerequisite for applications such as open-domain VQA.

Comparing the performance across different variant settings (I+Q+A/I+Q/I+A), we find that the model that re-

Models	All	VQAv2	VizWiz-VQA
SeqTR (I+Q+A)	**66.68**	**64.50**	**67.89**
SeqTR (I+Q)	62.04	58.46	64.02
SeqTR (I+A)	63.27	58.03	66.17
SEEM (I+Q+A)	**53.77**	**50.67**	**55.49**
SEEM (I+Q)	45.17	44.65	45.46
SEEM (I+A)	52.10	46.83	55.03
UNINEXT (I+Q+A)	**53.76**	**42.73**	**59.88**
UNINEXT (I+Q)	50.51	40.96	55.81
UNINEXT (I+A)	52.76	41.60	58.95

Table 5: Performance of models for predicting all answer groundings per visual question.

[5] ViLT is pretrained on GCC, SBU, COCO, and VG datasets and mPLUG-Owl is pretrained on LAION-400M, COYO-700M, Conceptual Captions and MSCOCO.

[6] We do not benchmark answer grounding models [25, 32, 24] since these show weak performance on existing challenges (e.g., [3].)

[7] Due to space constraints, we report overall model performance with respect to IoU-PQ in the supplementary materials. The scores align closely with IoU.

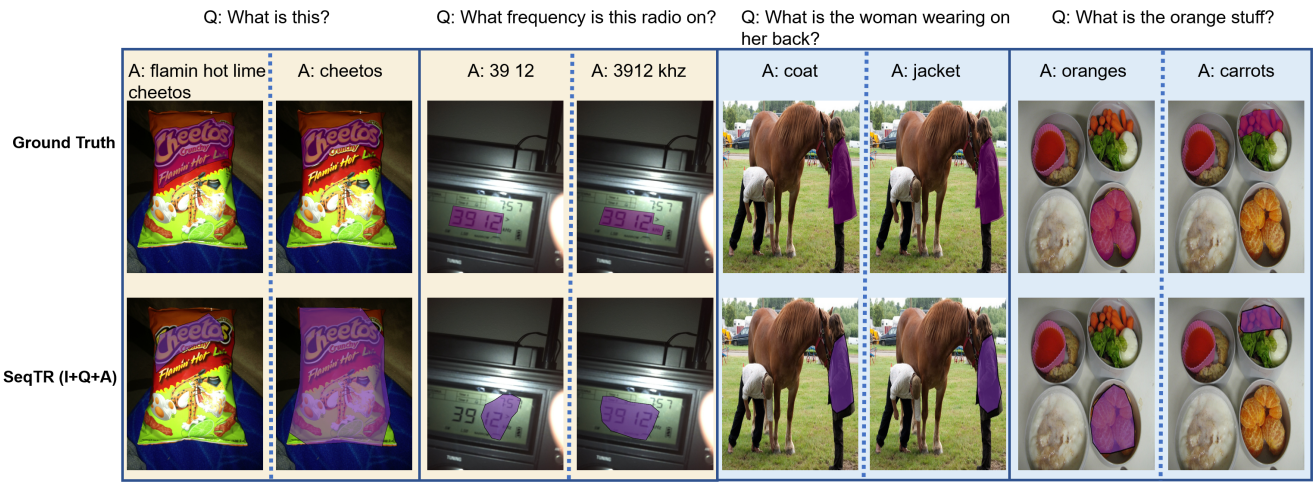

Figure 8: Qualitative results from SeqTR (I+Q+A) for visual questions coming from VizWiz-VQA (yellow background) and VQAv2 (blue background).

Models	Single	Multiple
SeqTR (All)	**71.69**	**43.66**
SEEM (All)	60.45	22.75
UNINEXT (All)	60.47	23.08
SeqTR (VQAv2)	**65.56**	**49.66**
SEEM (VQAv2)	51.65	31.70
UNINEXT (VQAv2)	43.79	24.15
SeqTR (VizWiz-VQA)	**75.94**	**42.66**
SEEM (VizWiz-VQA)	66.56	21.27
UNINEXT (VizWiz-VQA)	72.06	22.90

Table 6: Performance of models at localizing all answer groundings with respect to IoU-PQ scores. They struggle most for visual questions with multiple answer groundings.

ceives the most information as input (I+Q+A) performs best, which aligns with our intuition. We show the qualitative results for SeqTR (I+Q+A) model in Figure 8. We observed models often fail for vision questions with multiple answer groundings that require recognizing text. In contrast, models often perform well for visual questions that identify common objects.

Analysis With Respect to Single vs Multiple Answer Groundings. Table 6 presents the IoU-PQ scores for visual questions with respect to visual questions with a single answer grounding and multiple answer groundings. We use the settings of (I+Q+A) for each model to reveal the upper bound of what is possible from top-performing models.

We observe that the top-performing model, SeqTR, largely lacks the ability to predict multiple answer groundings. This suggests modern models are designed based on an incorrect assumption that only one answer grounding is needed for a visual question. Still, SeqTR significantly outperforms SEEM (All) and UNINEXT (All), highlighting a potential benefit of a modest amount for fine-tuning models for our target task.

Delving into the data based on VQA sources, a compelling pattern emerges. All models consistently deliver superior performance for visual questions with multiple answers groundings on VQAv2 compared to VizWiz-VQA. Conversely, performance for visual questions with a single answer grounding is worse on VQAv2 than for VizWiz. One potential factor leading to this outcome may stem from VQAv2 having a higher prevalence of complex scenes and so presenting a greater difficulty for grounding answers when only a single grounding is needed.

5. Conclusions

This work acknowledges a fundamental challenge that visual questions can have multiple valid answers. We support further exploration of this fact by introducing a new dataset, which we call VQA-AnswerTherapy, that provides a grounding for every valid answer to each visual question. We also propose two novel challenges of (1) predicting whether a visual question has a single answer grounding (versus multiple answer groundings) and (2) locating all answer groundings for a given visual question. Our algorithm benchmarking results reveal that modern methods perform poorly for these tasks, especially when a visual question has multiple answer groundings. We share our dataset and crowdsourcing source code to facilitate future extensions of this work.

Acknowledgments. This work was supported with funding from Microsoft AI4A and Amazon Mechanical Turk.

References

[1] 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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023.

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

[29] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.