Why Does a Visual Question Have Different Answers? - Supplementary Information

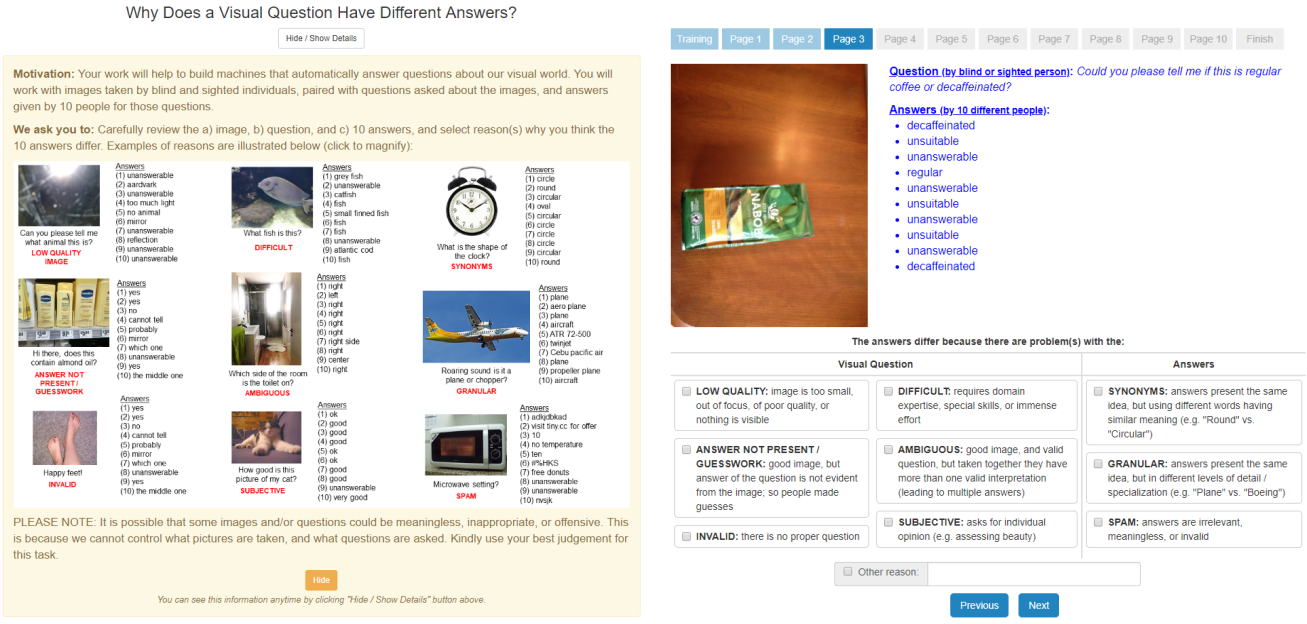
Nilavra Bhattacharya*, Qing Li+, Danna Gurari*
* University of Texas at Austin, + University of California, Los Angeles

https://vizwiz.org

1. Crowdsourcing Task (supplements section 3 of the main paper)

1.1. User Interface

The crowdsourcing user interface is shown in Figure 1.



(a) (b)

Figure 1: (a) Task instructions to train crowd workers about the reasons that can lead to different answers. (b) The interface crowd workers used to choose why different answers are observed for a given QI pair with its 10 corresponding answers.

1.2. Quality Control

We included a training example that each crowd worker had to complete prior to completing our task. The authors identified the correct labels beforehand for this example. For each HIT posted to Amazon Mechanical Turk, the worker had to select these correct labels in order to proceed to the actual task.

2. Dataset Analysis (supplements section 4 of the main paper)

2.1. Inter-Annotator Agreement for Reasons Labels

We examine inter-annotator agreement among crowd workers. To do so, we measure the Worker-Worker Similarity (WWS) as the pairwise annotation similarity between two workers across all the VQAs they have annotated in common. The WWS measure indicates how close a worker performs to the group of workers who have solved the same task. We calculate WWS between two crowd workers w_i and w_j using three approaches: (a) number of common labels selected, (b) cosine similarity, and (c) Cohen's κ [1].

WWS - Common Labels
This metric is defined as

$$wws(w_i, w_j) = \frac{\sum_{t \in T_{i,j}} numCommonLabels(w_i, w_j, t)}{\sum_{t \in T_{i,j}} numAnnotations(w_i, t)}$$

where $T_{i,j}$ is the subset of all VQA tasks T annotated by both workers; $numCommonLabels(w_i, w_j, t)$ is the number of identical labels selected by both workers w_i and w_j on a VQA task t; and $numAnnotations(w_i, t)$ is the total number of labels selected by a worker w_i for a single VQA task t.

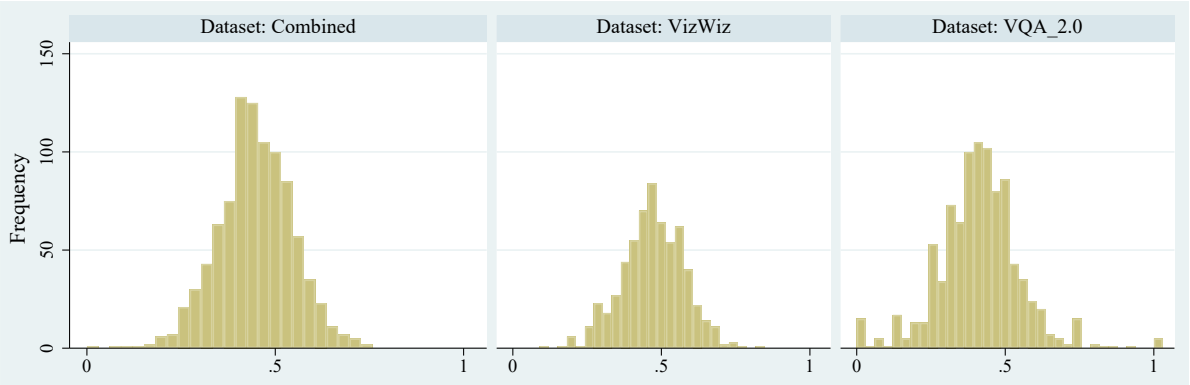


Figure 2: Distribution of 'WWS - Common Labels' for all crowd workers across both datasets alone as well as combined.

WWS - Cosine Similarity
This metric is defined as follows:

$$avg(\cos(V_{t,w_i}, V_{t,w_j})) \ \ \forall \ \text{worker } j, j \neq i$$

where V_{t,w_i} is the 'Task Vector' of worker w_i annotating VQA task t. A Task Vector for a worker annotating a VQA task is defined as a vector whose length is 10 (i.e. equal to the number of labels available), and whose individual elements are either 0 or 1, depending on whether the worker selected the label or not. E.g. if a worker selects the labels LQI, AMB, and SBJ, and the ordering of the labels in the Task Vector are LQI, IVE, INV, DFF, AMB, SBJ, SYN, GRN, SPM, OTH, then the Task Vector becomes: `[1,0,0,0,1,1,0,0,0,0]`.

WWS - Cohen's κ
This metric is defined as follows:

$$wws_\kappa = avg(\kappa)$$

where κ is the Cohen's kappa coefficient [1] used to measure inter-rater agreement.

Figures 2, 3, and 4 show the distribution of the three WWS metrics for the 934 distinct crowdworkers who provided annotations for our dataset, averaged for each worker. Among them, 615 distinct workers annotated VQAs from the *VizWiz* dataset, while 928 distinct workers annotated the *VQA_2.0* dataset.

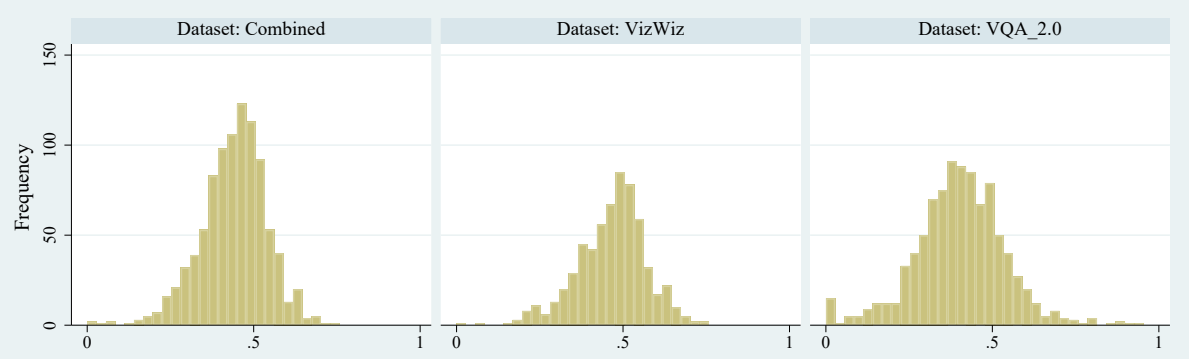


Figure 3: Distribution of 'WWS - Cosine Similarity' for all crowd workers across both datasets alone as well as combined.

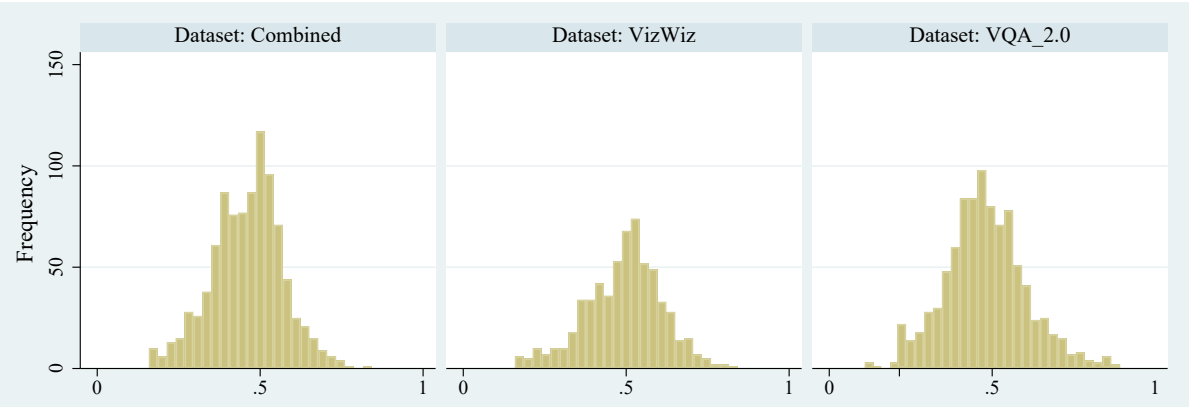
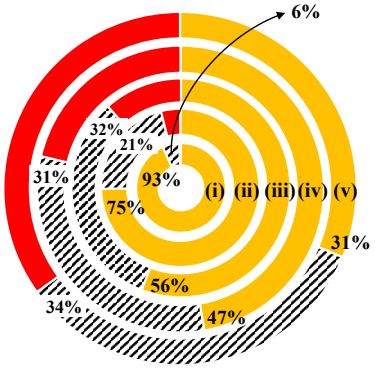


Figure 4: Distribution of 'WWS - Cohen's κ' for all crowd workers across both datasets alone as well as combined.

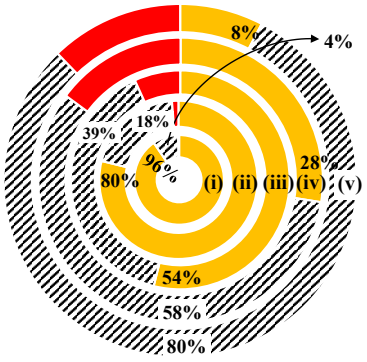
All the distributions assume an approximately normal form, with peaks at 0.5. This suggests that most workers agreed with 50% of the other workers with whom they shared common annotation tasks.

In the case of *VQA_2.0*, there seems to be a small yet distinct percentage of workers who did not agree with anyone. This is characterized by a small lump near the 0 value in the plots for *VQA_2.0*, of all the three WWS metrics (Figures 2, 3, & 4).

2.2. Analysis Using All Validity Thresholds



(a) VizWiz



(b) VQA_2.0

Answer difference due to issues with:

QI & A | QI | A

Validity Threshold	Count of VQAs	
	VizWiz	VQA_2.0
(i) 1 Person	29,921	15,034
(ii) 2 Person	29,879	15,020
(iii) 3 Person	28,030	13,853
(iv) 4 Person	20,181	9,035
(v) 5 Person	9,932	2,775

Figure 5: Relative proportion of the various sources of answer disagreement (augmented from Figure 2 in the paper).

We tallied the number of reasons leading to answer differences for each VQA, employing various levels of trust in crowd workers: from 1 person threshold to 5 person thresholds.

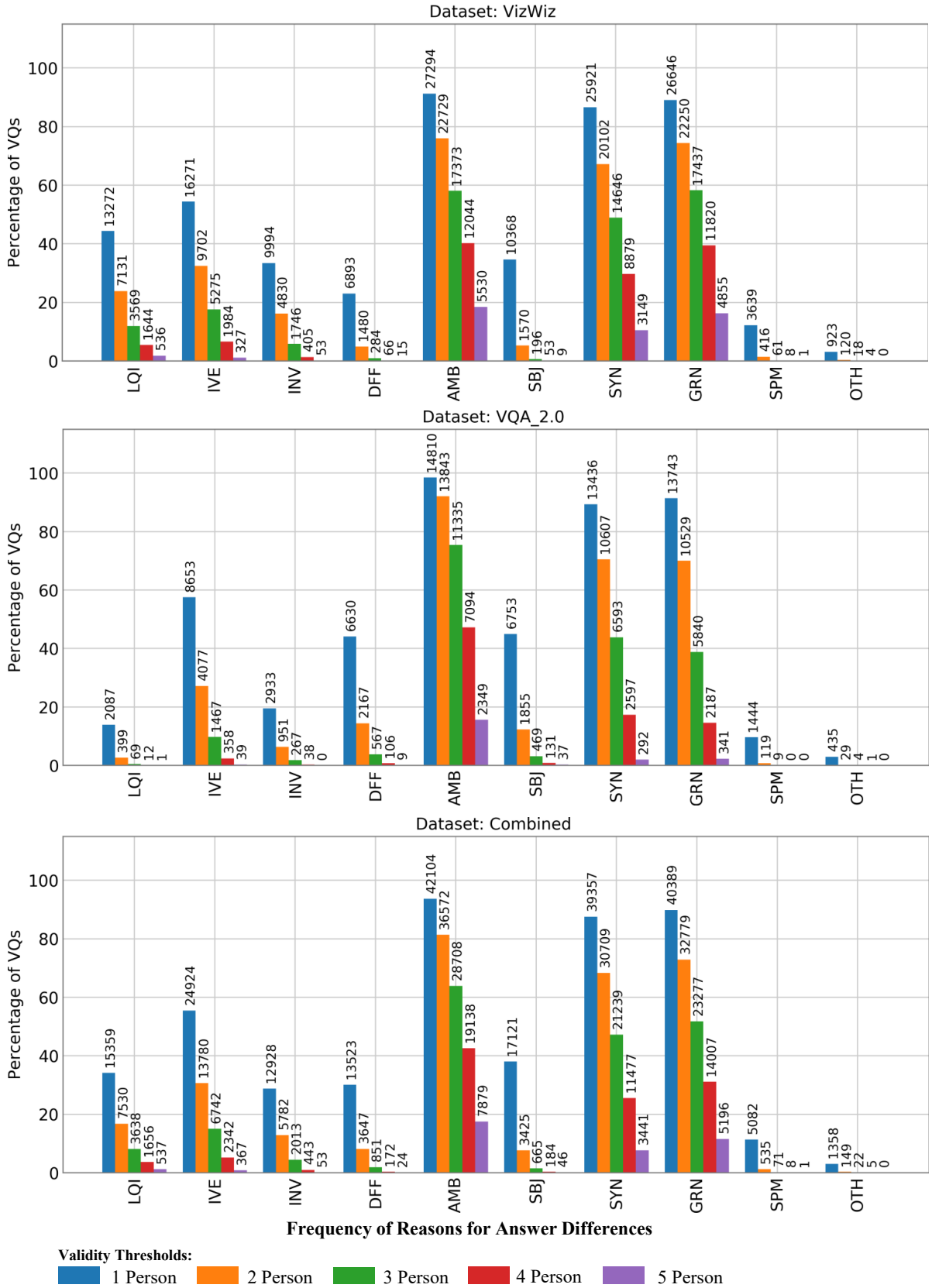


Figure 6: Histograms showing the frequency of each reason leading to answer differences (augmented from Figure 3 in the paper). Data labels show counts of VQAs matching the validity threshold.

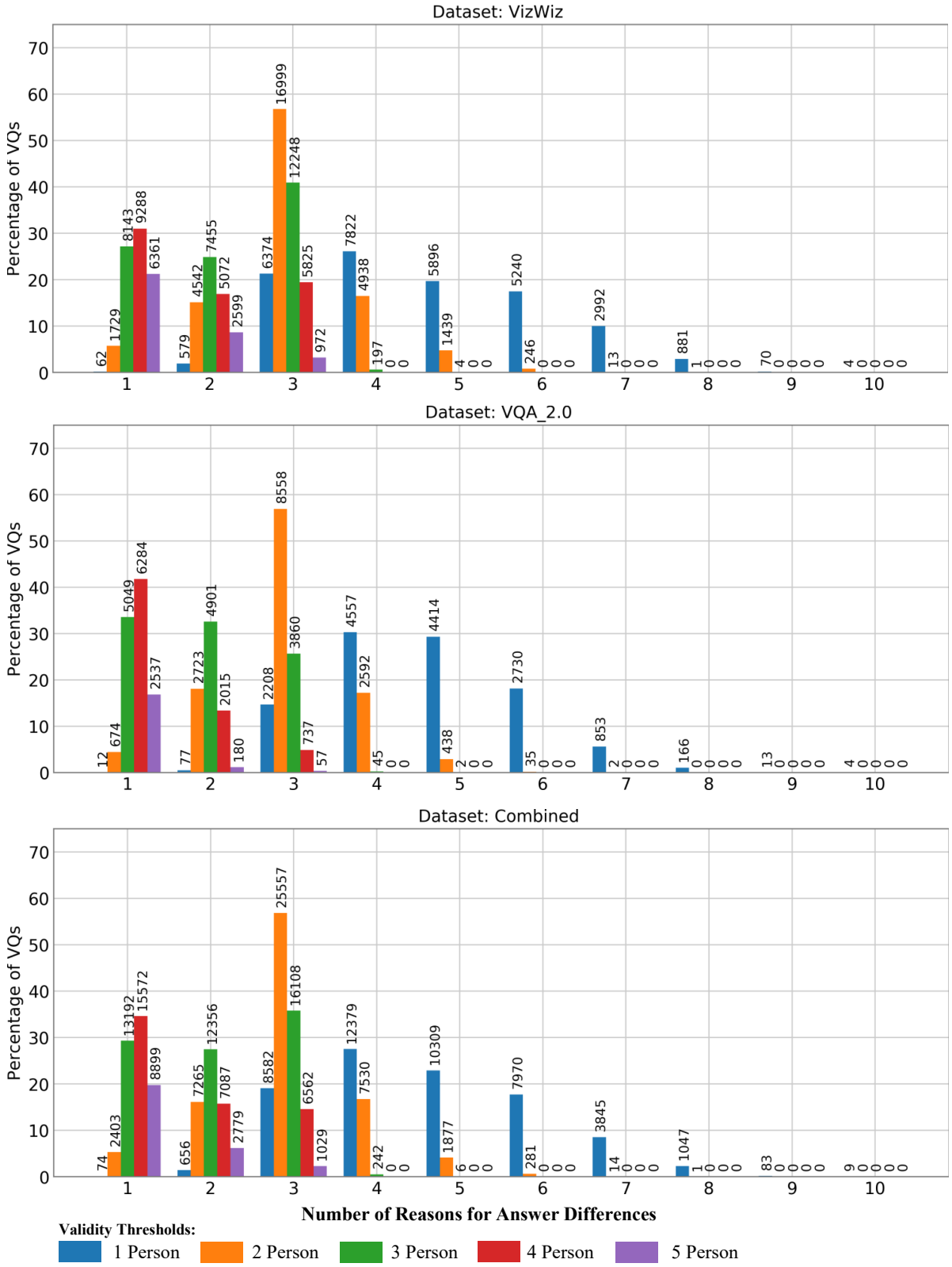


Figure 7: Histograms showing the number of unique reasons of answer differences identified for each visual question. Data labels show counts of VQAs matching the validity threshold.

Figure 5 shows the percentage of visual questions, where answer differences arise due to issues with both the QI pair and the 10 answers (**QI & A**, yellow), issues with the QI pair only (**QI**, striped), or issues with the 10 answers only (**A**, red), for the **(a)** *VizWiz* and **(b)** *VQA_2.0* datasets. Results are shown with respect to different levels of trust in the crowd workers: **(i)** *Trust All*: only one worker has to select the reason (1 person validity threshold); **(ii)** *Trust Any Pair*: at least two workers must agree on the reason (2 person validity threshold); **(iii)** *Trust Majority*: at least three workers must agree on the reason (3 person validity threshold); **(iv)**: at least four workers must have to select the reason (4 person validity threshold); and **(v)** *Trust Consensus*: all five workers must agree on the reason (5 person validity threshold).

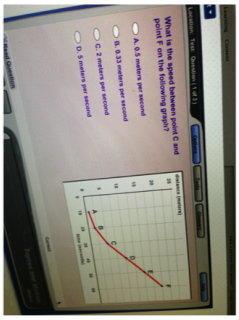
Figure 6 shows histograms of the frequency of each reason leading to answer differences for (a) 29,921 visual questions asked by blind people (*VizWiz*), (b) 15,034 visual questions asked by sighted people (*VQA_2.0*), and (c) combination of the previous two. The plots are computed based on increasing thresholds of inter-worker agreement required to make a reason valid, ranging from requiring only one worker selecting it (1 person validity threshold) up to all workers agreeing (5 person threshold). The most popular reasons are ambiguous visual questions (AMB), synonymous answers (SYN), and varying answer granularity (GRN) whereas the most rare are spam (SPM) and other (OTH).

Figure 7 shows the summary of how many unique reasons are identified as the sources of answer differences for 29,921 VQs asked by blind people (*VizWiz*), 15,034 VQs asked by sighted people (*VQA_2.0*), and their combination. Across both datasets, most commonly there are three unique reasons for answer differences. Visual inspections show that these are the three most popular reasons: 'ambiguous', 'synonyms', and 'granularity'.

3. Prediction Model Analysis (supplements section 5 of the main paper)

Table 1: Average precision for predicting why answers to visual questions will differ for the VQA_2.0 and VizWiz datasets when we exclude the "spam" category for training the models.

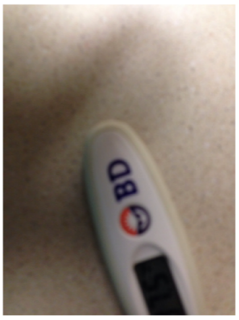
	Model	Overall	LQI	IVE	INV	DEF	AMB	SBJ	SYN	GRN	OTH
VQA_2.0	Random	33.54	3.71	22.43	15.09	14.62	95.19	14.18	64.99	69.42	**2.25**
	QI-Relevance [3]	35.76	4.01	43.16	15.09	14.62	94.11	14.18	64.99	69.42	2.25
	I	35.38	3.66	29.55	10.06	17.04	93.04	18.29	74.50	72.09	0.18
	Q	48.11	7.91	59.23	43.43	28.02	**96.70**	23.69	88.85	84.78	0.36
	Q+I	47.87	9.18	58.83	40.83	28.18	96.48	24.20	88.39	84.37	0.37
	Q+I+A	48.05	7.09	59.46	**45.18**	27.99	96.60	21.89	88.97	85.05	0.26
	Q+I+A_FT	**48.89**	**9.11**	**59.78**	44.99	**30.11**	96.52	**24.33**	**89.49**	**85.39**	0.25
	Q+I+A_GT	48.96	8.65	60.30	46.00	28.91	96.63	24.13	89.73	86.00	0.26
VizWiz	Random	33.35	23.59	33.69	18.15	5.70	74.70	5.14	66.61	71.94	0.62
	QI-Relevance [3]	38.76	30.56	40.52	18.15	5.7	76.53	5.14	66.61	71.94	0.62
	Unanswerable [2]	43.26	44.82	58.63	18.15	5.7	80.14	5.14	66.61	71.94	0.62
	I	44.79	55.23	50.38	29.85	8.17	83.42	9.19	79.96	86.34	0.62
	Q	44.76	35.38	54.43	38.91	13.59	84.44	10.59	79.68	85.15	0.65
	Q+I	50.59	56.54	61.91	45.25	13.80	87.55	**11.55**	85.97	91.42	**1.36**
	Q+I+A	55.18	**65.51**	**77.36**	**55.76**	10.38	**89.77**	10.83	**90.39**	**95.50**	1.14
	Q+I+A_FT	**55.35**	65.30	77.18	54.19	**14.24**	89.60	11.32	89.99	95.24	1.07
	Q+I+A_GT	55.97	66.03	77.80	56.55	12.94	90.03	12.51	90.41	95.51	1.97



What is this?
AMB, SYN, GRN

What are these?
GRN, AMB, SYN

Which color is this pen?
AMB, SBJ, SYN, GRN

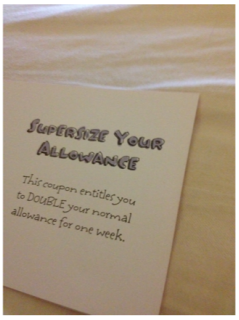


What does this say?
LQI, IVE, AMB, SYN, GRN

What kind of K cup?
LQI, IVE, AMB, GRN

What fragrance is this?
LQI, GRN, IVE

What's the length of a cane?
IVE, GRN, AMB



what does this paper say?
SYN, GRN, AMB

Figure 8: Qualitative examples of our prediction system (Q+I+A_FT). Green denotes correct prediction, red denotes wrong prediction, and turquoise denotes missing prediction.

References

[1] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 2

[2] 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. 6

[3] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. *EMNLP*, 2017. 6