Assessing Image Quality Issues for Real-World Problems - Supplementary Materials

Tai-Yin Chiu, Yinan Zhao, Danna Gurari
University of Texas at Austin

This document supplements Sections 3 and 4 of the main paper. In particular, it includes the following:

- Details and motivation for quality flaw interrelation index (supplements **Section 3.2**).
- Result of quality flaw prediction (supplements **Section 4.2**).
- Figures illustrating the crowdsourcing interface used to curate our labels (supplements **Section 3.1**), diversity of resulting unrecognizable images (supplements **Section 3.2**), performance of our prediction system in classifying unrecognizable images (supplements **Section 4.2**), and performance of the prediction of the reason for unanswerable questions (supplements **Section 5.2**).
- Clarification about baselines used for **Section 4.3**.

1. Quality flaw interrelation index

Details and motivation The most straightforward way to explore the relation of two quality flaws A and B is to look at their the co-occurrence or their joint probability $P(A, B)$. However, $P(A, B)$ cannot really capture the interrelation between quality flaws. For instance, we cannot say that the relation between DRK and FRM is stronger than the one between DRK and OBS simply because of $P(\text{DRK}, \text{FRM}) \gg P(\text{DRK}, \text{OBS})$. The reason for $P(\text{DRK}, \text{FRM}) \gg P(\text{DRK}, \text{OBS})$ is actually due to $P(\text{FRM}) = 55.6\% \gg P(\text{OBS}) = 3.6\%$ but has nothing to do with the interrelation of quality flaws.

Consequently, we introduce a new measure which we call interrelation index $I(A, B)$, which is defined as follows:

$$I(A, B) = \frac{P(B|A)}{P(B)} - \frac{P(B|\bar{A})}{P(B)}. \tag{1}$$

There are several advantages of this measure:

1. It measures causality from A to B: we can show that if $P(A)$ and $P(B)$ are both greater than zero, either $P(B|A) \geq P(B) \geq P(B|\bar{A})$ or $P(B|A) < P(B) < P(B|\bar{A})$ holds. Therefore, if $I(A, B) > 0$, then the existence of A must trigger B to happen more (i.e., $P(B|A) \geq P(B)$) and the inexistence of A must make B happen less (i.e., $P(B) \geq P(B|\bar{A})$), and vice versa.
2. It measures co-occurence of A and B: We can show that if $P(B|A) \geq P(B) \geq P(B|\bar{A})$, then $P(A|B) \geq P(A) \geq P(A|\bar{B})$ (it is also true for $<$ sign). Hence, we have $I(A, B) > 0 \Leftrightarrow I(B, A) > 0$. In other words, if A makes B happen more often, then B must make A happen more as well, and vice versa.
3. It avoids the aforementioned problem of using joint probability. That is, if $P(A) \gg P(B)$, it is very likely $P(A, C) \gg P(B, C)$. However, which of the values of $I(A, C)$ and $I(B, C)$ is greater and how greater it is cannot be told from $P(A) \gg P(B)$.

Co-occurrence of DRK and BRT. Since the values $I(\text{DRK}, \text{BRT}) = 74$ and $I(\text{BRT}, \text{DRK}) = 73$ are both greater than zero, it means that the quality flaws of DRK and BRT tend to co-occur despite their contradictory concepts. Nevertheless, the examples of such images in Figure 2 explain why this phenomenon happens. The main reason for this phenomenon is when blind people take pictures in places with poor lighting, they are not aware that the flashlights on mobile devices are turned on automatically, and therefore pictures taken are usually of dark surroundings and a bright spot. Note that this phenomenon is not captured by the joint probability of DRK and BRT, since $P(\text{DRK}, \text{BRT}) = 0.53\%$ is an extremely small value which does not manifest too much.

Co-occurrence of quality flaws. We exemplify the co-occurrence of other pairs of quality flaws in Figure 3.

2. Quality flaw prediction

Performance of quality flaw classification is shown in Table 1. We can tell that the Xception model outperforms the random guessing baseline for each quality flaw, with respect to precision, recall, and f1 score. Furthermore, Xception

		NON	BLR	BRT	DRK	OBS	FRM	ROT	OTH
Xception	precision	72.9	80.1	62.9	58.5	53.6	77.0	72.6	60.0
	recall	79.0	80.1	49.8	57.3	39.7	82.4	69.8	9.1
	f1 score	75.8	80.1	55.6	57.9	45.6	79.6	71.2	15.8
Random guessing	precision	48.6	40.5	4.9	7.2	4.0	55.0	15.6	0.0
	recall	50.5	40.3	4.3	6.7	4.0	54.3	15.7	0.0
	f1 score	49.5	40.4	4.5	7.0	4.0	54.6	15.6	0.0

Table 1: Performance of quality flaw prediction

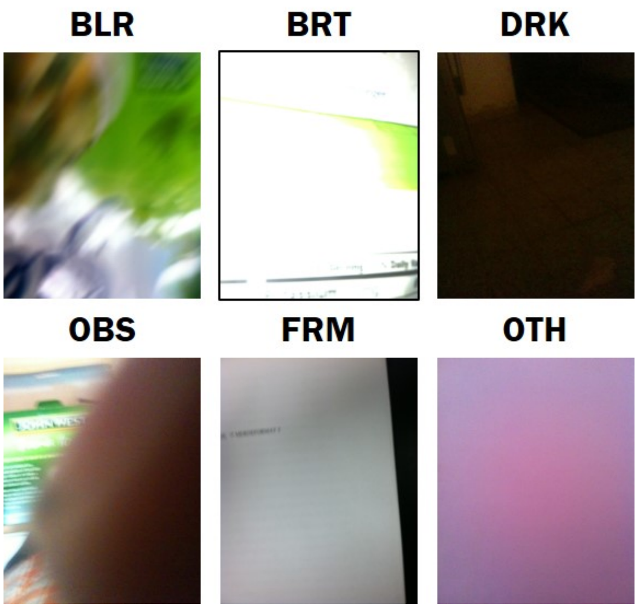

Figure 1: Unrecognizable images due to different quality flaws.

predicts much better in NON, BLR and FRM flaws for large portions of the dataset. On the other hand, quality flaws that represent small portions of the dataset are prone to few-shot learning, and so learning to predict them is harder. In the extreme case of OTH, with it representing 0.8% of the data, the Xception model yields very poor scores of 9.1 and 15.8 for recall and f1 score, respectively.

3. Miscellaneous

- Figure 1 illustrates the diversity of unrecognizable images that can arise from different quality flaws.
- Figure 4 shows a screen shot of the crowdsourcing interface used to collect the labels for the dataset.
- Figure 5 shows the examples of unrecognizability prediction by the Xception model.
- Figure 6 shows the examples of the prediction of the reason for unanswerable questions. The prediction model used is “TD+sigmoid” model.

4. Section 4.3: Clarification about Baselines

The two baselines, “random flag” and “perfect flag”, use the same number of images from the captioning-training-set as our method for algorithm training. That count is determined by our predictor, specifically the number of images that remain after removing all images that are deemed to be unrecognizable. “Random flag” chooses a random sample from the captioning-training-set. “Perfect flag” chooses images based on a ranking of images based on how many crowdworkers flag the images as unrecognizable, with selection starting from those where all five crowdworkers agreed the image is unrecognizable.

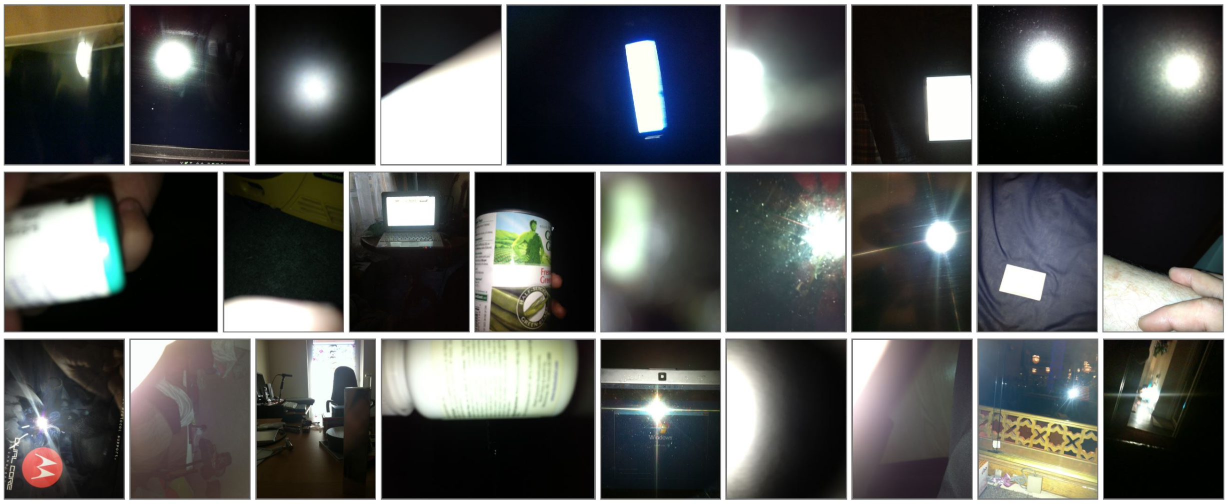

Figure 2: Examples of images that are both too dark and too bright. Note that both recognizable and unrecognizable images can appear here, since quality flaws do not necessarily render an image unrecognizable.

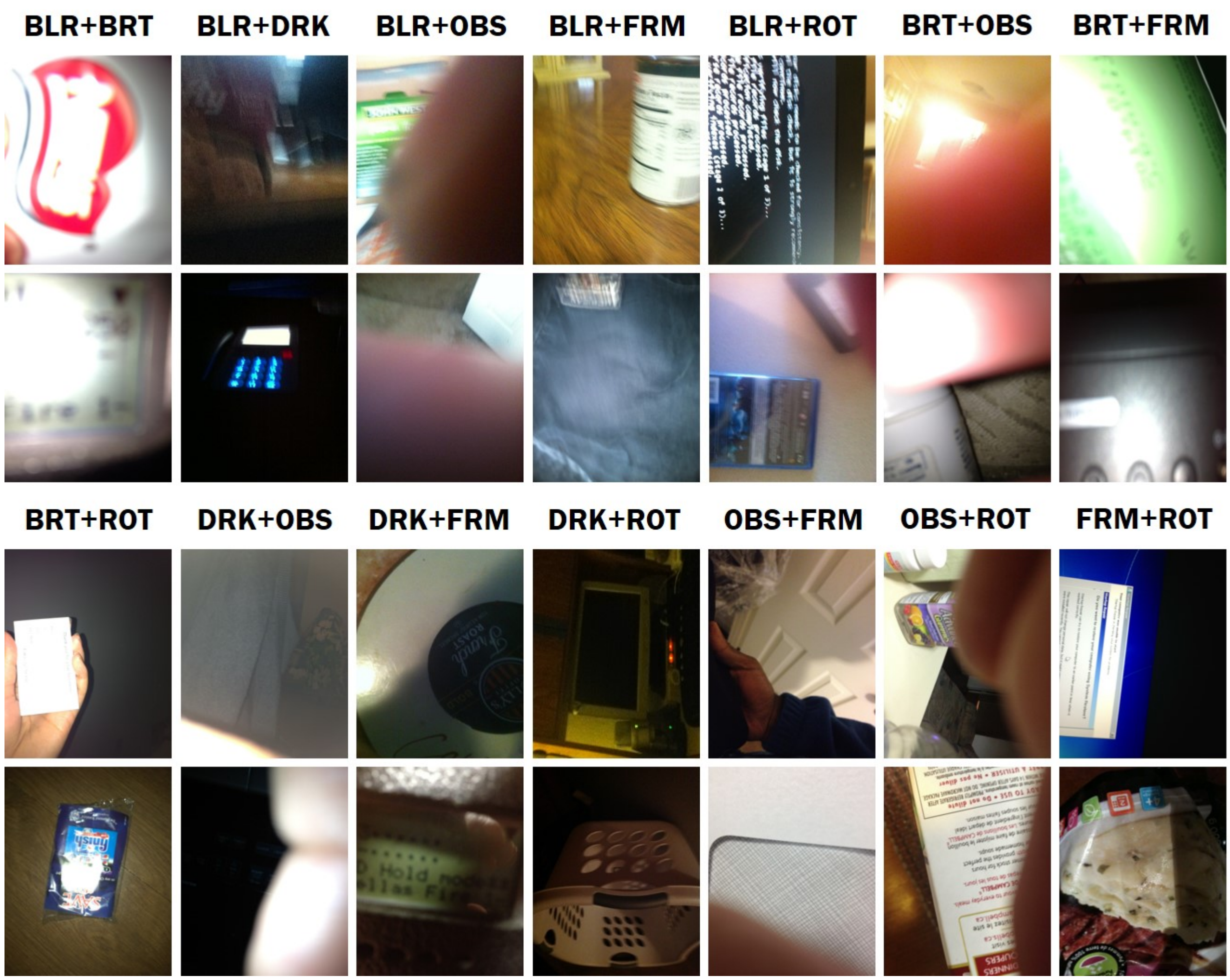

Figure 3: Examples of the co-occurrence of all quality flaw pairs. Again we obsere both recognizable and unrecognizable images appear since quality flaws do not necessarily render an image unrecognizable.

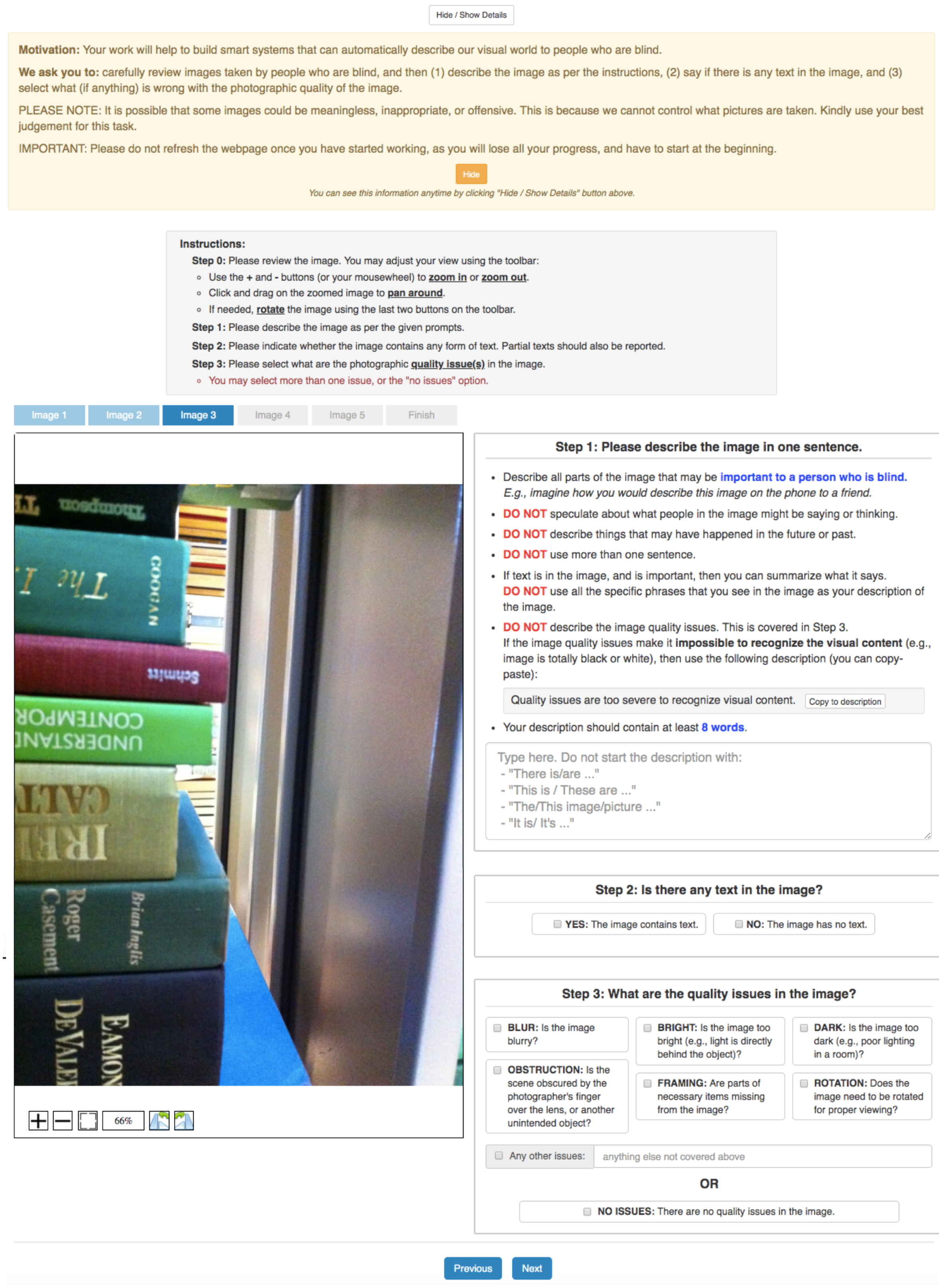

Figure 4: Interface used to crowdsource the collection of image captions.

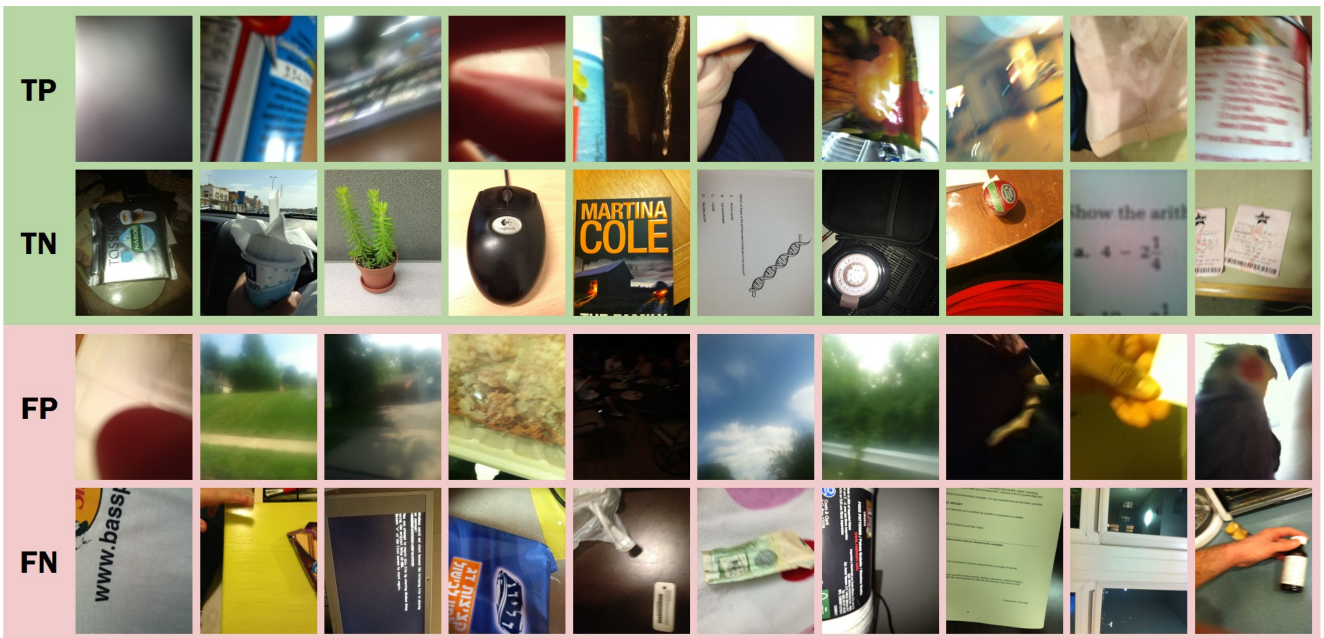

Figure 5: Examples of true-positives (TP), true-negatives (TN), false-positives (FP), and false-negatives (FN) in unrecognizability prediction. **TP**: unrecognizable images predicted to be unrecognizable. **TN**: recognizable images predicted to be recognizable. **FP**: recognizable images predicted to be unrecognizable. **FN**: unrecognizable images predicted to be recognizable.

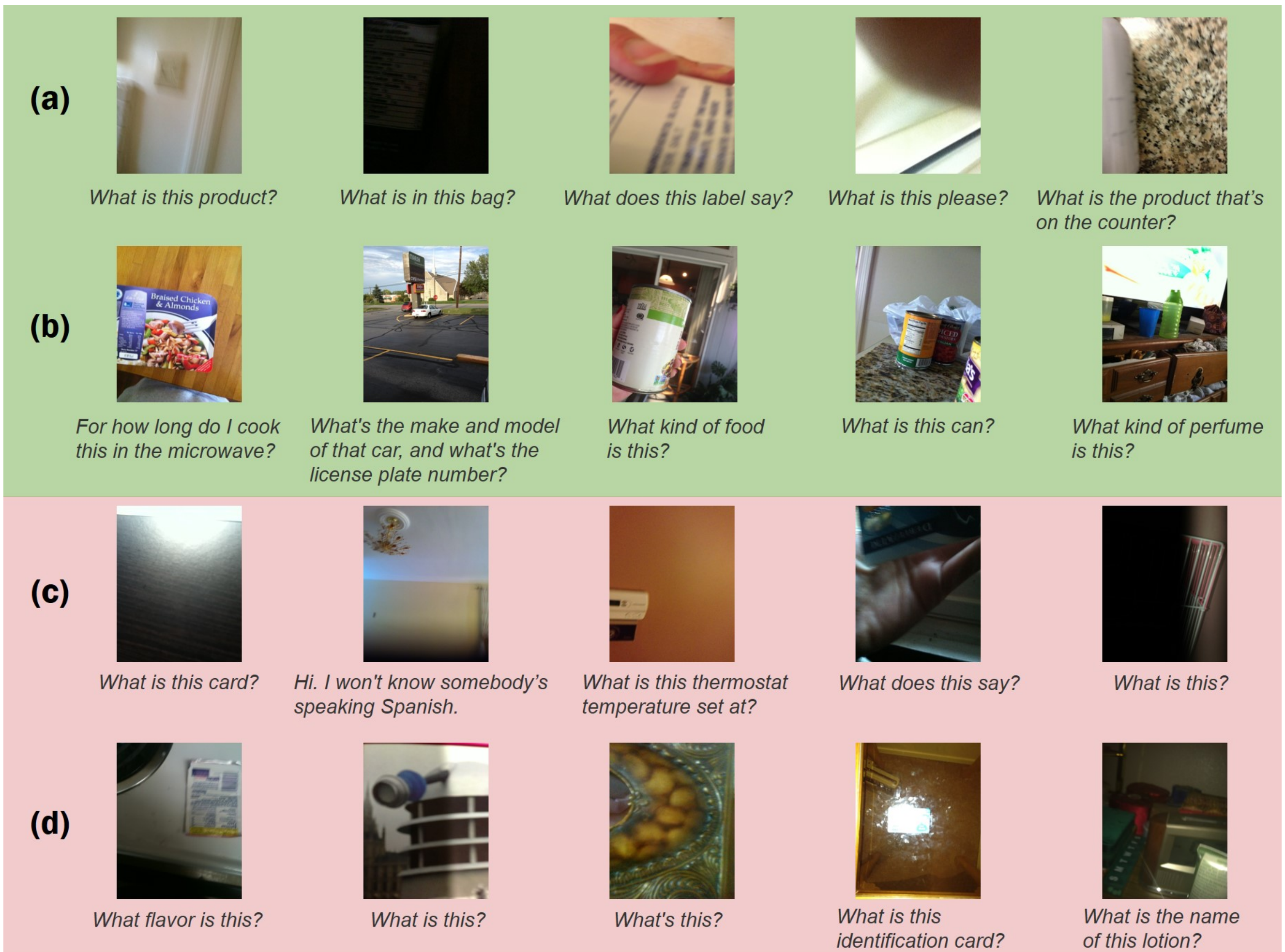

Figure 6: Prediction of the reason for unanswerable questions. Note that each visual question pair here is unanswerable. (a) Unanswerable questions are due to unrecognizable images and so are predictions. (b) Unanswerable questions are due to insufficient content and so are predictions. (c) Unanswerable questions are due to insufficient content but predicted to be due to unrecognizable images. (d) Unanswerable questions are due to unrecognizable images but predicted to be due to insufficient content.