Captioning Images Taken by People Who Are Blind - Supplementary Materials

Danna Gurari, Yinan Zhao, Meng Zhang, Nilavra Bhattacharya

University of Texas at Austin

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

- Implementation description of the crowdsourcing system (supplements **Section 3.1**)
- Analysis of the consistency of captions collected from different crowd workers for each image (supplements **Section 3.1**)
- Examples of images that are deemed insufficient or low quality for captioning (supplements **Section 3.2**)
- Visualizations and quantitative analysis demonstrating the diversity of content in VizWiz-Captions and how it compares to that in MSCOCO-Captions and the image classification datasets (supplements **Section 3.2**)
- Algorithm performance when using data augmentation during training by blurring images (supplements **Section 4**)

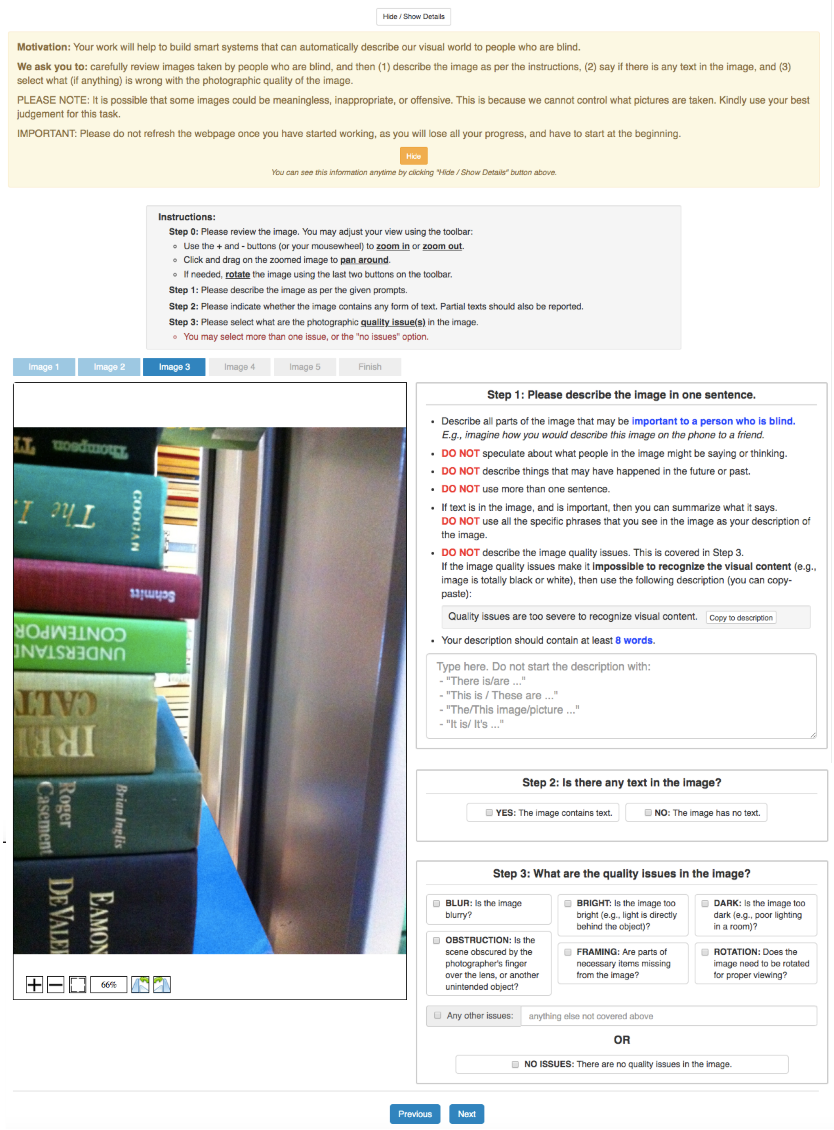
1 Dataset Creation (supplements Section 3.1)

1.1 Crowdsourcing Task Design

A screen shot of our crowdsourcing interface is shown in Figure 1. The interface prevented the crowdworker from proceeding to the next image (for the sequential set of five images) or submitting the work until the following criteria was met for each image description:

- Contains at least eight words (to encourage rich content)
- Contains only one period followed by a space (to restrict crowd worker to one-sentence descriptions)
- Sentences do not begin with the following prefixes (to discourage uninformative content): "There is", "There are", "This is", "These are", "The image", "The picture", "This image", "This picture", "It is", and "It's"

We collected all the annotations over five batches of Human Intelligence Tasks (HITs) in order to minimize the impact of inadequate workers. After each batch, we identified workers who we considered to be problematic and blocked them from participating in subsequent batches. To identify problematic workers, the authors reviewed a subset of the crowdworkers' results. The following mechanisms were used to determine which workers' captions to review:



Motivation: Your work will help to build smart systems that can automatically describe our visual world to people who are blind.

We ask you to: carefully review images taken by people who are blind, and then (1) describe the image as per the instructions, (2) say if there is any text in the image, and (3) select what (if anything) is wrong with the photographic quality of the image.

PLEASE NOTE: It is possible that some images could be meaningless, inappropriate, or offensive. This is because we cannot control what pictures are taken. Kindly use your best judgement for this task.

IMPORTANT: Please do not refresh the webpage once you have started working, as you will lose all your progress, and have to start at the beginning.

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

Instructions:

Step 0: Please review the image. You may adjust your view using the toolbar:

- Use the **+** and **-** buttons (or your mousewheel) to **zoom in** or **zoom out**.
- Click and drag on the zoomed image to **pan around**.
- If needed, **rotate** the image using the last two buttons on the toolbar.

Step 1: Please describe the image as per the given prompts.

Step 2: Please indicate whether the image contains any form of text. Partial texts should also be reported.

Step 3: Please select what are the photographic **quality issue(s)** in the image.

- You may select more than one issue, or the "no issues" option.

Step 1: Please describe the image in one sentence.

- Describe all parts of the image that may be **important to a person who is blind.** *E.g., imagine how you would describe this image on the phone to a friend.*
- **DO NOT** speculate about what people in the image might be saying or thinking.
- **DO NOT** describe things that may have happened in the future or past.
- **DO NOT** use more than one sentence.
- If text is in the image, and is important, then you can summarize what it says. **DO NOT** use all the specific phrases that you see in the image as your description of the image.
- **DO NOT** describe the image quality issues. This is covered in Step 3. If the image quality issues make it **impossible to recognize the visual content** (e.g., image is totally black or white), then use the following description (you can copy-paste):

Quality issues are too severe to recognize visual content.

- Your description should contain at least **8 words**.

Type here. Do not start the description with:
- "There is/are ..."
- "This is / These are ..."
- "The/This image/picture ..."
- "It is/ It's ..."

Step 2: Is there any text in the image?

□ **YES:** The image contains text. □ **NO:** The image has no text.

Step 3: What are the quality issues in the image?

□ **BLUR:** Is the image blurry?

□ **BRIGHT:** Is the image too bright (e.g., light is directly behind the object)?

□ **DARK:** Is the image too dark (e.g., poor lighting in a room)?

□ **OBSTRUCTION:** Is the scene obscured by the photographer's finger over the lens, or another unintended object?

□ **FRAMING:** Are parts of necessary items missing from the image?

□ **ROTATION:** Does the image need to be rotated for proper viewing?

□ Any other issues: anything else not covered above

OR

□ **NO ISSUES:** There are no quality issues in the image.

Fig. 1: Interface used to crowdsource the collection of image captions.

– Workers who were a statistical outlier in time-to-submit (taking either too little or too much time) by 1.95 times the standard deviation for all the results
– Workers who used CAPS LOCK for more than 50% of the caption text
– Workers who used the canned text (”Quality issues are too severe to recognize visual content”) for more than 50% of the images that they captioned

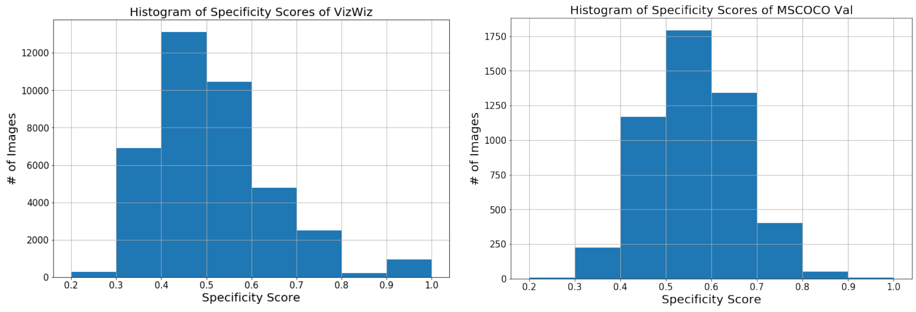
- Workers who were the only one to either use or not use the canned text (”Quality issues are too severe to recognize visual content”) for an image
- Workers who used words like ”quality”, ”blur” and ”blurry” (but not the canned text), and so were not focusing on content in the image
- Random sample from all results

We also included numerous additional quality control mechanisms. First, crowdworkers could not submit their results until their work passed an automated check that verified they followed a number of the task instructions, including writing at least 8 words, providing only one sentence, and not starting the description with “There is...” or other unsubstantial starting phrases. We also only accepted crowdworkers who previously had completed over 500 HITs with at least a 95% acceptance rate. We will publicly-share the crowdsourcing code to support reproducibility of this interface.

1.2 Caption Post-processing

We employed Microsoft Azure’s spell-checking API[1] to find and correct misspelled words in the submitted captions. We chose this approach because we found from initial testing that it outperforms other tested methods, including because it can recognize brand names (which are common in our dataset). It also does a good job of correcting grammar and capitalizing words when appropriate (e.g. changing ”dell” to ”Dell”). When spell-checking all captions which are neither canned text nor from blocked workers (i.e., 169,073 captions), 14% (i.e, 23,424) were flagged as containing unknown ”tokens” (aka - words). We replace each unknown “token” with the most confidently recommended word suggested by the Azure API.

[1] https://azure.microsoft.com/en-us/services/cognitive-services/spell-check/



(a) VizWiz-Captions

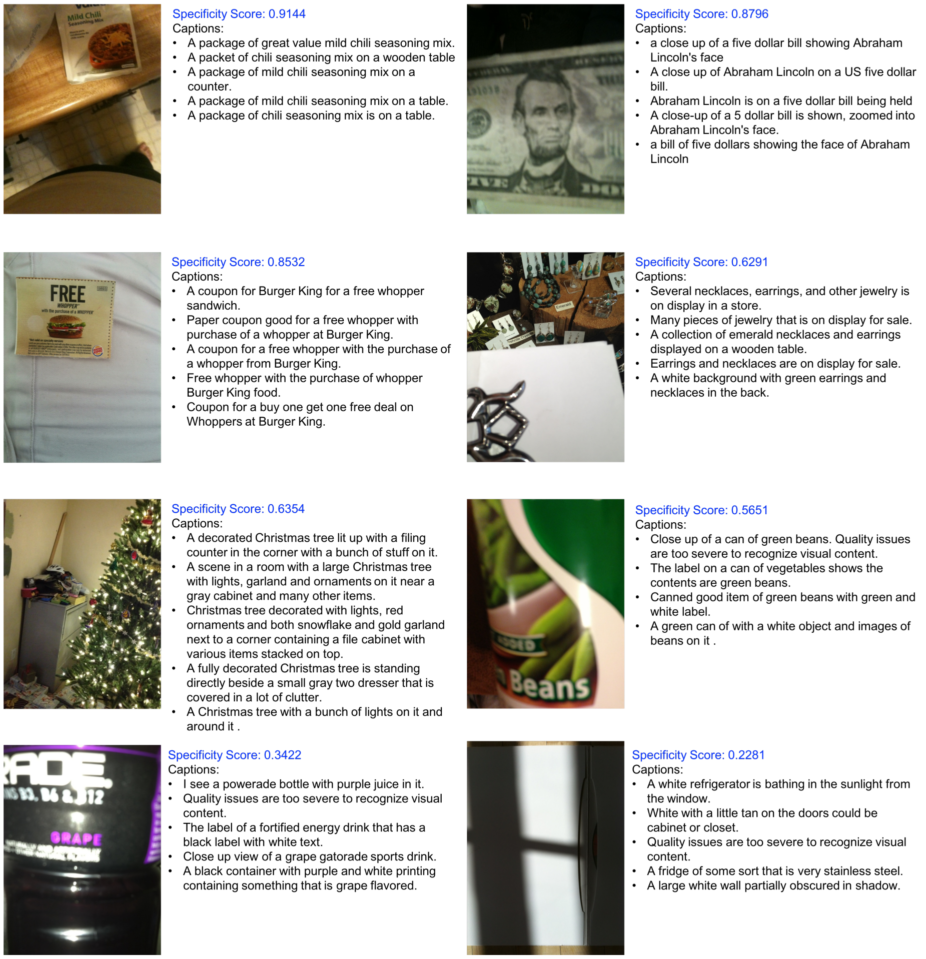
(b) MSCOCO-Captions

Fig. 2: Distribution of image specificity scores for images in (a) VizWiz-Captions and (b) MSCOCO-Captions.

2 Caption Consistency (supplements Section 3.2)

We report the distributions of specificity scores that indicate the similarity between the five captions per image generated by different humans for all images in our VizWiz-Captions dataset and the MSCOCO-Captions validation set independently. Scores range between 0 and 1, with numbers closer to 1 indicating greater consistency between the five captions per image. Results are shown in Figure 2.

While the distributions are similar overall, we observe scores are skewed more towards 0 for VizWiz-Captions. We attribute these slightly greater annotation differences to annotators providing different level of detail and different types



Specificity Score: 0.9144
Captions:
- A package of great value mild chili seasoning mix.
- A packet of chili seasoning mix on a wooden table
- A package of mild chili seasoning mix on a counter.
- A package of mild chili seasoning mix on a table.
- A package of chili seasoning mix is on a table.

Specificity Score: 0.8796
Captions:
- a close up of a five dollar bill showing Abraham Lincoln's face
- A close up of Abraham Lincoln on a US five dollar bill.
- Abraham Lincoln is on a five dollar bill being held
- A close-up of a 5 dollar bill is shown, zoomed into Abraham Lincoln's face.
- a bill of five dollars showing the face of Abraham Lincoln

Specificity Score: 0.8532
Captions:
- A coupon for Burger King for a free whopper sandwich.
- Paper coupon good for a free whopper with purchase of a whopper at Burger King.
- A coupon for a free whopper with the purchase of a whopper from Burger King.
- Free whopper with the purchase of whopper Burger King food.
- Coupon for a buy one get one free deal on Whoppers at Burger King.

Specificity Score: 0.6291
Captions:
- Several necklaces, earrings, and other jewelry is on display in a store.
- Many pieces of jewelry that is on display for sale.
- A collection of emerald necklaces and earrings displayed on a wooden table.
- Earrings and necklaces are on display for sale.
- A white background with green earrings and necklaces in the back.

Specificity Score: 0.6354
Captions:
- A decorated Christmas tree lit up with a filing counter in the corner with a bunch of stuff on it.
- A scene in a room with a large Christmas tree with lights, garland and ornaments on it near a gray cabinet and many other items.
- Christmas tree decorated with lights, red ornaments and both snowflake and gold garland next to a corner containing a file cabinet with various items stacked on top.
- A fully decorated Christmas tree is standing directly beside a small gray two dresser that is covered in a lot of clutter.
- A Christmas tree with a bunch of lights on it and around it .

Specificity Score: 0.5651
Captions:
- Close up of a can of green beans. Quality issues are too severe to recognize visual content.
- The label on a can of vegetables shows the contents are green beans.
- Canned good item of green beans with green and white label.
- A green can of with a white object and images of beans on it .

Specificity Score: 0.3422
Captions:
- I see a powerade bottle with purple juice in it.
- Quality issues are too severe to recognize visual content.
- The label of a fortified energy drink that has a black label with white text.
- Close up view of a grape gatorade sports drink.
- A black container with purple and white printing containing something that is grape flavored.

Specificity Score: 0.2281
Captions:
- A white refrigerator is bathing in the sunlight from the window.
- White with a little tan on the doors could be cabinet or closet.
- Quality issues are too severe to recognize visual content.
- A fridge of some sort that is very stainless steel.
- A large white wall partially obscured in shadow.

Fig. 3: Examples of images that lead to a range of specificity scores when analyzing the captions collected from different crowdworkers.

of detail, as exemplified in Figure 3 in the main paper. We show examples of the diversity of captions that arise from different crowdworkers for a range of specificity scores in Figure 3.

3 Dataset Analysis (supplements Section 3.2)

3.1 Insufficient Quality Images

Figure 4 exemplifies images that were deemed insufficient or lower quality for captioning based on the agreement of five crowdworkers.

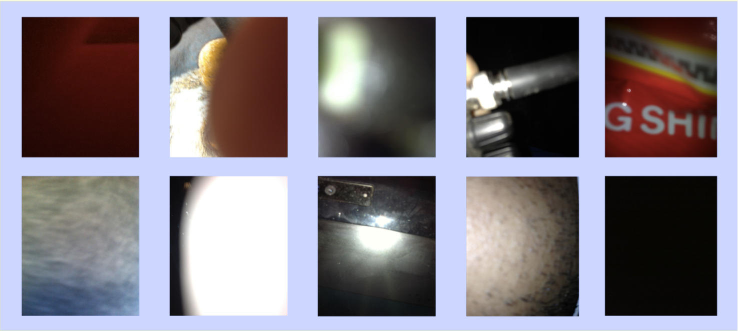
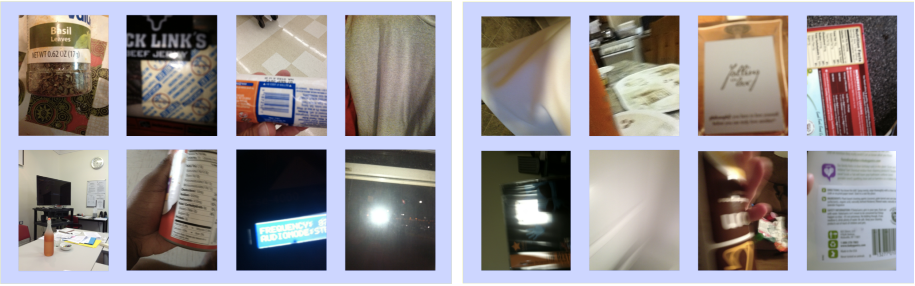


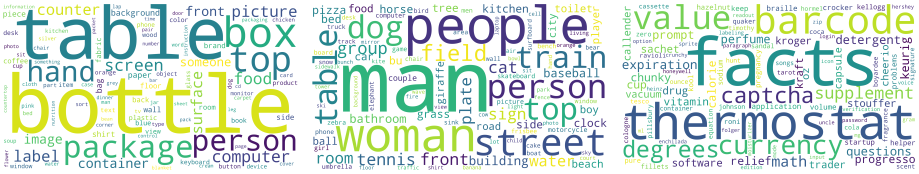
Fig. 4: Examples of images that are unanimously labeled as insufficient quality to be captioned.

We show examples of images that we deem medium or high difficulty based on the agreement of five crowdworkers who indicated the image is insufficient quality to generate a meaningful caption (i.e., 1-2 for medium and 3-4 for high difficulty) in Figure 5.



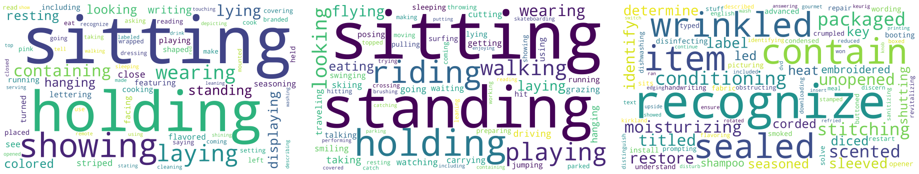
(a) Medium difficulty for image captioning (b) High difficulty for image captioning

Fig. 5: Examples of images that are labeled as insufficient quality by (a) 1-2 crowdworkers and (b) 3-4 crowdworkers.



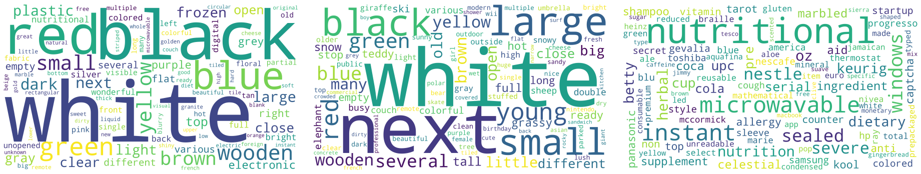
(a) VizWiz-Captions (b) MSCOCO-Captions (c) VizWiz-Captions only

Fig. 6: Wordclouds for the most popular 100 nouns across all captions that are in (a) VizWiz-Captions, (b) MSCOCO-Captions, and (c) in VizWiz-Captions but not in MSCOCO-Captions.



(a) VizWiz-Captions (b) MSCOCO-Captions (c) VizWiz-Captions only

Fig. 7: Wordclouds for the most popular 100 verbs across all captions that are in (a) VizWiz-Captions, (b) MSCOCO-Captions, and (c) in VizWiz-Captions but not in MSCOCO-Captions.



(a) VizWiz-Captions (b) MSCOCO-Captions (c) VizWiz-Captions only

Fig. 8: Wordclouds for the most popular 100 adjectives across all captions that are in (a) VizWiz-Captions, (b) MSCOCO-Captions, and (c) in VizWiz-Captions but not in MSCOCO-Captions.

3.2 Caption Characterization

We visualize the most popular words included in the captions for each of the following word types analyzed in Table 1 of the main paper: nouns, verbs, and adjectives. We do so both for VizWiz-Captions and MSCOCO-Captions to support comparison to today's mainstream captioning dataset. For each word type we show the most common 100 words in VizWiz-Captions and MSCOCO-Captions separately as well as the most common 100 words that are in VizWiz-Captions but not found in MSCOCO-Captions. Results for nouns, verbs, and adjectives are shown in Figures 6, 7, and 8 respectively. We observe in Figure 6 that many popular words focus on items from daily living such as 'table', 'person', 'box', 'food', and 'monitor'. Nouns that are absent from MSCOCO-Captions and popular in VizWiz-Captions largely focus on text and numbers (Figure 6c), including 'expiration', 'captcha', 'thermostat', 'password', and 'currency'.

We also quantify the extent to which people are present in the images, given the high prevalence in many mainstream vision datasets. When tallying how many images are captioned using words related to people (i.e., people, person, man, woman, child, hand, foot, torso), the result is 27.6% (i.e., 10,805/39,181). When applying a person detector [2][2], people are detected for 29.4% (i.e., 11,499/39,181) of images. We suspect this latter result is slightly larger than observed for captions because of person detections on background objects, such as newspaper or TV screens. We suspect crowdworkers found such person detections to be insufficiently salient to be described as part of the captions. Altogether, the relatively low prevalence of humans may in part be attributed to the fact that any images showing people's faces were filtered from the publicly-shared dataset to preserve privacy. We suspect that the presence of people in our VizWiz-Captions compared to popular vision datasets will differ in that either (1) only parts of people appear in our images, such as only hands, legs, and torsos or (2) when the full body is shown, often it is because the person is on the cover of media (book, magazine, cd, dvd) or a product box.

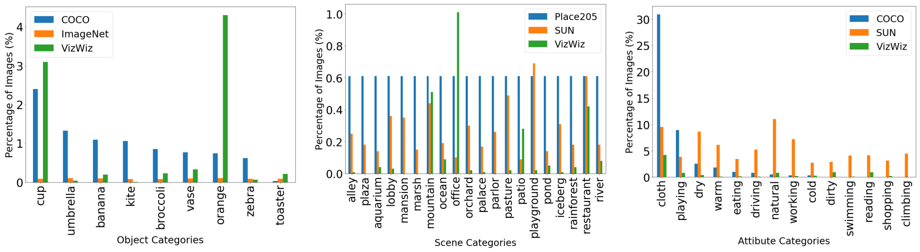
We next report the percentage of overlap between the most common 3,000 words in VizWiz-Captions and the most common 3,000 words in MSCOCO-Captions for all words as well as with respect to each of the following word types: nounds, adjectives, and verbs. Results are shown in Table 1. The higher percentage across all words than for the different word types is likely because there are many common stopwords that are shared across both datasets that do not belong to each word type.

words	nouns	adj	verbs
54.4%	45.1%	31.6%	42.8%

Table 1: Percentage of overlap between most common 3,000 words in VizWiz-Captions and the most common 3,000 words in MSCOCO-Captions.

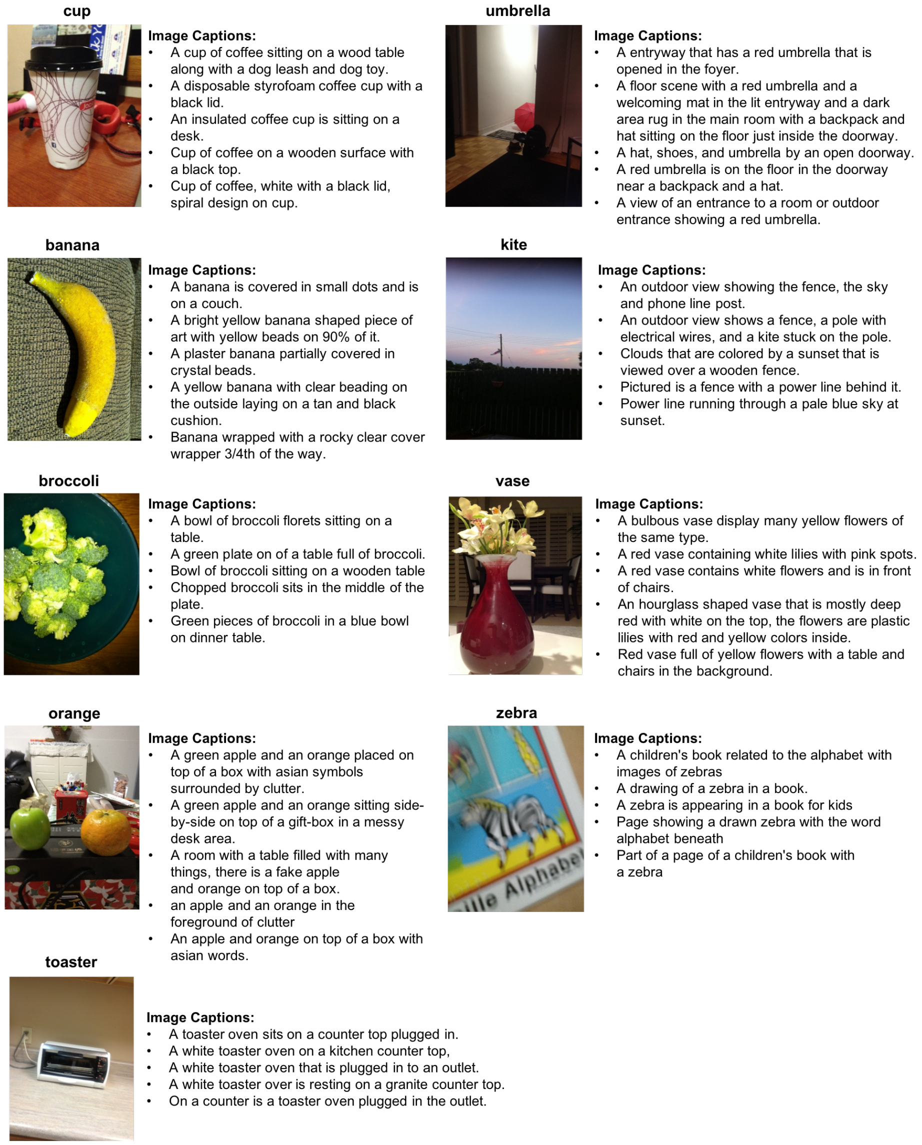
[2] We employed a faster-rcnn model pretrained on COCO. It has a ResNeXt-101 as the backbone and Feature Pyramid Network (FPN) to deal with objects of different scales. We only used the "person" category out of the 80 categories. We filtered the detections with a threshold of 0.3, meaning we only counted a detection as valid if the confidence score is above 0.3.

Finally, we provide histograms showing the relative prevalence of categories found in the mainstream computer vision datasets versus our dataset for three image classification tasks: recognizing objects, scenes, and attributes. Results are shown in Figure 9, complementing those shown in the main paper in Figure 2. Exemplar images in our dataset that show concepts overlapping with those in the mainstream computer vision datasets are shown in Figures 10, 11, and 12.



(a) (b) (c)

Fig. 9: Parallel results to those in Figure 2 of the main paper, showing the fraction of all images in our VizWiz-Captions and popular vision datasets that contain each category for the following vision problems: (a) object recognition, (b) scene classification, and (c) attribute recognition.



cup

Image Captions:
- A cup of coffee sitting on a wood table along with a dog leash and dog toy.
- A disposable styrofoam coffee cup with a black lid.
- An insulated coffee cup is sitting on a desk.
- Cup of coffee on a wooden surface with a black top.
- Cup of coffee, white with a black lid, spiral design on cup.

umbrella

Image Captions:
- A entryway that has a red umbrella that is opened in the foyer.
- A floor scene with a red umbrella and a welcoming mat in the lit entryway and a dark area rug in the main room with a backpack and hat sitting on the floor just inside the doorway.
- A hat, shoes, and umbrella by an open doorway.
- A red umbrella is on the floor in the doorway near a backpack and a hat.
- A view of an entrance to a room or outdoor entrance showing a red umbrella.

banana

Image Captions:
- A banana is covered in small dots and is on a couch.
- A bright yellow banana shaped piece of art with yellow beads on 90% of it.
- A plaster banana partially covered in crystal beads.
- A yellow banana with clear beading on the outside laying on a tan and black cushion.
- Banana wrapped with a rocky clear cover wrapper 3/4th of the way.

kite

Image Captions:
- An outdoor view showing the fence, the sky and phone line post.
- An outdoor view shows a fence, a pole with electrical wires, and a kite stuck on the pole.
- Clouds that are colored by a sunset that is viewed over a wooden fence.
- Pictured is a fence with a power line behind it.
- Power line running through a pale blue sky at sunset.

broccoli

Image Captions:
- A bowl of broccoli florets sitting on a table.
- A green plate on of a table full of broccoli.
- Bowl of broccoli sitting on a wooden table
- Chopped broccoli sits in the middle of the plate.
- Green pieces of broccoli in a blue bowl on dinner table.

vase

Image Captions:
- A bulbous vase display many yellow flowers of the same type.
- A red vase containing white lilies with pink spots.
- A red vase contains white flowers and is in front of chairs.
- An hourglass shaped vase that is mostly deep red with white on the top, the flowers are plastic lilies with red and yellow colors inside.
- Red vase full of yellow flowers with a table and chairs in the background.

orange

Image Captions:
- A green apple and an orange placed on top of a box with asian symbols surrounded by clutter.
- A green apple and an orange sitting side-by-side on top of a gift-box in a messy desk area.
- A room with a table filled with many things, there is a fake apple and orange on top of a box.
- an apple and an orange in the foreground of clutter
- An apple and orange on top of a box with asian words.

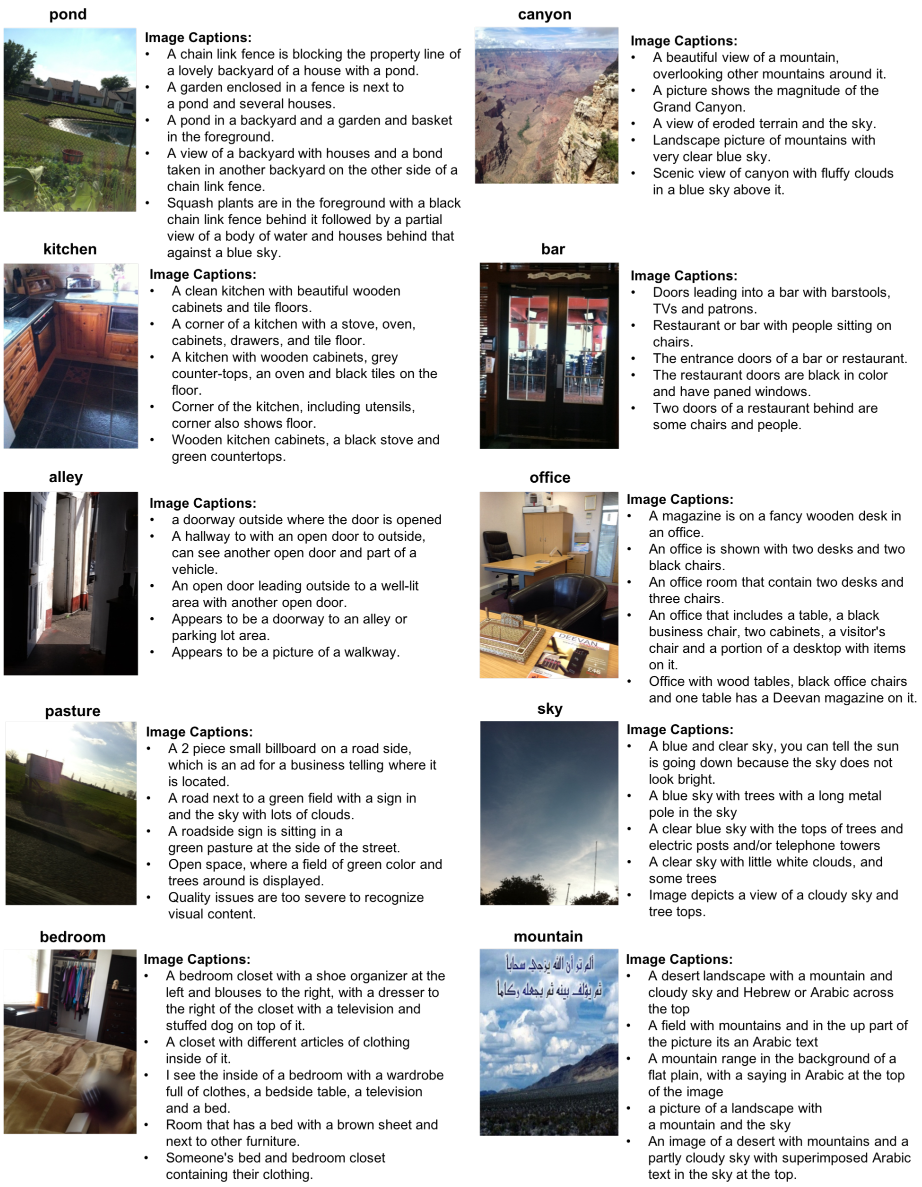
zebra

Image Captions:
- A children's book related to the alphabet with images of zebras
- A drawing of a zebra in a book.
- A zebra is appearing in a book for kids
- Page showing a drawn zebra with the word alphabet beneath
- Part of a page of a children's book with a zebra

toaster

Image Captions:
- A toaster oven sits on a counter top plugged in.
- A white toaster oven on a kitchen counter top,
- A white toaster oven that is plugged in to an outlet.
- A white toaster over is resting on a granite counter top.
- On a counter is a toaster oven plugged in the outlet.

Fig. 10: Examples of images showing visual concepts that are common in existing computer vision datasets for the object recognition task.



pond

Image Captions:
- A chain link fence is blocking the property line of a lovely backyard of a house with a pond.
- A garden enclosed in a fence is next to a pond and several houses.
- A pond in a backyard and a garden and basket in the foreground.
- A view of a backyard with houses and a bond taken in another backyard on the other side of a chain link fence.
- Squash plants are in the foreground with a black chain link fence behind it followed by a partial view of a body of water and houses behind that against a blue sky.

canyon

Image Captions:
- A beautiful view of a mountain, overlooking other mountains around it.
- A picture shows the magnitude of the Grand Canyon.
- A view of eroded terrain and the sky.
- Landscape picture of mountains with very clear blue sky.
- Scenic view of canyon with fluffy clouds in a blue sky above it.

kitchen

Image Captions:
- A clean kitchen with beautiful wooden cabinets and tile floors.
- A corner of a kitchen with a stove, oven, cabinets, drawers, and tile floor.
- A kitchen with wooden cabinets, grey counter-tops, an oven and black tiles on the floor.
- Corner of the kitchen, including utensils, corner also shows floor.
- Wooden kitchen cabinets, a black stove and green countertops.

bar

Image Captions:
- Doors leading into a bar with barstools, TVs and patrons.
- Restaurant or bar with people sitting on chairs.
- The entrance doors of a bar or restaurant.
- The restaurant doors are black in color and have paned windows.
- Two doors of a restaurant behind are some chairs and people.

alley

Image Captions:
- a doorway outside where the door is opened
- A hallway to with an open door to outside, can see another open door and part of a vehicle.
- An open door leading outside to a well-lit area with another open door.
- Appears to be a doorway to an alley or parking lot area.
- Appears to be a picture of a walkway.

office

Image Captions:
- A magazine is on a fancy wooden desk in an office.
- An office is shown with two desks and two black chairs.
- An office room that contain two desks and three chairs.
- An office that includes a table, a black business chair, two cabinets, a visitor's chair and a portion of a desktop with items on it.
- Office with wood tables, black office chairs and one table has a Deevan magazine on it.

pasture

Image Captions:
- A 2 piece small billboard on a road side, which is an ad for a business telling where it is located.
- A road next to a green field with a sign in and the sky with lots of clouds.
- A roadside sign is sitting in a green pasture at the side of the street.
- Open space, where a field of green color and trees around is displayed.
- Quality issues are too severe to recognize visual content.

sky

Image Captions:
- A blue and clear sky, you can tell the sun is going down because the sky does not look bright.
- A blue sky with trees with a long metal pole in the sky
- A clear blue sky with the tops of trees and electric posts and/or telephone towers
- A clear sky with little white clouds, and some trees
- Image depicts a view of a cloudy sky and tree tops.

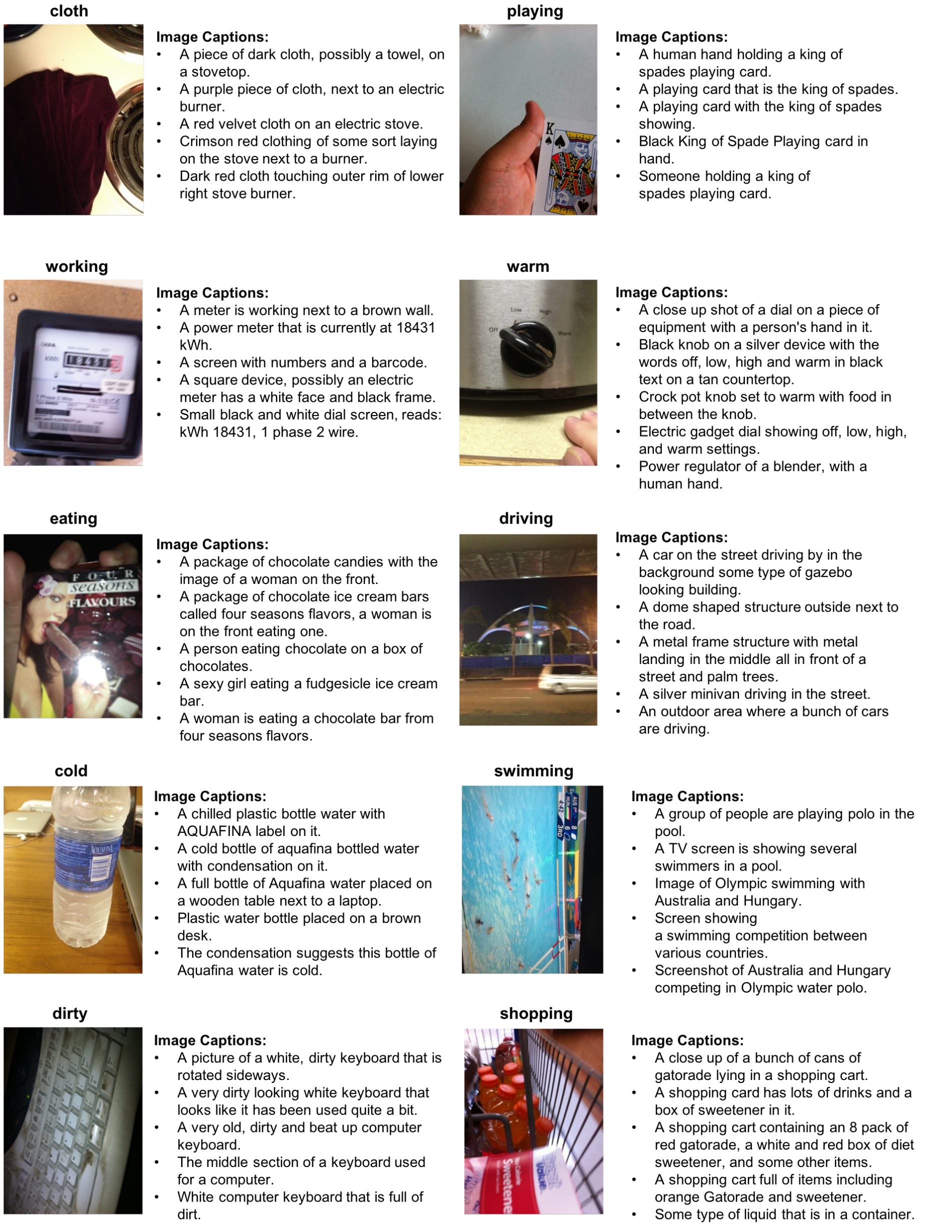
bedroom

Image Captions:
- A bedroom closet with a shoe organizer at the left and blouses to the right, with a dresser to the right of the closet with a television and stuffed dog on top of it.
- A closet with different articles of clothing inside of it.
- I see the inside of a bedroom with a wardrobe full of clothes, a bedside table, a television and a bed.
- Room that has a bed with a brown sheet and next to other furniture.
- Someone's bed and bedroom closet containing their clothing.

mountain

Image Captions:
- A desert landscape with a mountain and cloudy sky and Hebrew or Arabic across the top
- A field with mountains and in the up part of the picture its an Arabic text
- A mountain range in the background of a flat plain, with a saying in Arabic at the top of the image
- a picture of a landscape with a mountain and the sky
- An image of a desert with mountains and a partly cloudy sky with superimposed Arabic text in the sky at the top.

Fig. 11: Examples of images showing visual concepts that are common in existing computer vision datasets for the scene recognition task.



cloth

Image Captions:
- A piece of dark cloth, possibly a towel, on a stovetop.
- A purple piece of cloth, next to an electric burner.
- A red velvet cloth on an electric stove.
- Crimson red clothing of some sort laying on the stove next to a burner.
- Dark red cloth touching outer rim of lower right stove burner.

playing

Image Captions:
- A human hand holding a king of spades playing card.
- A playing card that is the king of spades.
- A playing card with the king of spades showing.
- Black King of Spade Playing card in hand.
- Someone holding a king of spades playing card.

working

Image Captions:
- A meter is working next to a brown wall.
- A power meter that is currently at 18431 kWh.
- A screen with numbers and a barcode.
- A square device, possibly an electric meter has a white face and black frame.
- Small black and white dial screen, reads: kWh 18431, 1 phase 2 wire.

warm

Image Captions:
- A close up shot of a dial on a piece of equipment with a person's hand in it.
- Black knob on a silver device with the words off, low, high and warm in black text on a tan countertop.
- Crock pot knob set to warm with food in between the knob.
- Electric gadget dial showing off, low, high, and warm settings.
- Power regulator of a blender, with a human hand.

eating

Image Captions:
- A package of chocolate candies with the image of a woman on the front.
- A package of chocolate ice cream bars called four seasons flavors, a woman is on the front eating one.
- A person eating chocolate on a box of chocolates.
- A sexy girl eating a fudgesicle ice cream bar.
- A woman is eating a chocolate bar from four seasons flavors.

driving

Image Captions:
- A car on the street driving by in the background some type of gazebo looking building.
- A dome shaped structure outside next to the road.
- A metal frame structure with metal landing in the middle all in front of a street and palm trees.
- A silver minivan driving in the street.
- An outdoor area where a bunch of cars are driving.

cold

Image Captions:
- A chilled plastic bottle water with AQUAFINA label on it.
- A cold bottle of aquafina bottled water with condensation on it.
- A full bottle of Aquafina water placed on a wooden table next to a laptop.
- Plastic water bottle placed on a brown desk.
- The condensation suggests this bottle of Aquafina water is cold.

swimming

Image Captions:
- A group of people are playing polo in the pool.
- A TV screen is showing several swimmers in a pool.
- Image of Olympic swimming with Australia and Hungary.
- Screen showing a swimming competition between various countries.
- Screenshot of Australia and Hungary competing in Olympic water polo.

dirty

Image Captions:
- A picture of a white, dirty keyboard that is rotated sideways.
- A very dirty looking white keyboard that looks like it has been used quite a bit.
- A very old, dirty and beat up computer keyboard.
- The middle section of a keyboard used for a computer.
- White computer keyboard that is full of dirt.

shopping

Image Captions:
- A close up of a bunch of cans of gatorade lying in a shopping cart.
- A shopping card has lots of drinks and a box of sweetener in it.
- A shopping cart containing an 8 pack of red gatorade, a white and red box of diet sweetener, and some other items.
- A shopping cart full of items including orange Gatorade and sweetener.
- Some type of liquid that is in a container.

Fig. 12: Examples of images showing visual concepts that are common in existing computer vision datasets for the attribute classification task.

4 Algorithm Performance with Data Augmentation (supplements Section 4)

Given the need for improved algorithm performance for low quality images, we examined the potential for using data augmentation during training to help the models better cope with low quality images at test time. For this analysis, we chose the AoANet algorithm since it has the best performance when training from scratch. We re-trained this algorithm from scratch again using VizWiz-Captions, but this time augmented a copy of all the training images after blurring them using a 15x15 averaging filter kernel on all the training images.

Results are shown in Table 3. We observe that the performance is worse with data augmentation; e.g., with respect to the CIDEr score, overall performance falls from 60.5 to 56.2. We suspect this performance drop is because artificial image distortions are unsuitable for mimicking real-world quality issues [1], and thus distract from model training.

		B@1	**B@2**	**B@3**	**B@4**	**METEOR**	**ROUGE**	**CIDEr**	**SPICE**
AOANet	All	66.4	47.0	32.3	22.1	20.0	46.5	56.2	13.8
	Easy	69.7	50.6	35.5	24.6	21.2	49.1	60.3	14.2
	Medium	63.3	42.7	27.7	18.2	18.5	43.2	50.8	13.6
	Difficult	36.3	19.8	11.0	6.4	12.3	29.7	40.3	11.3

Table 2: Analysis of the top-performing image captioning algorithm when trained from scratch using data augmentation of blurred images. Results are shown on all test images as well as only the subsets deemed easy, medium and difficult. (B@ = BLEU-)

We also report human performance based on the same set of evaluation metrics. To do so, we only consider images with five valid captions (i.e., "easy" images). We randomly choose one caption per image as the prediction, and use the remaining four captions for evaluation. Results are shown in Table 3.

	B@1	**B@2**	**B@3**	**B@4**	**METEOR**	**ROUGE**	**CIDEr**	**SPICE**
Easy	60.3	40.7	27.7	18.8	22.0	43.4	83.5	17.5

Table 3: Human performance on all test images deemed easy. (B@ = BLEU-)

References

1. T.-Y. Chiu, Y. Zhao, and D. Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656, 2020.
2. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.