Supplementary Material: Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies

In this supplementary material we study additional quantitative (see Sec. 1) and qualitative (see Sec. 2) results. We detail our experimental settings (see Sec. 3). Finally, we provide the code we use to perform the experiments (see Sec. 4).

1 Quantitative Evaluation

To study robustness, in Tab. S1, we provide analysis for an ensemble of different models on VQA-CPv2. For each model, we used different values of $\lambda$, and different regularization, i.e., Eq. (17) and Eq. (18). Using our Top-3 ensemble, we improved the accuracy by ~2% compared to our single model (56.74% vs. 54.55%), and ~4% compared to the previous single model baseline (56.74% vs. 52.05%).

In Tab. S2 we show a per question type comparison on VQA-CPv2. Our method improved significantly the “how many people are” question type that requires counting (19.98% vs. 42.81%). However, for the “is there” question type which requires image detection, our method’s performance dropped in a 19.6% accuracy (22.4% vs. 42%).

Table S1: A comparison of different ensemble models on VQA-CPv2. We underline the best single-model and bold the best ensemble model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>Answer type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes/No</td>
</tr>
<tr>
<td>Eq.(17)</td>
<td>54.01</td>
<td>73.02</td>
</tr>
<tr>
<td>Eq.(18)</td>
<td>54.55</td>
<td>74.03</td>
</tr>
<tr>
<td>Top-3</td>
<td>56.74</td>
<td>77.70</td>
</tr>
<tr>
<td>Top-4</td>
<td>56.41</td>
<td>77.34</td>
</tr>
<tr>
<td>Top-5</td>
<td>56.54</td>
<td><strong>77.48</strong></td>
</tr>
</tbody>
</table>

Table S2: Comparison between VQA-CPv2 models in the paper and an ensemble of three models. We show the differences between the state-of-the-art [1] and our method in parenthesis.

<table>
<thead>
<tr>
<th>Question-type</th>
<th>LMH</th>
<th>Eq.(18)</th>
<th>Eq.(17)</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>are there</td>
<td>35.4</td>
<td>28.89 (-6.51)</td>
<td>32.22 (-3.18)</td>
<td>32.86 (-2.54)</td>
</tr>
<tr>
<td>what brand</td>
<td>26.75</td>
<td>27.35 (0.6)</td>
<td>21.84 (-4.91)</td>
<td>22.76 (-3.99)</td>
</tr>
<tr>
<td>what room is</td>
<td>92.57</td>
<td>93.22 (0.65)</td>
<td>92.66 (0.09)</td>
<td>93.95 (1.38)</td>
</tr>
<tr>
<td>what color is</td>
<td>77.18</td>
<td>75.53 (-1.65)</td>
<td>78.03 (0.85)</td>
<td>78.35 (1.17)</td>
</tr>
<tr>
<td>is</td>
<td>77.21</td>
<td>69.78 (-7.43)</td>
<td>78.74 (1.53)</td>
<td>75.13 (-2.08)</td>
</tr>
<tr>
<td>are they</td>
<td>66.78</td>
<td>62.0 (-4.78)</td>
<td>61.67 (-5.11)</td>
<td>67.33 (0.55)</td>
</tr>
</tbody>
</table>

*itaigat@technion.ac.il

<table>
<thead>
<tr>
<th>Question</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>what number is</td>
<td>3.28</td>
<td>4.22</td>
<td>3.07</td>
<td>2.88</td>
</tr>
<tr>
<td>what does the</td>
<td>17.67</td>
<td>17.54</td>
<td>17.49</td>
<td>18.33</td>
</tr>
<tr>
<td>is this person</td>
<td>73.15</td>
<td>68.21</td>
<td>74.89</td>
<td>71.87</td>
</tr>
<tr>
<td>is the</td>
<td>36.83</td>
<td>40.32</td>
<td>36.61</td>
<td>37.91</td>
</tr>
<tr>
<td>what is the man</td>
<td>56.14</td>
<td>58.22</td>
<td>57.21</td>
<td>58.53</td>
</tr>
<tr>
<td>what kind of</td>
<td>42.83</td>
<td>45.7</td>
<td>44.26</td>
<td>45.75</td>
</tr>
<tr>
<td>does this</td>
<td>68.91</td>
<td>60.54</td>
<td>65.7</td>
<td>62.23</td>
</tr>
<tr>
<td>is there a</td>
<td>40.38</td>
<td>46.92</td>
<td>49.62</td>
<td>52.31</td>
</tr>
<tr>
<td>is he</td>
<td>78.14</td>
<td>77.3</td>
<td>83.11</td>
<td>78.97</td>
</tr>
<tr>
<td>what</td>
<td>34.95</td>
<td>36.75</td>
<td>35.03</td>
<td>36.19</td>
</tr>
<tr>
<td>does the</td>
<td>72.75</td>
<td>83.36</td>
<td>75.84</td>
<td>81.49</td>
</tr>
<tr>
<td>is the person</td>
<td>72.8</td>
<td>78.0</td>
<td>68.96</td>
<td>79.24</td>
</tr>
<tr>
<td>where is the</td>
<td>26.15</td>
<td>28.07</td>
<td>24.47</td>
<td>28.55</td>
</tr>
<tr>
<td>what animal is</td>
<td>32.81</td>
<td>39.45</td>
<td>36.83</td>
<td>34.37</td>
</tr>
<tr>
<td>how</td>
<td>17.38</td>
<td>19.79</td>
<td>17.69</td>
<td>18.82</td>
</tr>
<tr>
<td>what is the woman</td>
<td>45.83</td>
<td>46.62</td>
<td>58.51</td>
<td>61.55</td>
</tr>
<tr>
<td>what is this</td>
<td>59.05</td>
<td>60.47</td>
<td>77.65</td>
<td>81.7</td>
</tr>
<tr>
<td>which</td>
<td>31.68</td>
<td>32.79</td>
<td>30.82</td>
<td>32.34</td>
</tr>
<tr>
<td>where are the</td>
<td>35.22</td>
<td>33.85</td>
<td>32.83</td>
<td>36.11</td>
</tr>
<tr>
<td>are the</td>
<td>44.31</td>
<td>42.94</td>
<td>44.31</td>
<td>46.15</td>
</tr>
<tr>
<td>how many people are</td>
<td>19.98</td>
<td>42.81</td>
<td>35.98</td>
<td>37.08</td>
</tr>
<tr>
<td>what is on the</td>
<td>35.28</td>
<td>36.62</td>
<td>36.82</td>
<td>38.37</td>
</tr>
<tr>
<td>has</td>
<td>68.72</td>
<td>73.53</td>
<td>77.65</td>
<td>81.7</td>
</tr>
<tr>
<td>was</td>
<td>61.11</td>
<td>61.67</td>
<td>60.28</td>
<td>63.89</td>
</tr>
<tr>
<td>what type of</td>
<td>47.71</td>
<td>49.6</td>
<td>48.14</td>
<td>50.38</td>
</tr>
<tr>
<td>is this an</td>
<td>58.79</td>
<td>48.83</td>
<td>53.9</td>
<td>46.67</td>
</tr>
<tr>
<td>do</td>
<td>37.75</td>
<td>39.25</td>
<td>42.0</td>
<td>39.5</td>
</tr>
<tr>
<td>can you</td>
<td>64.0</td>
<td>81.84</td>
<td>70.13</td>
<td>80.9</td>
</tr>
<tr>
<td>who is</td>
<td>25.08</td>
<td>25.19</td>
<td>25.49</td>
<td>25.17</td>
</tr>
<tr>
<td>are these</td>
<td>48.59</td>
<td>51.77</td>
<td>50.91</td>
<td>51.95</td>
</tr>
<tr>
<td>do you</td>
<td>68.97</td>
<td>82.41</td>
<td>71.36</td>
<td>81.88</td>
</tr>
<tr>
<td>what time</td>
<td>20.36</td>
<td>24.13</td>
<td>20.04</td>
<td>23.25</td>
</tr>
<tr>
<td>is the woman</td>
<td>73.26</td>
<td>76.16</td>
<td>84.11</td>
<td>80.82</td>
</tr>
<tr>
<td>is this a</td>
<td>85.34</td>
<td>90.17</td>
<td>87.38</td>
<td>92.25</td>
</tr>
<tr>
<td>what are the</td>
<td>44.86</td>
<td>45.01</td>
<td>45.27</td>
<td>46.26</td>
</tr>
<tr>
<td>what color are the</td>
<td>67.6</td>
<td>64.67</td>
<td>67.35</td>
<td>68.83</td>
</tr>
<tr>
<td>why</td>
<td>18.0</td>
<td>17.98</td>
<td>18.37</td>
<td>18.83</td>
</tr>
<tr>
<td>none of the above</td>
<td>49.5</td>
<td>51.22</td>
<td>50.64</td>
<td>57.1</td>
</tr>
<tr>
<td>what is the person</td>
<td>61.41</td>
<td>62.08</td>
<td>61.52</td>
<td>63.96</td>
</tr>
<tr>
<td>how many people are in</td>
<td>48.16</td>
<td>52.78</td>
<td>54.3</td>
<td>58.36</td>
</tr>
<tr>
<td>is this</td>
<td>81.59</td>
<td>84.46</td>
<td>82.0</td>
<td>88.46</td>
</tr>
<tr>
<td>why is the</td>
<td>16.78</td>
<td>16.92</td>
<td>17.7</td>
<td>17.63</td>
</tr>
<tr>
<td>what is the color of the</td>
<td>78.81</td>
<td>75.87</td>
<td>75.86</td>
<td>80.08</td>
</tr>
<tr>
<td>what is</td>
<td>35.54</td>
<td>37.64</td>
<td>36.58</td>
<td>39.03</td>
</tr>
<tr>
<td>what are</td>
<td>54.15</td>
<td>56.32</td>
<td>54.03</td>
<td>55.57</td>
</tr>
<tr>
<td>is that a</td>
<td>58.76</td>
<td>54.87</td>
<td>55.75</td>
<td>58.85</td>
</tr>
<tr>
<td>what is in the</td>
<td>36.98</td>
<td>38.77</td>
<td>39.35</td>
<td>40.76</td>
</tr>
<tr>
<td>what sport is</td>
<td>89.17</td>
<td>91.95</td>
<td>91.02</td>
<td>91.64</td>
</tr>
<tr>
<td>how many</td>
<td>51.27</td>
<td>49.08</td>
<td>57.49</td>
<td>56.32</td>
</tr>
<tr>
<td>what is the</td>
<td>41.55</td>
<td>42.8</td>
<td>41.18</td>
<td>44.11</td>
</tr>
<tr>
<td>is it</td>
<td>68.75</td>
<td>71.49</td>
<td>66.26</td>
<td>69.75</td>
</tr>
<tr>
<td>is the man</td>
<td>53.54</td>
<td>50.3</td>
<td>52.13</td>
<td>52.56</td>
</tr>
<tr>
<td>what is the name</td>
<td>14.76</td>
<td>15.82</td>
<td>14.37</td>
<td>15.95</td>
</tr>
<tr>
<td>is there</td>
<td>42.0</td>
<td>22.4</td>
<td>37.2</td>
<td>36.0</td>
</tr>
<tr>
<td>what color is</td>
<td>63.27</td>
<td>64.85</td>
<td>64.29</td>
<td>67.89</td>
</tr>
<tr>
<td>what color</td>
<td>70.19</td>
<td>70.5</td>
<td>70.44</td>
<td>73.08</td>
</tr>
<tr>
<td>are</td>
<td>45.37</td>
<td>45.83</td>
<td>59.44</td>
<td>58.58</td>
</tr>
</tbody>
</table>

**Overall**                        | 52.05   | 54.01   | 54.55   | 56.74   |
2 Qualitative Evaluation

In this section, we qualitatively study success and failure cases using the VQA-CPv2 dataset. In Fig. 2 of the main paper, we show the proportions of information in each modality. We note that our regularization permits higher utilization of the visual modality. At the end of the document, we present success (see Fig. S1-S16) and failure cases (see Fig. S17-S32).

3 Experimental Settings

In the following, for each experiment, we describe its settings.

3.1 Colored MNIST

Recent works introduce different variants to Colored MNIST [2][3][4]. The main difference is the way bias is incorporated into the MNIST dataset. For instance, in [2] the authors use $\sigma = 0.1$ and evaluate their method on the gray-scale test set while in [4] the authors use $\sigma = 0.02$ and evaluate on a uniformly-colored test set. We follow the experimental settings introduced in REPAIR [2].

Model: We use LeNet [5] to encode the shape, then we concatenate the color representation to the last linear layer of LeNet and pass it through another linear layer with a bias term for classification. We use cross-entropy loss.

Hyper-parameters considered: We consider $\lambda$ in the range of $1e^{-9}$ to $1e^{-11}$. We report our results with $\lambda = 1e^{-10}$. Note, the values of the regularization term (i.e., the functional Fisher information) are approximately $1e9$. Therefore, the $\lambda$ hyperparameter acts as a normalizer that scales down the bias term to the $[0,10^2]$ range.

Computing infrastructure used: We use a single RTX2080Ti GPU. The average runtime for each epoch is one minute. The model converges in 35 epochs.

3.2 VQA-CPv2


Model: We add our regularization to the current state-of-the-art [1]. The current state-of-the-art model is based on “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering” [8].

Hyper-parameters considered: We consider $\lambda$ in the range of $1e^{-10}$ to $1e^{-20}$. We report our results with $\lambda = 4e^{-17}$. Note, the values of the regularization term (i.e., the functional Fisher information) are approximately $1e15$. Therefore, the $\lambda$ hyperparameter acts as a normalizer that scales down the bias term to the $[0,10^2]$ range.

Computing infrastructure used: We use two RTX2080Ti GPUs. The average runtime for each epoch is 5 minutes. The model converges in 15 epochs.

3.3 SocialIQ

SocialIQ [9] input data is constructed from a tuple of a video, a question about the social situation in the video, and an answer to that question. Tiven the tuple, the task is to predict whether the answer is correct or not.

Model: The textual inputs are encoded with BERT [10]. The visual input is encoded using a VGG16 [11]. From BERT we extract the last hidden layer’s representation. From VGG16, we extract the layer before the final fully-connected representation of four frames from the video (picked uniformly over the video). Then, we forward all representations through a linear layer. Subsequently, the concatenation of the representations is passed through an MLP consisting of two linear layers for classification. We use binary cross-entropy loss. We use a ReLU [12] activation between all layers and dropout [13] with a dropout rate of 0.2.
Hyper-parameters considered: We consider $\lambda$ in the range of $1e^{-9}$ to $1e^{-11}$. We report our results with $\lambda = 3e^{-10}$. Note, the values of the regularization term ($i.e.$, the functional Fisher information) are approximately $1e9$. Therefore, the $\lambda$ hyperparameter acts as a normalizer that scales down the bias term to the $[0, 10^2]$ range.

Computing infrastructure used: For the baseline, we use a single RTX2080Ti GPU, the average runtime for each epoch is one minute. The model converges after 20 epochs.

3.4 Dogs and Cats

In this task, we classify whether a given image shows a dog or a cat. This task was first presented as a Kaggle competition[^2]. The authors of “learning not to learn”[^4] introduce a modified version. They collected one split to hold bright dogs and dark cats (TB1) and another split that contains the opposite, i.e., dark dogs and bright cats.

Model: We follow the setting of Kim et al. [4]: The authors use a ResNet-18[^14] that was pre-trained on ImageNet[^15]. We fine-tuning the last fully-connected layer.

Hyper-parameters considered: In this dataset, we consider $\lambda$ in the range of $1e^{-10}$ to $1e^{-14}$. We report our results with $\lambda = 2e^{-10}$. Note, the values of the regularization term ($i.e.$, the functional Fisher information) are approximately $1e9$. Therefore, the $\lambda$ hyperparameter acts as a normalizer that scales down the bias term to the $[0, 10^2]$ range.

Computing infrastructure used: For the baseline, we use five RTX2080Ti GPUs. The average runtime for each epoch is 3 minutes. The model converges after 5 epochs.

4 Code

We use the PyTorch[^16] framework to conduct all of our experiments. We also provide a Python package which we wrote to calculate the regularization term. Further explanations of how to integrate our regularization to any multi-modal problem are in the README file.

[^2]: https://www.kaggle.com/c/dogs-vs-cats
Question type: are they

- Are they in the front or back yard?
  - GT: front, Ours: front
- Are they on a couch or a bed?
  - GT: bed, Ours: bed
- Are they on grass or dirt?
  - GT: dirt, Ours: dirt

Question type: what number is

- What number is the small hand on?
  - GT: 12, Ours: 12
- What number is on the sidewalk?
  - GT: no number, Ours: 0

Question type: what sport is

- What sport is being played?
  - GT: frisbee, Ours: frisbee
- What sport is this?
  - GT: baseball, Ours: baseball
- What sport is this man playing?
  - GT: baseball, Ours: baseball

Question type: is this person

- Is this person homeless?
  - GT: no, Ours: no
- Is this person wearing appropriate protective gear for skateboarding?
  - GT: no, Ours: no
- Is this person an adult?
  - GT: no, Ours: no
Question type: is the

Question: Is the horse one or multiple colors?
GT: 1, Ours: 1

Question: Is the adult a male or female sheep?
GT: female, Ours: female

Question: Is the child standing or sitting?
GT: sitting, Ours: sitting

Question type: what is the name

Question: What is the name of the animal?
GT: horse, Ours: horse

Question: What is the name of the clothing type she is wearing?
GT: dress, Ours: dress

Question: What is the name of the piece of furniture the stuffed animal is sitting on?
GT: chair, Ours: chair

Question type: how many

Question: How many giraffes are on the grass?
GT: 5, Ours: 5

Question: How many kids are here?
GT: 1, Ours: 1

Question: How many giraffes have their head down?
GT: 1, Ours: 1

Question type: does this

Question: Does this appear to be a noisy environment?
GT: no, Ours: no

Question: Does this appear to be in the United States?
GT: no, Ours: no

Question: Does this appear to be a noisy environment?
GT: no, Ours: no
Question type: is there a

- Question: Is there a man or woman on the train?
  - GT: man, Ours: man

- Question: Is there a male or female on the right side?
  - GT: male, Ours: male

- Question: Is there a man or woman standing directly behind the stop sign?
  - GT: man, Ours: man

Question type: is that a

- Question: Is that a cat or a dog?
  - GT: dog, Ours: dog

- Question: Is that a gas or electric stove?
  - GT: electric, Ours: electric

- Question: Is that a fork or a spoon?
  - GT: fork, Ours: fork

Question type: can you

- Question: Can you rent surfboards on this beach?
  - GT: yes, Ours: yes

- Question: Can you see grass?
  - GT: yes, Ours: yes

- Question: Can you turn left at this intersection?
  - GT: yes, Ours: yes

Question type: what

- Question: What room of a house would you find all of these items?
  - GT: kitchen, Ours: kitchen

- Question: What season is it?
  - GT: winter, Ours: winter

- Question: What structure is in the background?
  - GT: fence, Ours: fence
Question type: does the

Question: Does the man have a beard?
GT: yes, Ours: yes

Question: Does the person have any facial hair?
GT: yes, Ours: yes

Question: Does the fruit appear to be ripe?
GT: yes, Ours: yes

Question type: is the person

Question: Is the person blurry?
GT: yes, Ours: yes

Question: Is the person a good surfer?
GT: yes, Ours: yes

Question: Is the person flipping a skateboard?
GT: yes, Ours: yes

Question type: do you

Question: Do you think the bird is beautiful?
GT: yes, Ours: yes

Question: Do you get milk from these cows?
GT: yes, Ours: yes

Question: Do you see any phones?
GT: yes, Ours: yes

Question type: where is the

Question: Where is the woman and dog?
GT: beach, Ours: beach

Question: Where is the television?
GT: wall, Ours: on wall

Question: Where is the fridge?
GT: kitchen, Ours: kitchen
Question type: which

- Which sport is this?  
  GT: tennis, Ours: tennis
- Which room is it?  
  GT: kitchen, Ours: kitchen
- Which game are they playing?  
  GT: baseball, Ours: baseball

Question type: where are the

- Where are the motorcycles parked?  
  GT: street, Ours: street
- Where are the giraffes?  
  GT: field, Ours: field
- Where are the giraffes?  
  GT: zoo, Ours: inside

Question type: are the

- Are the buildings on the other side of the water near or far?  
  GT: far, Ours: far
- Are the people adults or children?  
  GT: adults, Ours: adults
- Are the men at the top or bottom of the mountain?  
  GT: bottom, Ours: bottom

Question type: how many people are

- How many people are standing up?  
  GT: 2, Ours: 2
- How many people are wearing glasses?  
  GT: 2, Ours: 2
- How many people are on the bike?  
  GT: 2, Ours: 2
Question type: what is on the ground?

Question: What is on the ground?
GT: snow, Ours: snow

Question: What is on the table?
GT: tv, Ours: tv

Question: What is on the ground?
GT: snow, Ours: snow

Question type: has

Question: Has this photo been enhanced?
GT: yes, Ours: yes

Question: Has the food been cooked already?
GT: yes, Ours: yes

Question: Has any of the pizza shown been eaten?
GT: yes, Ours: yes

Question type: was

Question: Was this a summer gathering or winter gathering?
GT: summer, Ours: summer

Question: Was this me at home or in a restaurant?
GT: home, Ours: home

Question: Was the photo taken on a farm or in the city?
GT: farm, Ours: farm

Question type: what type of

Question: What type of gate is this?
GT: wooden, Ours: wooden

Question: What type of building is the couple standing in front of?
GT: church, Ours: church

Question: What type of room is the girl in?
GT: living room, Ours: living room
Question type: what is the

- What is the boy holding?
  GT: bat, Ours: bat

- What is the player's position behind the batter?
  GT: catcher, Ours: catcher

- What is the girl wearing on her head?
  GT: hat, Ours: hat

Question type: what are the

- What are the colors of the chairs?
  GT: brown, Ours: brown

- What are the military men cutting?
  GT: cake, Ours: cake

- What are the zebras doing?
  GT: standing, Ours: standing

Question type: what color are the

- What color are the walls?
  GT: white, Ours: white

- What color are the man's pants?
  GT: white, Ours: white

- What color are the wheels?
  GT: green, Ours: green

Question type: is he

- Is he wearing a helmet?
  GT: no, Ours: no

- Is he playing baseball?
  GT: no, Ours: no

- Is he holding fishes?
  GT: no, Ours: no
Question type: why

- Question: Why is this man slightly squatting?
  GT: skateboarding, Ours: skateboarding
- Question: Why are they laying down?
  GT: resting, Ours: resting
- Question: Why are the Zebras in the water?
  GT: drinking, Ours: drinking

Question type: what is this

- Question: What is this animal bending down for?
  GT: food, Ours: food
- Question: What is this woman holding?
  GT: frisbee, Ours: frisbee
- Question: What is this a photo of?
  GT: food, Ours: food

Question type: how many people are in

- Question: How many people are in the photo?
  GT: 1, Ours: 1
- Question: How many people are in the photo?
  GT: 3, Ours: 3
- Question: How many people are in the photo?
  GT: 1, Ours: 1

Question type: what color

- Question: What color of gloves do they have on?
  GT: white, Ours: white
- Question: What color pants is the person wearing?
  GT: black, Ours: black
- Question: What color ribbon is in the girl's hair?
  GT: white, Ours: white
Question type: is this an

- Is this an Apple product?
  - GT: no, Ours: no
- Is this an old structure?
  - GT: no, Ours: yes
- Is this an English newspaper?
  - GT: no, Ours: no

Question type: what is in the

- What is in the mirror?
  - GT: dog, Ours: dog
- What is in the hanging bags?
  - GT: fruit, Ours: fruit
- What is in the photo?
  - GT: kitchen, Ours: kitchen

Question type: what does the

- What does the woman hold in her hands?
  - GT: ski poles, Ours: ski poles
- What does the girl have on her head?
  - GT: hair, Ours: hair
- What does the man have in his hand?
  - GT: surfboard, Ours: surfboard

Question type: what kind of

- What kind of fence is between us and the zebra?
  - GT: wood, Ours: wood
- What kind of meat is here?
  - GT: chicken, Ours: chicken
- What kind of fruit is in the basket?
  - GT: bananas, Ours: bananas
Question type: is it

Question: Is it winter or fall?
GT: winter, Ours: winter

Question: Is it day or night?
GT: day, Ours: day

Question: Is it daytime or nighttime?
GT: daytime, Ours: daytime

Question type: is the man

Question: Is the man skiing or snowboarding?
GT: snowboarding, Ours: snowboarding

Question: Is the man throwing or catching the frisbee?
GT: catching, Ours: catching

Question: Is the man wearing shorts or long pants?
GT: shorts, Ours: shorts

Question type: is there

Question: Is there water or oil on the windows?
GT: water, Ours: water

Question: Is there more pavement or grass?
GT: grass, Ours: grass

Question: Is there beer or cola on the door of the refrigerator?
GT: beer, Ours: beer

Question type: what color is the

Question: What color is the house?
GT: blue, Ours: blue

Question: What color is the sky?
GT: blue, Ours: blue

Question: What color is the train?
GT: blue, Ours: blue
Question type: what color is

Question: What color is this person's jacket?
GT: yellow, Ours: green

Question: What color is her shirt?
GT: white, Ours: white

Question: What color is his uniform?
GT: white, Ours: white

Question type: are

Question: Are all the vehicles automobiles?
GT: no, Ours: no

Question: Are all players playing for the same team?
GT: yes, Ours: no

Question: Are all the children smiling?
GT: yes, Ours: no

Question type: are there

Question: Are there more water glasses or wine glasses?
GT: wine, Ours: wine

Question: Are there more mashed potatoes or broccoli on the plate?
GT: broccoli, Ours: broccoli

Question: Are there more buses or cars in this photo?
GT: cars, Ours: cars

Figure S16
References


Question type: what brand

- What brand is the man's phone? (GT: apple, Ours: motorola)
- What brand is the motorcycle? (GT: ducati, Ours: yamaha)
- What brand of computer is the partially closed one? (GT: apple, Ours: dell)

Question type: what room is

- What room is this? (GT: office, Ours: dining room)
- What room is this? (GT: store, Ours: living)
- What room is the man probably in? (GT: office, Ours: living room)

Question type: what is the man

- What is the man holding? (GT: skis, Ours: ski poles)
- What is the man doing to the man in white? (GT: walking, Ours: catching)
- What is the man doing? (GT: watching tv, Ours: playing wii)

Question type: is

- Is anybody in the photo facing the camera? (GT: no, Ours: yes)
- Is everyone wearing a helmet? (GT: no, Ours: yes)
- Is today the dogs birthday? (GT: no, Ours: birthday)

Figure S17
Question type: are they

- Are they kite surfing or taking pictures?
  - GT: kitesurfing, Ours: flying kite
- Are they going uphill or downhill?
  - GT: downhill, Ours: uphill
- Are they playing singles or doubles tennis?
  - GT: singles, Ours: doubles

Question type: what number is

- What number is on the destination sign on the bus?
  - GT: 9, Ours: 38
- What number is on the back of the red shirt?
  - GT: 28, Ours: 0
- What number is on the bike?
  - GT: 9, Ours: 4

Question type: what sport is

- What sport is being played?
  - GT: football, Ours: frisbee
- What sport is the bat used in?
  - GT: baseball, Ours: frisbee
- What sport is she playing?
  - GT: baseball, Ours: tennis

Question type: is this person

- Is this person really skinny?
  - GT: no, Ours: yes
- Is this person and athlete?
  - GT: no, Ours: yes
- Is this person waterskiing?
  - GT: no, Ours: yes
Question type: is the

Question: Is the H2O above or below this man?
GT: below, Ours: water

Question: Is the H2O above or below this man?
GT: below, Ours: top

Question: Is the snowboarder facing towards or away from the camera?
GT: away, Ours: towards

Question type: what is the name

Question: What is the name of the hitter?
GT: batter, Ours: baseball

Question: What is the name of the hat he is wearing?
GT: cowboy, Ours: hat

Question: What is the name on of any road signs shown?
GT: sergio dukes ave, Ours: main street

Question type: how many

Question: How many dogs?
GT: 1, Ours: 2

Question: How many vehicles are on the road?
GT: 0, Ours: 1

Question: How many sprinkles are on this donut?
GT: 0, Ours: lot

Question type: does this

Question: Does this young man have on proper safety equipment for the sport?
GT: no, Ours: yes

Question: Does this food have a vegetable?
GT: no, Ours: yes

Question: Does this belong in a bathroom?
GT: no, Ours: yes
Question type: is there a

Question: Is there a paved street or a dirt trail in this photo?
GT: paved, Ours: road

Question: Is there a painting or a mirror in the frame?
GT: mirror, Ours: picture

Question: Is there a standard or international outlet installed?
GT: international, Ours: no

Question type: is that a

Question: Is that a limb or stick?
GT: limb, Ours: rope

Question: Is that a hot or cold beverage?
GT: hot, Ours: cold

Question: Is that a man or a statue of a man?
GT: statue, Ours: woman

Question type: can you

Question: Can you see birds?
GT: yes, Ours: no

Question: Can you name an advertiser on the fence?
GT: hood, Ours: bank of america

Question: Can you see any stars in the sky?
GT: yes, Ours: no

Question type: what

Question: What shape is the red sign?
GT: star, Ours: octagon

Question: What country does this appear to be in?
GT: india, Ours: usa

Question: What breed of dog are these?
GT: golden retriever, Ours: terrier
Question type: does the

Question: Does the woman look like she is happy to be on a date?  
GT: yes, Ours: no

Question: Does the suitcase match the color of his pants?  
GT: yes, Ours: no

Question: Does the room have a patio or balcony?  
GT: yes, Ours: table

Question type: is the person

Question: Is the person in the image a kid or a midget?  
GT: kid, Ours: child

Question: Is the person using this kitchen organized?  
GT: yes, Ours: no

Question: Is the person on the snow?  
GT: yes, Ours: no

Question type: do you

Question: Do you see a serving knife?  
GT: yes, Ours: no

Question: Do you see an ice cream?  
GT: yes, Ours: no

Question: Do you have to carry this bag or does it roll?  
GT: carry, Ours: luggage

Question type: where is the

Question: Where is the bus?  
GT: parked, Ours: on road

Question: Where is the train?  
GT: station, Ours: on tracks

Question: Where is the man and woman riding an elephant?  
GT: forest, Ours: outside
Question type: what animal is

- What animal is on top of the monitor?
  GT: dragon, Ours: cat

- What animal is black and white?
  GT: panda, Ours: bear

- What animal is that?
  GT: ram, Ours: sheep

Question type: how

- How long has this person been on the bench?
  GT: hour, Ours: 1 hour

- How fast is the speed limit?
  GT: 25, Ours: fast

- How old is the elephant?
  GT: old, Ours: young

Question type: is this a

- Is this a sandwich?
  GT: yes, Ours: no

- Is this a Nintendo Wii controller?
  GT: yes, Ours: wii

- Is this a black and white photo?
  GT: yes, Ours: black and white

Question type: none of the above

- Have the walkways been shoveled?
  GT: yes, Ours: snow

- At what approximate point on its body do this animal’s stripes stop?
  GT: stomach, Ours: head

- Can both of the planes go the same speed?
  GT: yes, Ours: no
Question type: which

Question: Which would most likely be eaten last?
GT: cake, Ours: food

Question: Which arm is he wearing a brace on?
GT: left, Ours: right

Question: Which toilet is hanged higher?
GT: left, Ours: right

Question type: where are the

Question: Where are the people in the picture?
GT: zoo, Ours: outside

Question: Where are the people playing?
GT: field, Ours: park

Question: Where are the bikes parked?
GT: street, Ours: parking lot

Question type: are the

Question: Are the giraffes enclosed or roaming free?
GT: enclosed, Ours: zoo

Question: Are the dishes clean or dirty?
GT: dirty, Ours: clean

Question: Are the three signs pointing right or left?
GT: left, Ours: right

Question type: how many people are

Question: How many people are sitting on the elephant?
GT: 2, Ours: 4

Question: How many people are on the elephant?
GT: 2, Ours: 3

Question: How many people are watching the player?
GT: 7, Ours: 4
Question type: what is on the

Question: What is on the suitcase?
GT: jacket, Ours: clothes

Question: What is on the clock?
GT: sticker, Ours: clock

Question: What is on the mountains?
GT: snow, Ours: rocks

Question type: has

Question: Has this suitcase been on a trip in the past?
GT: yes, Ours: no

Question: Has the wine been opened?
GT: yes, Ours: no

Question: Has he or she had wine?
GT: yes, Ours: neither

Question type: was

Question: Was the bread grilled or toasted?
GT: toasted, Ours: fried

Question: Was this picture taken in the northern or the southern hemisphere?
GT: northern, Ours: south

Question: Was this taken by a human or a camera trap?
GT: camera trap, Ours: man

Question type: what type of

Question: What type of weather is occurring?
GT: rain, Ours: rainy

Question: What type of picture do you call this?
GT: selfie, Ours: color

Question: What type of footprints are in the sand?
GT: dog, Ours: brown

29
Question type: do

Question: Do the bathroom fixtures appear contemporary or outdated?
GT: contemporary, Ours: real

Question: Do the boys look like they are jumping up or falling down?
GT: jumping up, Ours: standing

Question: Do these umbrella's have a solid color or are they multicolored?
GT: multicolored, Ours: striped

Question type: what is the person

Question: What is the person sitting on?
GT: atv, Ours: motorcycle

Question: What is the person wearing?
GT: shorts, Ours: wetsuit

Question: What is the person painted on the side of the van holding in his hands?
GT: umbrella, Ours: graffiti

Question type: who is

Question: Who is wearing a blue shirt?
GT: skateboarder, Ours: girl

Question: Who is in the water?
GT: people, Ours: surfer

Question: Who is the baby elephant hiding from?
GT: no one, Ours: elephant

Question type: are these

Question: Are these ducks or geese?
GT: geese, Ours: sheep

Question: Are these people in a field or on a street?
GT: street, Ours: ground

Question: Are these fruits or vegetables?
GT: vegetables, Ours: fruit
Question type: what is the woman doing?

- **Picture 1**: Woman brushing hair. GT: brushing hair, Ours: eating
- **Picture 2**: Woman standing near a cow. GT: nothing, Ours: sunglasses
- **Picture 3**: Woman talking on a phone. GT: listening to music, Ours: talking

Question type: what are these people wearing on their faces?

- **Picture 4**: Three people wearing masks. GT: masks, Ours: sunglasses
- **Picture 5**: Person holding a toothbrush. GT: brushing teeth, Ours: cutting
- **Picture 6**: Man looking down. GT: looking down, Ours: eating

Question type: what time is it?

- **Picture 7**: Clock showing 10:05. GT: 10:05, Ours: 1:10
- **Picture 8**: Clock showing 10:05. GT: 10:05, Ours: 3:00
- **Picture 9**: Clock showing 1:55. GT: 1:55, Ours: 3:00

Question type: is the woman...

- **Picture 10**: Woman playing tennis. GT: no, Ours: yes
- **Picture 11**: Woman holding a towel. GT: no, Ours: yes
- **Picture 12**: Woman wearing a hat. GT: no, Ours: yes
Question: What is the metal that most likely makes up the top of this tower? GT: gold, Ours: clock

Question: What is the little girl holding? GT: bunny, Ours: baby

Question: What is the fence made of? GT: wood, Ours: metal

Question: What are the words on the screen? GT: k roberts photography, Ours: unknown

Question: What are the donuts from? GT: starbucks, Ours: dunkin donuts

Question: What are the people holding? GT: signs, Ours: nothing

Question: What color are the tiles on the floor? GT: white, Ours: gray

Question: What color are the letters on the board? GT: white, Ours: silver

Question: What color are the players' shoes? GT: white, Ours: blue

Question: Is he a beginner surfer? GT: no, Ours: yes

Question: Is he bunting or swinging? GT: bunting, Ours: baseball

Question: Is he right handed? GT: no, Ours: yes
Question type: why

- Question: Why does he seem suspended in the air?
  - GT: jumping, Ours: skateboarding
- Question: Why is there a chair in the bathroom?
  - GT: to sit, Ours: yes
- Question: Why did the rider stop?
  - GT: rest, Ours: yes

Question type: what is this

- Question: What is this man's job?
  - GT: farmer, Ours: police
- Question: What is this?
  - GT: stove, Ours: kitchen
- Question: What is this man waiting for?
  - GT: food, Ours: bus

Question type: how many people are in

- Question: How many people are in the room?
  - GT: 4, Ours: 8
- Question: How many people are in the picture?
  - GT: 10, Ours: 15
- Question: How many people are in this photo?
  - GT: 4, Ours: 3

Question type: what color

- Question: What color of shirt is the man on the elephant wearing?
  - GT: white, Ours: pink
- Question: What color are this person's sunglasses?
  - GT: white, Ours: black
- Question: What color stands out in this picture?
  - GT: red, Ours: pink
Question type: is this

GT: yes, Ours: no

Question: Is this inside?

GT: lush and green, Ours: trees

Question: Is this area desert-like or lush and green?

GT: yes, Ours: no

Question: IS this outside?

GT: festival, Ours: traffic

Question: Why is the road ahead closed?

GT: no, Ours: yes

Question: Why is the picture funny?

GT: its moving, Ours: tired

Question: Why is the cat’s face blurred?

GT: white, Ours: pink

Question: What is the color of the bus?

GT: green and yellow, Ours: yellow

Question: What is the color of the vehicles?

GT: red, Ours: blue

Question: What is the color of the catcher’s hat?

GT: window, Ours: train

Question: What is clear in the background?

GT: fireplace, Ours: tv

Question: What is below the mirror?

GT: trash can, Ours: desk

Question: What is white and under the desk?

GT: yes, Ours: no
Question type: is this an affluent neighborhood?
GT: no, Ours: yes

Question type: is this an adult's bed?
GT: no, Ours: yes

Question type: is this an aerial view?
GT: no, Ours: yes

Question type: what is in the picture?
GT: bottle, Ours: vase

Question type: what is in the sky?
GT: nothing, Ours: flowers

Question type: what is in the glass?
GT: oil, Ours: beer

Question type: what does the under of the clock read?
GT: melrose, Ours: clock

Question type: what does the rack on top of the vehicle hold?
GT: trash, Ours: nothing

Question type: what does the writing say?
GT: slow, Ours: nothing

Question type: what kind of vehicle is this?
GT: jet ski, Ours: boat

Question type: what kind of animals are in the picture?
GT: horse and cow, Ours: cows

Question type: what kind of shorts is he wearing?
GT: khaki, Ours: jeans
Question type: is it

Is it warm or cold where this photo was taken?  
GT: warm, Ours: cold

Is it cold or warm?  
GT: warm, Ours: cold

Is it raining or snowing?  
GT: raining, Ours: snowing

Question type: is the man

Is the man or woman driving?  
GT: woman, Ours: man

Is the man standing straight or leaning on the surfboard?  
GT: leaning, Ours: neither

Is the man with the brown hair wearing a tight shirt or a loose shirt?  
GT: loose, Ours: no

Question type: is there

Is there an odd number of picnic tables or an even number?  
GT: odd, Ours: 2

Is there an even number of donuts or an odd number?  
GT: odd, Ours: 0

Is there lots of water or mainly land in the picture?  
GT: both, Ours: ocean

Question type: what color is the

What color is the wall in the picture? White and ____?  
GT: red, Ours: white

What color is the woman's shirt?  
GT: gray, Ours: green

What color is the chair?  
GT: gray, Ours: brown
Question type: what color is

![Image 1](image1.jpg)  Question: What color is her hair?  GT: red, Ours: blonde

![Image 2](image2.jpg)  Question: What color is her hair?  GT: red, Ours: brown

![Image 3](image3.jpg)  Question: What color is his luggage?  GT: red and black, Ours: red

Question type: are

![Image 4](image4.jpg)  Question: Are both elbows bent?  GT: yes, Ours: no

![Image 5](image5.jpg)  Question: Are all the people drinking coffee?  GT: no, Ours: coffee

![Image 6](image6.jpg)  Question: Are both pizzas the same?  GT: yes, Ours: unknown

Question type: are there

![Image 7](image7.jpg)  Question: Are there most likely a couple eating together or a group?  GT: couple, Ours: yes

![Image 8](image8.jpg)  Question: Are there two cats or just one looking in a mirror?  GT: 1, Ours: cat

![Image 9](image9.jpg)  Question: Are there bushes or trees?  GT: bushes, Ours: trees

Figure S32