Supplementary Material for
"Multi-task Learning of Hierarchical Vision-Language Representation"

Duy-Kien Nguyen[1] and Takayuki Okatani[1,2]

[1]Graduate School of Information Sciences, Tohoku University
[2]RIKEN Center for Advanced Intelligence Project
{kien, okatani}@vision.is.tohoku.ac.jp

This document contains the following: i) more details of setup of the experiments reported in the main paper (Sec.A); ii) additional results of the selection of optimal layers for the task-specific decoders (Sec.B); iii) additional results of ICR on MSCOCO 5,000 testing images (Sec.C); and iv) more visualization of the proposed network for a variety of images including failure cases of VQA and ICR (Sec.D and E).

A. More Details of the Experimental Setup

In all the experiments reported in this study, images and sentences (i.e., questions or captions) were preprocessed as follows. We used Faster-RCNN [9] to extract the bottom-up features from each image, which yields from 10 to 100 features (also referred to as regions in this study), i.e., $T \in [10, 100]$. Questions and captions were tokenized using Python Natural Language Toolkit (nltk) [2]. We used the vocabulary provided by the CommonCrawl-840B GloVe model for English word vectors [8], and set out-of-vocabulary words to *unk*.

As is mentioned in Sec.4.2 of the main paper, we conducted a hyper-parameter search on several training parameters including the layers used for the task-specific decoders by training the network on each individual task. The parameters thus determined are shown in Table 1.

We provide below additional details of the training procedures of the three tasks. We used the cross-entropy loss for all the three tasks.

Image Caption Retrieval This task consists of two subtasks; one is to retrieve relevant images given a query caption (image retrieval) and the other is to retrieve relevant captions given a query image (image annotation). In the training, given pairs of image-caption (I, C)'s, where I and C are an image and a caption, respectively, we compute the losses for the two subtasks for each pair as follows. For image retrieval, we randomly sample $F - 1$ images that are different from I, and compute the loss for F images including I, in which the label for the ground truth image (I) is set to 1 and those for the others are all 0.

Table 1: Hyperparameters determined by training on individual tasks and then used for joint training (# step: step size of learning rate decay, # iter: total of training iterations, K=1,000 units).

Task	Level	# step	# iter	Batch size	Cycle (C)
VQA	5	8K	20K	400	1
ICR	3	12K	30K	64	1
VG	2	4K	10K	64	1
VQA + VG	5, 2	12K	30K	400, 64	3
ICR + VG	3, 2	16K	40K	64, 64	4
VQA + ICR	5, 3	20K	50K	400, 64	5
VQA + ICR + VG	5, 3, 2	24K	60K	400, 64, 64	6

For image annotation, we randomly sample $F - 1$ captions that are not the ground truth captions corresponding to the image (I) and compute the loss of image annotation for F captions including C, in which the label for the ground truth caption (C) is 1 and those for the others are 0. We minimize the sum of the two losses. We used $F = 26$ for all the experiments.

Visual Question Answering We followed the procedure of [9]. We treat VQA as a multi-label classification task, where each training question is associated with one or several answers with soft accuracy label(s) in [0, 1]. Multiple answers appear in the case of disagreement among human annotators. The scores of answers are computed as in the original paper [1], that is,

$$\text{score}_i = \min\left(\frac{\text{\# humans that provide the } i\text{-th answer}}{3}, 1\right)$$

where score_i is the score of i-th answer in the predefined answer set.

Visual Grounding The dataset provides a set of samples, each of which is built upon a pair of an image and its caption. Each sample consists of a set of phrases in the caption and the corresponding box(es) in the image. We label each phrase with its corresponding box(es) as 1 and with other boxes as 0. These boxes are obtained in the aforementioned pre-processing using the pre-trained Faster-RCNN. The loss is the sum over all possible phrase-region pairs in the image and caption.

B. Additional Results of Layer Selection for the Three Tasks

As mentioned in Sec.5.2 of the main paper, we train our network on each individual task to choose the layer of the shared encoder fit for each task. Table 2 shows the results. Based on these, we determined $l_R = 3$ (image caption retrieval), $l_Q = 5$ (VQA), and $l_G = 2$ (visual grounding), as reported in the main paper. However, while it is simple and can be performed efficiently, this method may not provide optimal choice of layers for the three tasks, as it does not consider interactions among the three tasks.

Table 2: Performance of the proposed network trained and tested on the same individual task. These are used to determine the layer of the shared encoder for each of the three tasks.

Task \ Layer	1	2	3	4	5	6
VQA	64.72	65.21	65.34	65.35	**65.50**	65.27
ICR	56.45	56.78	**57.15**	54.18	-	-
	46.30	46.00	**48.05**	42.64	-	-
VG	57.74	**58.09**	57.80	-	-	-

Thus, we also tested another method for choosing the layers that is based on joint-training of the three tasks. Table 3 shows the results, which were obtained by the following procedure. Initially, we determine the order of the three tasks in terms of the level in the hierarchy of the shared representation. Based on the above results, we determine their (descending) order as follows: VQA, ICR, and VG. We first determine the optimal layer for VQA by training the network on VQA alone, which is the same as the first row (VQA) of Table 2; this results in $l_{\mathrm{Q}} = 5$. Next, we determine the optimal layer for ICR. To do this, we train the network on VQA+ICR for different choice of the layer for ICR (i.e., $l_{\mathrm{R}} = 1, \ldots$) while fixing the layer for VQA (i.e., $l_{\mathrm{Q}} = 5$). We evaluate the performance for different l_{R}'s on VQA, ICR (image annotation) and ICR (image retrieval). The second to forth rows of Table 3 show the performance on VQA, image annotation, and image retrieval, respectively. From this, we choose $l_{\mathrm{R}} = 3$. Finally, we determine the layer for VG. To do this, we train the network on VQA+ICR+VG for different layer l_{G} $(= 1, 2, 3)$ for VG. As above, we evaluate the performance on VQA, image annotation, and image retrieval, and VG, which are shown in the fifth to eighth rows of Table 3, respecrively. From this, we choose $l_{\mathrm{G}} = 2$.

In short, we obtain the same results as the first method based on individual task training. This confirms the validity of our choice of the layers for the three tasks. In the above experiments, we set $F = 16$ in image caption retrieval for efficient computation; the reduction of F contributes the most to reducing necessary computational resource.

Table 3: Performance of the proposed network trained and tested on several combinations of tasks. Their combinations are created in a cumulative fashion, assuming the order of the three tasks to be VQA, ICR, and VG in terms of level in the representaion hierarchy in the shared encoder. ICR1 and ICR2 indicate image annotation and image retrieval, respectively.

Task \ Layer		1	2	3	4	5	6
Trained/tested on VQA alone	(VQA)	64.72	65.21	65.34	65.35	**65.50**	65.27
Trained/tested on VQA+ICR(2 subtasks)	(VQA)	65.80	66.00	**66.09**	65.90	-	-
for different layers for ICR	(ICR1)	58.01	58.63	**59.25**	56.37	-	-
and layer = 5 for VQA	(ICR2)	47.91	48.70	**49.03**	47.27	-	-
Trained/tested on VQA+ICR(2)+VG	(VQA)	66.10	**66.15**	66.15	-	-	-
for different layers for VG	(ICR1)	**61.94**	61.88	61.90	-	-	-
and layer = 5 for VQA and 3 for ICR	(ICR2)	49.84	**50.81**	50.01	-	-	-
	(VG)	57.86	**58.17**	57.83	-	-	-

C. Comparisons on MS-COCO dataset of 5,000 testing images

As noted in the main paper, we conducted evaluation on MSCOCO 5,000 testing images. The results are shown in Table 4. Our method is comparable to the state-of-the-art method (S-E Model). It is noteworthy that our method provided only 50 mismatched pairs for each matched pair, while S-E Model provided 128.

Table 4: Results of image annotation and retrieval on the MSCOCO (5,000 testing) datasets.

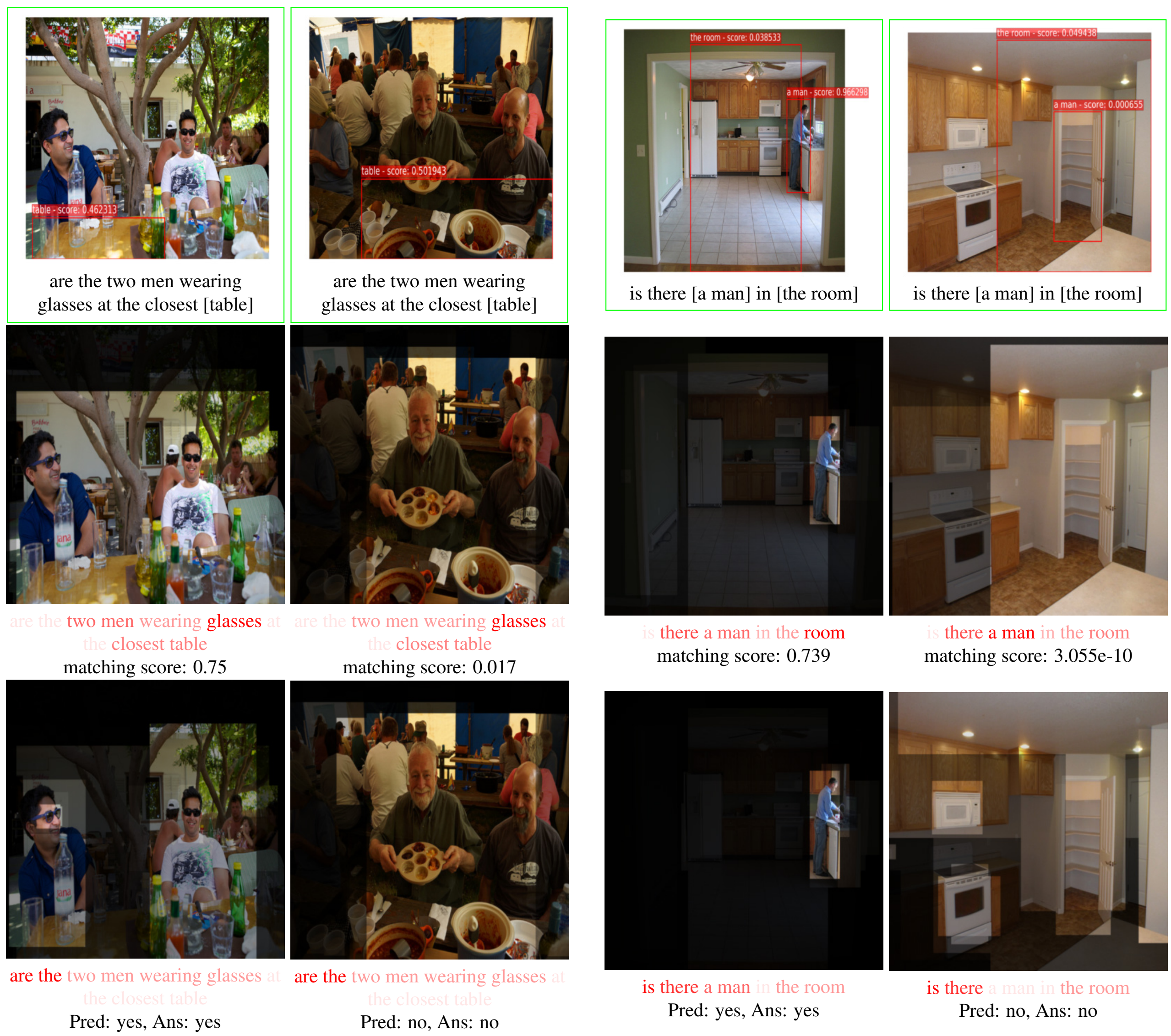
Method	Image Annotation			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [5]	11.8	32.5	45.4	8.9	24.9	36.3
FV [6]	17.3	39.0	50.2	10.8	28.3	40.1
OEM [10]	23.3	50.5	65.0	18.0	43.6	57.6
VQA [7]	23.5	50.7	63.6	16.7	40.5	53.8
VSE++ [3]	41.3	69.2	81.2	30.3	59.1	72.4
S-E Model [4]	**42.8**	**72.3**	**83.0**	33.1	62.9	75.5
Ours	42.2	69.1	80.6	**33.2**	**64.2**	**76.5**

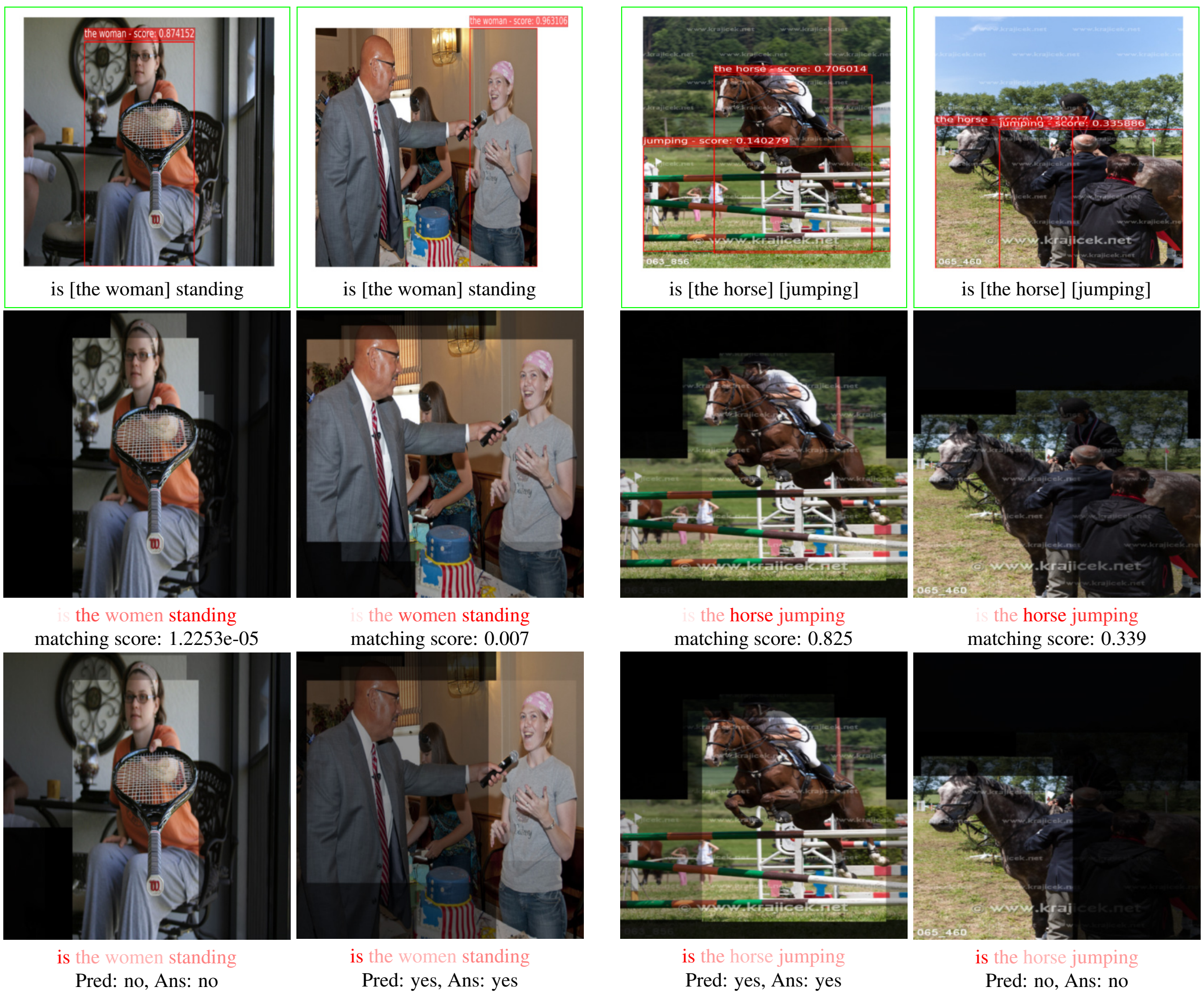
D. Visualization of Inference of Three Tasks on Visual Question Answering

In Fig. 2 of the main paper, we show a few examples of visualization of inference of VG, ICR, and VQA on complementary image-question pairs of VQA 2.0 (i.e., pairs of the same questions and different images). We show here more examples for success cases (Sec. D.1) as well as failure cases (Sec. D.2).

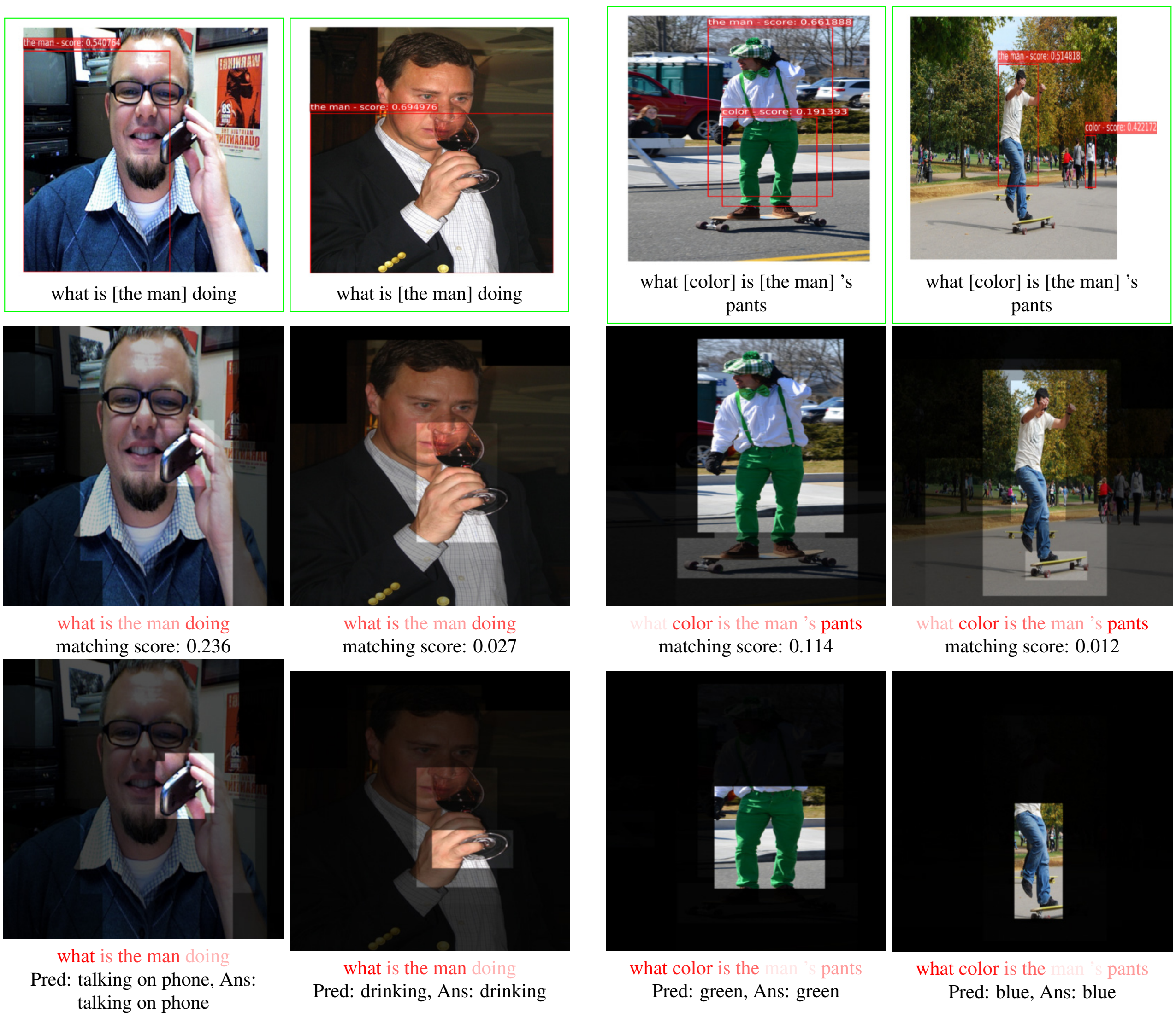
D.1. Success Cases

We first show visualization for success cases, i.e., image-question pairs for which our network provides the correct answers. As in Fig. 2 of the main paper, each of the left and right panels on each page shows visualization for a complementary image-question pair. The sames observation given in the main paper applies to theses examples.

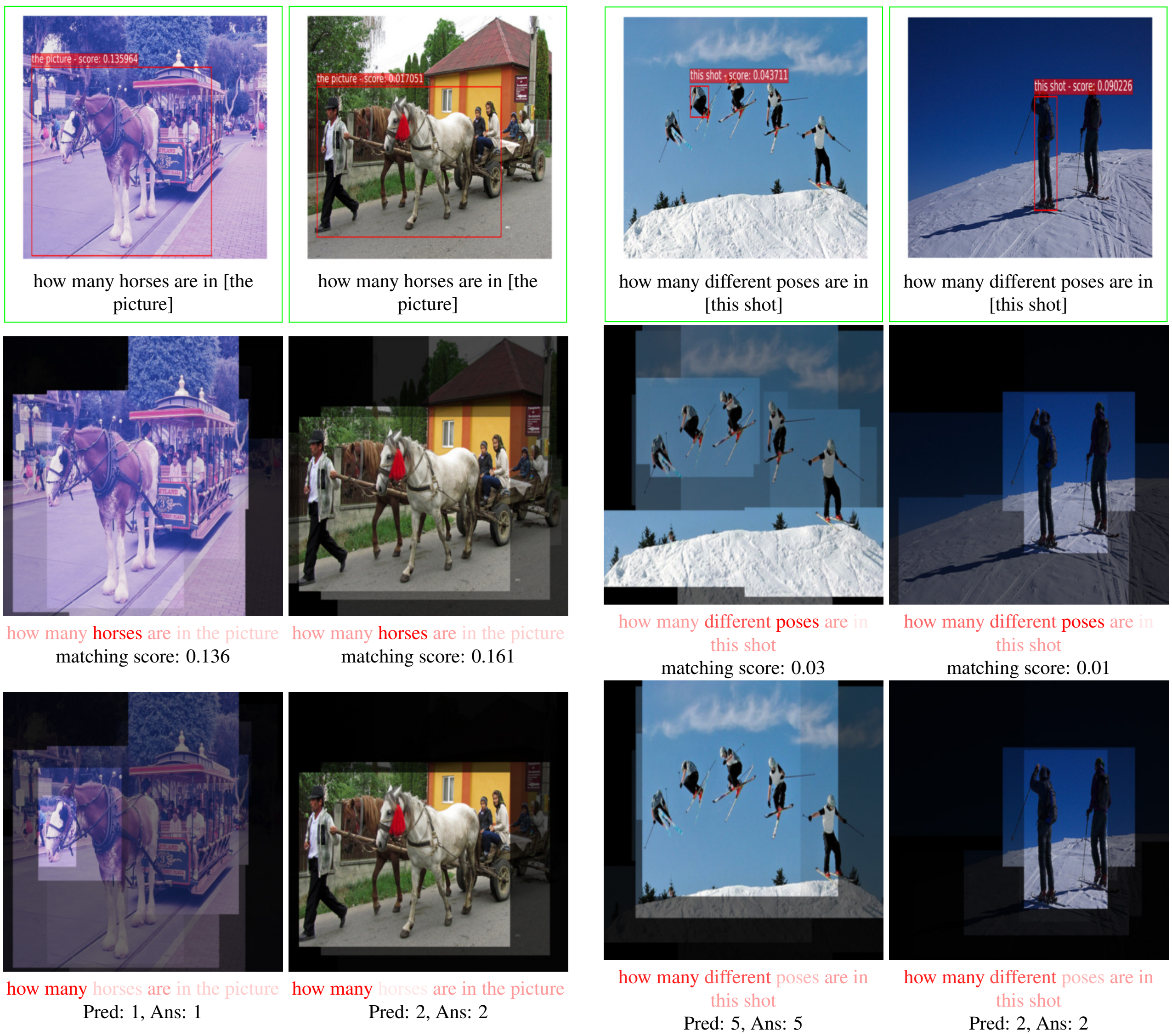


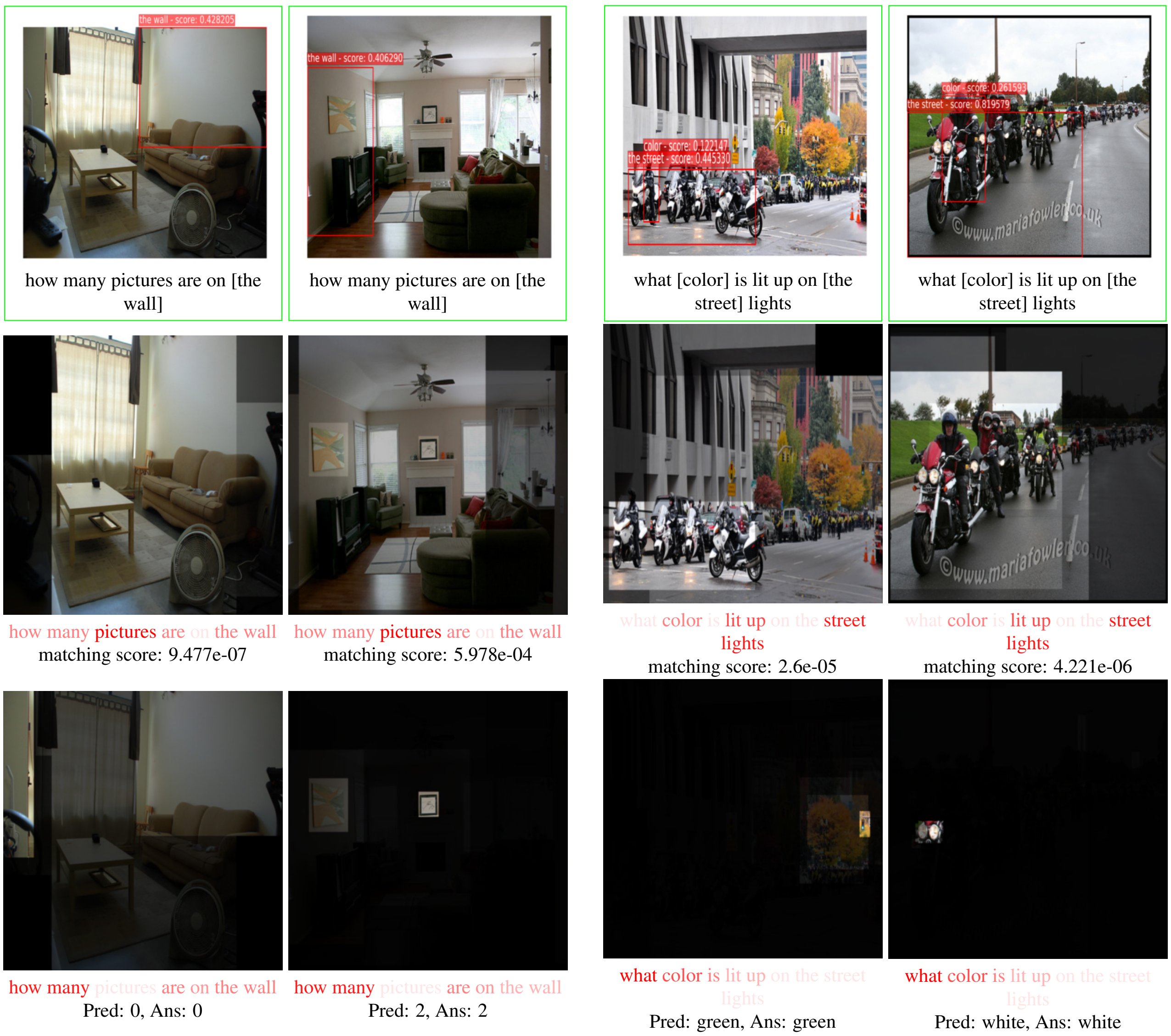


the woman - score: 0.874152
is [the woman] standing
the woman - score: 0.963106
is [the woman] standing
the horse - score: 0.706014
jumping - score: 0.140279
is [the horse] [jumping]
jumping - score: 0.335886
is [the horse] [jumping]
is the women standing
matching score: 1.2253e-05
is the women standing
matching score: 0.007
is the horse jumping
matching score: 0.825
is the horse jumping
matching score: 0.339
is the women standing
Pred: no, Ans: no
is the women standing
Pred: yes, Ans: yes
is the horse jumping
Pred: yes, Ans: yes
is the horse jumping
Pred: no, Ans: no



the man - score: 0.540764
what is [the man] doing
the man - score: 0.694976
what is [the man] doing
the man - score: 0.661888
color - score: 0.191393
what [color] is [the man] 's pants
the man - score: 0.514818
color - score: 0.422172
what [color] is [the man] 's pants
what is the man doing
matching score: 0.236
what is the man doing
matching score: 0.027
what color is the man 's pants
matching score: 0.114
what color is the man 's pants
matching score: 0.012
what is the man doing
Pred: talking on phone, Ans: talking on phone
what is the man doing
Pred: drinking, Ans: drinking
what color is the man 's pants
Pred: green, Ans: green
what color is the man 's pants
Pred: blue, Ans: blue


the picture - score: 0.135964
how many horses are in [the picture]
the picture - score: 0.017051
how many horses are in [the picture]
this shot - score: 0.043711
how many different poses are in [this shot]
this shot - score: 0.090226
how many different poses are in [this shot]
how many horses are in the picture
matching score: 0.136
how many horses are in the picture
matching score: 0.161
how many different poses are in this shot
matching score: 0.03
how many different poses are in this shot
matching score: 0.01
how many horses are in the picture
Pred: 1, Ans: 1
how many horses are in the picture
Pred: 2, Ans: 2
how many different poses are in this shot
Pred: 5, Ans: 5
how many different poses are in this shot
Pred: 2, Ans: 2



the wall - score: 0.428205
how many pictures are on [the wall]
the wall - score: 0.406290
how many pictures are on [the wall]
color - score: 0.122147
the street - score: 0.445330
what [color] is lit up on [the street] lights
color - score: 0.261593
the street - score: 0.819579
©www.mariafowler.co.uk
what [color] is lit up on [the street] lights
how many pictures are on the wall
matching score: 9.477e-07
how many pictures are on the wall
matching score: 5.978e-04
what color is lit up on the street lights
matching score: 2.6e-05
what color is lit up on the street lights
matching score: 4.221e-06
how many pictures are on the wall
Pred: 0, Ans: 0
how many pictures are on the wall
Pred: 2, Ans: 2
what color is lit up on the street lights
Pred: green, Ans: green
what color is lit up on the street lights
Pred: white, Ans: white

D.2. Failure Cases

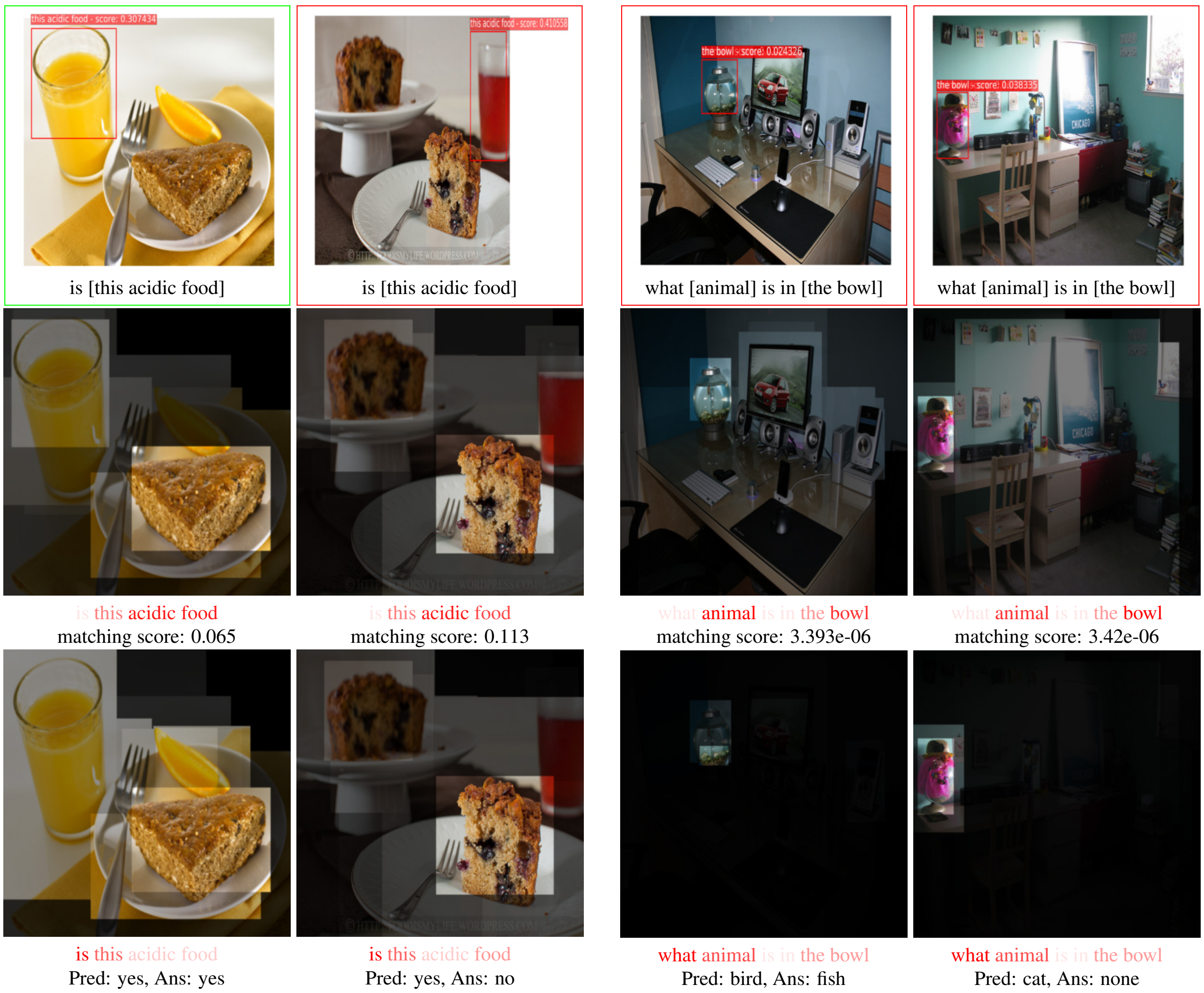
We next show failure cases, i.e., image-question pairs for which our network provides at least one wrong answer for the VQA task. The red bounding boxes indicate wrong answers and the green ones indicate correct answers. From the examples shown below, we can categorize failures for the VQA task into the following typical cases, for each of which we can explain why our network provides wrong answers and suggest possible solutions:

1) Although the VG and ICR decoders are able to correctly locate objects or concepts in the input image that appear in the input question, the VQA decoder fails to distinguish different objects or concepts that have similar appearance. This may be attributable to that the pretrained Faster R-CNN used for extracting image features is not trained to distinguish fine-grained concepts (e.g., "*terrier*" and "*lab*" (i.e., Labrador retriever), "*round*" and "*oval*"). It may help to train the Faster R-CNN with more fine-grained concepts.

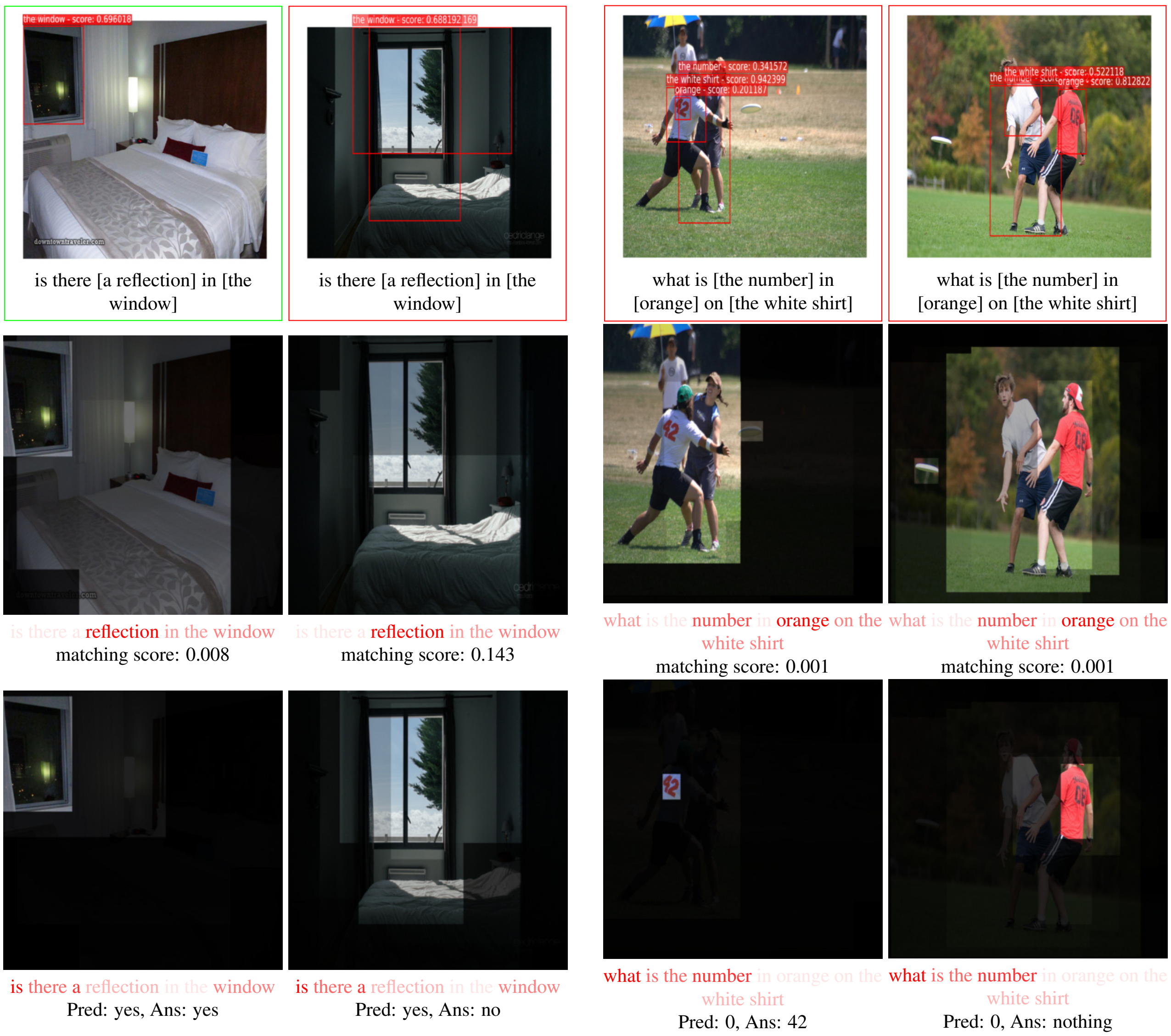
2) The network is unable to locate relevant image regions. This often occurs when the VG and ICR decoder at the lower layer of the network are unable to detect correct regions (e.g., "*the bottom corner*" or "*inside of the plane*"), which leads to the failure of the VQA decoder. This is mostly because the Faster R-CNN fails to extract right regions or extracts excessively large regions containing many objects.

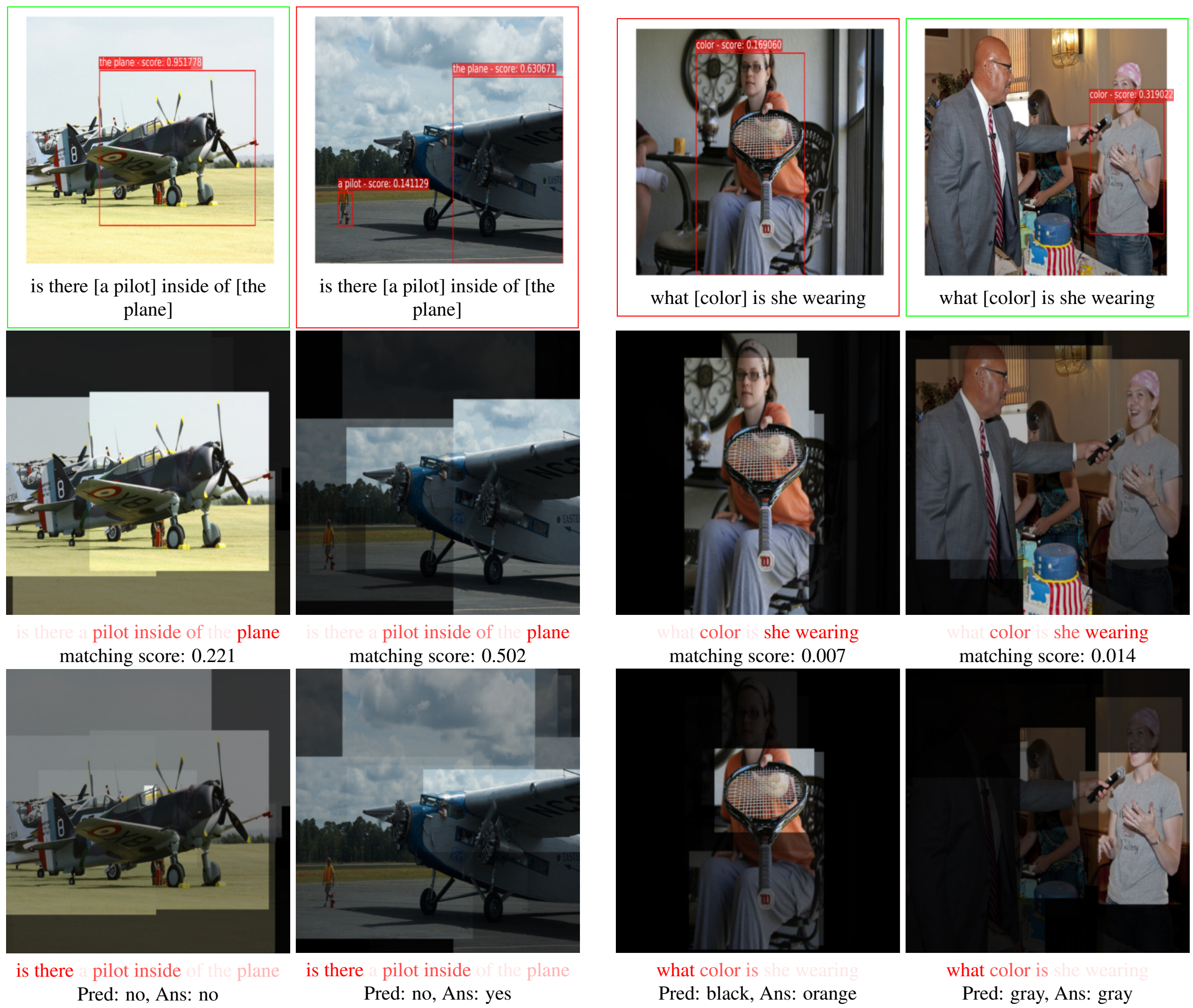
3) Questions require general knowledge that cannot be learned from only the training data (e.g., "*acidic food*" or "*reflection*"). For instance, for the question "*is this acidic food*", we can observe from the response of the VG and ICR decoders that the network recognizes all the food in the image as "*acidic food*"; for the question "*is there a reflection in the window*", the VG and ICR decoders give high confident scores for "*reflection*" even there is not.

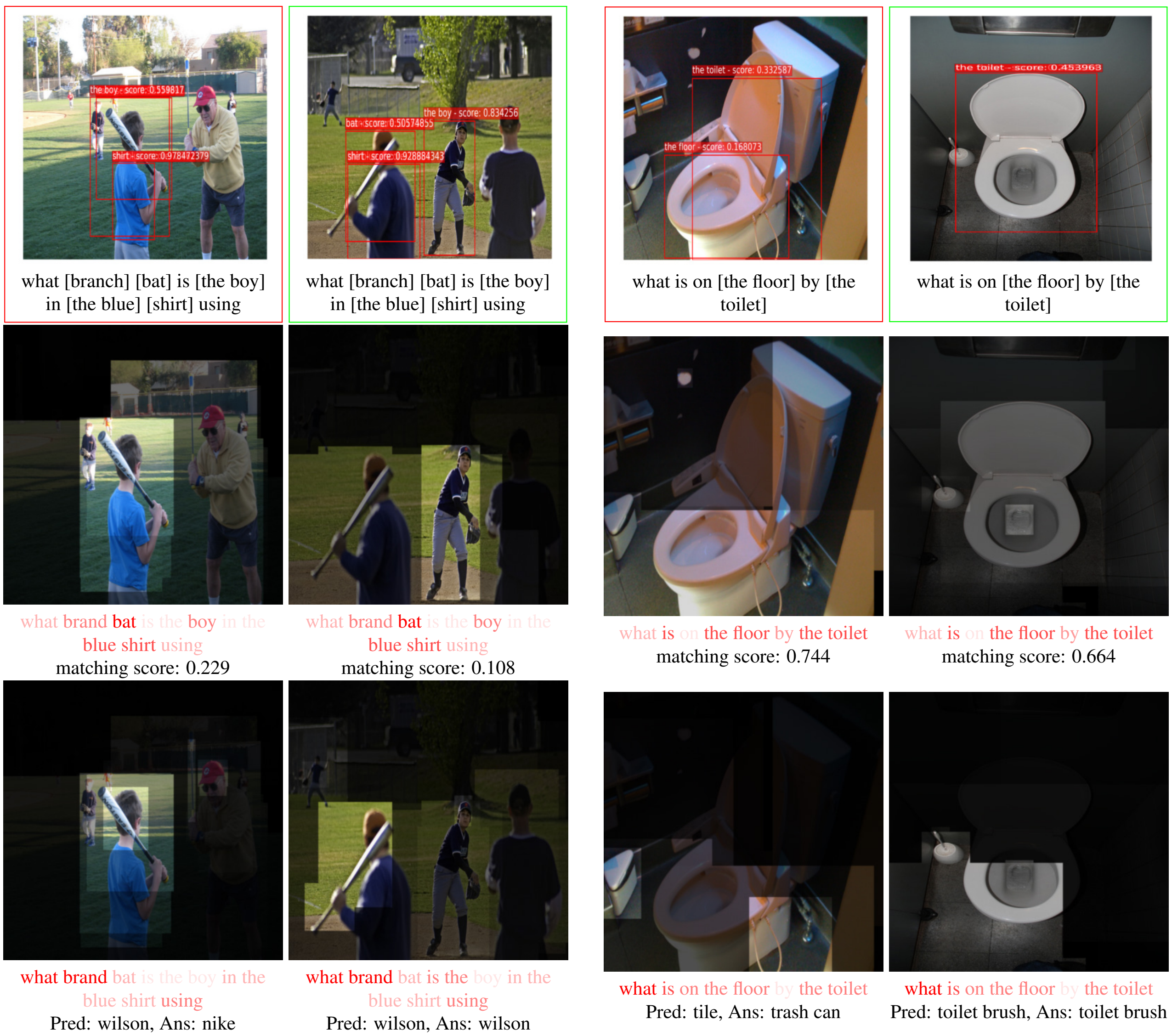
4) The answers given by the network are judged incorrect simply because they are not listed in the given set of correct answers in the dataset, but they are actually considered to be correct answers. For example, for the question "*what is on the floor by the toilet*", both of "*tile*" and "*trash can*" should be correct answers, but only the latter is listed in the correct answer set.

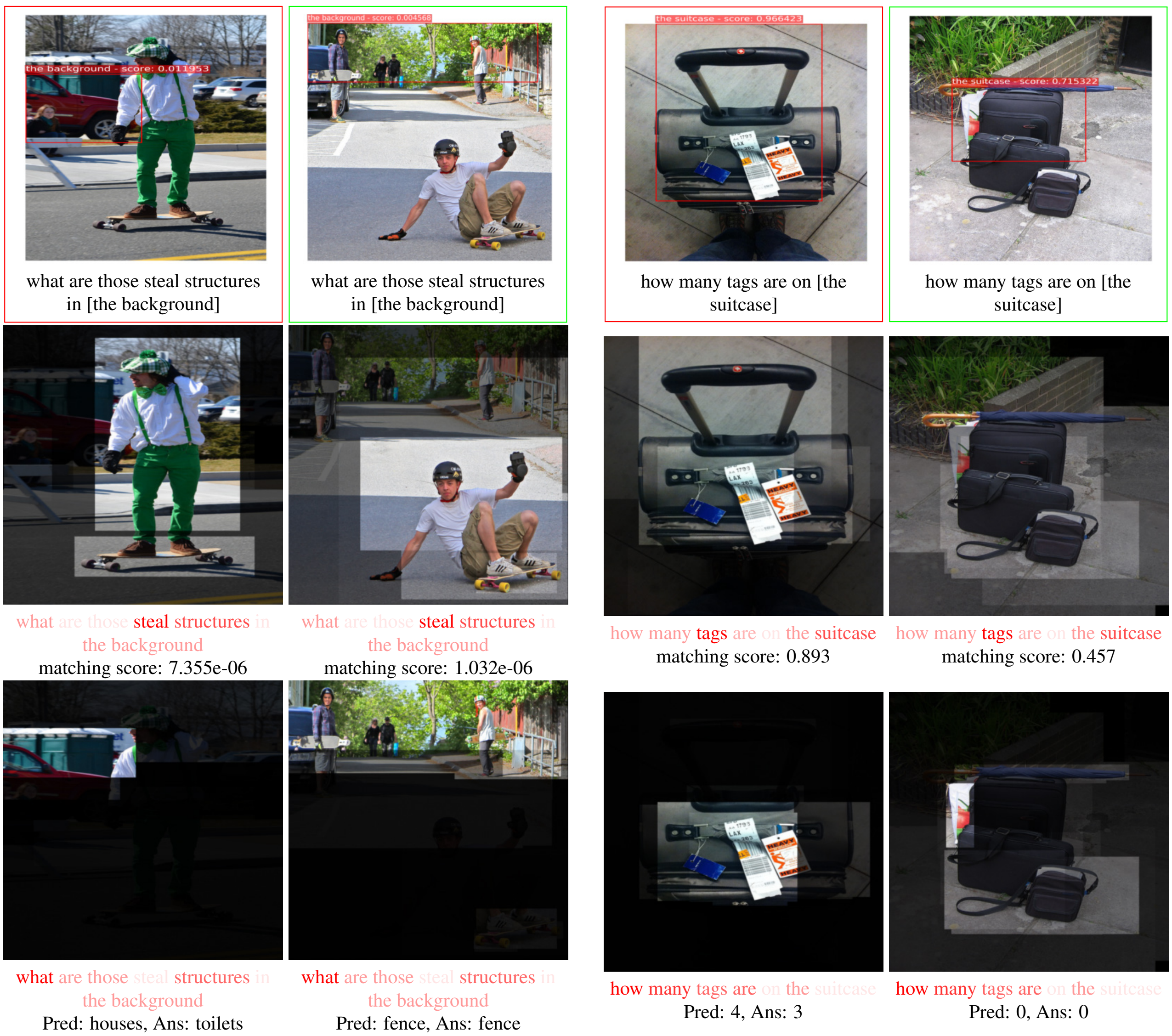


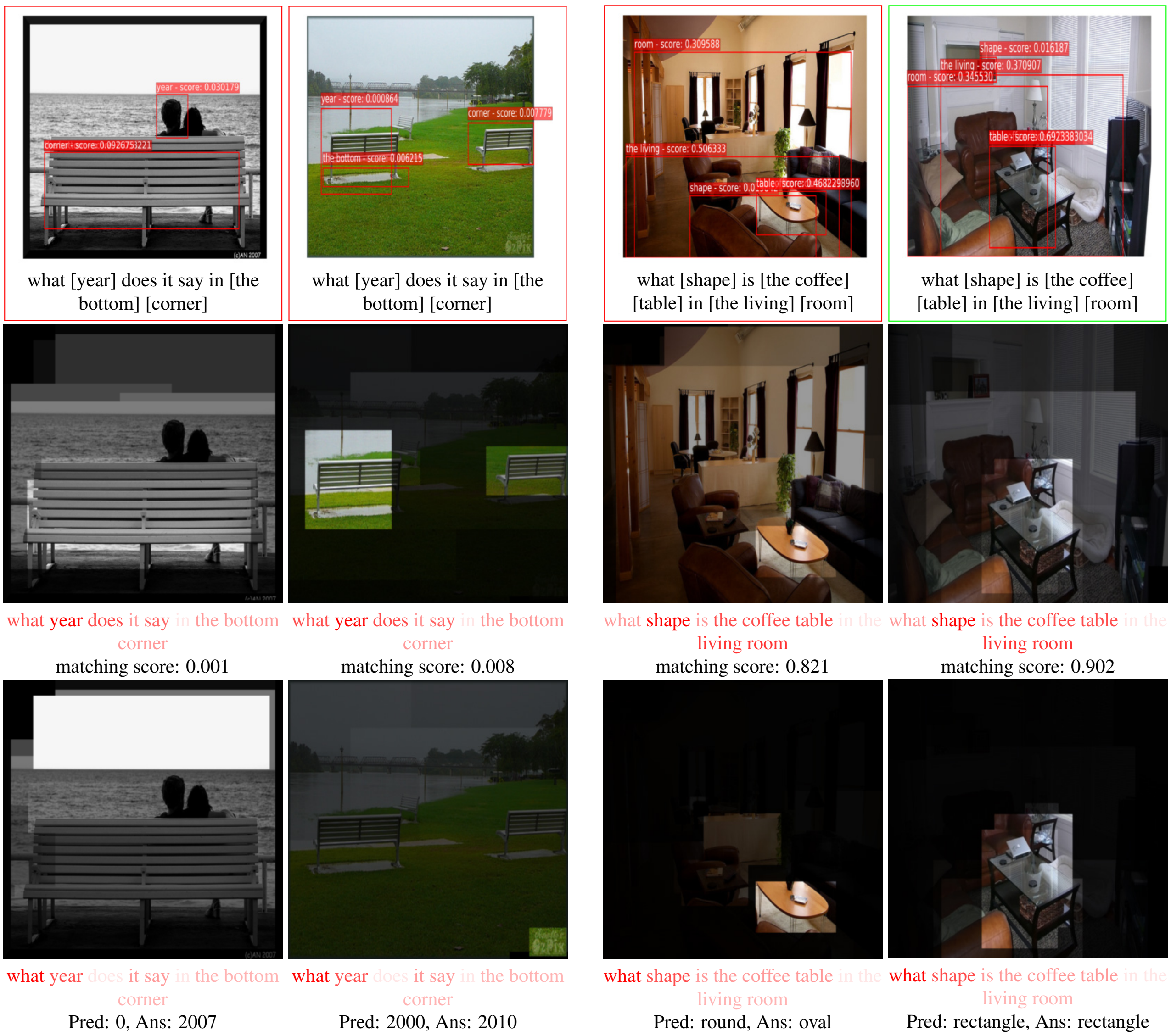
this acidic food - score: 0.307434
is [this acidic food]
this acidic food - score: 0.410558
is [this acidic food]
the bowl - score: 0.024326
what [animal] is in [the bowl]
the bowl - score: 0.038335
what [animal] is in [the bowl]
is this acidic food
matching score: 0.065
is this acidic food
matching score: 0.113
what animal is in the bowl
matching score: 3.393e-06
what animal is in the bowl
matching score: 3.42e-06
is this acidic food
Pred: yes, Ans: yes
is this acidic food
Pred: yes, Ans: no
what animal is in the bowl
Pred: bird, Ans: fish
what animal is in the bowl
Pred: cat, Ans: none

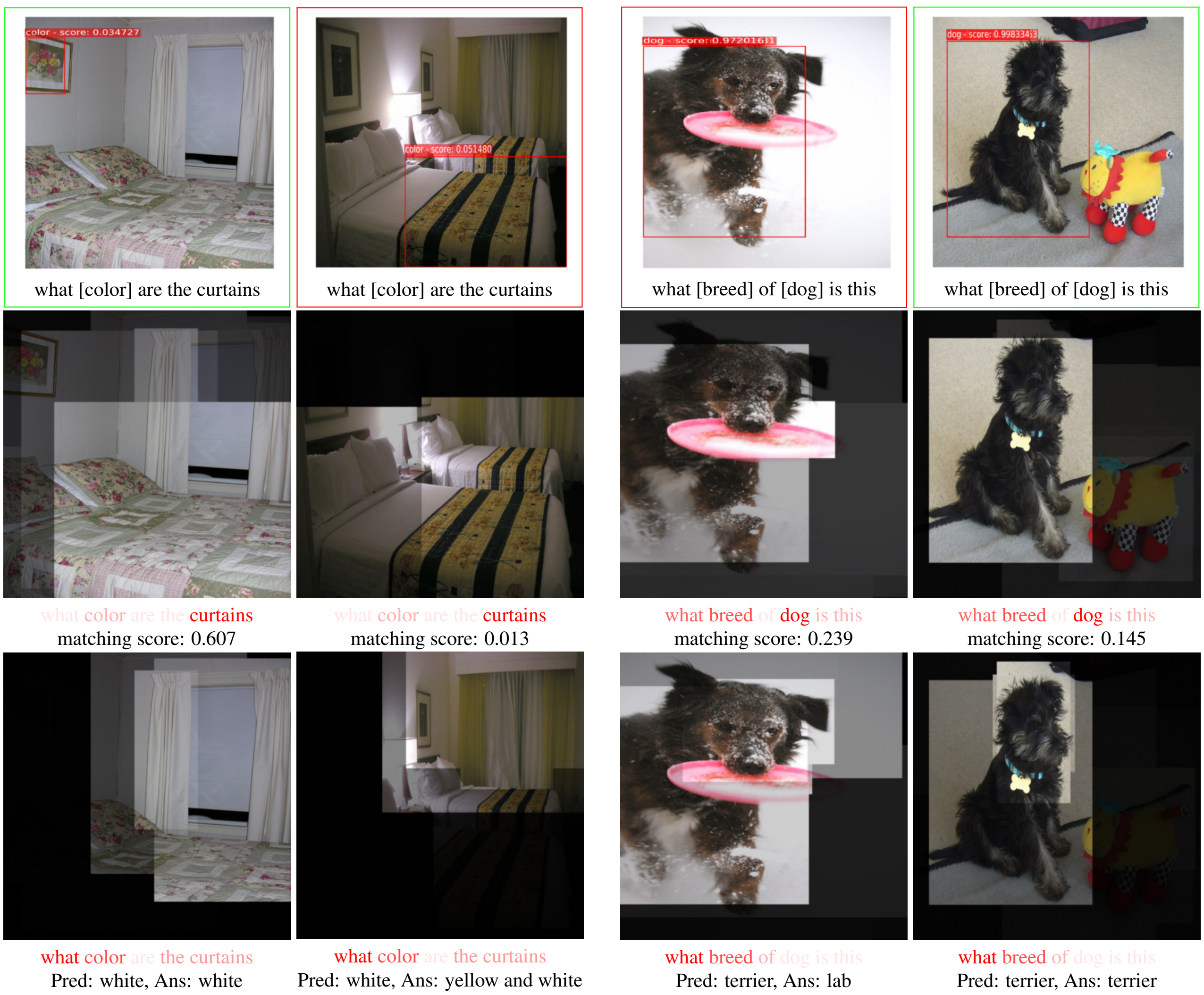

the window - score: 0.696018
is there [a reflection] in [the window]
the window - score: 0.688192169
is there [a reflection] in [the window]
the number - score: 0.341572
the white shirt - score: 0.942399
orange - score: 0.201187
what is [the number] in [orange] on [the white shirt]
the white shirt - score: 0.522118
orange - score: 0.812822
what is [the number] in [orange] on [the white shirt]
is there a reflection in the window
matching score: 0.008
is there a reflection in the window
matching score: 0.143
what is the number in orange on the white shirt
matching score: 0.001
what is the number in orange on the white shirt
matching score: 0.001
is there a reflection in the window
Pred: yes, Ans: yes
is there a reflection in the window
Pred: yes, Ans: no
what is the number in orange on the white shirt
Pred: 0, Ans: 42
what is the number in orange on the white shirt
Pred: 0, Ans: nothing


the plane - score: 0.951778
is there [a pilot] inside of [the plane]
the plane - score: 0.630671
a pilot - score: 0.141129
is there [a pilot] inside of [the plane]
color - score: 0.169060
what [color] is she wearing
color - score: 0.319022
what [color] is she wearing
is there a pilot inside of the plane
matching score: 0.221
is there a pilot inside of the plane
matching score: 0.502
what color is she wearing
matching score: 0.007
what color is she wearing
matching score: 0.014
is there a pilot inside of the plane
Pred: no, Ans: no
is there a pilot inside of the plane
Pred: no, Ans: yes
what color is she wearing
Pred: black, Ans: orange
what color is she wearing
Pred: gray, Ans: gray


what [branch] [bat] is [the boy] in [the blue] [shirt] using
what [branch] [bat] is [the boy] in [the blue] [shirt] using
what is on [the floor] by [the toilet]
what is on [the floor] by [the toilet]
what brand bat is the boy in the blue shirt using
matching score: 0.229
what brand bat is the boy in the blue shirt using
matching score: 0.108
what is on the floor by the toilet
matching score: 0.744
what is on the floor by the toilet
matching score: 0.664
what brand bat is the boy in the blue shirt using
Pred: wilson, Ans: nike
what brand bat is the boy in the blue shirt using
Pred: wilson, Ans: wilson
what is on the floor by the toilet
Pred: tile, Ans: trash can
what is on the floor by the toilet
Pred: toilet brush, Ans: toilet brush


the background - score: 0.011953
what are those steal structures in [the background]
the background - score: 0.004568
what are those steal structures in [the background]
the suitcase - score: 0.966423
how many tags are on [the suitcase]
the suitcase - score: 0.715322
how many tags are on [the suitcase]
what are those steal structures in the background
matching score: 7.355e-06
what are those steal structures in the background
matching score: 1.032e-06
how many tags are on the suitcase
matching score: 0.893
how many tags are on the suitcase
matching score: 0.457
what are those steal structures in the background
Pred: houses, Ans: toilets
what are those steal structures in the background
Pred: fence, Ans: fence
how many tags are on the suitcase
Pred: 4, Ans: 3
how many tags are on the suitcase
Pred: 0, Ans: 0


what [year] does it say in [the bottom] [corner]
what [year] does it say in [the bottom] [corner]
what [shape] is [the coffee] [table] in [the living] [room]
what [shape] is [the coffee] [table] in [the living] [room]
what year does it say in the bottom corner
matching score: 0.001
what year does it say in the bottom corner
matching score: 0.008
what shape is the coffee table in the living room
matching score: 0.821
what shape is the coffee table in the living room
matching score: 0.902
what year does it say in the bottom corner
Pred: 0, Ans: 2007
what year does it say in the bottom corner
Pred: 2000, Ans: 2010
what shape is the coffee table in the living room
Pred: round, Ans: oval
what shape is the coffee table in the living room
Pred: rectangle, Ans: rectangle


color - score: 0.034727
what [color] are the curtains
color - score: 0.051480
what [color] are the curtains
dog - score: 0.97201611
what [breed] of [dog] is this
what [breed] of [dog] is this
what color are the curtains
matching score: 0.607
what color are the curtains
matching score: 0.013
what breed of dog is this
matching score: 0.239
what breed of dog is this
matching score: 0.145
what color are the curtains
Pred: white, Ans: white
what color are the curtains
Pred: white, Ans: yellow and white
what breed of dog is this
Pred: terrier, Ans: lab
what breed of dog is this
Pred: terrier, Ans: terrier

E. Visualization of Inference of VG and ICR on Image Caption Retrieval

We show visualization of our network using an ICR dataset, MSCOCO. As success cases are not so informative, which are often shown in previous studies, we show only failure cases for the two subtasks of ICR, more specifically, the cases where the top-1 prediction is not correct for image annotation (Sec. E.1) and image retrieval (Sec. E.2). Each panel consists of 2×2 image-caption pairs; the first row shows VG and the second row shows ICR; the first column (with a green box) shows VG and ICR visualization for the ground-truth image-caption pair and the second column (with a red box) shows a pair of an input image and the predicted top-1 caption for image annotation and a pair of the predicted top-1 image and an input caption for image retrieval .

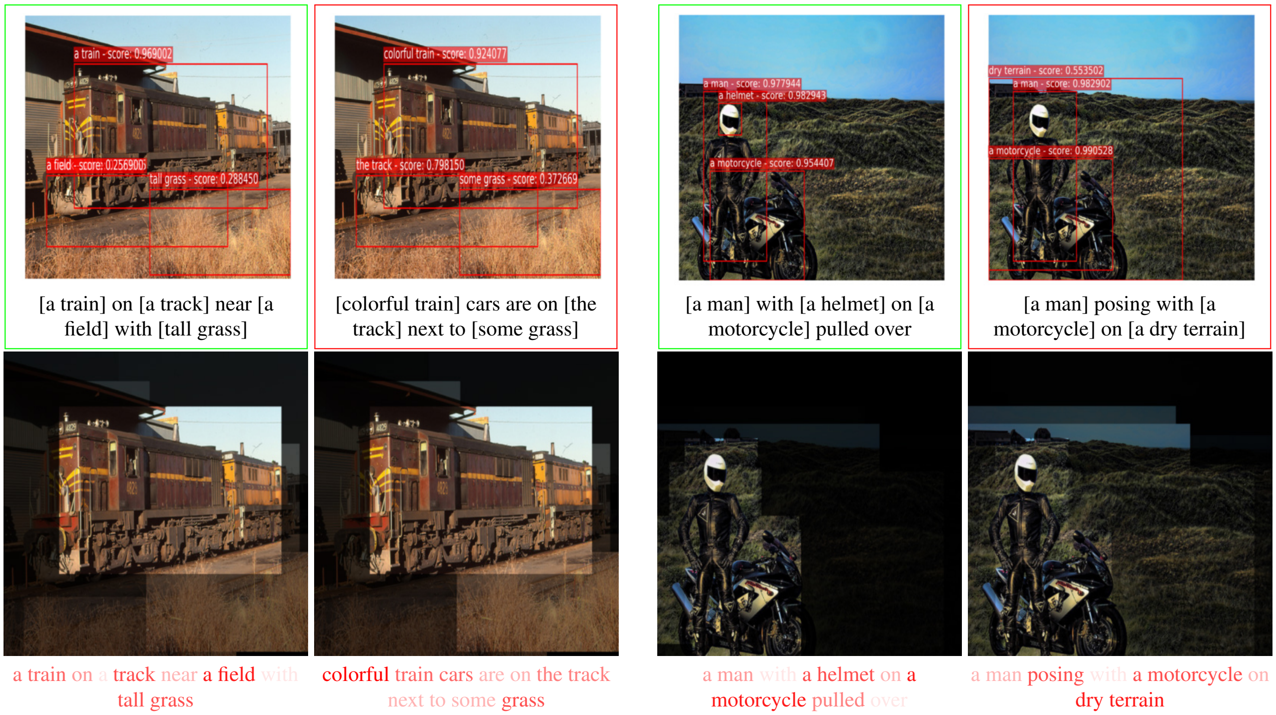
We think that the failure cases can be categorized into the following three types:

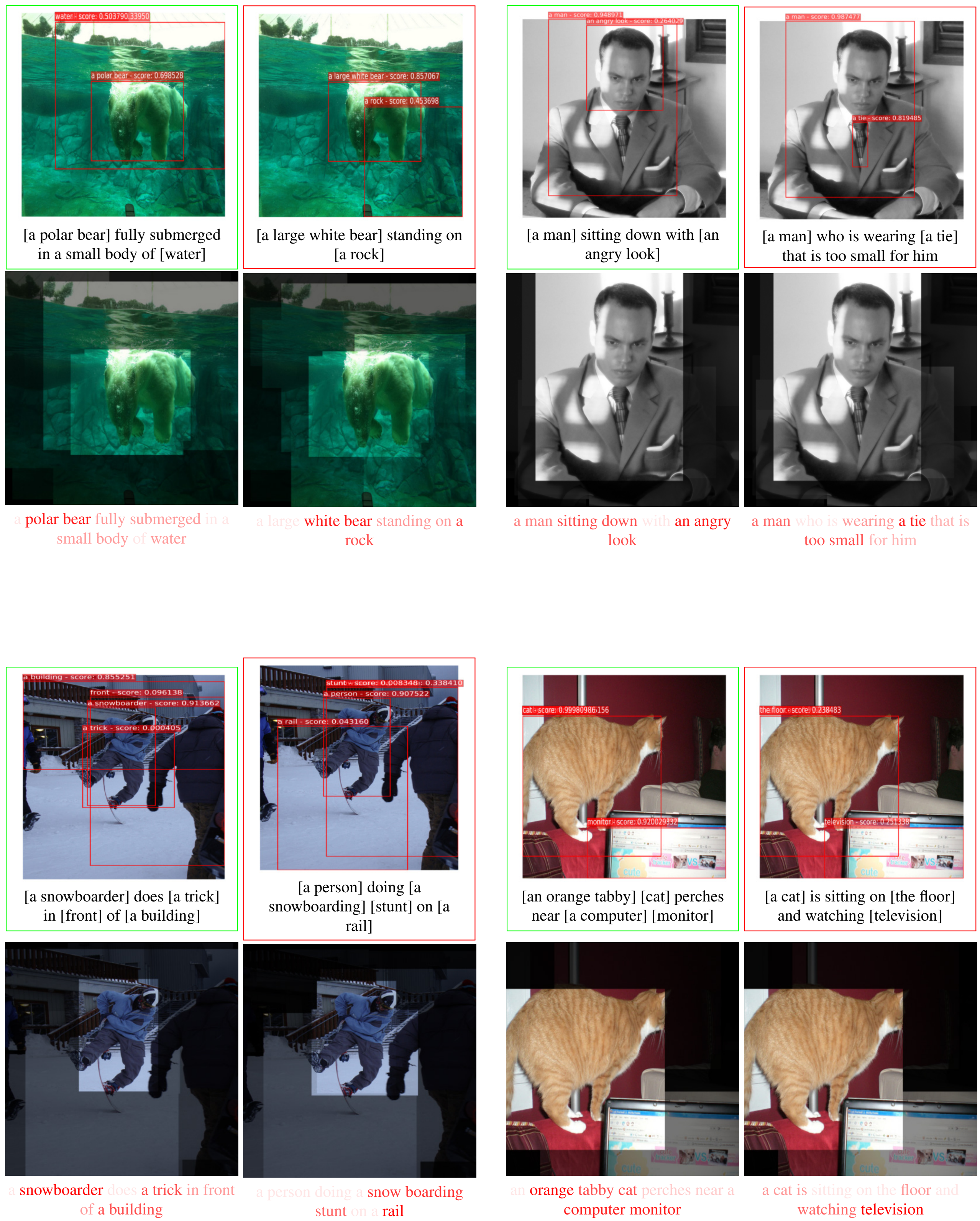
1) The network incorrectly recognizes objects or concepts that have similar appearance, since the Faster R-CNN features are not rich enough to distinguish them (e.g., "*computer monitor*" and "*television*"). Specifically, the VG decoder detects wrong objects, matching the input image with a wrong caption.

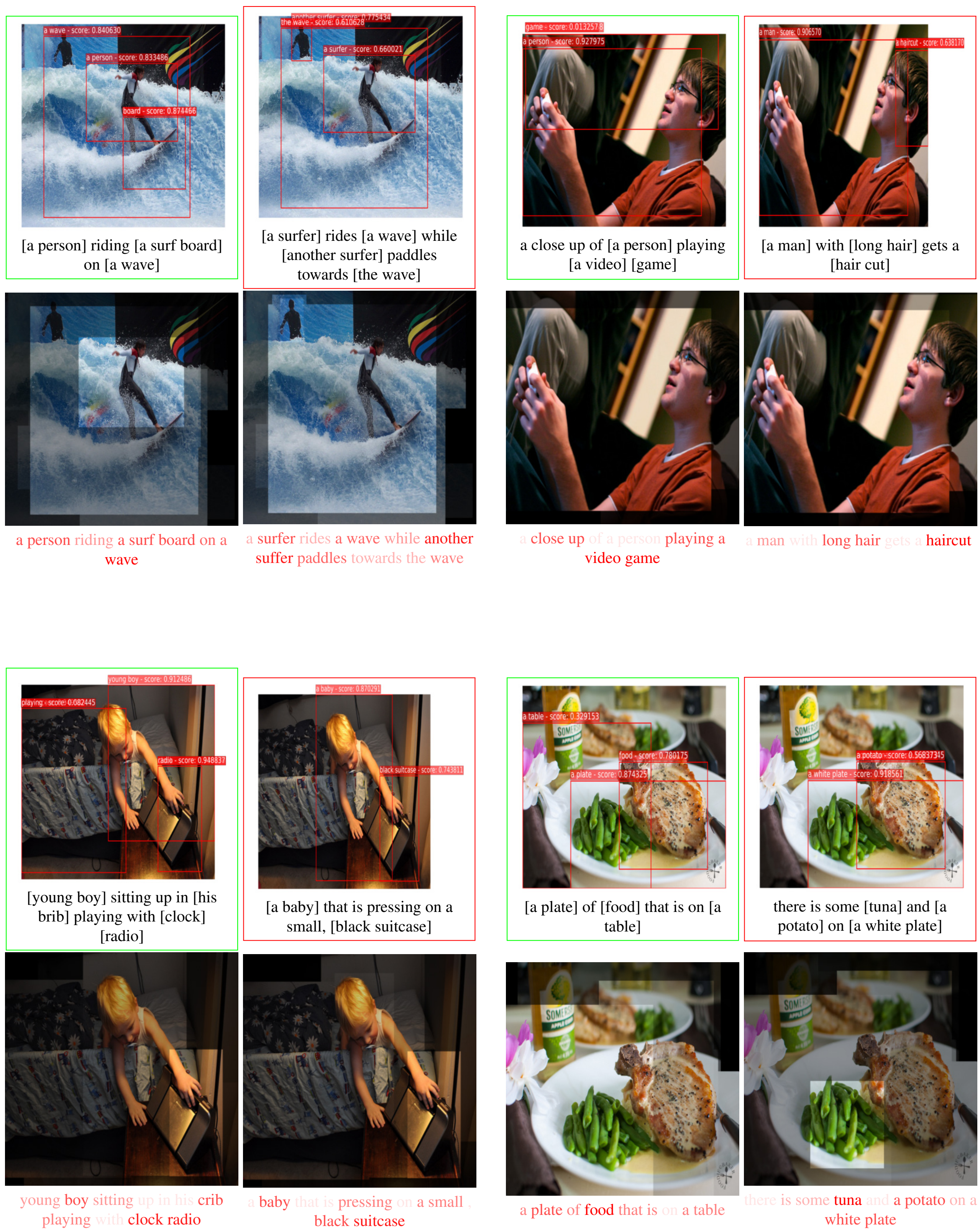
2) The VG and ICR decoders correctly align objects in the input caption with the corresponding image regions, but fail to recognize their actions (e.g., "*submerged in a small body of water*" vs. "*standing on a rock*"; "*perches*" vs. "*standing*"). Such failures may be eliminated by creating a dataset for the task of predicting such actions and using it in the joint training.

3) As in the case of VQA, although the top-1 caption or image predicted by our network matches well with the input image or caption, it is judged wrong because it is not listed in the set of correct answers. For image annotation, this happens because the same image content can be described in many ways (e.g., "*a train on a track near a field with tall grass*" or "*colorful train cars are on the track next to some grass*"); or because "correct" captions explain only one of multiple contents contained in input images (e.g., "*a man on the beach who is carrying a surfboard*" and "*multiple people are standing on the beach at the edge of the water*"). For image retrieval, the same often occurs when the input captions provide only too general explanation of a scene, most of which tend to be short simple captions, such as "*there are people flying kites in the park*". It should be noted that even if this is the case, the VG decoder is able to correctly detect objects in most cases.

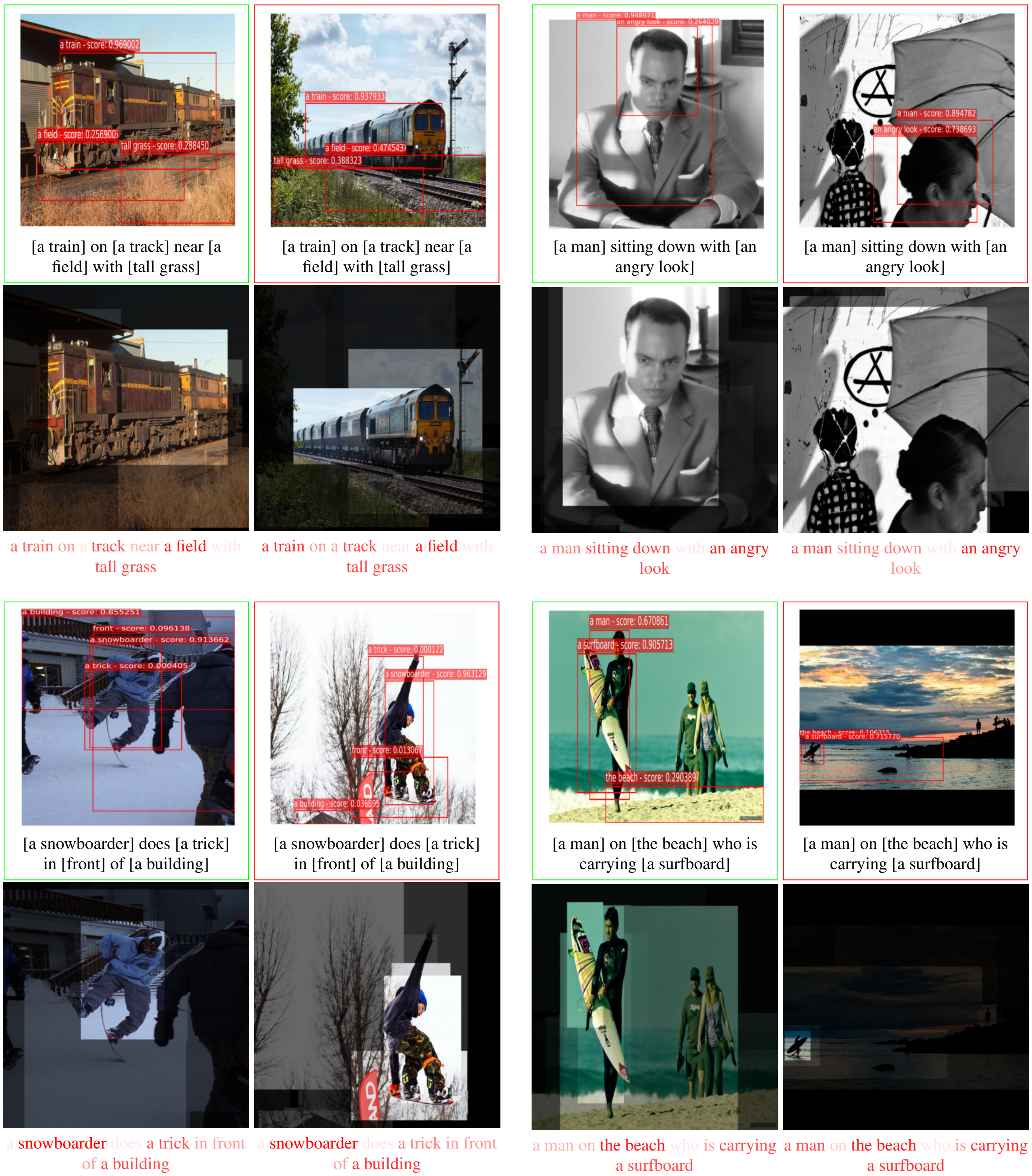
E.1. Image Annotation

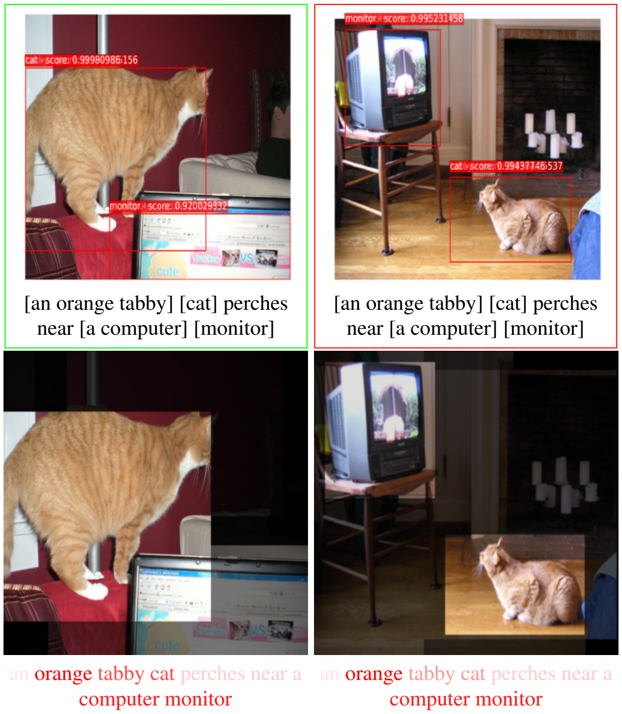



[a polar bear] fully submerged in a small body of [water]
[a large white bear] standing on [a rock]
[a man] sitting down with [an angry look]
[a man] who is wearing [a tie] that is too small for him
a polar bear fully submerged in a small body of water
a large white bear standing on a rock
a man sitting down with an angry look
a man who is wearing a tie that is too small for him
[a snowboarder] does [a trick] in [front] of [a building]
[a person] doing [a snowboarding] [stunt] on [a rail]
[an orange tabby] [cat] perches near [a computer] [monitor]
[a cat] is sitting on [the floor] and watching [television]
a snowboarder does a trick in front of a building
a person doing a snow boarding stunt on a rail
an orange tabby cat perches near a computer monitor
a cat is sitting on the floor and watching television

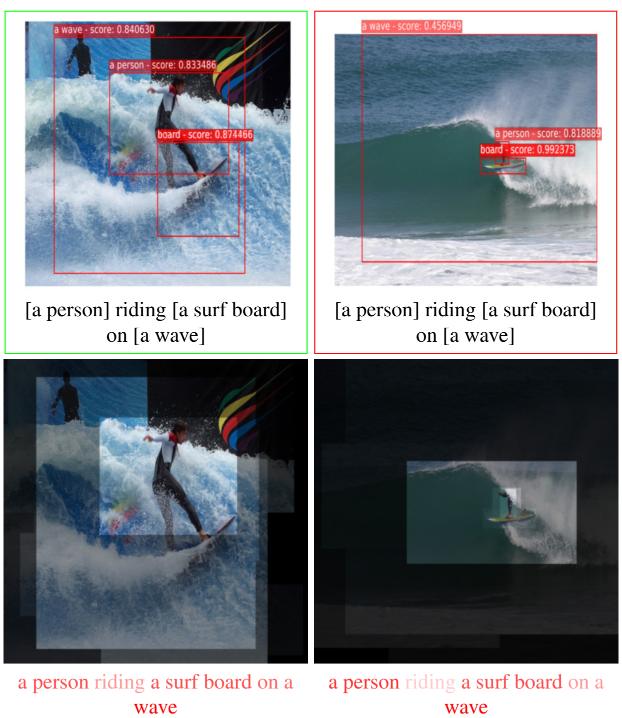

[a person] riding [a surf board] on [a wave]
[a surfer] rides [a wave] while [another surfer] paddles towards [the wave]
a close up of [a person] playing [a video] [game]
[a man] with [long hair] gets a [hair cut]
a person riding a surf board on a wave
a surfer rides a wave while another suffer paddles towards the wave
a close up of a person playing a video game
a man with long hair gets a haircut
[young boy] sitting up in [his brib] playing with [clock] [radio]
[a baby] that is pressing on a small, [black suitcase]
[a plate] of [food] that is on [a table]
there is some [tuna] and [a potato] on [a white plate]
young boy sitting up in his crib playing with clock radio
a baby that is pressing on a small , black suitcase
a plate of food that is on a table
there is some tuna and a potato on a white plate

E.2. Image Retrieval

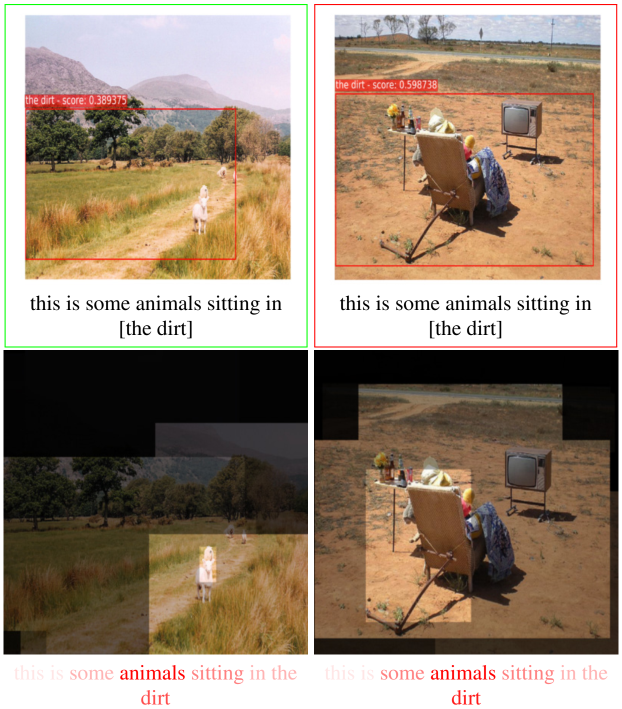



[an orange tabby] [cat] perches near [a computer] [monitor]
[an orange tabby] [cat] perches near [a computer] [monitor]
an orange tabby cat perches near a computer monitor
an orange tabby cat perches near a computer monitor

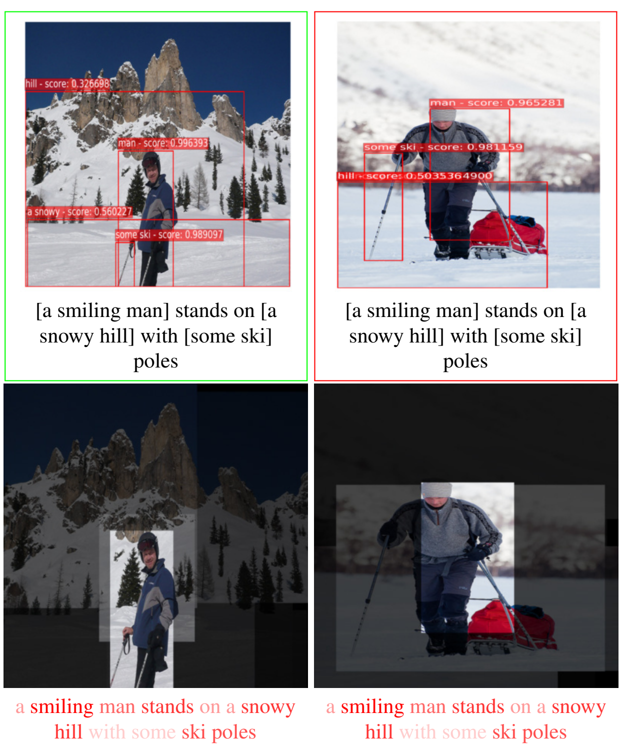

[a dude] on [a bile] rides quietly across [the place]
[a dude] on [a bile] rides quietly across [the place]
a dude on a bike rides quietly across the place
a dude on a bike rides quietly across the place


[a person] riding [a surf board] on [a wave]
[a person] riding [a surf board] on [a wave]
a person riding a surf board on a wave
a person riding a surf board on a wave


there are people flying kites in [the park]
there are people flying kites in [the park]
there are people flying kites in the park
there are people flying kites in the park


the dirt - score: 0.389375
the dirt - score: 0.598738
this is some animals sitting in [the dirt]
this is some animals sitting in [the dirt]
this is some animals sitting in the dirt
this is some animals sitting in the dirt


a little girl - score: 0.975363
a little girl - score: 0.984012
[a little girl] doing some arts and crafts
[a little girl] doing some arts and crafts
a little girl doing some arts and crafts
a little girl doing some arts and crafts


hill - score: 0.326698
man - score: 0.996393
a snowy - score: 0.560227
some ski - score: 0.989097
man - score: 0.965281
some ski - score: 0.981159
[a smiling man] stands on [a snowy hill] with [some ski] poles
[a smiling man] stands on [a snowy hill] with [some ski] poles
a smiling man stands on a snowy hill with some ski poles
a smiling man stands on a snowy hill with some ski poles


a man - score: 0.967699
woman - score: 0.933668
woman - score: 0.952911
a man - score: 0.945583
[a man] and [woman] look into each others eyes while getting married
[a man] and [woman] look into each others eyes while getting married
a man and woman look into each others eyes while getting married
a man and woman look into each others eyes while getting married

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[2] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009. 1

[3] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler. VSE++: improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 3

[4] Y. Huang, Q. Wu, C. Song, and L. Wang. Learning semantic concepts and order for image and sentence matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[5] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[6] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[7] X. Lin and D. Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[8] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 1

[9] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[10] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR)*, 2016. 3