Visual Madlibs: Fill in the blank Description Generation and Question Answering
Supplementary File

Licheng Yu, Eunbyung Park, Alexander C. Berg, Tamara L. Berg
Department of Computer Science, University of North Carolina, Chapel Hill
{licheng, eunbyung, aberg, tlberg}@cs.unc.edu

1. Results of filtered easy and hard tasks

We show the full tables of accuracies for the filtered easy and hard multiple-choice tasks in Table 1.

2. Quantitative analysis of Madlibs responses

In Section 5.1 of our paper, we analyzed the the phrasal structures of our collected Visual Madlibs descriptions for several of our fill-in-the-blank questions. Here, we show the relative frequencies for the top-5 most frequent templates used for all 12 Madlibs questions. In Fig. 1, it is observed that most of the distributions are concentrated on just a few choices, except for the future and past descriptions. One reason is that this question is more open ended, so Turkers are likely to write more lengthy descriptions, i.e., "One or two seconds before this picture was taken, ___". In this setup, annotators have great freedom in expressing their ideas.

We also show an analysis of answer consistency in Fig. 2. Here we compute a histogram of answer similarities for each question, where similarity is measured as the cosine similarity between the mean Word2Vec representations of the 3 collected answers. Most of the similarity histograms have a normal-like distribution with an extra peak around 1, which implies we get some very similar answers for a portion of questions. The exception is the answers for the questions about the past and future, where there is no peak at one for distribution of similarities. This indicates that there are fewer images for past/future predictions where people generate the exact same description, but for some images people do display more consistency than for other images. Also note, the mean for image's emotion is smaller compared with the others, perhaps indicating that this question is relatively more subjective.

Finally, we analyze the word distribution for each question type. In Fig. 3, we show the top 20 most frequently used words for all types of Madlibs questions. It is interesting to find that color words are more often used to describe object's attribute, while the entry-level category words ('woman', 'boy', 'girl', 'child') and the clothes are usually used to describe person's attribute. More than 20% of the images express positive emotions, i.e., 'happy' and 'excited'. Several words related to person are often used to indicate the relative position of object, e.g, 'hand', 'person', 'man', 'woman', etc. Perhaps it is due to the human-centric property of the Visual Madlibs images.

3. Additional examples of results

We also show additional examples of the two Visual Madlibs tasks: multiple-choice question answering and focused sentence generation. We first show some correct answers to multiple choice questions in Fig. 4, as well as some wrong answers in Fig. 5. All examples are from the hard version of our multiple-choice question answering task and answers are selected by the nCCA joint-embedding method. This task provides a more straightforward way to measure the quality of the learned joint embedding space in an application scenario.

Then, in Fig. 6, we show some focused sentence generation examples, generated by nCCA and CNN+LSTM. As observed, the nCCA can generate richer but sometimes unrelated sentences, while CNN+LSTM is able to generate relatively shorter but accurate sentences, which helps to achieve higher BLEU-1 and BLEU-2 scores.

Filtered Questions from **Easy Task**

	#Q	n-gram	CCA	nCCA	nCCA (place)	nCCA (bbox)	nCCA (all)	CNN+LSTM (madlibs)	CNN+LSTM(r) (madlibs)	Human
1. scene	5997	24.6%	77.4%	88.8%	87.4%	–	**89.7**%	76.1%	79.4%	96.4%
2. emotion	2663	27.5%	48.3%	58.8%	**59.7**%	–	51.0%	39.4%	49.0%	75.5%
3. past	4703	26.0%	62.8%	78.9%	73.9%	–	**81.7**%	50.5%	47.1%	96.8%
4. future	4495	28.7%	62.2%	79.4%	73.3%	–	**81.5**%	51.2%	51.4%	97.1%
5. interesting	4940	26.4%	67.6%	77.5%	72.9%	–	**79.8**%	56.1%	50.5%	96.8%
6. obj attr	6681	32.2%	45.1%	48.9%	45.8%	56.6%	52.4%	48.3%	**60.8**%	93.3%
7. obj aff	7043	31.0%	60.8%	74.3%	70.8%	73.5%	77.9%	–	**90.8**%	95.8%
8. obj pos	6906	27.0%	54.1%	67.3%	65.6%	60.3%	**71.0**%	54.9%	71.5%	94.9%
9. per attr	5753	27.3%	42.1%	50.5%	46.6%	**56.6**%	46.2%	37.2%	49.2%	92.2%
10. per act	6384	27.3%	70.6%	81.2%	77.4%	76.3%	**83.3**%	65.1%	69.5%	97.9%
11. per loc	6193	24.7%	72.0%	85.1%	85.2%	76.1%	**85.3**%	62.1%	73.6%	96.8%
12. pair rel	7206	29.5%	55.4%	64.4%	62.8%	65.6%	68.6%	–	**74.0**%	95.1%

Filtered Questions from **Hard Task**

	#Q	n-gram	CCA	nCCA	nCCA (place)	nCCA (bbox)	nCCA (all)	CNN+LSTM (madlibs)	CNN+LSTM(r) (madlibs)	Human
1. scene	4940	20.9%	70.4%	77.6%	**77.8**%	–	76.3%	69.4%	69.7%	89.5%
2. emotion	2052	27.7%	43.1%	49.0%	**49.5**%	–	43.8%	38.5%	43.0%	72.2%
3. past	3976	24.3%	51.0%	57.4%	53.8%	–	**59.4**%	42.8%	41.3%	86.6%
4. future	3820	25.7%	51.4%	**59.2**%	54.2%	–	58.3%	42.1%	41.7%	87.6%
5. interesting	4159	26.9%	56.1%	59.5%	55.1%	–	**61.3**%	47.8%	40.3%	89.5%
6. obj attr	5436	30.3%	45.3%	47.2%	44.7%	**54.6**%	42.8%	45.1%	46.3%	86.2%
7. obj aff	4581	28.0%	61.2%	71.0%	67.6%	70.5%	57.6%	–	**79.0**%	73.7%
8. obj pos	5721	29.1%	53.0%	**60.2**%	57.7%	54.6%	57.7%	48.8%	54.3%	84.5%
9. per attr	4893	25.7%	36.5%	42.4%	38.8%	**52.1**%	34.4%	36.1%	46.4%	88.2%
10. per act	5813	27.7%	62.0%	68.3%	65.3%	67.9%	**69.6**%	59.1%	55.3%	92.7%
11. per loc	5096	23.6%	63.1%	69.9%	**71.7**%	62.6%	70.0%	52.9%	60.6%	88.2%
12. pair rel	5981	28.5%	52.3%	57.6%	55.4%	**60.0**%	56.5%	–	57.4%	88.5%

Table 1: Accuracies computed for different approaches on the filtered multiple-choice questions of easy and hard task. CCA, nCCA, and CNN+LSTM are trained on the whole image representation for each type of question. nCCA(place) uses Places-CNN feature. nCCA(box) is trained and evaluated on ground-truth bounding-boxes from COCO segmentations. nCCA(all) trains a single embedding using all question types. CNN+LSTM(r) ranks the perplexity of {prompt+choice}.

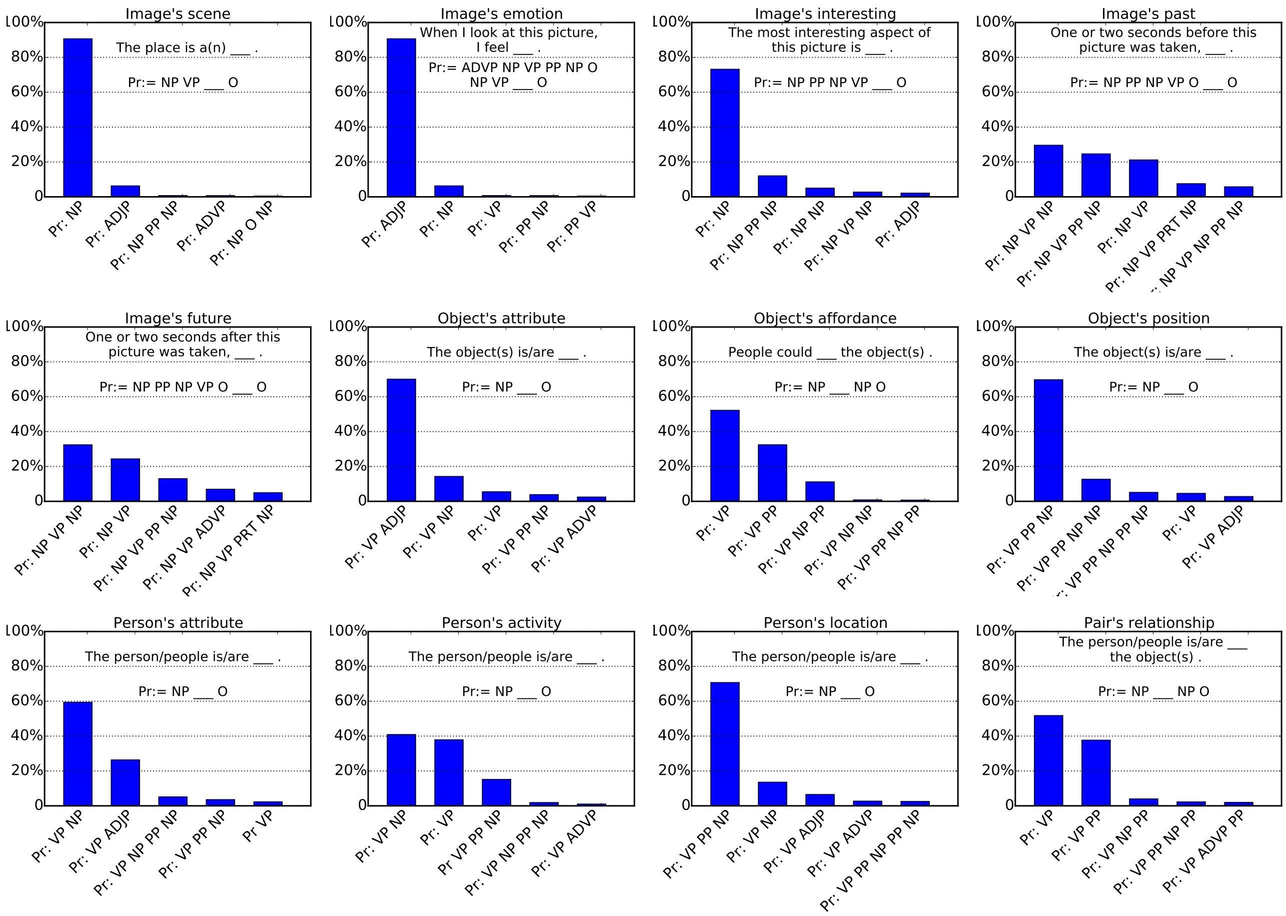


Figure 1: Top-5 most frequent phrase templates for 12 types of Madlibs questions.

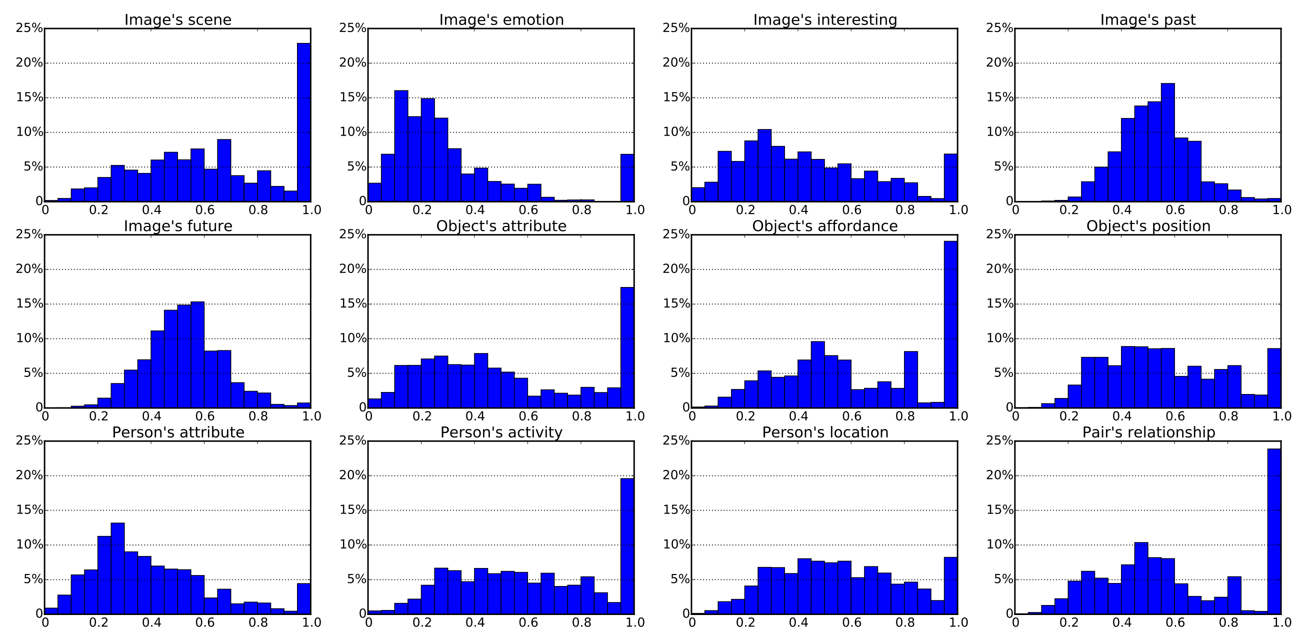


Figure 2: Histograms of similarity of answers for 12 types of Madlibs questions.

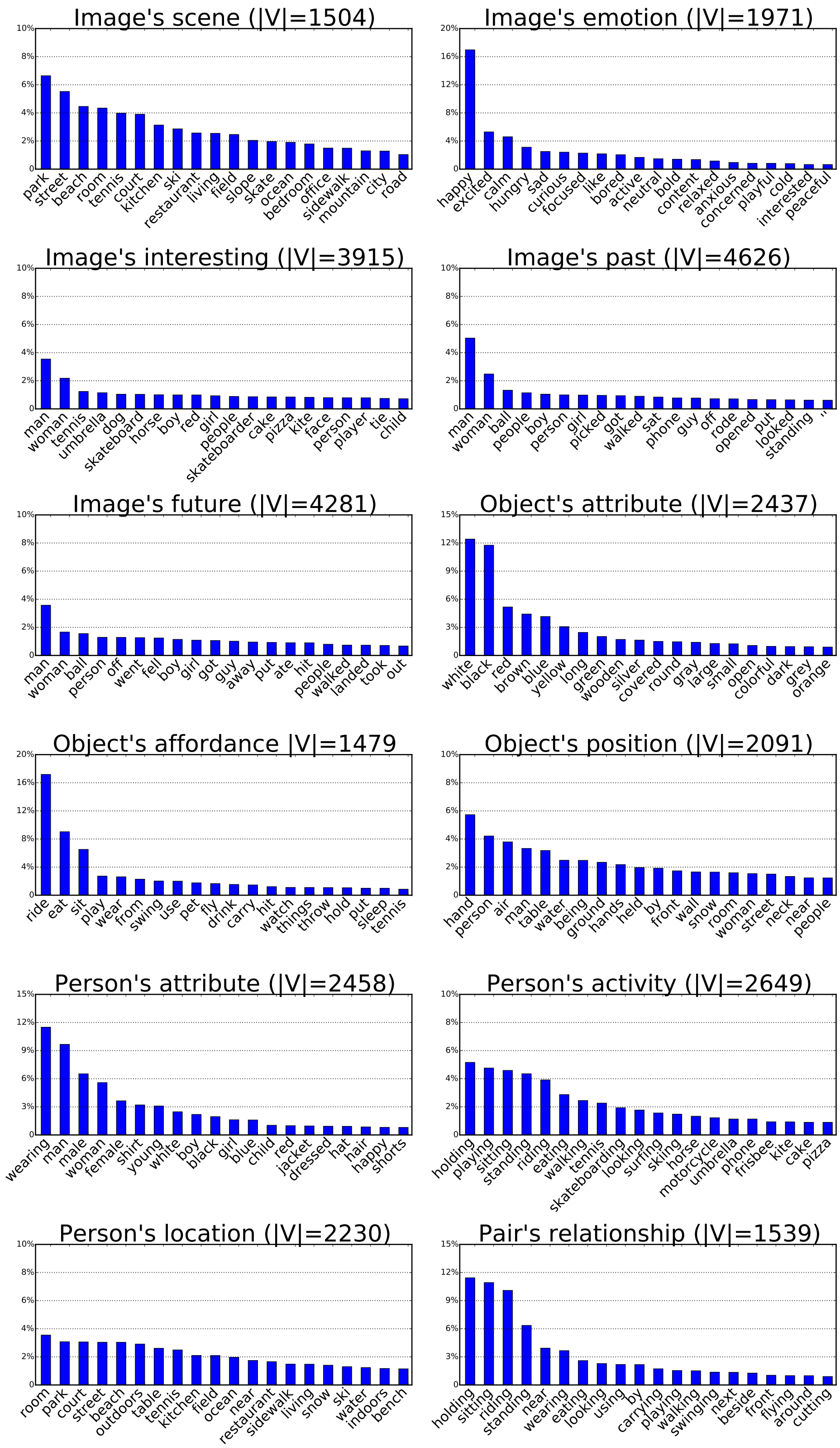


Figure 3: Top 20 words used in the answers for all 12 types of questions. $|V|$ denotes the vocabulary size. Note y-axis range differs across question types, depending on the highest bin of each.

1. image's scene

The place is a(n) ___ .
- [x] **mechanics shop**
- [ ] office
- [ ] tennis match
- [ ] living room

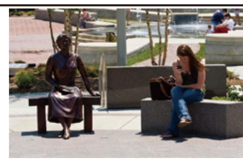
The place is a(n) ___ .
- [ ] track
- [ ] lawn
- [x] **park**
- [ ] tennis court

The place is a(n) ___ .
- [ ] office
- [ ] park
- [x] **tennis court**
- [ ] beach

The place is a(n) ___ .
- [x] **beach**
- [ ] ocean
- [ ] riverfront
- [ ] dock

2. image's emotion

When I look at this picture, I feel ___ .
- [x] **calm**
- [ ] happy
- [ ] gross
- [ ] stressed



When I look at this picture, I feel ___ .
- [x] **relaxed**
- [ ] affection
- [ ] happy
- [ ] rejected

When I look at this picture, I feel ___ .
- [ ] thirsty
- [ ] mad
- [x] **hungry**
- [ ] playful

When I look at this picture, I feel ___ .
- [ ] bored
- [ ] happy
- [ ] peace
- [x] **hot**

3. image's interesting

The most interesting aspect of the picture is ___ .
- [x] **the sheep with a pink head**
- [ ] dog
- [ ] the facial expressions
- [ ] the sheep's belly

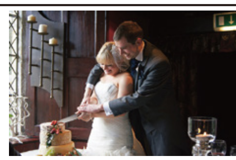
The most interesting aspect of the picture is ___ .
- [x] **horses racing**
- [ ] the rider
- [ ] the huge watermark
- [ ] the vinageness of this picture

The most interesting aspect of the picture is ___ .
- [ ] the boat in the background
- [x] **the blue water is gorgeous**
- [ ] the man is about to fall over
- [ ] the man

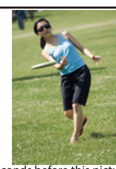
The most interesting aspect of the picture is ___ .
- [x] **the donuts**
- [ ] the girl's posture
- [ ] the texture of the donut and the skin
- [ ] the eye that can see through the the donuts

4. image's past

One or two seconds before this picture was taken, ___ .
- [x] **the man was surfing**
- [ ] the surfer gave the thumbs up
- [ ] the woman put on her sunglasses
- [ ] several people sat down



One or two seconds before this picture was taken, ___ .
- [ ] they cut the cake
- [ ] the person asked another person to take a picture.
- [x] **a man embraces a woman**
- [ ] tennis court

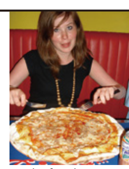


One or two seconds before this picture was taken, ___ .
- [x] **the person was holding the frisbee**
- [ ] he jumped up
- [ ] the family piled out of a truck
- [ ] a wave came ashore

One or two seconds before this picture was taken, ___ .
- [x] **the ball was hit to her side of the court**
- [ ] he picked up his racket
- [ ] she received a serve
- [ ] a man sat

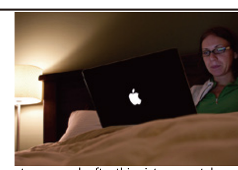
5. image's future

One or two seconds after this picture was taken, ___ .
- [ ] the guy was facing the camera
- [x] **the woman takes a bit out of the banana**
- [ ] the little boy knocked over the pile
- [ ] the man looked at the camera



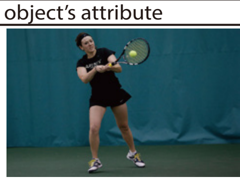
One or two seconds after this picture was taken, ___ .
- [ ] someone make a toast
- [ ] the man felt queasy
- [x] **she ate the pizza**
- [ ] the girl stopped smiling

One or two seconds after this picture was taken, ___ .
- [x] **a girl cuts a piece of cake**
- [ ] the man will pick up a cup
- [ ] the guy will get his own phone
- [ ] the man talks babytalk to the baby



One or two seconds after this picture was taken, ___ .
- [ ] the boy continued to smile
- [ ] the boys will eat carrots
- [ ] she will jump on the bed
- [x] **a woman closed a laptop**

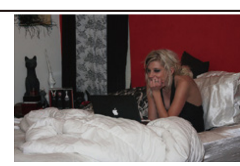
6. object's attribute



The tennis racket is ___ .
- [x] **black**
- [ ] red
- [ ] metal
- [ ] plastic

The skis is ___ .
- [ ] blue
- [ ] colorful
- [x] **long**
- [ ] short

The surfboard is ___ .
- [ ] long and slim
- [ ] red and black
- [x] **white and blue**
- [ ] yellow and red



The laptop is ___ .
- [x] **black**
- [ ] on
- [ ] gray
- [ ] silver

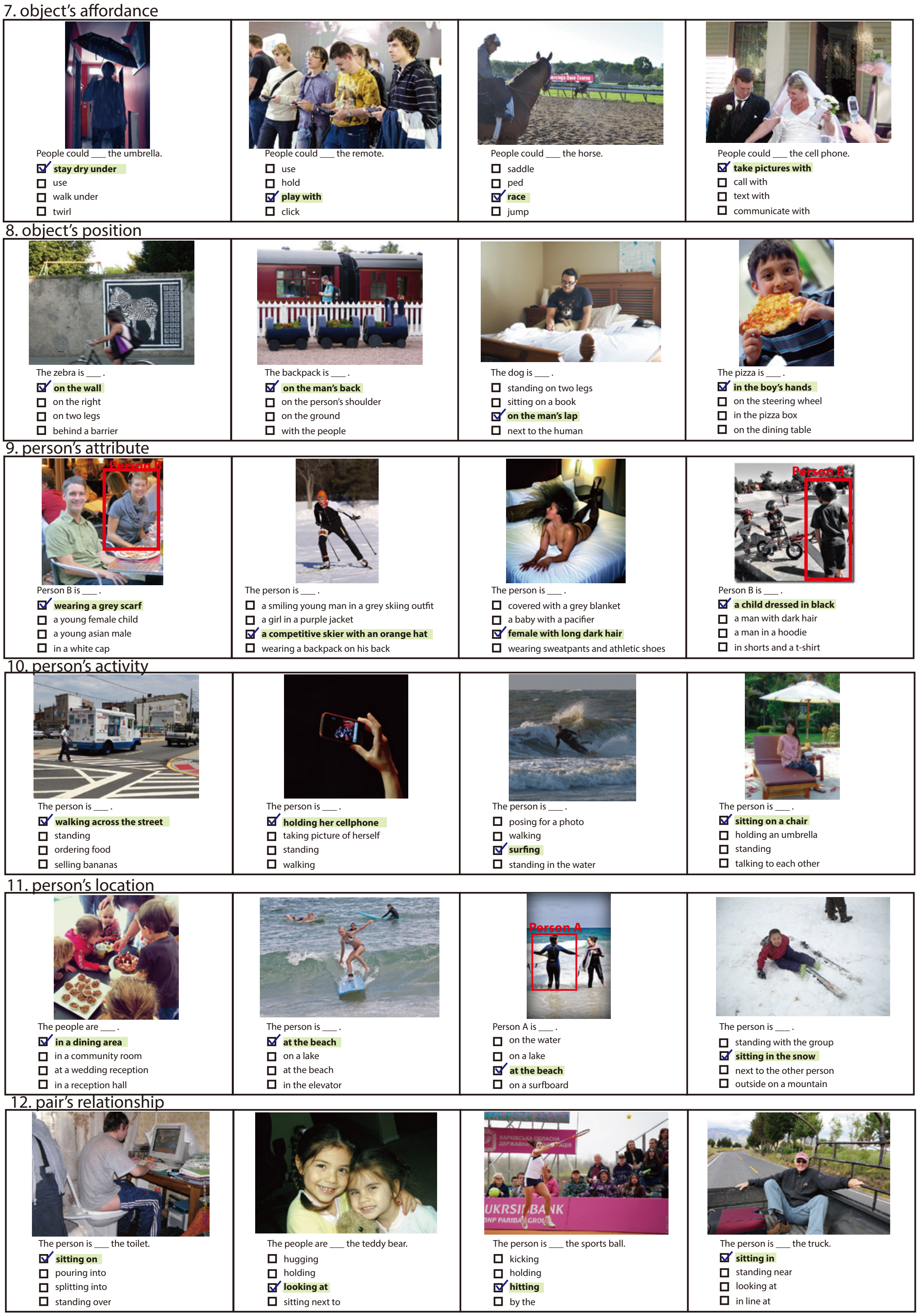


Figure 4: Examples of correct answers made by nCCA for 12 types of multiple-choice question-answering.

1. image's scene

This place is a(n) ___ .

- [ ] **court house**
- [x] bar
- [ ] wedding
- [ ] street

This place is a(n) ___ .

- [x] river
- [ ] hacienda
- [ ] **vacation spot**
- [ ] canal

This place is a(n) ___ .

- [x] airport
- [ ] street
- [ ] **subway**
- [ ] sidewalk

This place is a(n) ___ .

- [ ] **classroom**
- [ ] living room
- [ ] office
- [x] work place

2. image's emotion

When I look at this picture, I feel ___ .

- [x] hungry
- [ ] **serious**
- [ ] silly
- [ ] interested

When I look at this picture, I feel ___ .

- [x] despair
- [ ] concerned
- [ ] **natural**
- [ ] healthy

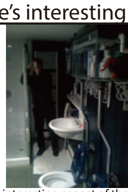
When I look at this picture, I feel ___ .

- [x] hungry
- [ ] annoyed
- [ ] **wierd**
- [ ] guarded

When I look at this picture, I feel ___ .

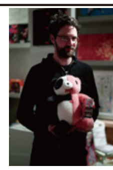
- [ ] **gentle**
- [ ] pretty
- [ ] tired
- [x] sad

3. image's interesting



The most interesting aspect of the picture is ___ .

- [ ] **the guy taking self inside a bathroom**
- [x] the toilet
- [ ] girls panties
- [ ] the person on the floor



The most interesting aspect of the picture is ___ .

- [x] the adorable children
- [ ] the injuries
- [ ] **the bear**
- [ ] facial expression

The most interesting aspect of the picture is ___ .

- [x] the wine
- [ ] the people who are celebrating
- [ ] **the remaining crist**
- [ ] the pizza

The most interesting aspect of the picture is ___ .

- [ ] **the pose of the person**
- [x] frisbee
- [ ] the jumping man
- [ ] peace sign

4. image's past

One or two seconds before this picture was taken, ___ .

- [ ] **two men grabbed their phones**
- [x] dad paid for the food
- [ ] the knife was down
- [ ] she finished eating

One or two seconds before this picture was taken, ___ .

- [x] she picked up the sandwich
- [ ] the phone was closed
- [ ] **the man held a cup**
- [ ] they labeled the donuts

One or two seconds before this picture was taken, ___ .

- [x] the man played tennis
- [ ] the woman in red scored a point
- [ ] **a ball flew by**
- [ ] he was smiling

One or two seconds before this picture was taken, ___ .

- [ ] **his phone rang**
- [ ] people rode motorcycles
- [ ] they stopped the scooters
- [x] the person walked towards the bus

5. image's future

One or two seconds after this picture was taken, ___ .

- [ ] **the two sisters fighted over the bear**
- [x] the child hugged the anmial
- [ ] the girl left for school
- [ ] the man made a missy face

One or two seconds after this picture was taken, ___ .

- [ ] the person fell to the ground
- [ ] **they will make their first throws**
- [ ] the person fell on the ground
- [x] the man is playing with balloon

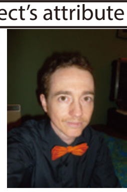
One or two seconds after this picture was taken, ___ .

- [x] the children finished their meal
- [ ] the man started to smile
- [ ] **all will take a sip**
- [ ] he blew out the candles

One or two seconds before this picture was taken, ___ .

- [ ] **the other person got on the bike**
- [ ] the guy was still posing for the camera
- [ ] the bank transported the bananas down the street
- [x] the cas was within inches of the motorcycle

6. object's attribute

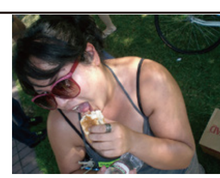


The tie is ___ .

- [ ] **orange**
- [x] black
- [ ] long
- [ ] checkered

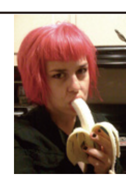
The suitcase is ___ .

- [x] black
- [ ] rectangle
- [ ] **big**
- [ ] colorful



The sandwich is ___ .

- [x] hot
- [ ] a reuben
- [ ] **partially eaten**
- [ ] very large



The banana is ___ .

- [ ] **opened**
- [ ] green
- [ ] picked
- [x] sliced

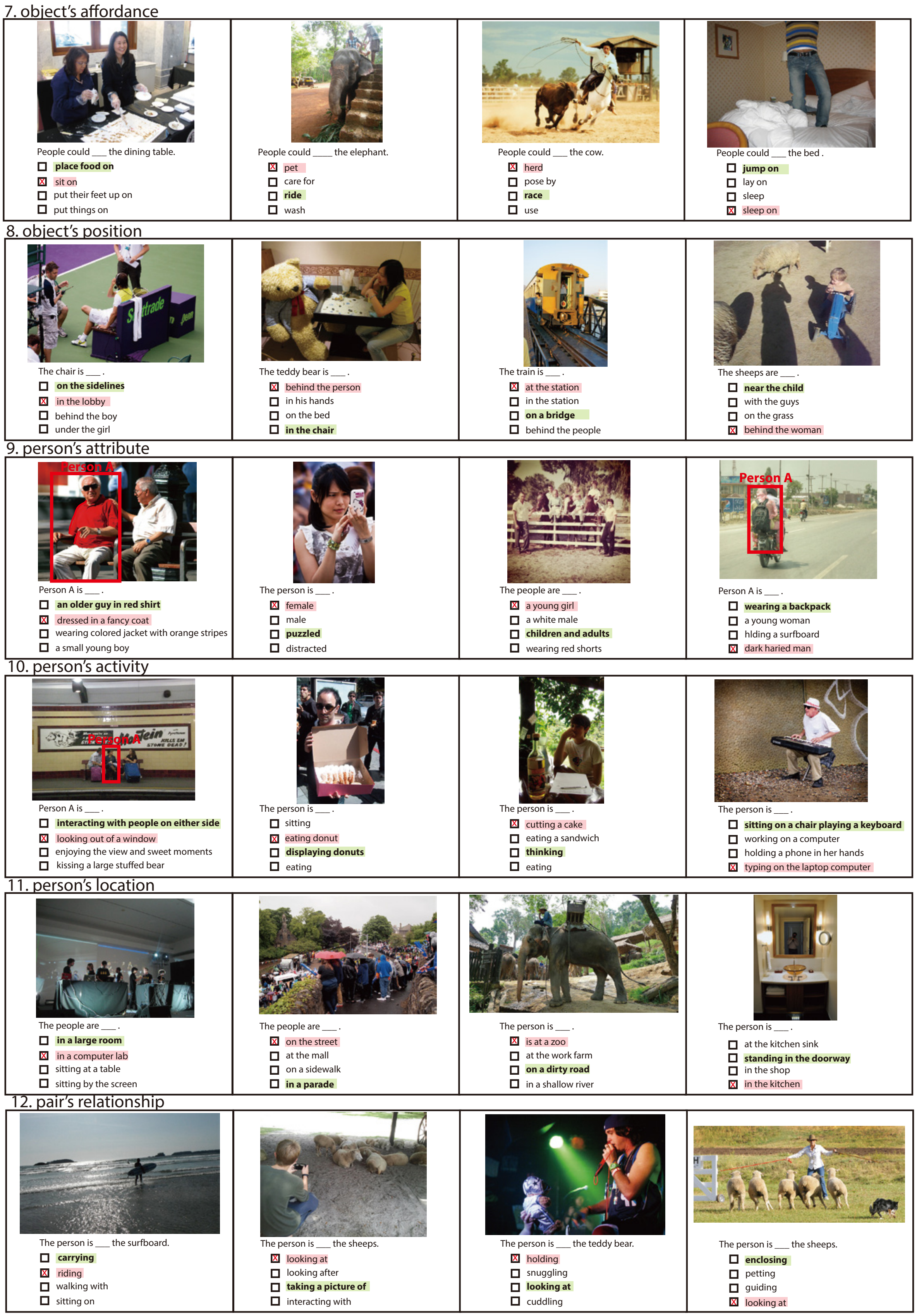
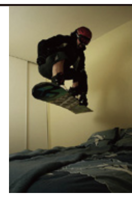


Figure 5: Examples of wrong answers made by nCCA for 12 types of multiple-choice question-answering.

1. image's scene

nCCA: This place is a(n) video recording.
CNN+LSTM: This place is a(n) living room.



nCCA: This place is a(n) bedroom.
CNN+LSTM: This place is a(n) bedroom.

nCCA: This place is a(n) restaurant.
CNN+LSTM: This place is a(n) restaurant.

nCCA: This place is a(n) parade on a street.
CNN+LSTM: This place is a(n) street.

2. image's emotion

nCCA: When I look at this picture, I feel active and excited.
CNN+LSTM: When I look at this picture, I feel excited.

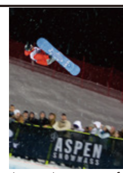
nCCA: When I look at this picture, I feel hungry.
CNN+LSTM: When I look at this picture, I feel hungry.

nCCA: When I look at this picture, I feel cold and free.
CNN+LSTM: When I look at this picture, I feel happy.

nCCA: When I look at this picture, I feel peaceful.
CNN+LSTM: When I look at this picture, I feel calm.

3. image's interesting

nCCA: The most interesting aspect of this picture is hat.
CNN+LSTM: The most interesting aspect of this picture is the man's face.



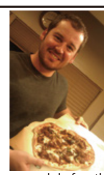
nCCA: The most interesting aspect of this picture is skateboard trick.
CNN+LSTM: The most interesting aspect of this picture is the skateboarder.

nCCA: The most interesting aspect of this picture is the cake.
CNN+LSTM: The most interesting aspect of this picture is the food.

nCCA: The most interesting aspect of this picture is a boy eating donuts.
CNN+LSTM: The most interesting aspect of this picture is the man's face.

4. image's past

nCCA: One or two seconds before this picture was taken he jumped on the skateboard.
CNN+LSTM: One or two seconds before this picture was taken, the man was on the ground.



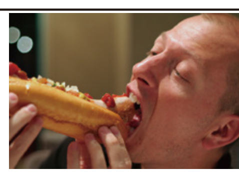
nCCA: One or two seconds before this picture was taken she sliced pizza.
CNN+LSTM: One or two seconds before this picture was taken, the woman was eating.

nCCA: One or two seconds before this picture was taken the ball was kicked.
CNN+LSTM: One or two seconds before this picture was taken, the ball was served.

nCCA: One or two seconds before this picture was taken they sliced the pizza.
CNN+LSTM: One or two seconds before this picture was taken, the woman picked up the pizza.

5. image's future

nCCA: One or two seconds after this picture was taken, the man boarded the bus.
CNN+LSTM: One or two seconds after this picture was taken, the man will move forward.



nCCA: One or two seconds after this picture was taken, the man swallowed his bite.
CNN+LSTM: One or two seconds after this picture was taken, the man will eat the food.

nCCA: One or two seconds after this picture was taken, he went skateboarding.
CNN+LSTM: One or two seconds after this picture was taken, the man will go down the slope.

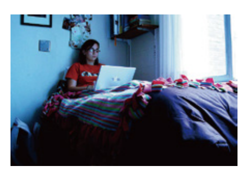
nCCA: One or two seconds after this picture was taken, the skier landed.
CNN+LSTM: One or two seconds after this picture was taken, he fell.

6. object's attribute

nCCA: The boat is filled with fruit.
CNN+LSTM: The boat is wooden.

nCCA: The sports ball is white and orange.
CNN+LSTM: The sports ball is white.

nCCA: The teddy bear is grey.
CNN+LSTM: The teddy bear is brown.



nCCA: The bed is blue.
CNN+LSTM: The bed is covered with a blanket.

Figure 6: Examples of focused sentence generation achieved by nCCA and CNN+LSTM.