Journal of Artificial Intelligence Research 47 (2013) 853-899 Submitted 03/13; published 08/13

Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics

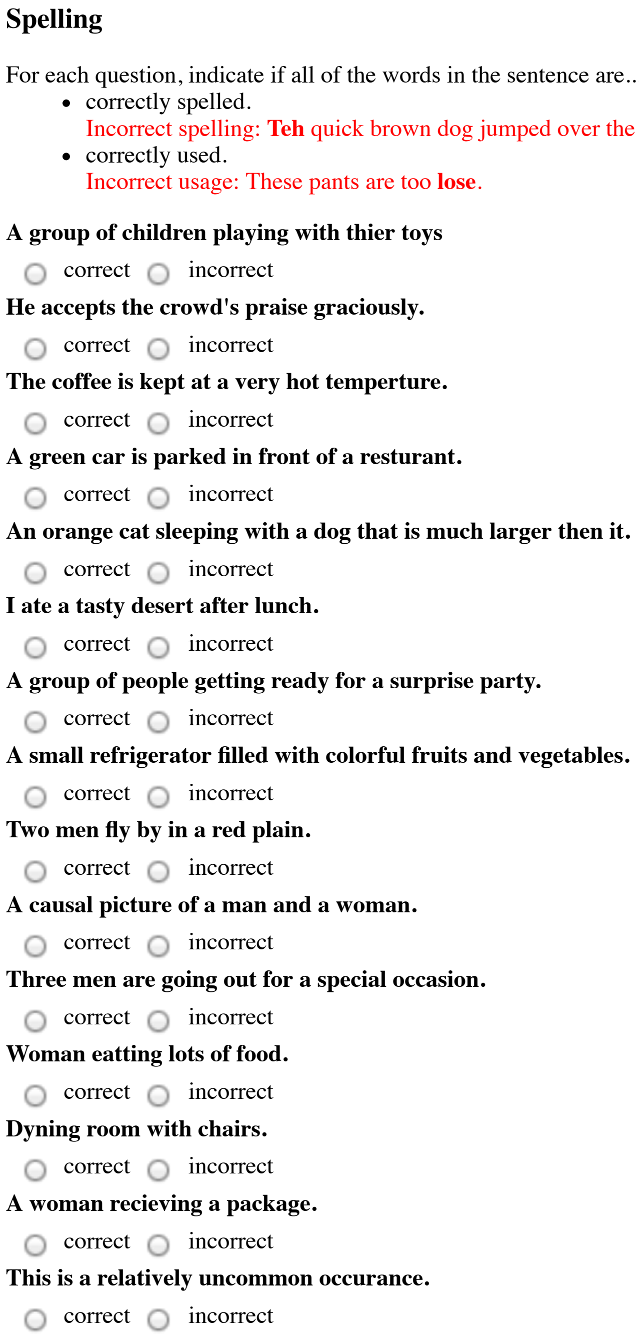
Online Appendix

Micah Hodosh MHODOSH2@ILLINOIS.EDU
Peter Young PYOUNG2@ILLINOIS.EDU
Julia Hockenmaier JULIAHMR@ILLINOIS.EDU
Department of Computer Science
University of Illinois
Urbana, IL 61801, USA

This appendix contains screenshots of the qualification test we designed for the image description task on Amazon Mechanical Turk (Section 1, p. 2–7 here; Section 2.3 of the main paper), the guidelines and one example of the image description task itself (Section 2, p. 8 here; Section 2.3 of the main paper), the instructions for the fine-grained ('expert') human evaluation task (Section 3, p. 9–18 here; Section 4.2.1 of the main paper), and the instructions and examples for the crowd-sourced human evaluation task (Section 4, p. 18–20 here; Section 4.3.2 of the main paper).

1. Image Annotation Qualification Test on Amazon Mechanical Turk

This test consists of three sections, covering spelling, grammar and judging image descriptions, and was a requirement to perform our image description task.



Spelling

For each question, indicate if all of the words in the sentence are...

- correctly spelled.
  Incorrect spelling: **Teh** quick brown dog jumped over the lazy foxes.
- correctly used.
  Incorrect usage: These pants are too **lose**.

A group of children playing with thier toys

○ correct ○ incorrect

He accepts the crowd's praise graciously.

○ correct ○ incorrect

The coffee is kept at a very hot temperture.

○ correct ○ incorrect

A green car is parked in front of a resturant.

○ correct ○ incorrect

An orange cat sleeping with a dog that is much larger then it.

○ correct ○ incorrect

I ate a tasty desert after lunch.

○ correct ○ incorrect

A group of people getting ready for a surprise party.

○ correct ○ incorrect

A small refrigerator filled with colorful fruits and vegetables.

○ correct ○ incorrect

Two men fly by in a red plain.

○ correct ○ incorrect

A causal picture of a man and a woman.

○ correct ○ incorrect

Three men are going out for a special occasion.

○ correct ○ incorrect

Woman eatting lots of food.

○ correct ○ incorrect

Dyning room with chairs.

○ correct ○ incorrect

A woman recieving a package.

○ correct ○ incorrect

This is a relatively uncommon occurance.

○ correct ○ incorrect

Grammar

For each question, indicate if the provided sentence or noun phrase uses correct English or not. For example:

- The train pulls out of the station.
  Very Good: This sentence contains no grammar or spelling errors.
- The train pulling out of the station.
  Good: This is correct English. For example, it could be a good answer to the question, "What do you see in this picture?".
- The train pull out of the station.
  Bad: Incorrect, because the verb does not agree with the noun.
- The train pulls away of the station.
  Bad: This is not good English because you either pull "away from" or "out of," but not "away of."

A man giving pose to camera.

○ correct ○ incorrect

The white sheep walks on the grass.

○ correct ○ incorrect

She is good woman.

○ correct ○ incorrect

He should have talk to him.

○ correct ○ incorrect

He has many wonderful toy.

○ correct ○ incorrect

He sended the children home to their parents.

○ correct ○ incorrect

The passage through the hills was narrow.

○ correct ○ incorrect

A sleeping dog.

○ correct ○ incorrect

The questions on the test was difficult.

○ correct ○ incorrect

In Finland, we are used to live in a cold climate.

○ correct ○ incorrect

Three white sheeps graze on the grassy field.

○ correct ○ incorrect

Between you and me, this is wrong.

○ correct ○ incorrect

They are living there during six months.

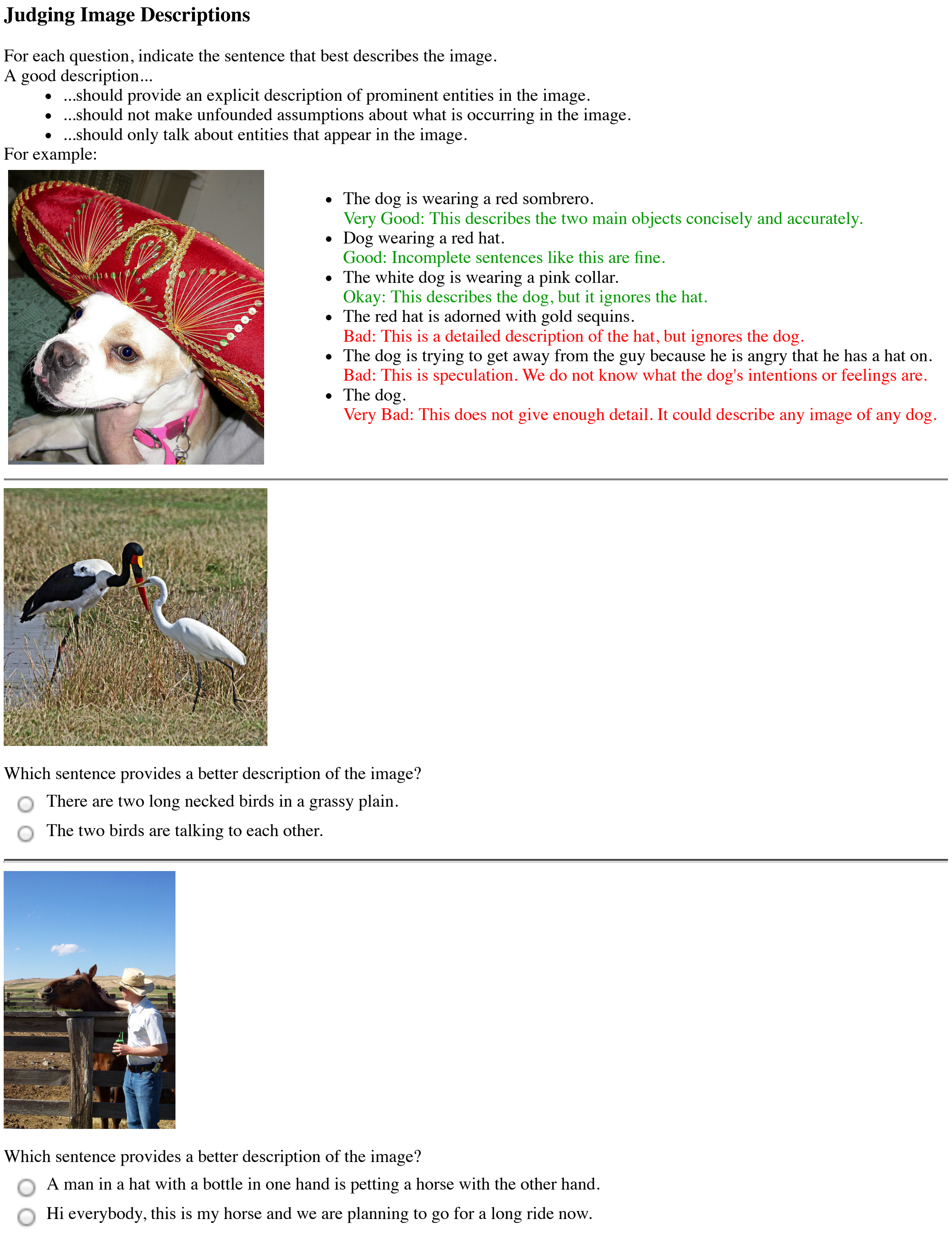
○ correct ○ incorrect

I was given lots of advices about buying new furnitures.

○ correct ○ incorrect

A horse being led back to it's stall.

○ correct ○ incorrect



Judging Image Descriptions

For each question, indicate the sentence that best describes the image.
A good description...

- ...should provide an explicit description of prominent entities in the image.
- ...should not make unfounded assumptions about what is occurring in the image.
- ...should only talk about entities that appear in the image.

For example:

- The dog is wearing a red sombrero.
  Very Good: This describes the two main objects concisely and accurately.
- Dog wearing a red hat.
  Good: Incomplete sentences like this are fine.
- The white dog is wearing a pink collar.
  Okay: This describes the dog, but it ignores the hat.
- The red hat is adorned with gold sequins.
  Bad: This is a detailed description of the hat, but ignores the dog.
- The dog is trying to get away from the guy because he is angry that he has a hat on.
  Bad: This is speculation. We do not know what the dog's intentions or feelings are.
- The dog.
  Very Bad: This does not give enough detail. It could describe any image of any dog.

Which sentence provides a better description of the image?

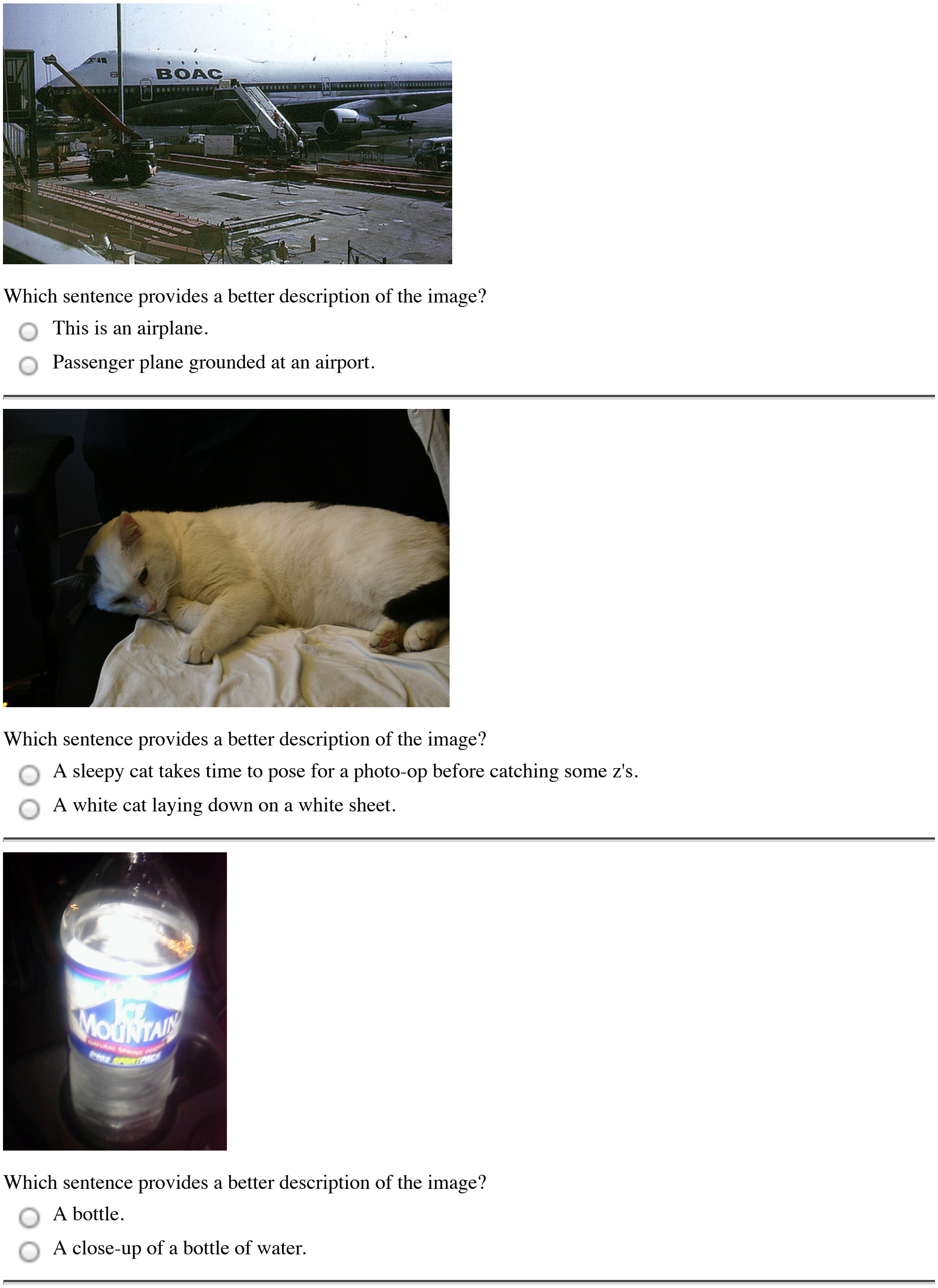
○ There are two long necked birds in a grassy plain.

○ The two birds are talking to each other.

Which sentence provides a better description of the image?

○ A man in a hat with a bottle in one hand is petting a horse with the other hand.

○ Hi everybody, this is my horse and we are planning to go for a long ride now.



Which sentence provides a better description of the image?

○ This is an airplane.

○ Passenger plane grounded at an airport.

Which sentence provides a better description of the image?

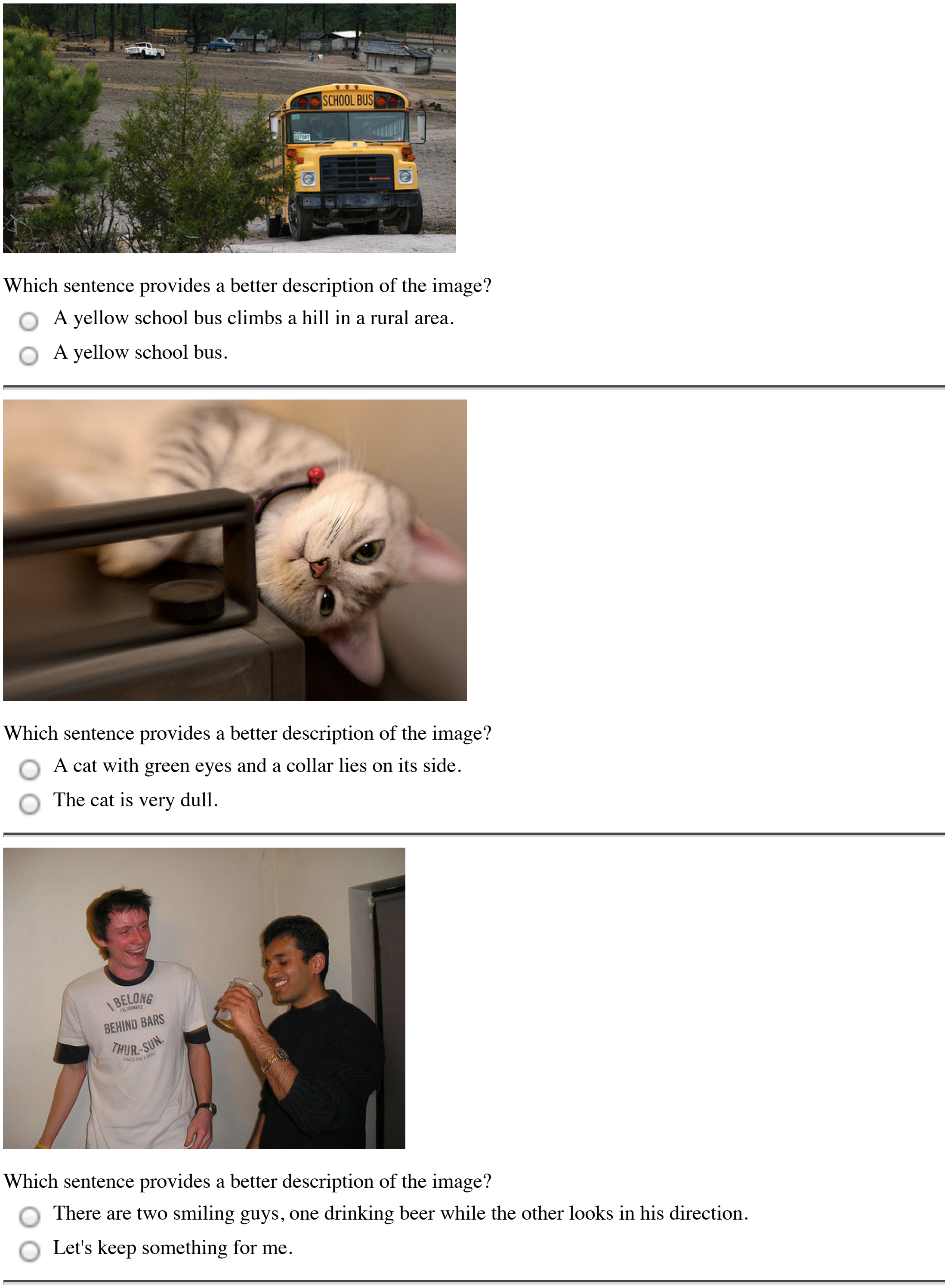
○ A sleepy cat takes time to pose for a photo-op before catching some z's.

○ A white cat laying down on a white sheet.

Which sentence provides a better description of the image?

○ A bottle.

○ A close-up of a bottle of water.



Which sentence provides a better description of the image?

○ A yellow school bus climbs a hill in a rural area.

○ A yellow school bus.

Which sentence provides a better description of the image?

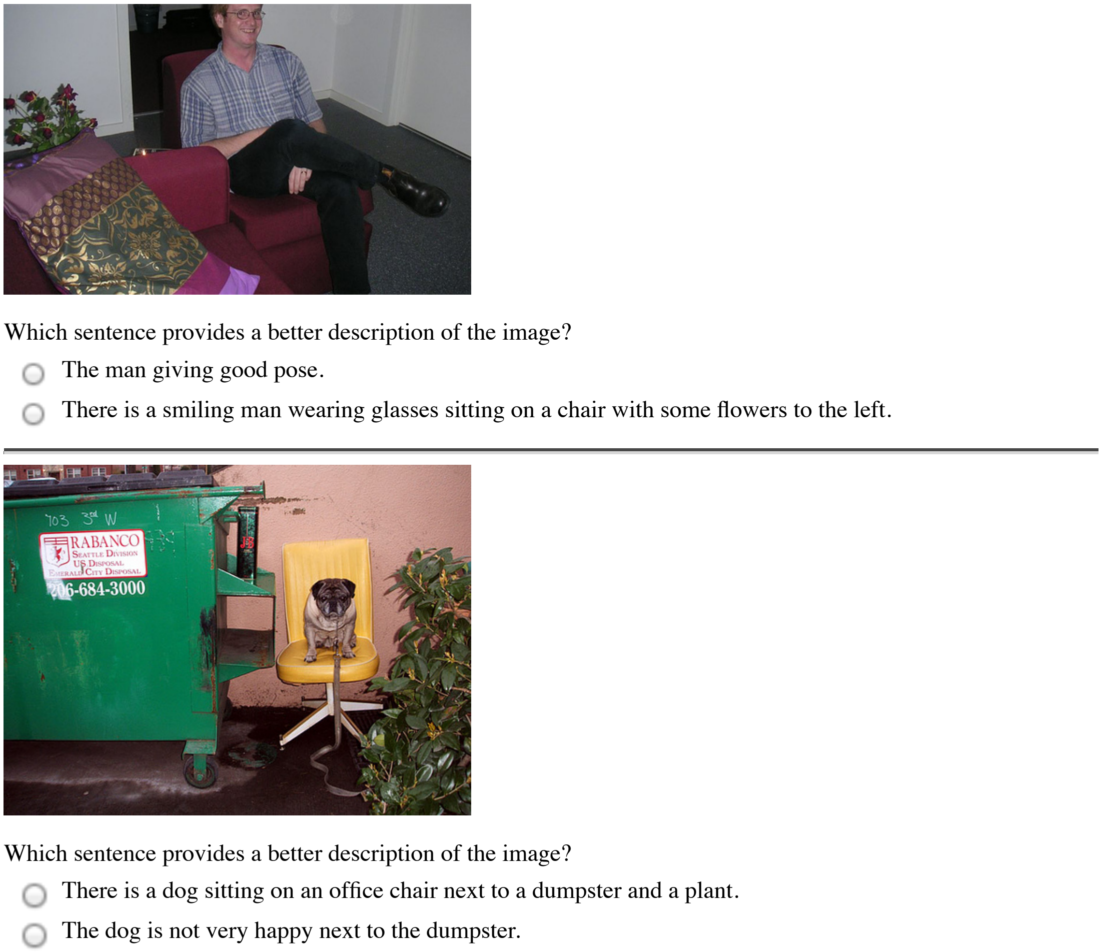
○ A cat with green eyes and a collar lies on its side.

○ The cat is very dull.

Which sentence provides a better description of the image?

○ There are two smiling guys, one drinking beer while the other looks in his direction.

○ Let's keep something for me.



Which sentence provides a better description of the image?

○ The man giving good pose.

○ There is a smiling man wearing glasses sitting on a chair with some flowers to the left.

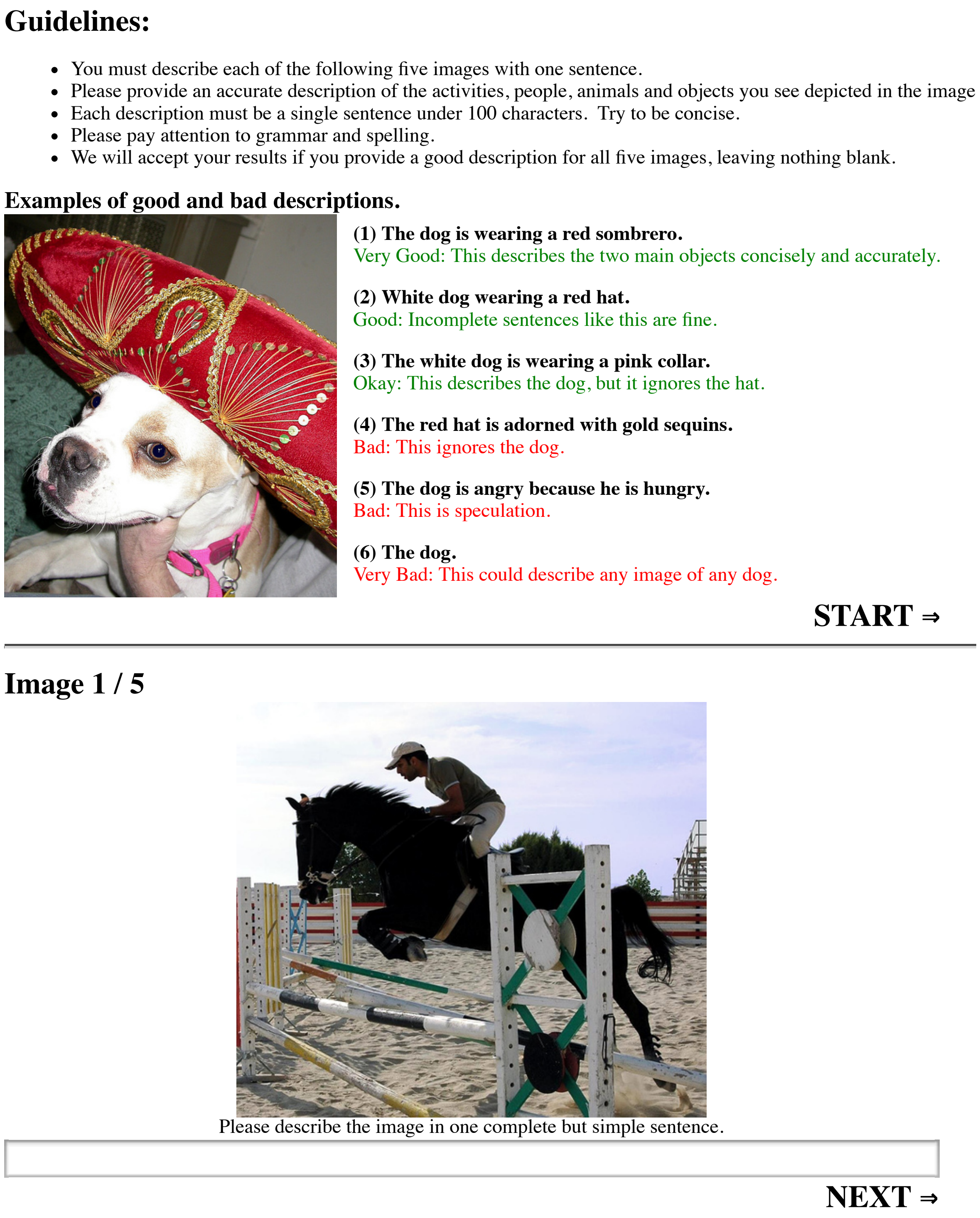
Which sentence provides a better description of the image?

○ There is a dog sitting on an office chair next to a dumpster and a plant.

○ The dog is not very happy next to the dumpster.

2. The Image Description Task on Amazon Mechanical Turk

The figure below shows the guidelines and a screenshot of the image description task.



Guidelines:

- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
Okay: This describes the dog, but it ignores the hat.

(4) The red hat is adorned with gold sequins.
Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
Bad: This is speculation.

(6) The dog.
Very Bad: This could describe any image of any dog.

Image 1 / 5

Please describe the image in one complete but simple sentence.

3. Instructions for the Fine-Grained ('Expert') Human Evaluation

This section contains the instructions used for our detailed ('expert') human evaluation ("Is this Sentence a Good Description of this Image?").

Instructions:

You will see a sequence of 50 image-sentence pairs. Your task is to judge how well each sentence describes its image. Please assign a score between 1 and 4 to each image-sentence pair, where:

4 = Sentence describes the image

- Sentence contains no errors.
- Everything described in the sentence appears in the image.

3 = Sentence *almost* describes the image

- Sentence contains only a single moderate error or a small number of minor errors.
- Major details described in the sentence appear in the image.

2 = Sentence *barely* describes the image

- Sentence contains a major error or many moderate or minor errors.
- Only some minor details described in the sentence appear in the image.

1 = Sentence does not describe the image at all

- No details described in the sentence appear in the image.
- Sentence is unrelated to image.

Image Descriptions

The sentence may describe the **events**, major **actors** or other **objects**, and the overall **scene**:

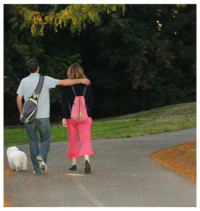
A white dog with a red leash carries a frisbee on the beach.

How well the sentence describes the image depends on the extent to which these correspond to the **events**, **actors**, **objects**, and **scenes** that appear in the image. Note that a sentence can be a correct description for an image even if it does not fully describe everything in the image. However, a correct description should not contain errors. We distinguish between major, moderate, and minor errors. The number and types of errors a sentence contains determines its score. We will first show you examples of these error types, and then explain the scores.

Actors

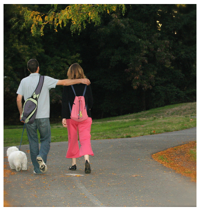
Actors perform or participate in the main action. Most sentences have a single major actor, but some may have more than one:

Good Descriptions of the Actors:



man and woman
guy and gal
couple
two adults
two people

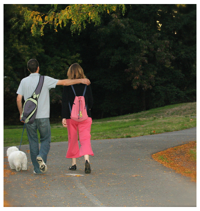
Major Errors Involving Actors



The major actor of the sentence is not the same as the major actor in the image:

A dog walking down the street.
(No mention of the man or woman.)
Two skiers walking a dog.
(Neither person is a skier.)

Moderate Errors Involving Actors



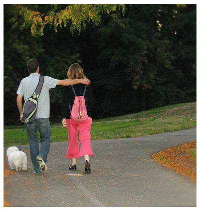
The major actor in the image is not the same as, but could be mistaken for the major actor of the sentence:

A boy and girl walking a dog.
(The man and woman are adults and should not be referred to as boy and girl.)

Not all of the major actors in the image are mentioned in the sentence:

A man walking a dog.
(The man is mentioned, but the woman is not.)

Minor Errors Involving Actors



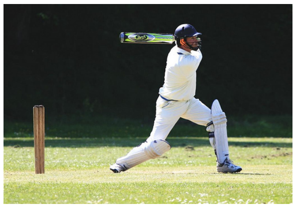
The description of a major actor contains details that are incorrect:

A couple walking a black dog.
(The dog is white.)
Three people walking a dog.
(There are only two people.)

Events

Most images depict a single action or event.

Good Descriptions of the Event:



swinging a bat
hitting a ball
playing cricket
batting
defending a wicket

Major Errors Involving Events

The major event of the sentence is not the same as the major event in the image:

A cricket batsman is sleeping.
(The batsman is swinging his bat, not sleeping.)
A man rides a bike across a grassy field.
(The man is batting in a cricket game, not riding a bike.)

Moderate Errors Involving Events

The major event in the image is not the same as, but could be mistaken for the major event of the sentence:

A man in a white uniform playing baseball.
(Cricket is not baseball, but is similar.)
A batsman sprinting across the field.
(The man is not running, but his posture is somewhat similar.)

Minor Errors Involving Events

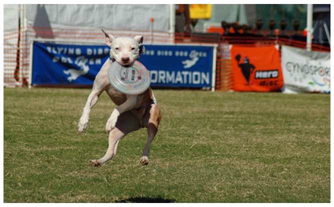
The description of a major event contains details that are incorrect:

A batsman swinging his bat over his right shoulder.
(The bat is over his left shoulder.)

Objects

Objects are other things or entities that appear in the image. The main action may or may not involve them.

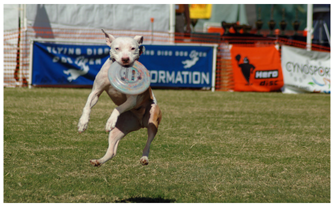
Objects in the Image:



frisbee, disc
fence
banner, advertisement

Major Errors Involving Objects
There are no major errors associated with objects.

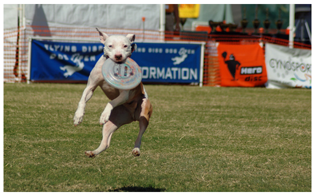
Moderate Errors Involving Objects



An object described in the sentence does not appear in the image.

A dog jumps a large rock to catch a Frisbee.
(There is no rock in the image.)
A dog holding a bone in his mouth.
(The dog is holding a Frisbee, not a bone.)

Minor Errors Involving Objects



The description of an object contains details that are incorrect:

A white dog catches a green Frisbee
(The Frisbee is grey.)

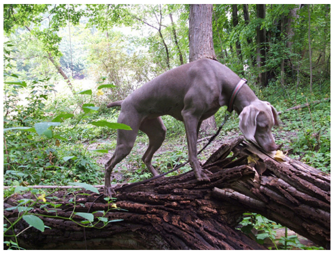
An object in the image is not the same as, but could be mistaken for an object described in the sentence:

A white dog catches a plastic dish.
(The Frisbee could be mistaken for a dish.)

Scene

The scene describes the background or setting in which the event occurs.

Good Descriptions of the Scene:



forest
woods
wooded area

Major Errors Involving the Scene
There are no major errors associated with scenes.

Moderate Errors Involving the Scene

The scene in the sentence is not the scene in the image:

A black dog walking down a city street.
(The dog is in a forest, not a city.)
A dog sniffing at something on a beach.
(The dog is not at the beach or by any sort of body of water.)

Minor Errors Involving the Scene

The description of the scene contains details that are incorrect:

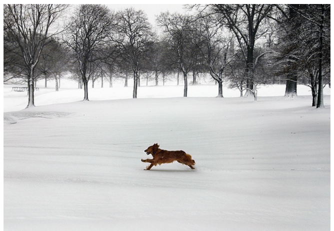
A black dog climbs a log in a desiccated forest.
(The forest appears to be fairly green.)

The image scene is not the same as, but could be mistaken for the scene described by the sentence:

A black dog climbs a log in a jungle.
(Jungles and forests are somewhat similar.)

Types of Errors

Major Errors A major error means that the image-sentence pair can score a 2 at best. Possible major errors:[1]



The major actor in the image is not the same as, but could be mistaken for the major actor of the sentence.

A man running through the snow.

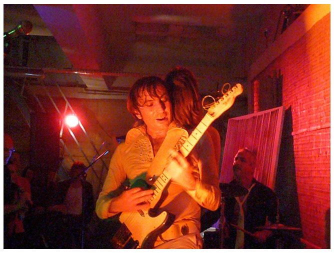
(It is clearly a dog running.)

The major event of the sentence is not the same as the major event in the image.

A dog eating snow.

(The dog is running, not eating.)

Moderate Errors A single moderate error means that the image-sentence pair can still score a 3. Additional errors are likely to mean that it is a 2 or lower, however. Possible moderate errors:



The major actor in the image is not the same as, but could be mistaken for the major actor of the sentence.

A woman playing the guitar in a dark room.

(The guitar player is male.)

The major event in the image is not the same as, but could be mistaken for the major event of the sentence.

A man playing the mandolin in a dark room.

(The man is playing a guitar, an instrument that is played in a similar manner.)

An object described in the sentence does not appear in the image.

A man wearing three piece suit playing the guitar.

(The man is wearing a white jumpsuit.)

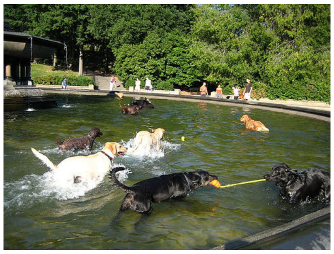
The scene in the sentence is not the scene in the image.

A man playing the guitar on an open air stage

(The man is not on an open air stage.)

[1] This instruction contained an error. It should have been "The major actor in the image is not the same as the major actor of the sentence."

Minor errors Minor errors, by themselves, will only lower the score of a image-sentence pair to a 3, unless there are a large number of minor errors. Possible minor errors:



The description of a major actor contains details that are incorrect.

Two dogs playing in a fountain.
(There are at least seven dogs in the fountain.)

The description of a major event contains details that are incorrect.

Dogs sadly playing in a fountain.
(The dogs look pretty happy.)

The description of an object or the scene contains details that are incorrect.

Dogs wade in blue waters.
(The water is green colored.)

The image scene is not the same as, but could be mistaken for the scene described by the sentence.

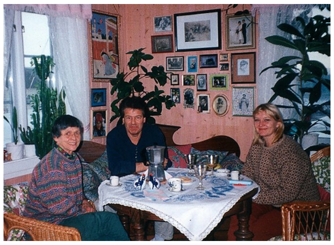
Dogs playing in a pool.
(The dogs are playing in a fountain, not a pool.)

An object in the image is not the same as, but could be mistaken for an object described in the sentence.

Dogs playing, one of them chasing a baseball.
(There is a tennis ball in the fountain.)

Examples and Scoring

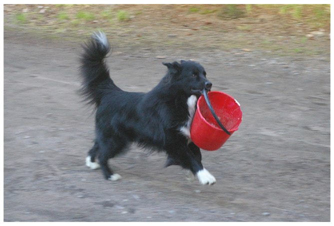
Correct score: 4 Sentence describes the image.
Everything described in the sentence appears in the image.



Three people are sitting at a table

Three people are sitting at a table covered in trophies inside a house

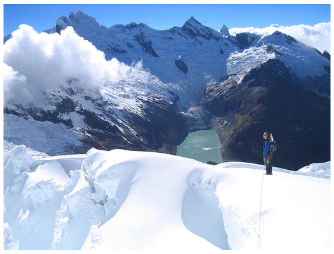
Correct score: 4 Sentence describes the image.
Although shorter than ideal, the vague captions are completely accurate.



A dog is running

A dog carrying a bucket

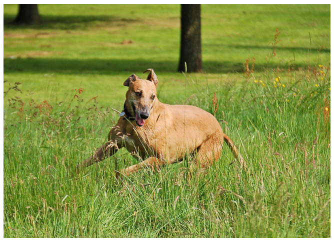
Correct score: 3 Sentence *almost* describes the image.
The sentence closely describes the image, except for a single moderate or minor error.



Two people are standing upon a snow covered mountain
Minor error: The number of people is incorrect.

A man is running on top of a mountain
Moderate error: A major action is incorrectly described, but standing, walking, and running are similar actions that can potentially be misinterpreted as each other.

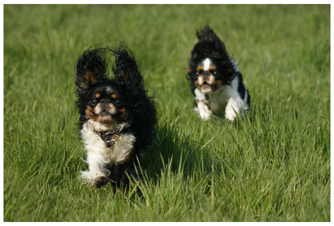
Correct score: 3 Sentence *almost* describes the image.
There are multiple errors in the captions, but they are minor.



Two dogs are in the brown grass
Minor error: The number of dogs is wrong.
Minor error: A word describing the grass is wrong.

A black dog is in the brown grass
Minor error: A word describing the dog is wrong.
Minor error: A word describing the grass is wrong.

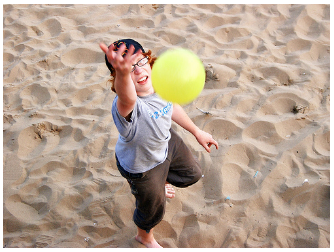
Correct score: 2 Sentence *barely* describes the image.
There are either multiple errors in the captions and some of them are not minor, or there is a major error.



One dog is running through the snow
Minor error: The number of dogs is wrong.
Moderate error: There is no snow.

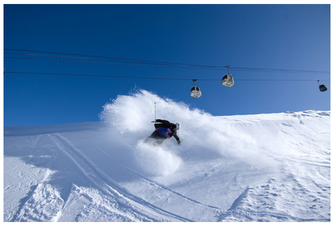
A man is running in the grass
Major error: There is no man.

Correct score: 2 Sentence *barely* describes the image.
There is a major error.



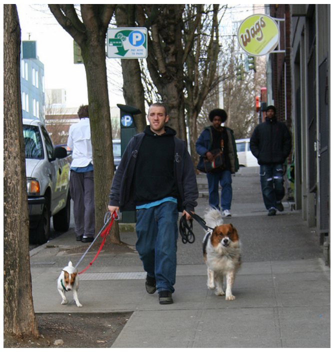
A dog is in the sand
Major error: There is no dog.

Correct score: 2 Sentence *barely* describes the image.
The sentence only describes something going on in the background.



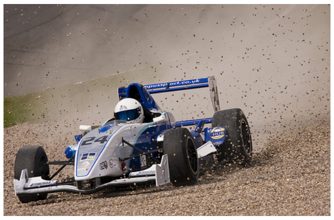
Cable cars going up and down a mountain
Major actor error: No mention of the skier.
Major event error: No mention of skiing.
Correct object and event: There are moving cable cars in the background.

Correct score: 1 Sentence does *not* describe the image at all.
No major elements of the image are mentioned in the sentence.



A truck driving on a dirt track
Major actor error: No mention of the man or the two dogs.
Major event error: No mention of dog walking.
Moderate scene error: This is a city/street scene, not a dirt track.
Correct object: There is a truck in the image, but it is not a major actor.

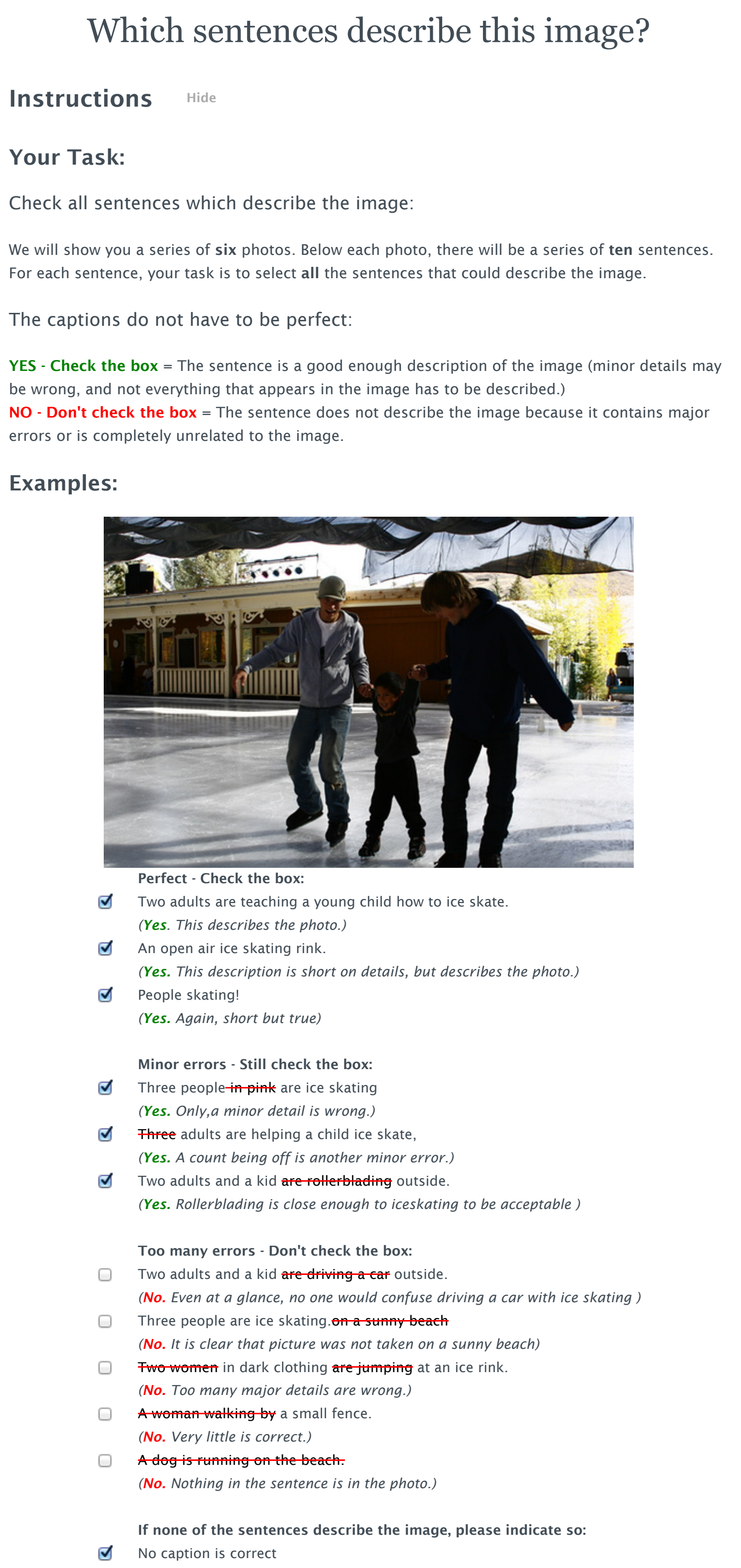
Correct score: 1
Sentence does not describe the image at all. Nothing in the caption is correct.



People are waiting in line at a restaurant

4. Instructions for the Crowd-Sourced Human Evaluation Task

The following two pages contain the instructions for the binary, crowd-sourced human evaluation task, as well as a few examples of the image-caption pairs that had to be judged.



Which sentences describe this image?

Instructions

Your Task:

Check all sentences which describe the image:

We will show you a series of **six** photos. Below each photo, there will be a series of **ten** sentences. For each sentence, your task is to select **all** the sentences that could describe the image.

The captions do not have to be perfect:

YES - Check the box = The sentence is a good enough description of the image (minor details may be wrong, and not everything that appears in the image has to be described.)
NO - Don't check the box = The sentence does not describe the image because it contains major errors or is completely unrelated to the image.

Examples:

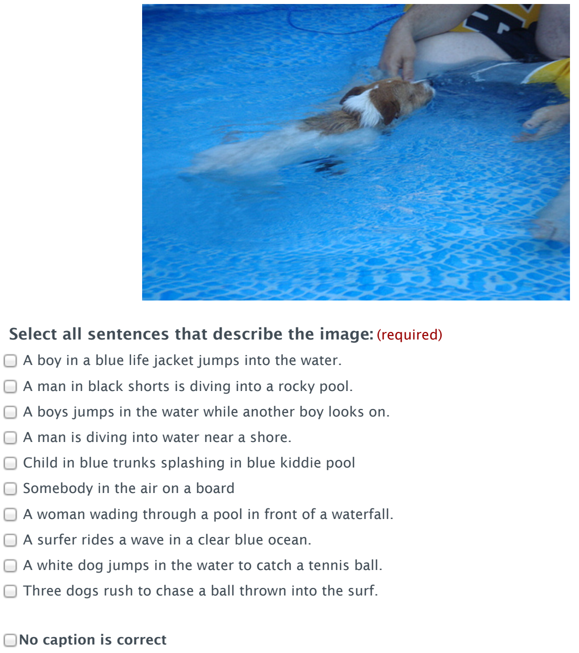
Perfect - Check the box:
☑ Two adults are teaching a young child how to ice skate.
(*Yes. This describes the photo.*)
☑ An open air ice skating rink.
(*Yes. This description is short on details, but describes the photo.*)
☑ People skating!
(*Yes. Again, short but true*)

Minor errors - Still check the box:
☑ Three people ~~in pink~~ are ice skating
(*Yes. Only,a minor detail is wrong.*)
☑ ~~Three~~ adults are helping a child ice skate,
(*Yes. A count being off is another minor error.*)
☑ Two adults and a kid ~~are rollerblading~~ outside.
(*Yes. Rollerblading is close enough to iceskating to be acceptable*)

Too many errors - Don't check the box:
□ Two adults and a kid ~~are driving a car~~ outside.
(*No. Even at a glance, no one would confuse driving a car with ice skating*)
□ Three people are ice skating.~~on a sunny beach~~
(*No. It is clear that picture was not taken on a sunny beach*)
□ ~~Two women~~ in dark clothing ~~are jumping~~ at an ice rink.
(*No. Too many major details are wrong.*)
□ ~~A woman walking by~~ a small fence.
(*No. Very little is correct.*)
□ ~~A dog is running on the beach.~~
(*No. Nothing in the sentence is in the photo.*)

If none of the sentences describe the image, please indicate so:
☑ No caption is correct

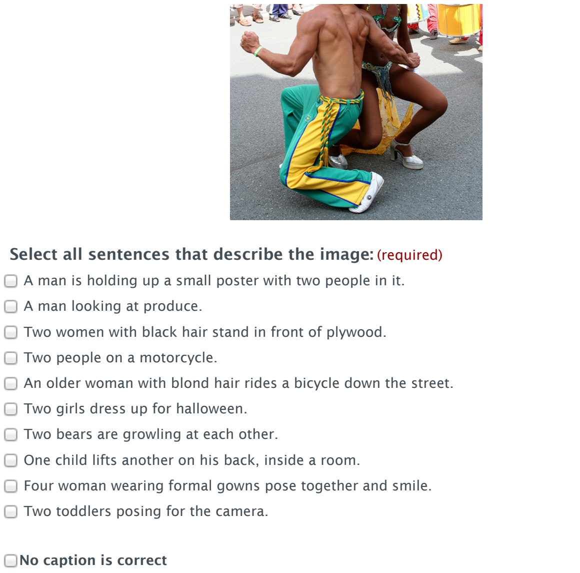
Figure 1: Instructions for the crowd-sourced human evaluation task



Select all sentences that describe the image: (required)

□ A boy in a blue life jacket jumps into the water.
□ A man in black shorts is diving into a rocky pool.
□ A boys jumps in the water while another boy looks on.
□ A man is diving into water near a shore.
□ Child in blue trunks splashing in blue kiddie pool
□ Somebody in the air on a board
□ A woman wading through a pool in front of a waterfall.
□ A surfer rides a wave in a clear blue ocean.
□ A white dog jumps in the water to catch a tennis ball.
□ Three dogs rush to chase a ball thrown into the surf.

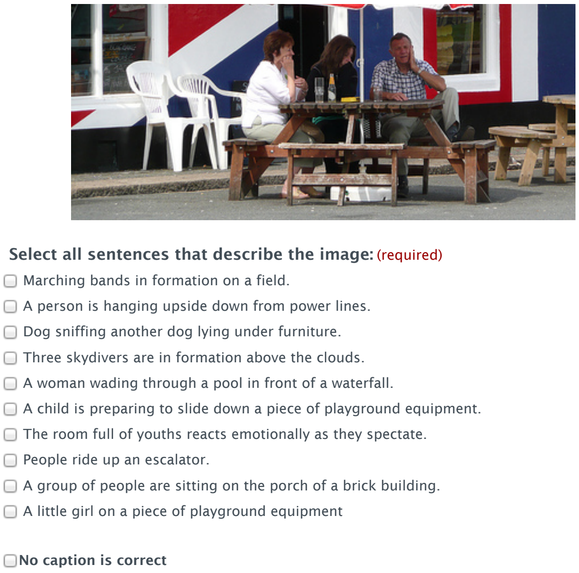
□ **No caption is correct**



Select all sentences that describe the image: (required)

□ A man is holding up a small poster with two people in it.
□ A man looking at produce.
□ Two women with black hair stand in front of plywood.
□ Two people on a motorcycle.
□ An older woman with blond hair rides a bicycle down the street.
□ Two girls dress up for halloween.
□ Two bears are growling at each other.
□ One child lifts another on his back, inside a room.
□ Four woman wearing formal gowns pose together and smile.
□ Two toddlers posing for the camera.

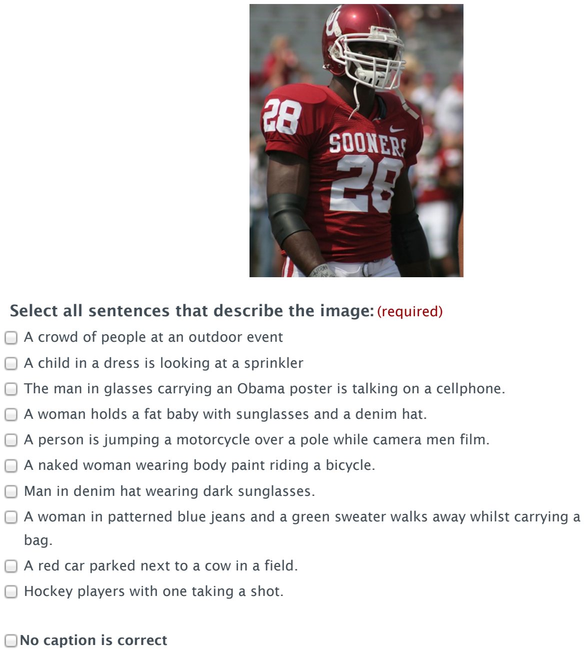
□ **No caption is correct**



Select all sentences that describe the image: (required)

□ Marching bands in formation on a field.
□ A person is hanging upside down from power lines.
□ Dog sniffing another dog lying under furniture.
□ Three skydivers are in formation above the clouds.
□ A woman wading through a pool in front of a waterfall.
□ A child is preparing to slide down a piece of playground equipment.
□ The room full of youths reacts emotionally as they spectate.
□ People ride up an escalator.
□ A group of people are sitting on the porch of a brick building.
□ A little girl on a piece of playground equipment

□ **No caption is correct**



Select all sentences that describe the image: (required)

□ A crowd of people at an outdoor event
□ A child in a dress is looking at a sprinkler
□ The man in glasses carrying an Obama poster is talking on a cellphone.
□ A woman holds a fat baby with sunglasses and a denim hat.
□ A person is jumping a motorcycle over a pole while camera men film.
□ A naked woman wearing body paint riding a bicycle.
□ Man in denim hat wearing dark sunglasses.
□ A woman in patterned blue jeans and a green sweater walks away whilst carrying a bag.
□ A red car parked next to a cow in a field.
□ Hockey players with one taking a shot.

□ **No caption is correct**