Supplementary Material: Convolutional Image Captioning

Jyoti Aneja*, Aditya Deshpande*, Alexander G. Schwing
University of Illinois at Urbana-Champaign
{janeja2, ardeshp2, aschwing}@illinois.edu

1. Additional Qualitative Results

In Figures 1, 2, 3, 4, 5 we provide additional qualitative results. We compare the captions predicted by our CNN+Attn to LSTM and ground-truth.

2. Qualitative Results for Attention

In Figures 6 to 10, we show qualitative results for our attention mechanism. The attention parameters are overlayed over the image. The bright regions denote the portions of the image that are attended to when predicting the word below. While, the darker regions recieve no attention.

* Denotes equal contribution.

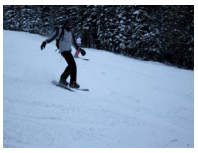
LSTM: a man and a woman are standing in the grass with a dog
CNN: a group of people standing next to a dog
GT: A group of teenage girls are walking a dog on the grass.

LSTM: a couple of people sitting on a couch with a remote
CNN: a pile of scissors and a pair of scissors
GT: A car seat covered in a piece of cake and frosting.

LSTM: a group of kites flying in the sky

CNN: a kite flying in the sky with a blue sky
GT: a kite flying high above in the sky

LSTM: a man is playing tennis on a tennis court
CNN: a man is playing tennis on a tennis court
GT: Two people on a tennis court swinging to hit a tennis ball.



LSTM: a person riding a snowboard down a snow covered slope
CNN: a person on a snowboard on a snowy slope
GT: A person in cold weather gear riding on a snowboard down a slope.

LSTM: a bowl of oranges and a bowl of fruit
CNN: a table with a bowl of orange and a half of orange

GT: Sliced oranges, a juicer, a glass and a paring knife used to make orange juice.

LSTM: a train is stopped at a train station

CNN: a train is stopped at a train station

GT: A train that is sitting on the tracks under wires.

LSTM: a dog is looking at a bird with its mouth
CNN: a couple of brown and white dogs standing next to each other

GT: Small kitten warming up to a shy dog.

LSTM: a red motorcycle parked on the side of the road
CNN: a motorcycle parked on the side of a street
GT: A couple of motorcycles parked next to each other.

LSTM: a boat is in the water with a boat in the background

CNN: a boat is on the water with a boat in the background

GT: a bunch of people are on a small boat

LSTM: a man riding a skateboard up the side of a ramp
CNN: a man riding a skateboard up the side of a ramp
GT: A man riding a skateboard down the side of a ramp.

LSTM: a knife and a knife on a cutting board
CNN: a close up of a knife and a knife
GT: A bunch of big chef knives by a small piece of broccoli.

LSTM: a bathroom with a sink and a mirror

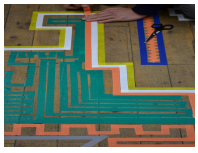
CNN: a bathroom with a large window and a sink
GT: A bathroom has a floating toilet and sink and a walk-in shower.

LSTM: a traffic light sitting on the side of a road
CNN: a street with cars and cars on the street
GT: cars that are stopped at a traffic light

LSTM: a computer monitor sitting on top of a desk
CNN: a computer monitor sitting on top of a desk
GT: A computer that is on a wooden desk.

LSTM: a man and a dog are walking down a street
CNN: a man and a dog are walking down a street
GT: a person standing in an alley with sheep

LSTM: a giraffe standing next to a tree in a forest
CNN: a giraffe standing next to a tree in a zoo
GT: A giraffe walking through a tree filled forest.



LSTM: a sign that is on a wall with a sign on it

CNN: a sign that is on a wall with a sign
GT: A person is making a pattern on the floor with tape.

LSTM: a man walking down a street next to a clock tower
CNN: a clock tower with a clock on it

GT: the people are walking about, there is a clock on the road

LSTM: a clock on a pole on a city street
CNN: a clock on a pole in front of a building
GT: Large four sided clock hangs on the corner of the building.

Figure 1: Comparison of captions predicted by LSTM, CNN and the ground-truth.

LSTM: a red fire hydrant sitting on a sidewalk

CNN: a red fire hydrant sitting on a dark street
GT: A red fire hydrant along the curb at night.

LSTM: a train is coming down the tracks in a city
CNN: a train is on the tracks at a train station
GT: Red stoplights that are next to a train.

LSTM: a bathroom with a toilet sink and mirror

CNN: a bathroom with a toilet and a sink

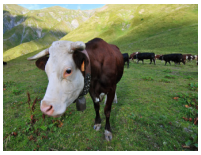
GT: A bathroom with white fixtures and green flooring

LSTM: a bathroom with a toilet sink and mirror

CNN: a bathroom with a toilet sink and mirror

GT: A blue bathroom has an updated sink and toilet.

LSTM: a woman is sitting on a horse in a field
CNN: a man is sitting on a horse in a field
GT: A BLACK AND WHITE PIC OF A MAN AND A HORSE



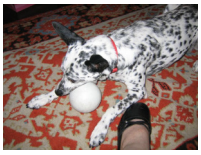
LSTM: a couple of cows are standing in a field
CNN: a couple of cows standing on top of a grass covered field

GT: A brown cow standing on a field with long horns.

LSTM: a man riding a horse on a street

CNN: a man on a skateboard in front of a building
GT: A group of people walking next to stairs.

LSTM: a group of people playing a game of soccer
CNN: a group of people playing a game of soccer
GT: a couple of players are out in a baseball field



LSTM: a dog is laying on a couch with a stuffed animal
CNN: a dog laying on a bed with a stuffed animal
GT: A spotted dog sits protectively with it's toy.

LSTM: a bathroom with a toilet and a sink

CNN: a bathroom with a toilet and a sink

GT: A bathroom filled with toilets and a tub next to a sink.

LSTM: a zebra standing in a field of grass

CNN: a zebra standing next to a tree in a field
GT: 2 Zebras standing next to each other in plaines

LSTM: a person on a snowboard on a snowy slope
CNN: a person on a snowboard is going down a hill
GT: A man riding skis down a snow covered slope.

LSTM: a row of urinals in a public restroom

CNN: a bathroom with three urinals and a wall

GT: four white urinals against a green wall with lines

LSTM: a row of motorcycles parked next to each other
CNN: a motorcycle parked next to a building with a crowd of people

GT: A group of motorcycles parked next to each other.

LSTM: a baseball player swinging a bat at a ball
CNN: a baseball player swinging a bat at a ball
GT: A baseball player holding a bat while standing in a field.

LSTM: a group of giraffes standing in a field

CNN: a group of giraffes standing in a field

GT: A giraffe grazing on a tree in the wilderness with other wildlife

LSTM: a man and a woman sitting at a table with a plate of food
CNN: a group of people sitting around a table with a cat
GT: Two people that are sitting on a table.

LSTM: a group of giraffes standing in a field

CNN: a giraffe and a giraffe in a field

GT: Two giraffe standing next to each other in a forest.

LSTM: a person walking down a street holding an umbrella
CNN: a woman holding an umbrella standing in the rain
GT: A man holding an umbrella next to a frozen over fire hydrant.

LSTM: a large jetliner sitting on top of an airport runway
CNN: a plane is taking off from a runway

GT: An airliner taking off from an airport runway.

Figure 2: Comparison of captions predicted by LSTM, CNN and the ground-truth.

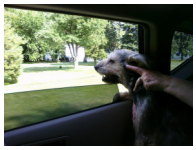
LSTM: a man riding a skateboard down a cement ramp
CNN: a man riding a skateboard down a ramp

GT: A person on a skateboard rides on a platform.

LSTM: a stuffed bear and a stuffed animal are sitting on a shelf

CNN: a stuffed bear is sitting next to a book
GT: A picture of children's toy reading story.

LSTM: a living room with a couch and a tv
CNN: a living room with a couch chair and television
GT: A living room filled with living room furniture and decor.



LSTM: a dog is looking out the window of a car
CNN: a dog is looking out the window of a car
GT: a dog with its head out of the window

LSTM: a sign on a building with a sign on it
CNN: a bicycle is attached to a building on a street
GT: a cell phone painted on a wall near a bike

LSTM: a street sign on a pole with a street sign
CNN: a street sign with a green light on it
GT: a green sign on a pole with a street light

LSTM: a hot dog with a pickle and a pickle on a plate

CNN: a hot dog with a pickle and a drink
GT: A hotdog with toppings served in a red basket

LSTM: two giraffes walking in a zoo enclosure
CNN: a giraffe standing next to a giraffe in a field
GT: A picture of two giraffes, fairly close to a road, with a bus traveling up it.

LSTM: a person holding a carrot in a bowl

CNN: a carrot is being eaten by a knife

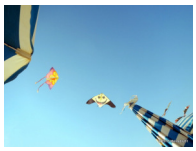
GT: a jar of sauce sits next to some carrots

LSTM: a man and a woman riding horses in a forest
CNN: a man and a dog are riding a horse
GT: A man riding on the back of a brown horse.

LSTM: a person on skis is standing in the snow
CNN: a man riding skis down a snow covered slope
GT: Woman cross country skiing alone on a trail in the woods.

LSTM: a clock tower with a weather vane on top
CNN: a clock tower with a clock on the top
GT: A large clock tower with an American flag flying from the top of it.

LSTM: a man riding a wave on top of a surfboard
CNN: a man riding a wave on top of a surfboard
GT: A man riding a wave on a surfboard.



LSTM: a group of kites flying in the sky

CNN: a group of kites flying in the sky

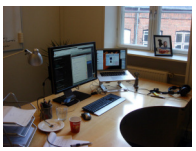
GT: Many kites can be seen in the air through umbrellas.

LSTM: a toilet with a white seat and a black seat
CNN: a toilet with a white seat on the floor
GT: A white toilet sitting in a bathroom on a tiled floor.

LSTM: a group of people sitting around a table

CNN: a group of people standing around a room with luggage
GT: A group of people walking on the side walk

LSTM: a large clock tower with a clock on its side
CNN: a large clock tower with a clock on it
GT: A IMAGE OF A TOWER CLOCK WITH THE CLOCK ON IT



LSTM: a desk with a computer and a laptop on it
CNN: a desk with a laptop computer and a laptop
GT: a desk with a cup plate laptop monitor and keyboard

LSTM: a computer keyboard sitting on top of a table
CNN: a close up of a keyboard on a desk
GT: I really cant see this image very well.

LSTM: a man is eating a hot dog with a pickle
CNN: a man holding a doughnut in his hand

GT: A man holding a half eaten hot dog and a dollar.

Figure 3: Comparison of captions predicted by LSTM, CNN and the ground-truth.

LSTM: a street sign that is on a pole

CNN: a street sign with a bunch of signs on it
GT: A crowded street has a number of street signs.

LSTM: a man riding a wave on top of a surfboard
CNN: a man riding a wave on top of a surfboard
GT: Two men can be seen out in the water and one is on a surf board.

LSTM: a man in a suit and tie is talking on a cell phone
CNN: a man in a suit and tie standing next to a man

GT: Men standing and one pointing to an object on a street.

LSTM: a man sitting on a couch with a laptop
CNN: a man sitting on a couch in a living room
GT: A man sitting on a top of a green couch.

LSTM: two dogs laying on a bed with a blanket
CNN: a black and white dog laying on top of a bed
GT: A cat and dog are close together on a bed.

LSTM: a man and a woman sitting on a couch with a laptop

CNN: a man and a woman sitting on a couch with a laptop

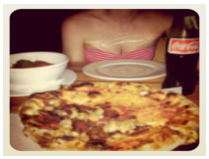
GT: A group of young people sitting around a piece of luggage.

LSTM: a laptop computer sitting on top of a wooden desk
CNN: a laptop computer sitting on top of a wooden desk
GT: A standard computer and laptop on a cluttered desk.

LSTM: a bathroom with a sink and a mirror

CNN: a bathroom with a sink and a mirror

GT: A clean bathroom is pictured in this image.



LSTM: a pizza sitting on top of a white plate
CNN: a pizza with a slice of pizza on it
GT: A woman with nice breast sitting at a table with a pizza.

LSTM: a bench sitting on a sidewalk next to a street
CNN: a broken broken broken suitcase on the side of a road
GT: a small pamphlet is sitting on a benches arm

LSTM: a bowl of food with a spoon and a bowl of salad

CNN: a bowl of food with a bowl of vegetables and a bowl of
GT: a bowl of food on a table near other plates

LSTM: a woman holding an umbrella in the rain

CNN: a woman in a black dress holding an umbrella
GT: a woman walking holding a pink umbrella near a train

LSTM: a herd of sheep grazing on a lush green field
CNN: a herd of sheep grazing on a dry grass field
GT: Several lambs and sheep standing on hay and eating it.

LSTM: a boat is sailing in the water with a person on it

CNN: a boat is sailing in the water near a city
GT: A couple of people on kayak boats in the middle of the ocean.

LSTM: a woman is playing tennis on a court

CNN: a woman in a blue shirt and a tennis ball
GT: A woman swinging a racket at a tennis ball.

LSTM: a train is stopped at a train station

CNN: a train is traveling down the tracks in a city
GT: A train traveling down train tracks near a train station.

LSTM: a toilet with a seat up in a bathroom
CNN: a toilet with a white seat in a bathroom
GT: A white toilet sitting in a bathroom stall next to a TP dispenser.

LSTM: a man riding a surfboard on a beach

CNN: a man riding a bike on a beach

GT: A group of men carrying surfboards on a beach.

LSTM: a group of people sitting at a table with umbrellas
CNN: a group of people sitting at a table under a tree
GT: people on the street near a sea with waters

LSTM: a zebra and a baby zebra are standing in a field
CNN: a group of zebras grazing in a field

GT: A couple of zebra standing on top of a grass field.

Figure 4: Comparison of captions predicted by LSTM, CNN and the ground-truth.

LSTM: a man riding a motorcycle down a street

CNN: a man riding a motorcycle down a street

GT: The motorcyclist has his hands at his side while riding swiftly down the road.

LSTM: a teddy bear is sitting in the snow

CNN: a teddy bear is sitting on a chair in a chair
GT: A basket with a stuffed teddy bear hangs outside.

LSTM: a group of teddy bears sitting on a pile of grass
CNN: a group of teddy bears sitting on a rock
GT: A group of stuffed animals sitting on top of a wall.

LSTM: a man flying through the air while riding a snowboard
CNN: a man flying through the air while riding a snowboard
GT: A skateboarder is performing an aerial manuever during a competition.

LSTM: two people are standing on a hill with a backpack
CNN: a man and a woman are skiing in the snow
GT: A couple of women standing next to a pair of skis.

LSTM: a group of people sitting on a couch playing a video game

CNN: a group of people standing around a living room
GT: A little boy is being held on a lap while eating.

LSTM: a woman sitting at a desk with a laptop
CNN: a young girl sitting at a desk with a laptop
GT: A man looks straight ahead sitting at a desk.

LSTM: a vase with a flower in it sitting on a table
CNN: a vase with a flower in it sitting on a table
GT: A brown vase filled with two purple flowers.

LSTM: a bathroom with a sink toilet and bathtub

CNN: a bathroom with a sink toilet and mirror

GT: A towel that is on a rack in a bathroom.

LSTM: a train is traveling down the tracks near a building
CNN: a train on a track near a forest

GT: a single red white and yellow train engine on the tracks

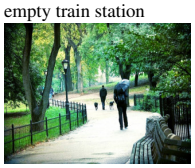
LSTM: a large building with a clock tower in the background
CNN: a large church with a clock tower and a clock
GT: A clock that is on the side of a tower.

LSTM: a dog is playing with a frisbee in its mouth
CNN: a dog with a frisbee in its mouth

GT: the dog has a green frisbee behind it's head

LSTM: a train is pulling into a station with people waiting
CNN: a train is pulling into a station with people on the platform

GT: a white yellow and blue train at an empty train station



LSTM: a couple of people walking down a path near a forest
CNN: a group of people walking down a path next to a fence

GT: A person with two dogs, and another with an umbrella, walking down a road in a park.

LSTM: a kitchen with a table and a stove

CNN: a kitchen with a lot of cabinets and a sink
GT: A kitchen filled with lots of counter space.

LSTM: two people riding horses on a beach near the ocean
CNN: a couple of people walking on a beach with surfboards
GT: a group of people walk on a beach with surf boards

LSTM: a bathroom with a sink toilet and bathtub

CNN: a bathroom with a sink toilet and tub

GT: The bathroom is equipped with several new appliances.

LSTM: a person holding a pair of scissors in a box
CNN: a person holding a knife and knife to cut a piece of paper
GT: A bloody arm with iv's inserted into an injury victim's arm are depicted in this graphic shot.

LSTM: a group of people riding motorcycles on a city street
CNN: a group of people riding motorcycles down a street
GT: A group of parked motorcycles sitting on the side of a road.

LSTM: a teddy bear sitting on a chair in a room
CNN: a teddy bear sitting on a chair in a chair
GT: A large teddy bear with a smaller teddy sitting in a rocking chair.

Figure 5: Comparison of captions predicted by LSTM, CNN and the ground-truth.

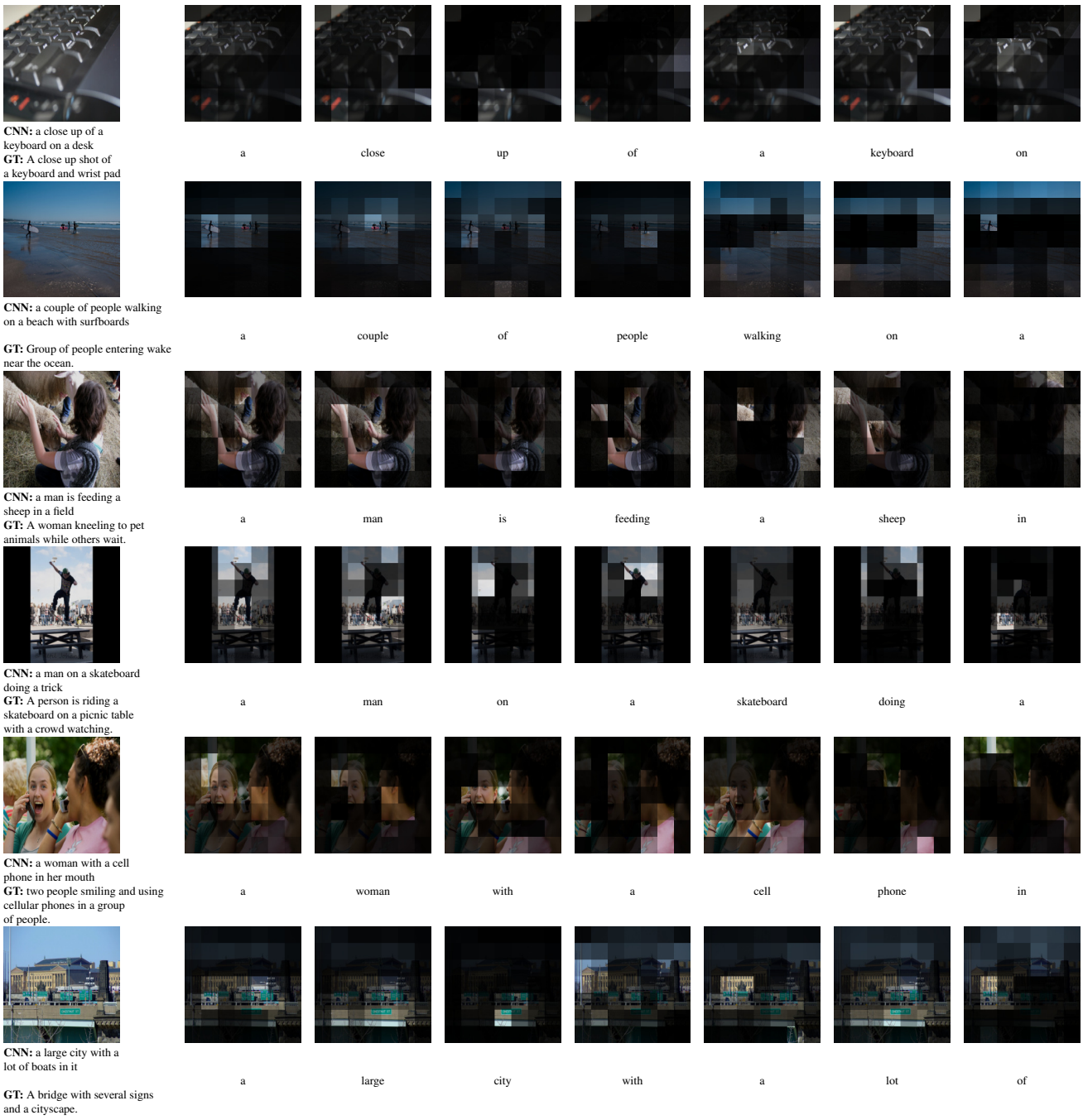


Figure 6: Qualitative results for attention, our attention parameters are overlayed on the image.

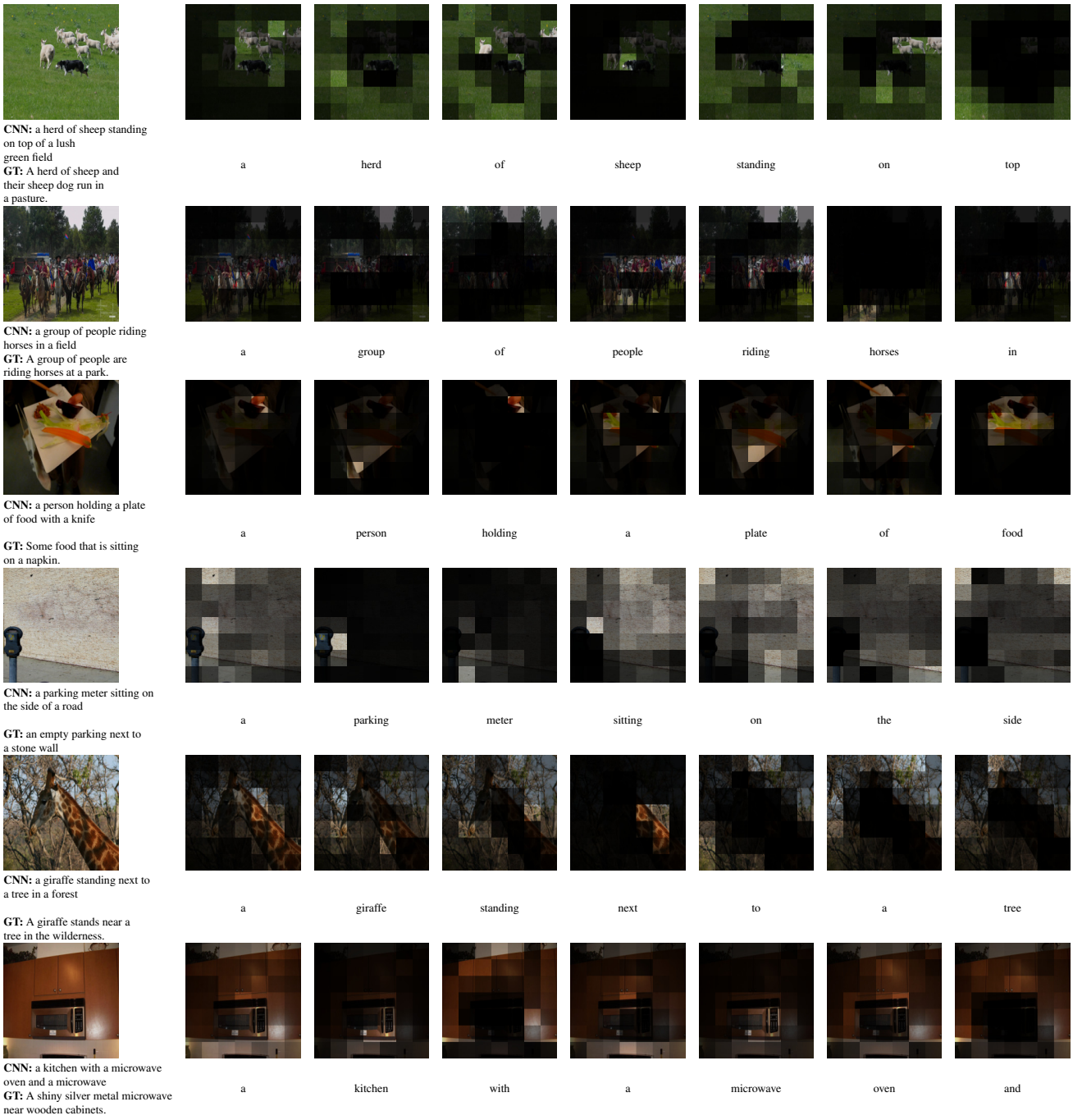


Figure 7: Qualitative results for attention, our attention parameters are overlayed on the image.

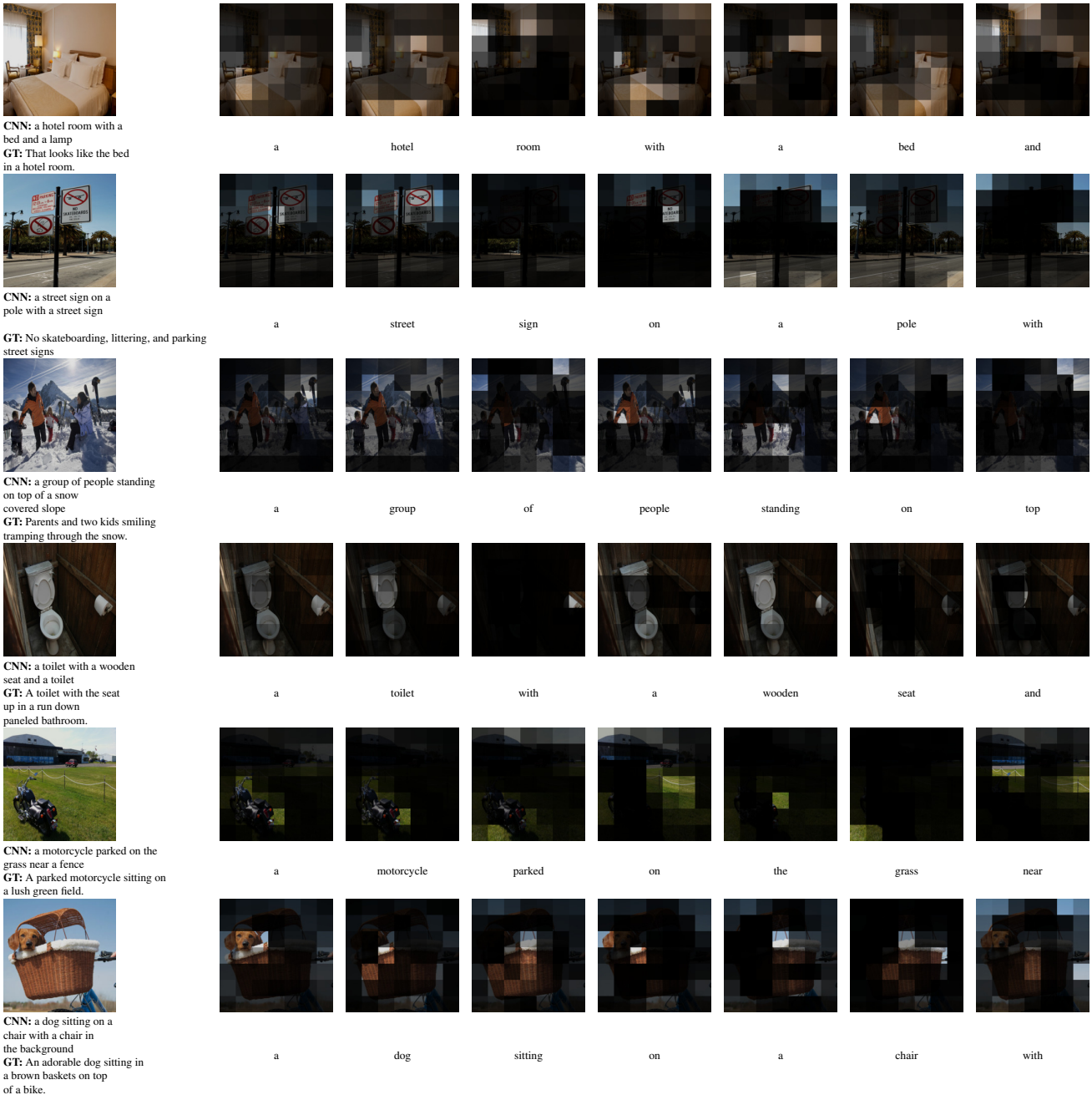


Figure 8: Qualitative results for attention, our attention parameters are overlayed on the image.

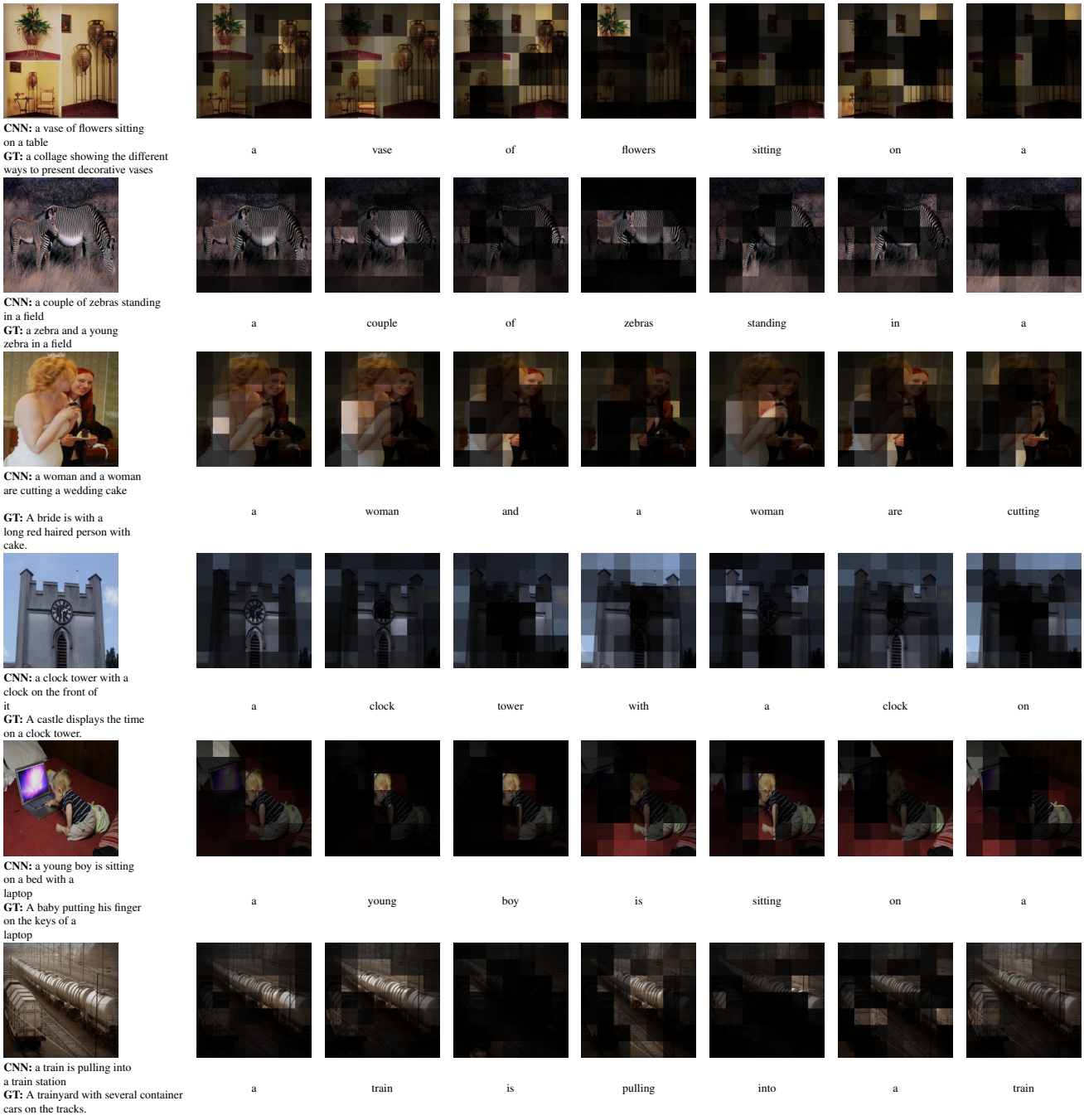


Figure 9: Qualitative results for attention, our attention parameters are overlayed on the image.

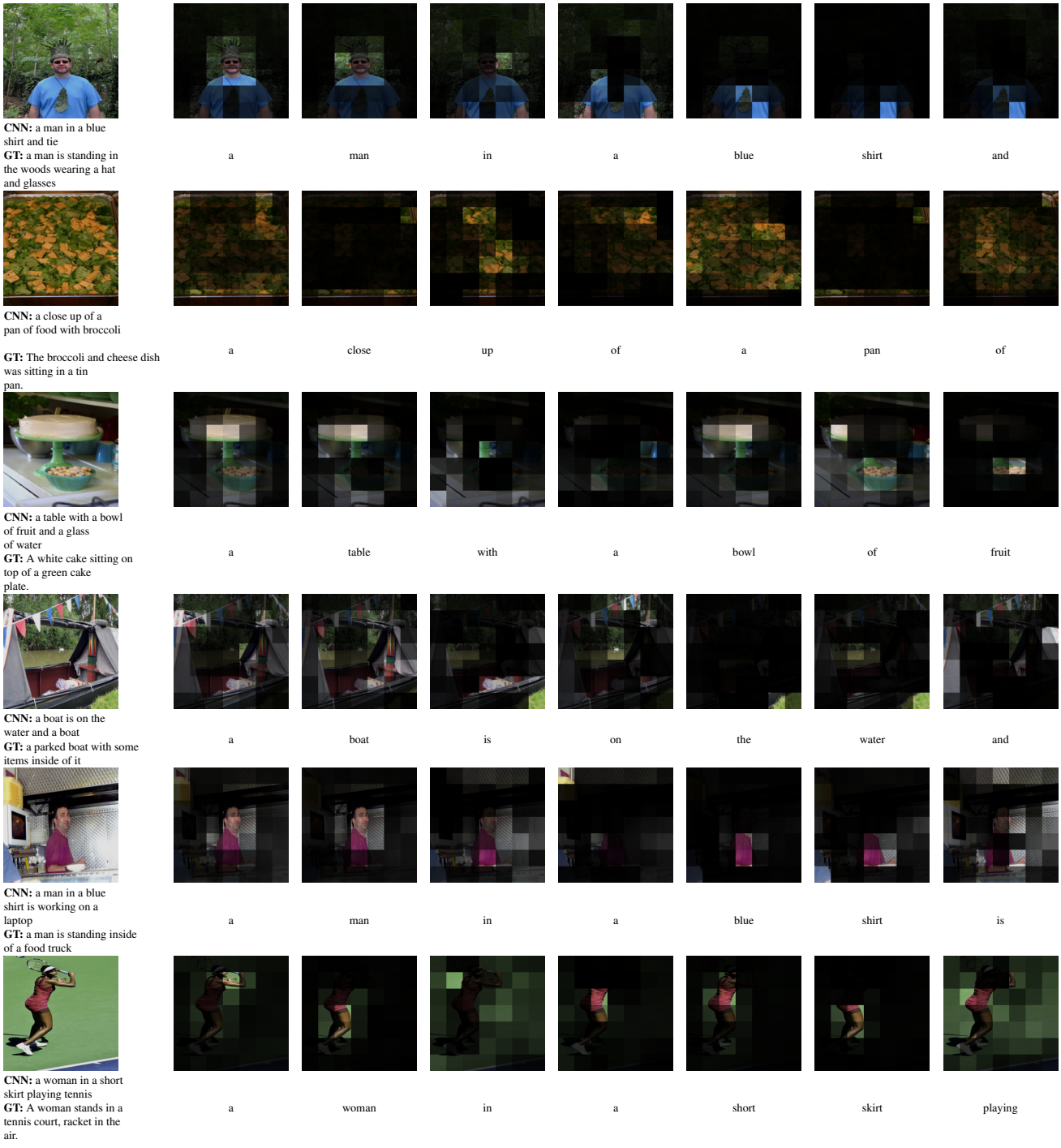


Figure 10: Qualitative results for attention, our attention parameters are overlayed on the image.