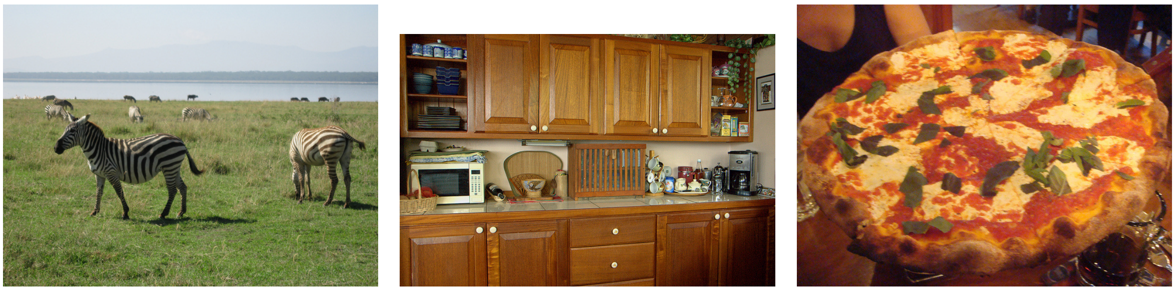
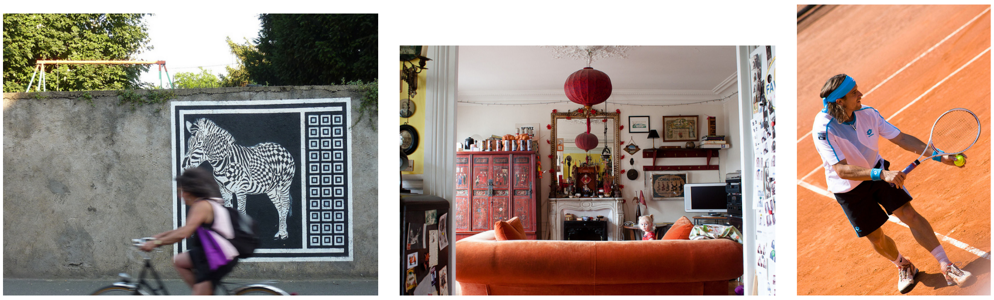


Base: A couple of animals that are standing in the grass.
Tags: field, zebra, grass, walking.
Base+T4: Two zebras walking through a grass covered field.

Base: A kitchen with wooden cabinets and wooden cabinets.
Tags: kitchen, wooden, microwave, oven.
Base+T4: A kitchen with wooden cabinets and a microwave oven.

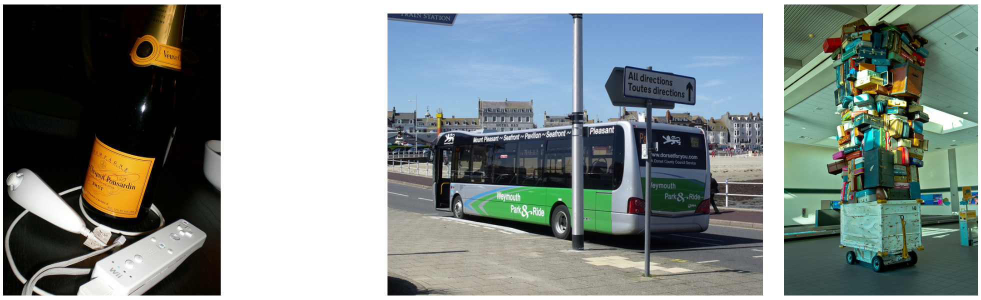
Base: A close up of a plate of food on a table.
Tags: pizza, top, pan, cheese.
Base+T4: A pizza pan with cheese on top of it.



Base: A young boy is playing with a frisbee.
Tags: young, outside, wall, man.
Base+T4: A young man is outside in front of a graffiti wall.

Base: A picture of a living room with a tv.
Tags: room, living, couch, window.
Base+T4: A living room with a couch and a window.

Base: A man swinging a tennis ball on a tennis court.
Tags: racket, tennis, court, hold.
Base+T4: A man holding a racket on a tennis court.

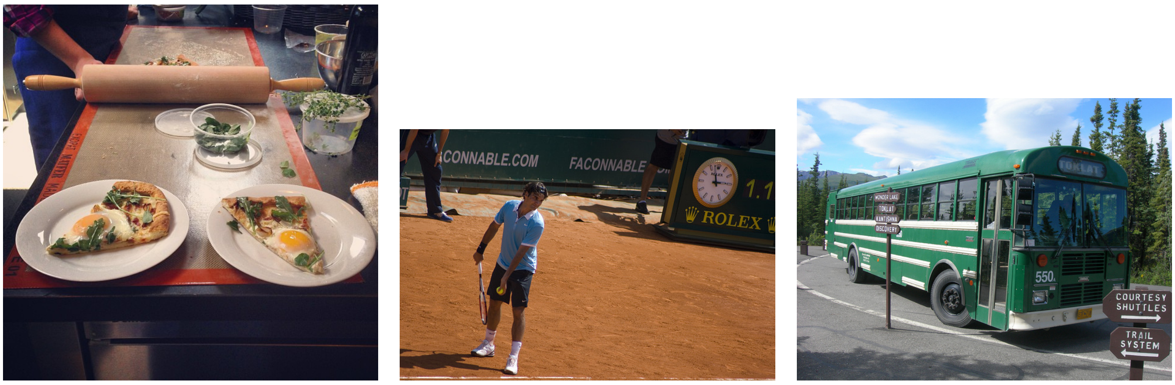


Base: A close up of a microphone on a table.
Tags: table, bottle, black, brown.
Base+T4: A brown and black bottle sitting on top of a table.

Base: A street sign on the side of the road.
Tags: bus, street, city, side.
Base+T4: A bus is parked on the side of a city street.

Base: There is a toy train that is on display.
Tags: luggage, cart, piled, bunch.
Base+T4: A toy cart with a bunch of luggage piled on top of it.

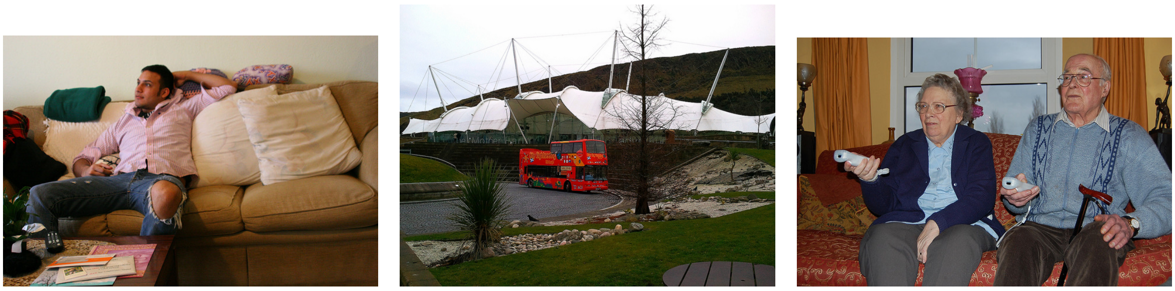
Figure 1: Additional examples of out-of-domain captions generated on MSCOCO using the base model (Base), and the base model constrained to include four predicted image tags (Base+T4). Words never seen in training captions are underlined.



Base: A couple of plates of food on a table.
Tags: table, food, meal, dinner.
Base+T4: A meal of food on a dinner table.

Base: A woman is playing tennis on the beach.
Tags: man, dirt, player, woman.
Base+T4: A man standing on the dirt with a womens player.

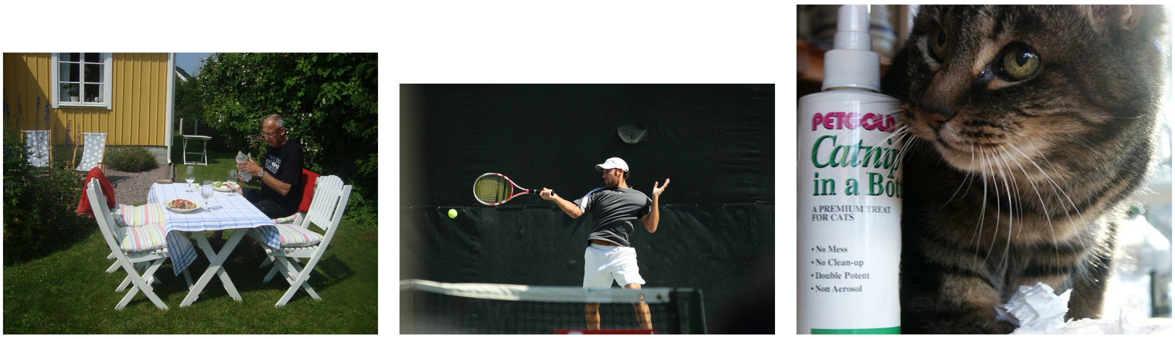
Base: A green truck driving down a street next to a building.
Tags: bus, street, green, road.
Base+T4: A green bus truck driving down a street road.



Base: A man and a woman sitting on a bed.
Tags: bed, couch, woman, girl.
Base+T4: A woman and a girl couch on a bed.

Base: A red fire hydrant sitting in the middle of a river.
Tags: train, grass, field, old.
Base+T4: An old red train traveling through a grass covered field.

Base: A man and a woman are playing a video game.
Tags: hold, man, couch, other.
Base+T4: A man holding a couch with other people.

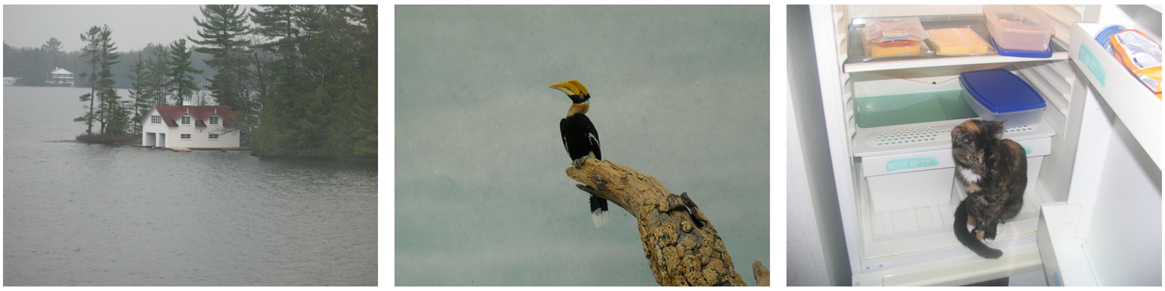


Base: A man sitting at a table with a laptop.
Tags: table, outside, white, grass.
Base+T4: A man sitting at a table outside with white grass.

Base: A man is playing a game of tennis.
Tags: racket, tennis, man, player.
Base+T4: A man with a racket in front of a tennis player.

Base: A cat that is sitting on top of a table.
Tags: cat, up, brown, close.
Base+T4: A close up of a cat on a brown and white cat.

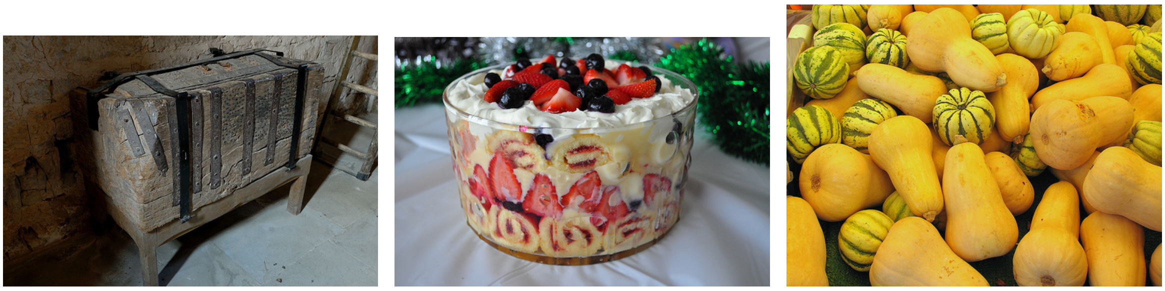
Figure 2: Additional examples of out-of-domain MSCOCO caption failure cases, illustrating the impact of multiple image tags relating to the same object (top row), poor quality tag predictions (middle row) and other captioning errors (bottom row). Words never seen in training captions are underlined.



Base: A large white boat on a body of water.
Synset: boathouse.
Base + Synset: A white and red boathouse on a lake.

Base: A bird perched on top of a tree branch.
Synset: hornbill.
Base + Synset: An hornbill bird perched on top of a tree branch.

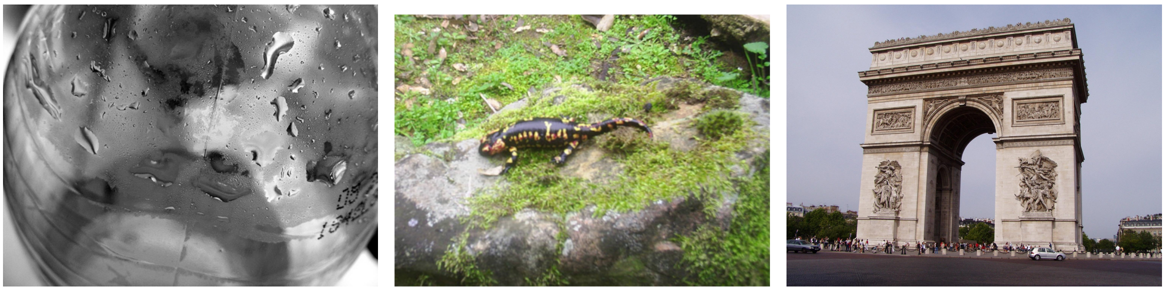
Base: A cat sitting inside of an open refrigerator.
Synset: refrigerator, icebox.
Base + Synset: A cat sitting inside of an open refrigerator.



Base: An old wooden suitcase sitting on a stone wall.
Synset: chest.
Base + Synset: An old wooden chest sitting next to a stone wall.

Base: A close up of a dessert on a table.
Synset: trifle.
Base + Synset: A close up of a trifle cake with strawberries.

Base: A large pile of yellow and yellow apples.
Synset: butternut squash.
Base + Synset: A pile of yellow and green butternut squash.

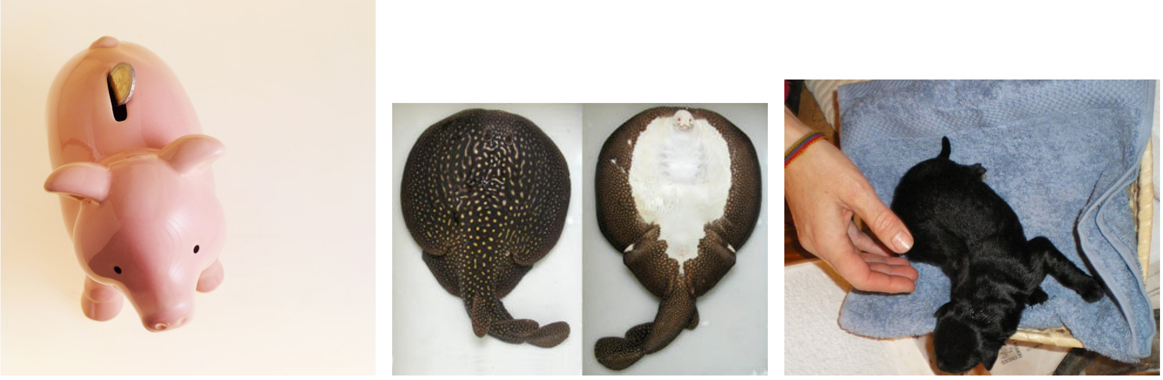


Base: A black and white photo of a glass of water.
Synset: water bottle.
Base + Synset: A black and white picture of a water bottle.

Base: A group of animals laying on the ground.
Synset: Salamandra salamandra, European fire salamander.
Base + Synset: A European fire salamander laying on the ground.

Base: A tall tower with a clock on it.
Synset: triumphal arch.
Base + Synset: A large stone building with a triumphal arch.

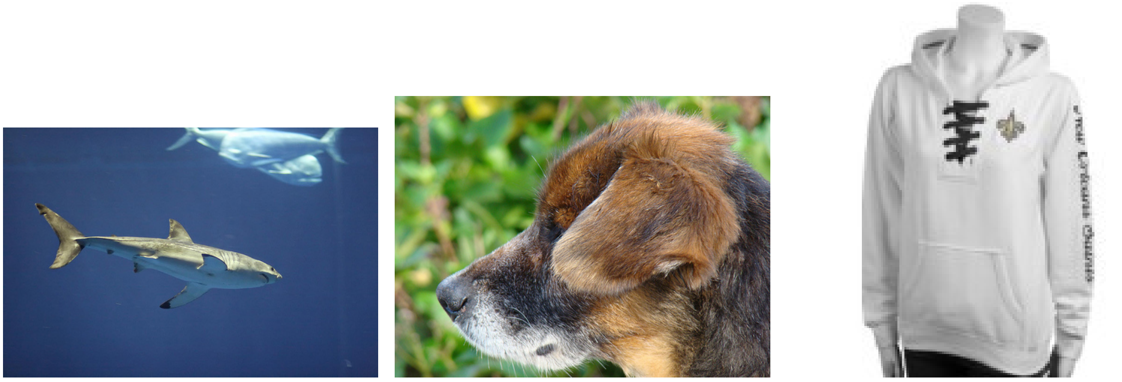
Figure 3: Additional examples of ImageNet captions generated by the base model (Base), and by the base model when constrained to include the ground-truth synset (Base+Synset). Occasionally the introduction of the ground-truth synset label has no effect (e.g. top right image). Words never seen in the combined MSCOCO / Flickr30k caption training set are underlined.



Base: A teddy bear sitting on top of a table.
Synset: piggy bank, penny bank.
Base + Synset: A teddy bear sitting on a penny bank.

Base: Two pictures of a dog and a dog.
Synset: electric ray, torpedo.
Base + Synset: Two pictures of a dog and a torpedo.

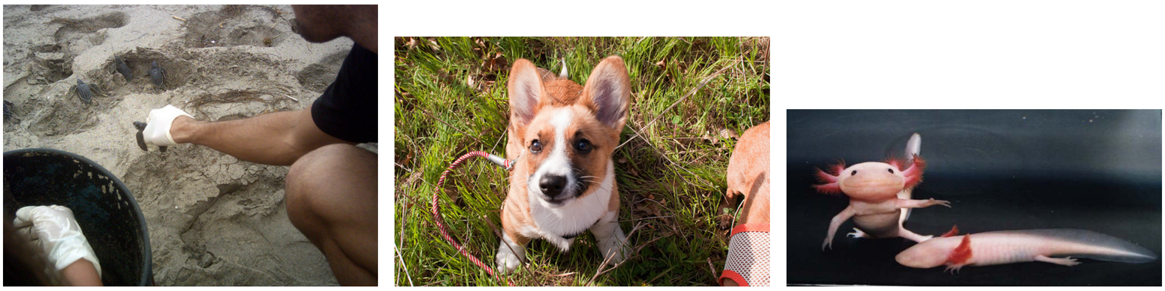
Base: A person is feeding a cat on the ground.
Synset: groenendael.
Base + Synset: A person holding a groenendael and a dog.



Base: A large jet flying through a blue sky.
Synset: great white shark, white shark, man-eater, Carcharodon carcharias, man-eating shark.
Base + Synset: A white shark flying through a blue sky.

Base: A brown dog is looking at the camera.
Synset: English foxhound.
Base + Synset: A English foxhound dog is looking at the camera.

Base: A man wearing a white shirt and tie.
Synset: sweatshirt.
Base + Synset: A man wearing a white sweatshirt and tie.



Base: A man is digging a hole in the sand.
Synset: leatherback turtle, leatherback, leathery turtle, Dermochelys coriacea.
Base + Synset: A man is digging a leatherback in sand.

Base: A brown and white dog laying on the grass.
Synset: Cardigan, Cardigan Welsh corgi.
Base + Synset: A Cardigan and white dog laying on the grass.

Base: A group of stuffed animals sitting on top of each other.
Synset: Ambystoma mexicanum, axolotl, mud puppy.
Base + Synset: The axolotl of two animals are on display.

Figure 4: Additional examples of ImageNet caption failure cases, including hallucinated objects (top), incorrect scene context (middle), and nonsensical captions (bottom). Words never seen in the combined MSCOCO / Flickr30k caption training set are underlined.