Supplementary Material: ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic

Yoad Tewel Yoav Shalev Idan Schwartz Lior Wolf
School of Computer Science, Tel Aviv University

This supplementary material describes our experimental setup (see Appendix A), provides additional ablation study (see Appendix B), provides additional qualitative results (see Appendix C), explores the limitations of our approach (see Appendix D), and discusses visual relation benchmark failure cases (see Appendix E).

A. Experimental Setup

As part of our experiments, we used COCO's validation set (Karpathy splits) for both qualitative and quantitative evaluations. We report the beam with the lowest CLIP loss score among the five beams. Our model has several hyper-parameters: (i) λ (see Eq. (2)), which was set to 0.2; (ii) τ_c (see Eq. (3)), which was set to 0.01; (iii) α (see Eq. (5)), which was set to 0.3; (iv) We decreased the likelihood of repeated tokens by a factor of two in order to mitigate repetitions. Based on a human assessment, these parameters produced concise, fluent, and image-related captions. We use the PyTorch framework [4].

Pre-trained models: As part of our approach, we use two large-scale pre-trained models: (i) GPT-2, using HuggingFace's gpt2-medium implementation[1], with 24 attention models and 345M trainable parameters. This model was trained on an 8M web-page dataset with a causal language modeling (CLM) objective; (ii) CLIP, trained on 400M (images, text) crawled from the web. We use the OpenAI implementation[2]. We employed a version of CLIP with a vision transformer image encoding architecture that is equivalent to ViT-B/32 [1].

Prompt engineering: Our method begins with an initial prompt. In the majority of our experiments, we used "Image of a". We determine the caption from the words generated after the initial prompt. We did observe that the prompt affected output results, e.g., "Image of text that says," is much better if the caption is intended for OCR.

[1] https://huggingface.co/transformers/model_doc/gpt2.html

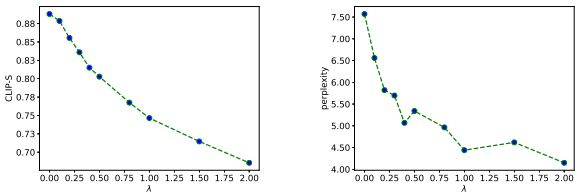
[2] https://github.com/openai/CLIP

B. Ablation Study

Effect of CLIP-based optimization: A further ablation was performed, in which CLIP's score is used directly to optimize the LM. In Fig. 1, we show two variants: (A1) selecting tokens one by one to maximize the CLIP score, and (A2) doing so on a score that combines CLIP score with an LM-score. Evidently, the captions are not competitive with our method. We also assessed the differences in language fluency (perplexity measured with GPT Neo) and image correspondence (measured with CLIP Score). Despite a higher CLIP score (Tab. 1), our method has improved language fluency. It is worth noting that higher CLIP doesn't necessarily translate to better wording.

A human study further supports this, conducted to determine which method is perceived as the best one. The study included 50 images randomly selected from COCO and 40 annotators. Our caption was selected 70.5% , (A1) 8.9%, and (A2) 20.6%.

Effect of regularizer coefficient: As shown below, an increase in the regularizer coefficient results in a decrease in the perplexity score measured with GPT Neo (*i.e.*, language fluency improves) while it decreases the clip similarity. We find $\lambda = 0.2$ to be a good trade-off point.



Human evaluation: We conducted an additional human study on 50 images. We picked the images from the web (*e.g.*, video-game screenshot, real-world knowledge; specifically, the subreddit 'i took a picture'). We asked the annotators to score between 1 to 5 two properties: human-like and visual grounding. We compared against a supervised method ClipCap. On human-like, our approach got 3.79 *vs.* 3.17 of ClipCap. On image grounding, our method got 3.98 *vs.* 3.21 of ClipCap.

C. Additional Qualitative Results

Image Captioning: In Fig. 6 (shown at the end of the document due to size), we present our results on 200 randomly-selected images along with baselines. For baselines, we use ClipCap [3], CLIP-VL [5], and VinVL [6]. Our method generates original captions that are completely different both in vocabulary and pattern from the baselines' captions.

D. Limitations

We detail both the caption quality issues and the biases resulting from the noisy web-scale data used to train CLIP and GPT-2 in the following sections.

Web-scale noise: The captions we generate are influenced by CLIP's training data. Due to its extraction from the web without special care, it contains noise. This leads to two undesirable outcomes: 1) Generating entities related to the data source (*e.g.*, Flickr) or irrelevant entities (*e.g.*, the name of the photographer). We solve this problem by adding a negative prior regularization to any capitalized subword. Consequently, a more generic caption will be created, but at the expense of world-knowledge capabilities. We show samples with and without the mechanism in Fig. 2; and 2) At times, the captions become irrelevant because they fail to remain focused. This can be controlled using two hyperparameters. We multiply the probability of the end token by a factor of f_e, starting from time-step t_e. In our method we used $f_e = 1.04$, and $t_e = 3$. In Fig. 3, several random examples are shown, and the length control mechanism is ablated.

Bias and Fairness: It is common for web-scale data to contain biased sources (*e.g.*, news), resulting in an unintended bias against some ethnic groups. In Fig. 4, an abstract illustration of a terrorist is described as Palestinian. Another example, racial characteristics are used to portray a child as an immigrant. Additionally, a caption implies homosexual orientation for an image of two men.

E. Visual Relations Benchmark Study

Our benchmark combines real-world knowledge with the ability to represent visual relationships. In Appendix E. we show at typical mistakes. Samples are referred to by their character counter: (a) Unpopular real-world knowledge. GPT-2 and CLIP training are based on web crawled data. Consequently, it may choose words based on popularity on the Internet. Sydney is a more popular city than Canberra worldwide (we validate this with Google Trends); (b) Synonyms. The relationship between the president and his or her country leads to "Canadian" rather than "Canada;" (c) Closely related. Rather than relating the pyramids to Egypt, this sample refers to Sinai, an area in Egypt; (d), (e),(f) Relation mistake. Subtracting Australia from Canberra conveys a relationship relevant to a university. It appears that adding the relationship to the UK led to 'Berkeley.' A 'Chinese university' is generated by adding it to China, and a 'German university' is generated by adding it to Germany. This might be due to Canberra being known for its university. Since we use the same relation (pair subtraction) for multiple triplet of images, inferring the wrong relation can lead to many errors in the benchmark.

Method	CLIP-S	Perplexity
A1	**0.98**	8.61
A2	0.91	6.04
Ours	0.87	**5.50**

Table 1. Comparison of our method with and without optimization. We show two variants: (A1) selecting tokens one by one to maximize the CLIP score, and (A2) doing so on a score that combines CLIP score with an LM-score.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

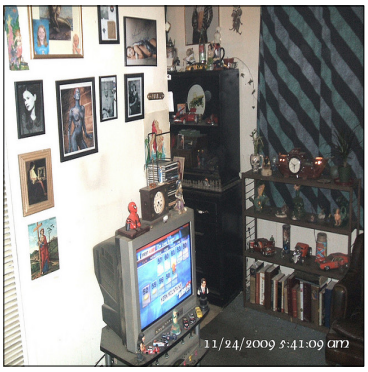
[3] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 6

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

[5] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2, 6

[6] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2, 6

A1: A mock cap 2013 Montreal and Leaf Blue.
A2: A baseball cap with mapleleaf stand blue.
Ours: A promotional cap from the Toronto Blue Jays 10/09 season.



A1: A space at home 3DO Studio located overlooking his gaming brazil.
A2: A computer games room at the House of Horror in 2001.
Ours: A room dedicated to games and other forms of entertainment that were popular in the late 90s.

A1: A real food model cake at Carpoolcar at Includes on San On.
A2: A train car from the Sain-Ollie and Beau-Niver.
Ours: Sean's truck cake.

Figure 1. Illustration of methods that employ CLIP directly without optimization to the LM. We show two variants: (A1) selecting tokens one by one to maximize the CLIP score, and (A2) doing so on a score that combines CLIP score with an LM-score.

With capital: A pizza with with and wine on Flickr license.
W/o capital: A pizza with wine.

With capital: A high school school dogwalking photo on Flickr showing the difference in behavior between two
W/o capital: A college dog in hand holding a leash.

With capital: A damaged suitcase on a a hillside in Kwa Zulune.
W/o capital: A damaged suitcase in the bush.

With capital: A recent skiing instruction program in Yosemite National Park website » The program is designed in
W/o capital: A group skiing pose.

Figure 2. The effect of our entity-control mechanism. With the mechanism (With Capital) and without the mechanism (W/O Capital).

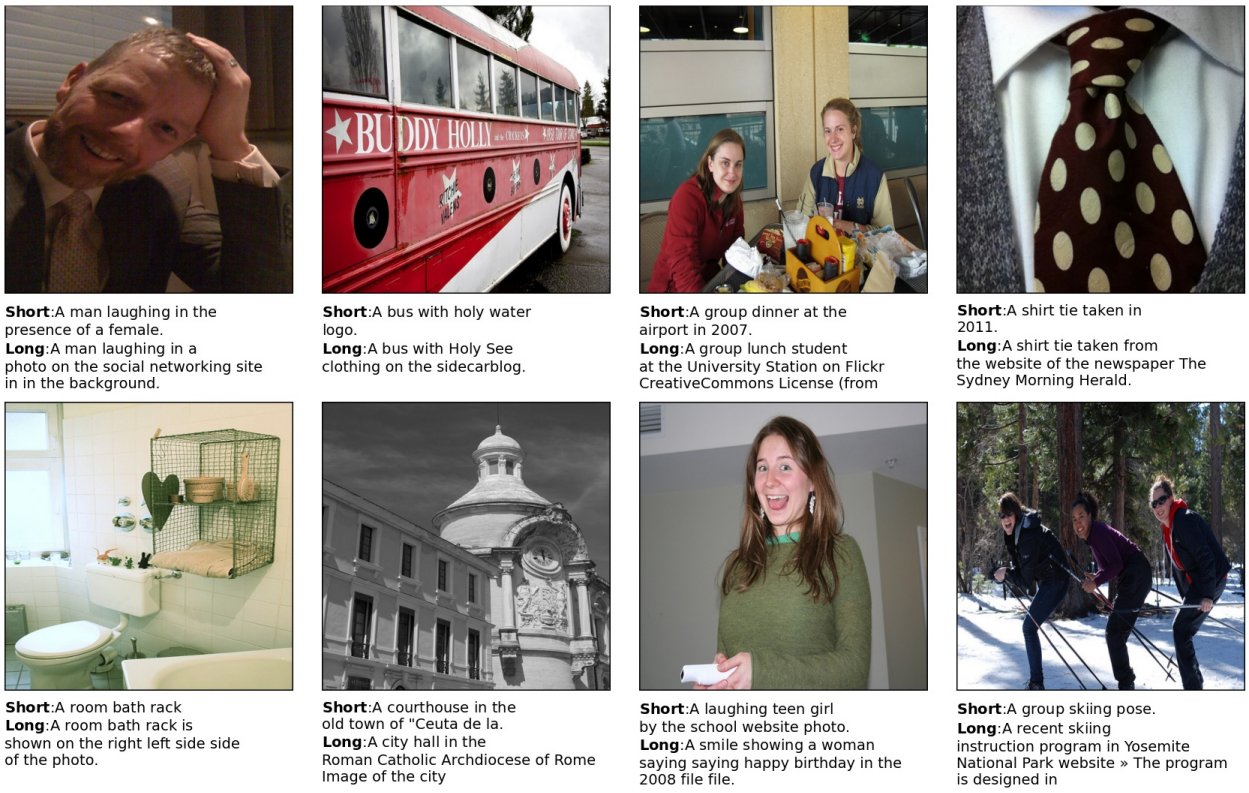
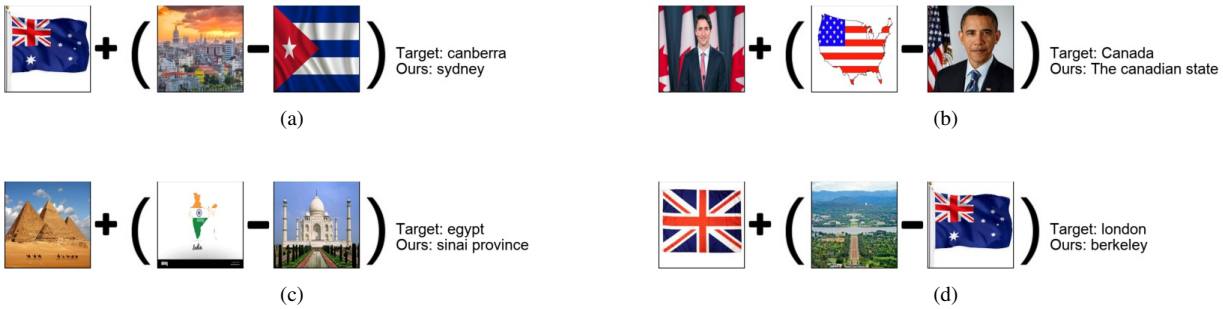
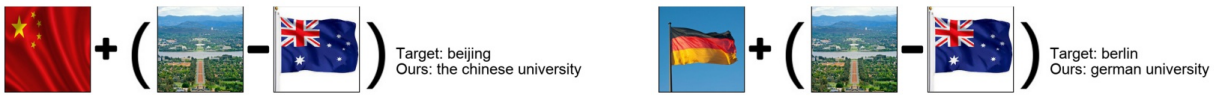


Figure 3. The effect of our length-control mechanism. With the mechanism (Short) and without the mechanism (Long).

Figure 4. Bias cases against distinct groups.





(e) (f)

Figure 5. Error analysis of the visual relations benchmark.

Figure 6. Generated captions by our method and by the baseline methods for images from the MS-COCO [2] test-set. CP=ClipCap [3], CVL=CLIP-VL [5], VVL=VinVL [6].

CP:a close up of a plate of food with broccoli on it
CVL:a plate of food with chicken and broccoli
VVL:a plate of food with meat and broccoli.
Ours:A typical meal served in the chicken and broccoli restaurant.

CP:A group of people standing around a table covered in food.
CVL:a group of people standing around a table with food
VVL:a group of people standing around a table with food.
Ours:A meeting in the city's community garden.

CP:a close up of a pizza on a pan on a stove
CVL:a pizza sitting on top of a wooden cutting board
VVL:a pizza sitting on a wooden cutting board next to a bottle of wine.
Ours:A pizza with wine.

CP:a close up of a cat wearing a tie
CVL:a cat wearing a tie laying on a blanket
VVL:a cat wearing a bow tie laying on a bed.
Ours:A cat dressed in business attire.

CP:A fighter jet flying over a forest filled with trees.
CVL:an airplane flying in the sky over trees
VVL:a large airplane flying over a tree.
Ours:A jet taking off in the jungle.

CP:The wing of an airplane flying in the sky.
CVL:a view of the clouds from an airplane window
VVL:the view of the wing of an airplane flying over the clouds.
Ours:A plane on the air.

CP:A man in a suit and tie holding a beer.
CVL:a man in a suit and tie holding a beer bottle
VVL:a man in a suit and tie holding a beer.
Ours:A guy in suits drinking a beer in the photo.

CP:A couple of cats laying on top of a bed.
CVL:three cats laying on top of a bed
VVL:two cats laying on a white bed.
Ours:A pair cat in the bed.

CP:A cow that is laying down in the street.
CVL:a brown cow laying on the ground next to a motorcycle
VVL:a brown cow laying on the ground next to a motorcycle.
Ours:A cow resting on the streets.

CP:A bunch of food that is laying on the ground.
CVL:a group of people sitting on the ground with food
VVL:a group of people sitting on the sand with food.
Ours:A makeshift kitchen picnic set in the summit finishway.

CP:A young boy doing a trick on a skateboard.
CVL:a man doing a trick on a skateboard in a skate park
VVL:a man doing a trick on a skateboard at a skate park.
Ours:A young skateater in the video.

CP:Four bowls filled with different types of fruits and vegetables.
CVL:three bowls of food with fruit and vegetables on a table
VVL:three bowls of fruit and salad in them on a table.
Ours:A healthy salad in various configurations.

CP:A cat sitting on top of a wooden bench.
CVL:a cat sitting on a wooden bench in a garden
VVL:a cat sitting on a wooden bench in a garden.
Ours:A cat perched in the city's garden.

CP:A man standing next to a row of parked motorcycles.
CVL:a man working on a dirt bike in a parking lot
VVL:a man on a motorcycle doing a trick in a parking lot.
Ours:A bike show at the mountain.

CP:A bathroom with a hole in the floor and a toilet in it.
CVL:a dirty tiles on the floor of a bathroom
VVL:a dirty bathroom with a urinal on the floor.
Ours:A toilet rupture in the home.

CP:A traffic light with two street signs on it.
CVL:a traffic light on a pole with a sign
VVL:a traffic light with a sign on it.
Ours:A pedestrian camera sign.

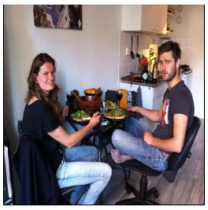
CP:A horse that is standing in the snow.
CVL:a black horse standing next to a fence in the snow
VVL:a couple of horses standing next to a fence in the snow.
Ours:A pony in the winter.

CP:A coffee mug sitting on top of a wooden table next to a pair of scissors.
CVL:a pot with three pairs of scissors on a wooden table
VVL:a pair of scissors and other utensils in a cup.
Ours:A metal dining utensil holder.

CP:A man riding on the back of a yellow motorcycle.
CVL:a man riding a yellow motorcycle on a highway
VVL:a man sitting on a yellow motorcycle in a parking lot.
Ours:A motorcycle loaded with high-speed cargo.

CP:A glass of beer and a slice of pizza on a table.
CVL:a beer and a pizza on a table
VVL:a glass of beer and a slice of pizza on a table.
Ours:A beer and football featuring a pizza.

CP:A group of people sitting on chairs under umbrellas.
CVL:a group of people sitting at tables under a umbrella
VVL:a group of people standing around a food stand.
Ours:A market in the village of al-Sakrab street.



CP:A man and a woman sitting at a table with plates of food.
CVL:a man and a woman sitting at a table eating food
VVL:a man and a woman sitting at a table eating food.
Ours:A dinner in the flat with a friend.

CP:A blue double decker bus parked in front of a building.
CVL:a purple double decker bus parked in a building
VVL:a purple double decker bus parked in front of a building.
Ours:A bus from the museum showing a number of different views.

CP:A clock hanging from the side of a building.
CVL:three clocks on the side of a building
VVL:a large clock on the side of a building.
Ours:A clock in the downtown area.

CP:A bathroom with a sink, toilet, and bathtub.
CVL:a bathroom with a sink and a window
VVL:a bathroom with a sink and a large mirror.
Ours:A typical bathroom in the hotel.

CP:A red motorcycle parked in front of a crowd of people.
CVL:a red motorcycle parked in a parking lot
VVL:a row of motorcycles parked next to each other.
Ours:A motorcycle displayed in 2015.

CP:A group of lawn chairs sitting on top of a sandy beach.
CVL:a group of chairs and umbrellas on a beach
VVL:a group of yellow umbrellas on a beach.
Ours:A beach patio in the northern city of Wisconsin.

CP:A row of trucks parked next to each other in a parking lot.
CVL:a group of trucks parked in a dirt field
VVL:a group of trucks parked in a parking lot.
Ours:A truck fleet gathering.

CP:Two dogs playing with a ball on the beach.
CVL:two dogs running on the beach in the water
VVL:a couple of dogs running on the beach.
Ours:A duel in the beach via @jenn_dog.

CP:A cruise ship docked at a pier in a city.
CVL:a boat in the water in front of a building
VVL:a large white boat in the water near a city.
Ours:A ferry in front of the hotel.

CP:A man riding a snowboard down the side of a snow covered slope.
CVL:a man flying through the air while riding a snowboard
VVL:a man riding a snowboard down a snow covered slope.
Ours:A man skiing on the lifts.

CP:A red motorcycle parked on top of a dirt field.
CVL:a motorcycle parked in front of a fence
VVL:a red and black motorcycle parked in front of a fence with sheep.
Ours:A motorcycle in the southern province of al-Jazirah.

CP:A couple of people standing next to a stop sign.
CVL:two people standing next to a stop sign
VVL:a couple of women standing in front of a stop sign.
Ours:A few young teens in the cold north.

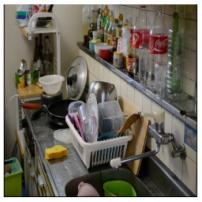
CP:A city street filled with lots of tall buildings.
CVL:a group of cars on a city street with a traffic light
VVL:a city street with tall buildings and cars.
Ours:A downtown street in the video.

CP:an elephant with a blanket on its back and a person standing next to it
CVL:two people riding on the back of an elephant
VVL:an elephant with a blanket on its back standing next to a tree.
Ours:A man sheltering an elephant in the compound.

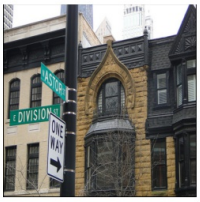
CP:A man in a suit and tie sitting on a bench.
CVL:a man wearing a suit and tie sitting in a chair
VVL:a man wearing a shirt and tie sitting in a chair.
Ours:A researcher smiling in his labelling study.

CP:A person riding on the back of a brown horse.
CVL:a group of men standing next to a horse in a field
VVL:a man standing next to a woman riding a horse in a field.
Ours:A small pony training with the trainer in a small pony training area.

CP:A couple of giraffe standing next to each other.
CVL:a zebra and a giraffe standing next to each other
VVL:a giraffe and a zebra eating grass in a zoo.
Ours:A mother feeding her young in the enclosure.



CP:A kitchen sink filled with dishes and dishes.
CVL:a kitchen counter with spices and spices on it
VVL:a kitchen sink with dishes and utensils in it.
Ours:A kitchen in the rebel-controlled city of the besieged eastern city of al-



CP:A street sign on a pole in front of a building.
CVL:two street signs on a pole in front of a building
VVL:a couple of street signs on a pole in front of a building.
Ours:A home in the city with signs.

CP:A man standing on top of a snow covered slope.
CVL:a man standing on a snowboard in the snow
VVL:a man standing on a snowboard in the snow.
Ours:A sunset ski instructor at the park.

CP:A man riding a skateboard down the side of a ramp.
CVL:a man riding a skateboard on a wall
VVL:a man doing a trick on a skateboard in a room.
Ours:A young and rebellious student jumping wall.

CP:A group of people standing around a brown dog.
CVL:a group of women standing in the grass with a dog
VVL:a group of women standing next to a dog.
Ours:A college dog in hand holding a leash.

CP:A cow that is standing in the grass.
CVL:a black and white cow drinking water in a field
VVL:a black and white cow drinking water from a pond.
Ours:A cow image cropped up 4/09.

CP:A market with a variety of fruits and vegetables.
CVL:a market with a bunch of fruits and vegetables
VVL:a bunch of fruits and vegetables on display at a market.
Ours:A bazaar in the city.

CP:A stove top oven sitting inside of a kitchen.
CVL:a microwave oven sitting next to a microwave
VVL:three microwaves sitting next to each other on a counter.
Ours:A standard kitchen microwave.

CP:A cat is sitting in front of a box of pizza.
CVL:a cat standing next to a pizza box with a pizza
VVL:a cat standing next to a box of pizza.
Ours:A pizza cat in the photo is a composite.

CP:a black and white photo of a street sign on a hill
CVL:a street sign on the side of a road
VVL:a street sign on the side of a road.
Ours:A 22 mph speed limit.

CP:A cat sitting on top of a table next to a bike.
CVL:a cat sitting on the ground in front of a window
VVL:a small cat sitting on the ground next to a fence.
Ours:A stray in the city.

CP:A pile of luggage sitting on top of a bed.
CVL:a suitcase filled with a laptop computer sitting on a bed
VVL:a suitcase and a laptop on a bed.
Ours:A laptop seized in the southern city of de la theo in 2011.

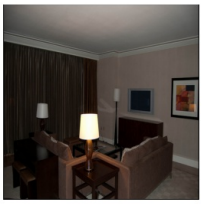
CP:A large brown dog sitting on top of a car seat.
CVL:a dog sitting in the back of a car
VVL:a dog sitting on the lap of a person in a car.
Ours:A dog driver in 2006.

CP:A group of men on a field playing baseball.
CVL:a group of baseball players standing on a field
VVL:a group of baseball players on a field.
Ours:A pitch at the 2011 season opener.

CP:A person holding a cell phone in their hand.
CVL:a person holding a cell phone in their hand
VVL:a person holding a cell phone in their hand.
Ours:A mobile phone in the user's hand taken by a third-party security system.

CP:A plate of food that is on a table.
CVL:a pizza on a white plate with a bowl of sauce
VVL:a pizza with beans and cheese on a blue plate.
Ours:A taco pizza with beans.

CP:A bathroom with three urinals mounted to the wall.
CVL:three urinals in a bathroom with a sink
VVL:two urinals and a sink in a bathroom.
Ours:A typical sink in the downtown campus.



CP:A living room filled with furniture and a flat screen TV.
CVL:a living room with a couch and a table
VVL:a living room with a couch and a piano.
Ours:A livingroom in the hotel.

CP:A man riding a wave on top of a surfboard.
CVL:a man riding a wave on a surfboard in the ocean
VVL:a man riding a wave on a surfboard in the ocean.
Ours:A man surfing in the area city.

CP:A brown and white cow standing on top of a grass covered field.
CVL:a brown and white cow standing in a field
VVL:a cow standing in a field of grass.
Ours:A cattle in the fields.

CP:A white car parked on top of a sandy beach.
CVL:a man standing next to a car on the beach
VVL:a man standing next to a car on a beach.
Ours:A vehicle lying on the desert with a person loading bag.

CP:A woman standing on top of a snow covered slope.
CVL:a woman standing on skis in the snow
VVL:a person standing on skis in the snow.
Ours:A group instructor at the first ski class of a 2010 season.

CP:A white plate topped with rice and vegetables.
CVL:a white plate of food with rice and vegetables
VVL:a white plate of food with rice and vegetables.
Ours:A healthy plate by courtesy of www.

CP:A man riding a bike past a stop sign.
CVL:a man riding a bike next to a stop sign
VVL:a man riding a bike down a road next to a stop sign.
Ours:A cyclist stopping at the trail.

CP:A couple of people on a small boat in the water.
CVL:a man riding in a boat in the water
VVL:a man sitting in a small boat in the water.
Ours:A boat on the canal.

CP:A plate of food with french fries and a glass of juice.
CVL:two plates of food on a table with orange juice
VVL:a plate of food on a table with a glass of orange juice.
Ours:A large breakfast at the popular restaurant in downtown and surrounding.

CP:A bus driving down a street at night.
CVL:a bus driving down a city street at night
VVL:a white bus driving down a city street at night.
Ours:A bus in the night.

CP:A man riding a skateboard up the side of a ramp.
CVL:a man doing a trick on a skateboard in a skate park
VVL:a man doing a trick on a skateboard at a skate park.
Ours:A high-kick flip.

CP:A wooden bench sitting in front of a garden.
CVL:a wooden bench sitting in front of a white house
VVL:a wooden bench in front of a white house.
Ours:A bench in the garden.

CP:A black and white photo of two boys sitting next to each other.
CVL:two young boys holding a stuffed animal
VVL:two young boys sitting on a couch with a teddy bear.
Ours:A 1950s pair of baby's in the family.

CP:A small black dog standing on top of a bath tub.
CVL:a black dog standing in the water
VVL:a black dog standing in a bath tub.
Ours:A pet's wet behavior.

CP:A group of cows that are standing in the dirt.
CVL:a group of cows eating hay in a pen
VVL:a couple of cows eating hay in a pen.
Ours:A cow stalls in the lab.

CP:A cat that is laying down in front of a book.
CVL:a cat sitting on top of a book shelf
VVL:a cat sleeping on top of a wooden shelf next to books.
Ours:A cat laughing from the library.

CP:a black and white photo of a street sign and some trees
CVL:a black and white photo of a street sign in front of a cemetery
VVL:a black and white photo of a street sign.
Ours:A road in the middle of a cemetery.

CP:A bus that is sitting on the side of the road.
CVL:a red bus parked on the side of the street
VVL:a red and green bus driving down a street next to flowers.
Ours:A bus in the gardens.

CP:An orange and white cat laying on top of a computer keyboard.
CVL:a cat sleeping on top of a laptop computer
VVL:a cat laying in front of a laptop computer.
Ours:A young cat using the laptop.

CP:A large brown dog sitting in the middle of a pile of clothes.
CVL:a dog sitting on a pile of clutter next to a bed
VVL:a brown dog sitting on a bed with clothes.
Ours:A homeless pit in the home.

CP:A herd of sheep standing on top of a lush green field.
CVL:a sheep and a baby sheep in a field
VVL:a couple of sheep standing in a field with a bird.
Ours:A sheep running in the background.

CP:A man in a suit holding a bicycle in front of a house.
CVL:a man in a suit standing next to a bike
VVL:a man in a suit standing next to a bike.
Ours:A bicyclist on display at the home of businessman and bicycle owner.

CP:A young boy riding skis down a snow covered slope.
CVL:a group of children on skis in the snow
VVL:a small child is on skis in the snow.
Ours:A child's ski touring.

CP:A woman sitting at a table in front of a plate of food.
CVL:a woman sitting at a table in a restaurant
VVL:a woman sitting at a table in a restaurant.
Ours:A woman in the restaurant environment.

CP:A giraffe standing on top of a lush green field.
CVL:a giraffe standing in front of trees
VVL:a giraffe walking in the dirt near trees.
Ours:A tall animal in the zoo.

CP:A white cat sitting in a bathroom sink.
CVL:a white cat sitting in a bathroom sink
VVL:a white cat standing on top of a bathroom sink.
Ours:A female cat being shampooed in the bathroom with shampoo on the sink.

CP:a close up of a slice of pizza on a table
CVL:a pizza in a box on a table
VVL:a large pizza sitting in a box on a table.
Ours:A pizza with chicken.

CP:A person holding a wine glass in their hand.
CVL:a person holding a glass of wine
VVL:a person holding a wine glass in their hand.
Ours:A glass of champagne being presented as a gift can be seen in the video.

CP:a desk with a keyboard a monitor and a mouse
CVL:a desk with a computer and a laptop on it
VVL:a desk with two computer monitors and a laptop.
Ours:A typical desktop setup in the city.

CP:A traffic light sitting on top of a pole under a blue sky.
CVL:a traffic light on a pole with a blue sky
VVL:a couple of traffic lights on a pole.
Ours:A traffic light in front of the city's new traffic commissioner.

CP:A bathroom with a walk in shower next to a toilet.
CVL:a bathroom with a toilet and a shower
VVL:a bathroom with a toilet and a sink and a shower.
Ours:A typical shower at the home.

CP:A brown teddy bear sitting in front of a book.
CVL:a teddy bear sitting on a couch holding a book
VVL:a teddy bear sitting on a chair holding a book.
Ours:A toy reading bear.

CP:A man riding on the back of a brown horse on top of a sandy beach.
CVL:a person riding a horse on the beach
VVL:a person riding a horse on the beach.
Ours:A man riding on the coast horse in a calm.

CP:A man and woman pose for a picture together.
CVL:a bride and groom posing for a picture
VVL:a bride and groom are posing for a picture.
Ours:A couple in red ties.

CP:A person in a field flying a kite.
CVL:a man sitting in a field flying a kite
VVL:a person flying a kite in a field with a dog.
Ours:A parking ground with dog.

CP:A man sitting in front of a group of children.
CVL:a group of children sitting in a library with a dog
VVL:a group of children petting a dog in a room.
Ours:A child care dog playing in a a classroom demonstration.

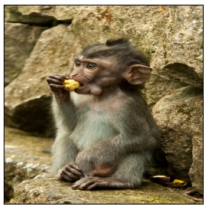
CP:a close up of a plate of food with broccoli on a table
CVL:three bowls of food on a table
VVL:a bowl of salad and a pan of food on a table.
Ours:A typical dinner made with greens.

CP:A street sign on a pole on a city street
CVL:a street sign on the side of a city street
VVL:a couple of street signs on a pole in a city.
Ours:A car zone sign in the city.

CP:A pizza sitting on top of a wooden cutting board.
CVL:a pizza sitting on top of a wooden cutting board
VVL:a pizza sitting on top of a wooden cutting board.
Ours:A pizza shown on the ad.

CP:A red and white street sign sitting on top of a flooded street.
CVL:a traffic light on a pole in the water
VVL:a traffic light and a street sign in the water.
Ours:A report pole in flood.

CP:A vase filled with lots of different colored flowers.
CVL:a green vase filled with red flowers on a table
VVL:a green vase filled with red flowers on a table.
Ours:A "virtual roses display" created by the artist in collaboration with the artist.



CP:A monkey sitting on a rock eating a banana.
CVL:a newborn baby sitting on rocks eating a piece of food
VVL:a small monkey sitting on a rock eating a piece of food.
Ours:A monkey eating bread.

CP:A bunch of bananas sitting on top of a table.
CVL:a bunch of bananas in a wooden crate
VVL:a bunch of bananas sitting in a wooden box.
Ours:A banana in front of a vendor stand.

CP:An airplane is parked at the gate at an airport.
CVL:a group of airplanes parked on the runway at an airport
VVL:a view of airplanes on the runway from an airport window.
Ours:A plane boarding at the gate.

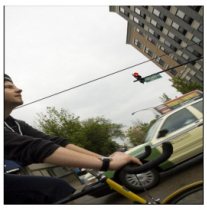
CP:A man riding a skateboard down the side of a ramp.
CVL:a man doing a trick on a skateboard
VVL:a man riding a skateboard in a tunnel.
Ours:A pedestrian skating on the banks.

CP:A flock of birds flying over a body of water.
CVL:a group of birds flying in the cloudy sky
VVL:a group of birds flying in the sky over a beach.
Ours:A flying birds in the background.

CP:A glass bowl filled with apples and bananas.
CVL:a bowl of apples and oranges on a table
VVL:a glass plate with fruit on a table.
Ours:A fruit table courtesy of the photographer and used with permission of the artist.

CP:A cut in half sandwich sitting on top of a wooden cutting board.
CVL:a sandwich on a cutting board with a knife
VVL:a sandwich on a cutting board on a table.
Ours:A small sandwich provided by the restaurant.

CP:A woman putting a turkey into an oven.
CVL:a woman taking a turkey out of an oven
VVL:a woman pulling a dish out of an oven.
Ours:A turkey preparing in the coopage.



CP:A man driving a car down a street next to tall buildings.
CVL:a man riding a bike with a traffic light
VVL:a man riding a bike at a traffic light.
Ours:A driver observing cyclists in front of the intersection.

CP:a bathroom with a sink and a toilet in it
CVL:a bathroom with a sink and a toilet
VVL:a bathroom with a toilet and a sink.
Ours:A baf [before pic] of the bathroom.

CP:Three birds perched on top of a wooden table.
CVL:two birds sitting on top of a table
VVL:two small birds standing on a table next to a knife.
Ours:A little birders' picnic.

CP:A group of people riding skis down a snow covered slope.
CVL:a group of people skiing down a snow covered slope
VVL:a group of people skiing down a snow covered slope.
Ours:A typical ski in the state ranges from snowy mountains that rise to the low hills

CP:A herd of sheep standing on top of a lush green field.
CVL:a herd of sheep laying in the snow
VVL:a herd of sheep grazing in the snow in front of a building.
Ours:A farm with sheep in the winter.

CP:A rusty fire hydrant sitting on the side of a road.
CVL:a green fire hydrant on the side of a street
VVL:a green fire hydrant sitting next to a yellow pole.
Ours:A pump attached to the curb.

CP:A tow truck towing a car in a parking lot.
CVL:a man standing in the back of a truck
VVL:a man standing in the back of a truck.
Ours:A man cleaning vehicles in the area.

CP:A clock on top of a pole in front of a building.
CVL:a clock on a pole in front of a building
VVL:a clock on a pole in front of some trees.
Ours:A clock on in the district.

CP:A vase sitting on top of a table filled with flowers.
CVL:a vase with flowers in it on a table
VVL:a orange vase with flowers on a table.
Ours:A typical lamp made from a plant of the orange flame-billed cherry.

CP:An airport with several planes parked on the tarmac.
CVL:an airplane parked on the runway at an airport
VVL:a plane is parked on the runway at an airport.
Ours:A runway deck view at the airport.

CP:A vase filled with yellow flowers on top of a table.
CVL:a bunch of bananas hanging from a store
VVL:a bunch of bananas hanging from a ceiling in a store.
Ours:A bar decorated banana arrangements.

CP:A red fire truck parked in front of a building.
CVL:a red fire truck parked in front of a building
VVL:a red fire truck parked in front of a building.
Ours:A fire engine on the roof.

CP:A man sitting on the ground next to an elephant.
CVL:a man sitting on a chair next to an elephant
VVL:a man sitting next to an elephant.
Ours:A farmer's elephant in the village.

CP:A woman in red shirt and black skirt playing a game of tennis.
CVL:a woman holding a tennis racket at a tennis ball
VVL:a woman serving a tennis ball on a tennis court.
Ours:A player in tennis is displayed.

CP:A bunch of pots that are sitting on the ground.
CVL:a cat sitting on top of a brick garden
VVL:a cat sitting on top of a chimney surrounded by plants.
Ours:A home garden in the village.

CP:A group of people riding skis down a snow covered slope.
CVL:a man riding a snowboard down a snow covered street
VVL:a man riding a snowboard down a snow covered street.
Ours:A snowy scene in the parking lot is shown running children.

CP:A traffic light sitting on the side of a road.
CVL:a traffic light with a street sign on a pole
VVL:a traffic light with a bicycle sign on it in front of a building.
Ours:A cyclist on the red signal.

CP:A black and white cat laying on top of a laptop computer.
CVL:a black and white cat laying on a laptop computer
VVL:a black and white cat laying on a desk next to a laptop.
Ours:A working with cat on the computer.

CP:A baseball player holding a bat next to home plate.
CVL:a baseball player swinging a bat at a ball
VVL:a baseball player swinging a bat at a ball.
Ours:A hitter striking out in the video.

CP:A couple of motorcycles parked on the side of a road.
CVL:a group of people riding motorcycles on a road
VVL:a group of people riding motorcycles on a road.
Ours:A motorcycle group hiking the summit.

CP:A group of people standing on top of a lush green field.
CVL:a group of people standing in a field with a frisbee
VVL:a group of people standing in a field with a frisbee.
Ours:A party at the park with a group of friends.

CP:A wooden bench sitting next to a brick wall.
CVL:a wooden bench sitting next to a brick wall
VVL:a wooden bench sitting next to a plant.
Ours:A bench at the courthouse in suburban.

CP:A little girl standing on top of a snow covered slope.
CVL:a young girl riding skis down a snow covered slope
VVL:a little girl standing on skis in the snow.
Ours:A child ski girl.

CP:A young man wearing glasses and a neck tie.
CVL:a man wearing glasses and a black shirt and a tie
VVL:a man wearing a black shirt and a green tie.
Ours:A guy dressed in geeky girl-dressing tie.

CP:An elephant with tusks walking through a field.
CVL:an elephant walking in the grass in a field
VVL:a large elephant walking through a dry grass field.
Ours:A elephant in the wild.

CP:a public transit bus on a city street
CVL:a bus driving down a city street with cars
VVL:a white bus driving down a city street.
Ours:A bus approaching on the sidewalk.

CP:A large passenger jet flying over a tall building.
CVL:an airplane flying in the sky over tall buildings
VVL:a plane flying in the sky over some tall buildings.
Ours:A flight in the city.

CP:A man taking a picture of himself in a train mirror.
CVL:a man taking a picture of himself in a mirror with a surfboard
VVL:a man taking a picture of himself in a mirror with a surfboard.
Ours:A tourist in the lift.

CP:A red, white and blue airplane flying in the sky.
CVL:a red and white plane flying in a blue sky
VVL:a red and white plane flying in the sky.
Ours:A plane in flight.

CP:A herd of sheep standing on top of a lush green field.
CVL:a group of sheep grazing in a field
VVL:a group of sheep grazing in a field.
Ours:A sheep in the field.

CP:A bunch of umbrellas that are in the grass.
CVL:a row of umbrellas sitting next to a house
VVL:a group of umbrellas are line up outside of a house.
Ours:A resort in the northern mountains.

CP:A young girl is eating a donut on a plate.
CVL:a little girl eating a piece of pizza
VVL:a young girl is eating a piece of pizza.
Ours:A child eating doughroll.

CP:Boats are docked at a pier on a cloudy day.
CVL:a group of boats are parked on the water
VVL:a group of boats are docked in the water.
Ours:A downtown dock showing the city.

CP:a close up of a cat wearing a tie
CVL:a cat wearing a tie laying on a blanket
VVL:a cat wearing a bow tie laying on a bed.
Ours:A cat dressed in business attire.

CP:A woman swinging a tennis racquet on top of a tennis court.
CVL:a woman holding a tennis racket on a court
VVL:a woman swinging a tennis racket on a tennis court.
Ours:A player wearing tennis shoes with a dress.

CP:A group of people standing around a pile of luggage.
CVL:a group of men standing in a room with luggage
VVL:a group of men sitting in front of a pile of luggage.
Ours:A crowded packing trip.

CP:A woman standing next to a red and white biplane.
CVL:a woman standing next to a small airplane
VVL:a woman standing next to a small plane.
Ours:A flight instructor in the promotional video.

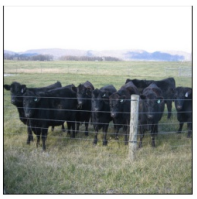
CP:A slice of pizza sitting on top of a white plate.
CVL:a slice of pizza on a yellow plate
VVL:a slice of pizza on a white plate with a knife.
Ours:A pizza from the restaurant in 2014.

CP:A white toilet sitting in a bathroom next to a wall.
CVL:a bathroom with a toilet and a bucket
VVL:a bathroom with a toilet and a black bucket on the wall.
Ours:A man-stall refurbishment.

CP:A red and yellow train traveling down train tracks.
CVL:a red and yellow train on the tracks next to a building
VVL:a yellow and red train is parked in front of a building.
Ours:A tram at the station.

CP:A group of people sitting on top of a couch.
CVL:a group of people sitting on a couch playing a video game
VVL:a couple of people sitting in a living room playing a video game.
Ours:A social gaming session.

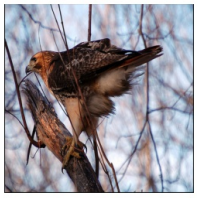
CP:A boat floating on top of a body of water.
CVL:a red and blue boat sitting in the water
VVL:a reflection of a red and blue boat in the water.
Ours:A boat reflection in the spring.



CP:A herd of black cows standing on top of a grass covered field.
CVL:a herd of black cows standing behind a fence
VVL:a herd of black cows standing behind a barbed wire fence.
Ours:A herd at the farm fence.

CP:A large brown teddy bear sitting on top of a couch.
CVL:a large teddy bear sitting on a couch in a living room
VVL:a large teddy bear sitting on top of a couch.
Ours:A large plush bear.

CP:A woman eating a banana in front of a window.
CVL:a woman eating a banana on a plate
VVL:a woman sitting at a table eating a banana.
Ours:A banana-eating Japanese woman.



CP:a close up of a bird on a branch of a tree
CVL:a bird sitting on top of a tree branch
VVL:a brown and white bird perched on a tree branch.
Ours:A hawk sniffing in the woods on a tree.

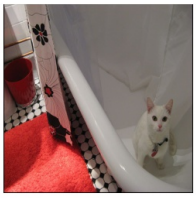
CP:A couple of cats laying on top of a bed.
CVL:three cats laying on a bed
VVL:a group of cats sleeping on a bed.
Ours:A trio sleeping on the sheets.

CP:A vase filled with flowers sitting on top of a wooden table.
CVL:two vases with flowers sitting on a wooden table
VVL:two vases of flowers sitting on a wooden table.
Ours:A few floral arrangement.

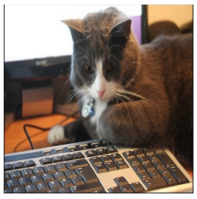
CP:An elephant is crossing a river with a baby elephant.
CVL:an elephant standing in the water in a river
VVL:a large elephant walking across a river.
Ours:A elephant in the river of northern central and southern western parts.

CP:A large brown horse standing on top of a lush green field.
CVL:a black horse standing in the grass in a forest
VVL:a black horse standing in the grass near trees.
Ours:A horse in the woodland area.

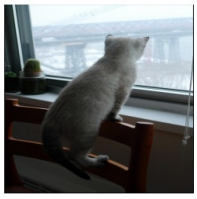
CP:A group of people riding skis down a snow covered slope.
CVL:a group of people on skis in the snow
VVL:a group of people on skis in the snow.
Ours:A massive climb accident in snow climbing.



CP:A black and white cat sitting in a bathroom sink.
CVL:a white cat sitting in a bath tub
VVL:a white cat sitting in a bath tub next to a shower curtain.
Ours:A cat spa in the bath.



CP:A cat laying on top of a computer keyboard.
CVL:a cat laying on top of a computer keyboard
VVL:a cat laying on top of a computer keyboard.
Ours:A cat hacking the keyboard.



CP:A cat sitting on a window sill looking out a window.
CVL:a cat sitting on top of a window sill
VVL:a cat sitting on a chair looking out a window.
Ours:A cat observing the weather.

CP:A black bird eating a bird feeder full of bird seed.
CVL:a black bird standing in a pot eating an apple
VVL:a black bird sitting in a flower pot eating food.
Ours:A crow feed in the garden.

CP:A couple of dirt bikes sitting on top of a table.
CVL:a motorcycle on display in a museum
VVL:a white motorcycle on display on a glass floor.
Ours:A 2009 model displayed in the local bike shop.

CP:A person cutting a pizza on top of a wooden cutting board.
CVL:a person cutting a pizza and a glass of wine
VVL:a person is making a pizza on a table with wine glasses.
Ours:A pizza with wine.

CP:A piece of luggage that is laying on the ground.
CVL:a suitcase laying on its side on the ground
VVL:a blue suitcase sitting on the ground next to grass.
Ours:A damaged suitcase in the bush.

CP:A group of seagulls standing on a line of wooden posts.
CVL:a group of birds sitting on posts in the water
VVL:a group of birds sitting on wooden posts in the water.
Ours:A flock waiting in line at a dock.

CP:A lit up sign on the side of a building.
CVL:a traffic light on a street sign on a building
VVL:a traffic light in front of a building with signs.
Ours:A bar with lights.

CP:A computer monitor sitting on top of a wooden desk.
CVL:a woman sitting at a desk with a computer
VVL:a woman sitting at a desk with a computer monitor.
Ours:A computer viewer in the lab is shown.

CP:A white truck driving down a street next to tall buildings.
CVL:a truck driving down a city street
VVL:a police truck parked in the middle of a city street.
Ours:A vehicle blocking traffic.

CP:A couple of men riding on the back of motorcycles.
CVL:a man wearing a helmet sitting on a motorcycle
VVL:a man riding a motorcycle down a street.
Ours:A cop riding in uniform.

CP:Two motorcycles parked on the side of the road.
CVL:a group of motorcycles parked in a parking lot
VVL:a couple of motorcycles parked in a parking lot.
Ours:A motorcycle rental in the mountainsview park in 2002.

CP:A group of people on mopeds on a city street.
CVL:a group of motorcycles parked on the side of a street
VVL:a group of people on motorcycles on a street.
Ours:A traffic on the street.

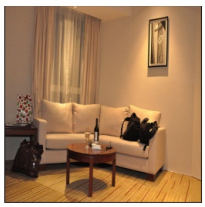
CP:a close up of a pizza on a pan on a stove
CVL:a pizza sitting on top of a tray on a table
VVL:a pizza sitting on top of a stove top.
Ours:A pizza recipe from the 2012 video.

CP:A brown horse running across a dry grass field.
CVL:two horses standing in a field of dry grass
VVL:a brown horse standing in a field with a desert background.
Ours:A horse roaming desert grass.

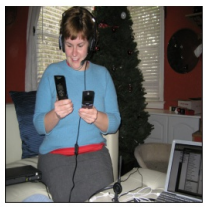
CP:A person riding a bike down a street in the rain.
CVL:a person walking in the rain with an umbrella
VVL:a person crossing a city street with an umbrella.
Ours:A city in winter rains.

CP:A woman riding on the back of a white horse.
CVL:a woman riding a white horse in a field
VVL:a person riding a horse in a field.
Ours:A horse rider in the forest.

CP:A blue fire hydrant sitting on the side of a road.
CVL:a fire hydrant sitting on the side of a street
VVL:a yellow fire hydrant sitting in front of a building.
Ours:A street laser sculpture.



CP:a living room with a couch a table and a tv monitor
CVL:a living room with a couch and a table
VVL:a living room with a couch and a table with a bottle of wine.
Ours:A lounge in the apartment.



CP:A woman sitting on a couch holding two cell phones.
CVL:a woman holding two cell phones in front of a christmas tree
VVL:a woman sitting on a couch holding up a cell phone.
Ours:A mobile gamegirl playing in a a holiday party.

CP:a close up of many oranges on a plate
CVL:a bunch of oranges sitting on a white plate
VVL:a group of oranges in a white bowl.
Ours:A healthy oranges in the table of contents.

CP:A man sitting at a table with a plate of donuts.
CVL:a person sitting at a table with a plate of donuts
VVL:a person sitting at a table with plates of food.
Ours:A typical breakfast challenge in the 1980s.

CP:A man taking a picture of himself in a mirror.
CVL:a man taking a picture of himself with his cell phone
VVL:a man in a suit taking a picture of himself in a mirror.
Ours:A friend in the lift.

CP:a close up of a cat laying on a luggage bag
CVL:a black cat laying on top of a suitcase
VVL:a black cat sleeping on top of a suitcase.
Ours:A suitcase cat.

CP:A group of people posing for a picture in front of a building.
CVL:a group of people posing for a picture in front of a hotel
VVL:a group of people posing for a picture in front of a building.
Ours:A group at the prom in 2003.

CP:A couple of people on skis in the snow.
CVL:two women standing in the snow with skis
VVL:a couple of women standing in the snow with skis.
Ours:A scene filming on the slopes in 2012.

CP:A woman laying on top of a bed next to a cat.
CVL:a woman sleeping on a bed with a cat
VVL:a woman laying in a chair with a cat.
Ours:A sleeping and cat.

CP:A group of people standing around a luggage carousel.
CVL:a group of people standing in a mall with luggage
VVL:a group of people waiting for their luggage at an airport.
Ours:A shopping queue in the arrivals department.

CP:A train pulling into a train station next to a platform.
CVL:a train on the tracks at a train station
VVL:a yellow and black train at a train station.
Ours:A train being approaching the station use white and with a yellow roof.