
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) $\lambda$ (see Eq. (2)), which was set to 0.2; (ii) $\tau_c$ (see Eq. (3)), which was set to 0.01; (iii) $\alpha$ (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.

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 VinvVL [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.

References


A2: A baseball cap with maple leaf 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.
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.

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.
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].
A horse that is standing in the snow.
A coffee mug sitting on top of a wooden table next to a pair of scissors.
A man riding on the back of a yellow motorcycle.
A glass of beer and a slice of pizza on a table.
A group of people sitting on chairs under umbrellas.
A man and a woman sitting at a table with plates of food.
A man and a woman sitting at a table eating food.
A blue double decker bus parked in front of a building.
A clock hanging from the side of a building.
A bathroom with a sink, toilet, and bathtub.
A man and a woman sitting at a table eating food.
A red motorcycle parked in front of a crowd of people.
A row of motorcycles parked next to each other.
A row of trucks parked next to each other in a parking lot.
Two dogs playing with a ball on the beach.
A cruise ship docked at a pier in a city.
A man riding a snowboard down a snow covered slope.
A red motorcycle parked on top of a dirt field.
A couple of horses standing next to a fence in the snow.
A pot with three pairs of scissors on a wooden table.
A man sitting on a yellow motorcycle in a parking lot.
A red motorcycle parked in front of a building.
A large white boat in the water near a city.
A large clock on the side of a building.
A large umbrella on a beach.
A motorcycle loaded with high-speed cargo.
A man riding a yellow motorcycle on a highway.
A beer and a pizza on a table.
A pony in the winter.
A metal dining utensil holder.
A dinner in the flat with a friend.
A motorcycle displayed in 2015.
A beach patio in the northern city of Wisconsin.
A group of lawn chairs sitting on top of a sandy beach.
A group of chairs and umbrellas on a beach.
A red and black motorcycle parked in front of a fence with sheep.
A truck fleet gathering.
A ferry in front of the hotel.
A man skiing on the lifts.
CP: A couple of people standing next to a stop sign.
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 zebra and a giraffe standing next to each other.
CVL: 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 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.
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 plane.
Ours: A living room 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 cow 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 white cat sitting in a bathroom sink.
**CV:** A white cat sitting in a bathroom sink
**VV:** 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 traffic light sitting on top of a pole under a blue sky.
**CV:** A traffic light on a pole with a blue sky
**VV:** A couple of traffic lights on a pole.
**Ours:** A traffic light in front of the city's new traffic commissioner.

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

**CP:** A street sign on a pole on a city street.
**CV:** A street sign on the side of a city street
**VV:** 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.
**CV:** A pizza sitting on top of a wooden cutting board
**VV:** A pizza sitting on top of a wooden cutting board.
**Ours:** A pizza shown on the ad.

**CP:** A person holding a wine glass in their hand.
**CV:** A person holding a glass of wine
**VV:** 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 and a mouse.
**CV:** A desk with a computer and a laptop on it
**VV:** A desk with two computer monitors and a laptop.
**Ours:** A typical desktop setup in the city.

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

**CP:** A close up of a plate of food with broccoli on a table.
**CV:** Three bowls of food on a table
**VV:** A bowl of salad and a pan of food on a table.
**Ours:** A typical dinner made with greens.

**CP:** A red and white street sign sitting on top of a flooded street.
**CV:** A traffic light on a pole in the water
**VV:** 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.
**CV:** A green vase filled with red flowers on a table
**VV:** 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 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: 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 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 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 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 ticking the summit.

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: An 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 doughnut.

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 at 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 plane.
VVL: A woman standing next to a small airplane.
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 in a living room 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 barber 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 snatching 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: An 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 in 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 mountains view park in 2002.

CP: A group of people on mopeds on a city street.
CVL: a group of people on motorcycles on a street.
VVL: a group of people on motorcycles parked on the side of 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 running in a field
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 and a table.
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 gamer girl playing in 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 man sitting at a table with plates of food.
Ours: A typical breakfast challenge in the 1990s.

CP: A man taking a picture of himself in a mirror.
CVL: A man taking a picture of himself with a 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.