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

Abstract

Recent text-to-image matching models apply contrastive learning to large corpora of uncurated pairs of images and sentences. While such models can provide a powerful score for matching and subsequent zero-shot tasks, they are not capable of generating caption given an image. In this work, we repurpose such models to generate a descriptive text given an image at inference time, without any further training or tuning step. This is done by combining the visual-semantic model with a large language model, benefiting from the knowledge in both web-scale models. The resulting captions are much less restrictive than those obtained by supervised captioning methods. Moreover, as a zero-shot learning method, it is extremely flexible and we demonstrate its ability to perform image arithmetic in which the inputs can be either images or text and the output is a sentence. This enables novel high-level vision capabilities such as comparing two images or solving visual analogy tests. Our code is available at: https://github.com/YoadTew/zero-shot-image-to-text.

1. Introduction

Deep learning has led to at least three major revolutions in computer vision: (i) machines that achieve, in multiple domains, what is considered a human level of performance earlier than anticipated [40, 66], (ii) effective transfer learning, which supports rapid modeling of new domains [76], and (iii) a leap in unsupervised learning through the use of adversarial and self-supervised learning [9, 24].

A fourth revolution that is currently taking place is that of zero-shot learning. A seminal work by OpenAI presented the transformer-based [67] GPT-3 model [6]. This model is trained on extremely large text corpora and can then generate text given a prompt. If the prompt contains an instruction, GTP-3 can often carry it out. For example, given the prompt "Translate English to French: typical $\rightarrow$ typique, house $\rightarrow$ ... " would generate the word "maison."

Impressive zero-shot capability was later on demonstrated, also by OpenAI, in computer vision. While state-of-the-art computer vision models are often trained as task-specific models that infer a fixed number of labels, Radford *et al*. [57] have presented the CLIP image-text transformer model, which can perform tens of downstream tasks, without further training, with an accuracy comparable to the state of the art. This is done by selecting, given an image, the best match out of sentences of the form "This is an image of X." Subsequently, Ramesh *et al*. [59] presented a bimodal Transformer termed DALL-E, which generates images that match a given description in unseen domains with unprecedented performance.

In this work, we employ CLIP to perform the inverse task of DALL-E, namely zero-shot image captioning. Given an image, we employ CLIP together with the GPT-2 language model [58] (we do not have access to GPT-3) to generate a textual description of the input image. This adds a new image-analysis capability to CLIP, beyond the fixed-prompt zero-shot learning demonstrated by Radford *et al*.

As a zero-shot method, our approach does not involve any training. One can argue that the underlying CLIP model is trained with exactly the same type of supervision that image captioning methods [64, 78] are trained on, i.e., pairs of matching images and captions. However, image captioning methods are trained from curated sources, such as MS-COCO [44] or Visual Genome [39], while CLIP is trained on WebImageText (WIT), which is an automatically collected web-scale dataset. Previous attempts to train a captioning model on WIT have led to poor performance in recognizing the objects in the image, see Sec. 2.

As a result of the difference in both methodology and underlying data, the captions produced by our method are very different from those obtained by the supervised captioning methods. While supervised methods can mimic human annotators and provide similar sentences, in terms of conventional NLP metrics (such as BLEU [54]) to the ground truth sentences, our results exhibit much more freedom and match the image better in the visual-semantic CLIP embedding space (ours is optimized for this). Moreover, the semantic knowledge incorporated into CLIP and GPT-2 is manifested in the resulting caption, see Fig. 1.

In addition to the different nature of the obtained captions, our method is also more flexible, since all the computing occurs at inference time. Specifically, we show the

Figure 1. Our novel captioning method ZeroCap exhibits real-world knowledge, generates text that is more diverse and less scripted than existing methods, can address the written content of an image, and can perform visual-semantic arithmetic.

ability to perform semantic analysis in image space by using a new form of arithmetic. A well-known example for concept arithmetic in NLP is that of retrieving the word 'queen' as the closest word, in the embedding space, to the equation involving the embedding vectors associated with 'king,' 'man,' and 'woman,' after subtracting the 2nd from the 1st and adding the 3rd. We present the novel ability to do the same, only with images instead of words, such that the result is generated as a short sentences, and not just a word, see Fig. 1.

As a corollary, we can, for example, ask what the difference is between two scenes. This ability to compare two images semantically is a novel computer vision capability, which further demonstrates the power of zero-shot learning.

2. Related work

The first deep captioning methods applied RNNs to generate sequences of words [38, 49]. Attention was added to identify relevant salient objects [61, 63, 73]. Graph neural networks and scene graphs incorporated spatial as well as semantic relationships between objects [37, 74, 75]. Subsequently, Transformers modeled interactions among all image elements with self-attention [17, 52, 62, 67]. On the text modeling side of the problem, language models (LMs) have also advanced with the development of LSTMs [16, 71], CNNs [4] and Transformers [28, 30, 48]. Language improvements include devising better image grounding [47], decoding non-visual words (e.g., 'the,' 'and') [46], generating fine, novel and diverse sentences [7,25,72], and incorporating information from different semantic taggers [12, 34].

In recent years, significant improvements have been achieved by utilizing large-scale vision-language data sets. The unsupervised data is used as a pre-training phase, to initialize models with image-text correspondence [15, 42, 78]. With this technique, millions of image and text pairs from the web can be adopted. Nevertheless, in previous work we are aware of, all captioning models employ human-annotated datasets, such as MS-COCO or the Visual Genome, in the last stage of training.

It is likely impossible to construct a database of curated captions that is large enough, to describe even a modestly large fraction of plausible images and objects. This results in biases [21, 22, 70]. Several approaches focused on describing novel objects by conditioning the model on external unsupervised data during training [1, 29, 69]. Alternatively, external object taggers can be used during different phases (e.g., pre-training, training, or inference) [3, 19, 32, 43, 47]. Semi-supervised methods are also available [36]. Unsupervised approaches can be achieved by training with a visual concept detector or by learning a joint image-language embedding space [20, 41]. In contrast, our method makes use of an existing image-text alignment score to direct an existing large-scale LM toward a given image without training.

CLIP is trained on 400M images/sentence pairs from the web [57], resulting with a powerful text-image matching score. Originally CLIP's authors explored training an image-to-caption language model with this training set, but found that it struggled with zero-shot transfer. In a 16 GPU-day experiment, a language model only achieved 16% accuracy on ImageNet [14]. CLIP achieves the same level of accuracy roughly 10x faster.

Using prompts, it is possible to imitate some capabilities of text generation. For example, CLIP-based applications exhibit zero-shot solving capabilities in various scenarios never seen before. With careful engineering of the prompt, one can, for example, improve detection of unseen objects [27]. Zero-shot prompt engineering has also been used for higher-level tasks (e.g., VQA), but it is nowhere near the level of supervised methods [64].

CLIP also provides powerful means for supporting text-driven image manipulation with Generative Adversarial Networks (GANs) or other generative models [8, 56, 60]. Our work explores the other direction: generating text us-

ing an image, by guiding a large-scale LM with CLIP.

Guided language modeling has become a primary challenge, as researchers strive to tune prior knowledge within large-scale LMs, such as GPT-2 [58]. Fine-tuning is often accomplished by employing Reinforcement Learning [79] or GANs [77] for each attribute separately. Disentangling the latent representations into style and content is also relevant in terms of text style transfer [33, 65]. A controllable LM can also be formed using fixed control codes [35]. Ideally, conditioning should be applied directly to the existing large-scale LM, without the need for fine-tuning. Several studies have explored the idea of steering an LM using small neural networks [10, 26]. Following that, PPLM [13] demonstrated that a simple attribute classifier could steer a model without any further training. With our work, we present a novel visual LM guidance from visual cues.

3. Method

Visual captioning is the process of generating a descriptive sentence for an image. It can be formalized as a sequence generation problem given an input image I, i.e., as a conditional probability inference for the i-th word x_i of the sentence, i.e., $p(x_i|[x_t]_{t<i}, I)$.

This is typically accomplished in a supervised manner, by optimizing weights to reproduce ground truth sentences. However, since carefully curated datasets are small, and cannot adequately describe all images, the sentences generated often describe the content at the basic level of the objects present in the scene and sound artificial. Such problems can be mitigated with the use of web-scale datasets. We present a zero-shot method for guiding large-scale language models with a large-scale text-image alignment model.

Overview Our approach uses a transformer-based LM (e.g., GPT-2) to infer the next word from an initial prompt, such as "Image of a," as illustrated in Fig. 2. To incorporate image-related knowledge to the auto-regression process, a calibrated CLIP loss $\mathcal{L}_{\text{CLIP}}$ stimulates the model to generate sentences that describe a given image. An additional loss term $\mathcal{L}_{\text{CE}}$ is used to maintain the next token distribution similar to the original language model. Optimization occurs during auto-regression, and repeated for each token.

Furthermore, the flexibility of our method enables the capturing of semantic relations through simple arithmetic of visual cues in CLIP's embedding space. Finally, combining multi-modal encoders with our method allows knowledge to be extracted in a new way that mixes between text and images.

.

Language models In recent years, LMs have improved significantly and are getting closer to AI-complete capabilities, including broad external knowledge and solving a wide variety of tasks with limited supervision. A Transformer-based LM typically models interactions between the generated token and past tokens at each time-step.

Recall that the transformer block has three embedding functions K, Q, V [67]. The first two, K, Q, learn the token interactions that determine the distribution over V. The attention mechanism pools values based on the similarity between queries and keys. Specifically, the pooled value for each token i depends on the query associated with this token Q_i, which is computed using the function Q over the current embedding of this token. The result is obtained as the weighted average of the value vectors, based on the cosine similarity between Q_i and the keys associated with all tokens K_j.

While K and V are functions, the obtained key and values K_j and V_j are used repeatedly when generating text, one word at a time. K_j and V_j can therefore be stored in what is called a context cache, in order to keep track of past embedding outputs of K and V. The sequence generation process can then be written as

$$x_{i+1} = \text{LM}\left(x_i, [(K_j^l, V_j^l)]_{j<i, 1\leq l\leq L}\right), \tag{1}$$

where x_i is the i-th word of the generated sentence, K_j^l, V_j^l are the context transformer's key and value of the j-th token, and l indicates the index of the transformer layers, out of a total of L layers. Our method employs GPT-2, which has $L = 24$ layers.

We next describe how we align our LM with the input image. We do so by modifying, during inference, the values of the context cache $C_i = [(K_j^l, V_j^l)]_{j<i, 1\leq l\leq L}$ leaving the LM unchanged.

CLIP-Guided language modelling Our goal is to guide the LM towards a desired visual direction with each generation step. The guidance we propose has two primary goals: (i) alignment with the given image; and (ii) maintaining language attributes. The first goal is obtained through CLIP, which is used to assess the relatedness of a token to an image and adjust the model (or, rather, the cache) accordingly. For the second goal, we regularize the objective to be similar to the original target output, i.e., before it was modified.

The solved optimization problem adjusts the context cache C_i at each time point and is formally defined as $\arg\min_{C_i} \mathcal{L}_{\text{CLIP}}(\text{LM}\left(x_i, C_i\right), I)$

$$+ \lambda \mathcal{L}_{\text{CE}}\left(\text{LM}\left(x_i, C_i\right), \hat{x}_{i+1}\right), \tag{2}$$

where $\hat{x}_{i+1}$ is the token distribution obtained using the original, unmodified, context cache. The second term employs CE loss to ensure that the probability distribution across words with the modified context is close to the one of the original LM. The hyperparameter λ balances the two loss terms. It was set to 0.2 early on in the development process and was unmodified since. Next, we explain how the CLIP loss term is calculated.

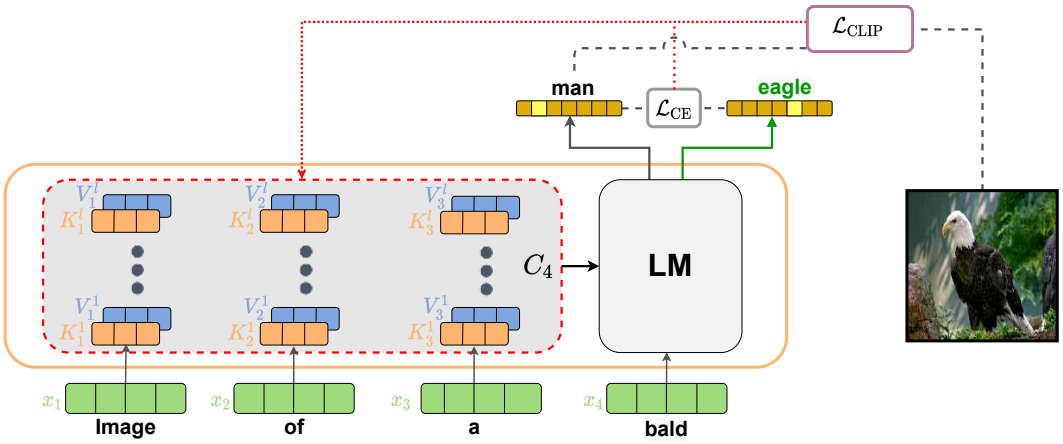


Figure 2. An overview of our approach. We guide the model towards the phrase 'eagle' instead of 'man'. We do this by adjusting the context (C_4), using the gradients of CLIP loss ($\mathcal{L}_{\text{CLIP}}$) illustrated with a red arrow. To maintain language attributes, we optimize the minimum distance to the original distribution of the LM ($\mathcal{L}_{\text{CE}}$).

CLIP loss We calculate image relevance for the possible tokens at time i. It is sufficient to compute potentials for the top 512 token candidates and set the rest to zero potential for efficiency. To this end, the corresponding candidate sentence $s_i^k = (x_1, ..., x_{i-1}, x_i^k)$ for the k-th candidate token is matched against the image I.

The clip potential of the k-th token is computed as

$$p_i^k \propto \exp(\mathrm{D}_{\text{CLIP}}(E_{\text{Text}}(s_i^k), E_{\text{Image}}(I))/\tau_c)), \qquad (3)$$

where D_{CLIP} is the cosine distance between CLIP's embeddings of the text (i.e., E_{Text}) and the image (i.e., E_{Image}), and $\tau_c > 0$ is a temperature hyperparameter that controls the sharpness of the target distribution. In all our experiments, we set τ_c to 0.01.

The CLIP loss is defined as the cross-entropy loss between the clip potential distribution and the target distribution of the next token x_{i+1} obtained by the language model:

$$\mathcal{L}_{\text{CLIP}} = \mathrm{CE}(p_i, x_{i+1}). \qquad (4)$$

This loss encourages words that lead to higher CLIP matching scores between the image and the generated sentence.

Inference As a zero-shot method, no training takes place. At inference time one optimizes the problem in Eq. (2), which we denote as $p(x_{i+1}|C_i)$, by conducting five steps of gradient descent, i.e.,

$$C_i \longleftarrow C_i + \alpha \frac{\nabla_{C_i} p(x_{i+1}|C_i)}{\|\nabla_{C_i} p(x_{i+1}|C_i)\|^2}. \qquad (5)$$

This update rule is simplified for brevity. With each newly-generated token, the optimization is re-done. In our implementation, the gradients are normalized with Euclidean normalization before each step, separately for each transformer layer. We set the learning rate α to 0.3.

Beam search The byte-level tokenizer used employs 256 bytes of base tokens to represent every word in existence [58]. Any word can also be split into more than one subwords, e.g., the word 'zebra' is tokenized as 'zeb' and 'ra'. As a result, we found that images of zebras are described as striped animals, since the token 'zeb' is not picked. Beam search inference helps solve this problem by enabling the search to be conducted in a less myopic way.

4. Visual-Semantic Arithmetic

Recent studies suggested that CLIP multi-modal representation holds an elaborate concept taxonomy [23]. In accordance with this intuition, we find that our method can express CLIP's embedding in a textual way. For instance, subtracting between CLIP-encoded images and applying our method transcribes a relationship between the two images. Furthermore, by summing vectors we can steer the generated caption towards a conceptual direction.

To perform arithmetic in CLIP's embedding space, we first encode the image/text using CLIP's image/text encoder. For instance, let I_1, I_2 be two images. We encode the images with CLIP's encoder, i.e., $E_{\text{image}}(I_1), E_{\text{image}}(I_2)$. Next, we carry out the desired arithmetic, e.g., addition with $E_{\text{image}}(I_1) + E_{\text{image}}(I_2)$. Finally, we use the obtained result instead of the image encoding $E_{\text{image}}(I)$ within Eq. (3) to steer the generated sentence.

Consequently, we can generate detailed knowledge of the external world by moving in conceptual directions. This way, our method can answer questions expressed visually, for example, "who is the president of Germany?" To achieve this, we subtract "America's flag" from an image of "Obama" and obtain a presidential-direction, to which we can then add the image of a second country's flag.

Our approach extends beyond visual interactions alone. Using CLIP's textual encoder, interaction with a natural lan-

	Supervised Metrics					Diversity Metrics		Unsupervised Metric
Method	B@4	M	C	S	CLIP-S^{Ref}	Vocab	%Novel	CLIP-S
ClipCap [51]	32.15	27.1	108.35	20.12	0.81	1650	66.4%	0.77
CLIP-VL [64]	40.2	29.7	134.2	23.8	0.82	2464	85.1%	0.77
VinVL [78]	**41.0**	**31.1**	**140.9**	**25.2**	**0.83**	1125	77.9%	0.78
Ours	2.6	11.5	14.6	5.5	0.79	**8681**	**100%**	**0.87**

Table 1. For each method, we report supervised metrics (i.e., ones requiring human references): B@1 = BLEU-1, M = METEOR, C = CIDEr, S = SPICE. We also report diversity metrics, which measures the vocabulary size (Vocab), and the number of novel sentences w.r.t the training set (%Novel). Finally, we report semantic relatedness to the image (CLIP-S), and to the human references (CLIP-S^{Ref}) based on CLIP's embeddings.

guage is possible. In this case, one performs arithmetic operations in the embedding space such that the expression contains both image- and text-embeddings.

5. Experiments

For all the results reported in this section, we used a strategy for reducing repetitions, in which the probability for generating tokens that were generated at the last four time-steps was decreased by a factor of two. We also incorporated a mechanism that directly controls the length of the generated text by multiplying the probability of the end token by a factor of f_e, starting from time-step t_e. We use $f_e = 1.04$ and $t_e = 3$ for image captioning, and $f_e = 1.06$ and $t_e = 1$ for image arithmetic. On a single Titan X GPU, five beams and 512 candidate tokens can be generated in three seconds. Inference time is proportional to the number of candidates and beams.

5.1. Image Captioning

We begin by studying our zero-shot method for caption generation. Notably, we find our captions to exhibit human-like characteristics, such as generating diverse captions, reading, exploiting a wide range of external knowledge, and coping with abstract concepts. In Tab. 1, we present our results for COCO's test set [44]. Two recent baselines that use CLIP's embedding are compared to: ClipCap [51] and CLIP-VL [64]. In ClipCap, the image is encoded using CLIP and the representation is transferred and plugged as a token into a fine-tuned GPT-2. CLIP-VL incorporates spatial grid features from CLIP into a transformer network. Another method, VinVL [78] is a state-of-the-art technique.

We first consider supervised metrics, i.e., metrics requiring human references. These metrics include the BLEU [53], METEOR [5], CIDEr [68], SPICE [2], and CLIPScoreRef that we discuss below. As can be seen, our method lags in these metrics in comparison to the supervised captioning methods. Since the ground truth human annotation is obtained similarly to the training set, with the same group of annotators using similar terms, there is a clear advantage for methods trained on COCO annotations.

We next consider diversity metrics. Our vocabulary over COCO's test set is significantly larger than previous approaches (8681 vs. 2464). In addition, none of the generated sentences appear in the training set of COCO (100% on %Novel).

CLIPScore [31] is a reference-free method for evaluating relatedness between an image and its caption, using CLIP's alignment score. Evidently, our method is much better in this metric than the supervised method (87% vs. 77%). As an alternative to exact correspondence with human reference, we use CLIPScoreRef to measure the semantic distance from the references. Although supervised methods outperform our method in this score (similarity in the vocabulary and the sentence style still provide an advantage), the gap is narrower than in other supervised metrics.

Qualitative Analysis Fig. 3 compares our zero-shot approach with other baselines, demonstrating that our method can generate human-like captions, i.e., textually richer, better at image reasoning, and more effective at grounding objects. We discuss each image from left to right. First, as opposed to CLIP-VL, which assumes a toilet is in the bathroom, and VinVL, which disregards the background buildings and presumes it is on a sidewalk, our method determines it is on a rooftop. Next, our method attempts to generate the written text on a boat's side. The following image describes a flight meal as a regular tray of food with the baselines, whereas our method describes it as a flight meal. We accurately describe the next image as a bar restroom with portraits and not a bathroom. Our method and VinVL specify specific birds in the following photo (red falcons and hawks are hard to tell apart). Next, the baselines repeat the same sentence, while our method mentions an interesting mesh tile pattern. In the next photo, our method identifies a family rather than a general group. Last, our method accurately describes a room's interior, such as a bedroom with posters, and deduce that the posters depict bands. Note that the baselines' captions are generally of the same pattern, while our method generates novel sentences. Also, note that the images are taken from COCO dataset,

CLIP-VL: A toilet sitting on a ledge in a bathroom. **VinVL:** A white toilet sitting on a sidewalk next to a wall. **Ours:** A new rooftop toilet being installed on a residential street.

CLIP-VL: A boat with a picture on it in the water. **VinVL:** A sail boat is tied up in the water. **Ours:** A boat with message on the side saying "token".

CLIP-VL: A tray of food on a table with bread and drinks. **VinVL:** A tray of food on a table with a lot of items. **Ours:** A flight meal being served on the Boeing 707 aircraft.

CLIP-VL: A bathroom with a sink and a mirror. **VinVL:** A bathroom sink with a mirror and pictures. **Ours:** A bar restroom with portraits of the late 19th and 20th century.

CLIP-VL: A bird sitting on top of a window. **VinVL:** A hawk perched on top of a window sill.. **Ours:** A Red Falcon in the exhibit.

CLIP-VL: A bathroom with a sink and a mirror. **VinVL:** A bathroom with a sink and a mirror. **Ours:** A bathroom with a mesh tile pattern.

CLIP-VL: A group of people standing in the snow with skis. **VinVL:** A group of people standing in the snow with skis. **Ours:** Winter activities of parents with children.

CLIP-VL: A mirror on the wall in a bedroom with a bed. **VinVL:** A bedroom with a bed and posters on the wall. **Ours:** A bedroom with posters on the wall that shows bands.

Figure 3. Examples of our zero-shot image captioning compared against supervised captioning methods.

The president Kennedy's death. | Welcome. | The University of Stanford. | Central PARK.

Figure 4. Examples of OCR capabilities.

The President Trump smiles. | The Simpsons. | Manhattan on the iconic island. | The Avengers poster. | Mario driving. | Stanford University.

Figure 5. Examples of real-world knowledge.

which was used to fine-tune CLIP-VL and VinVL.

OCR The ability of CLIP to classify text within an image from a closed set of possible prompts is impressive [57]. We show in Fig. 4 that these capabilities can be exploited in a generative manner. To accomplish this, we change the prefix prompt we use in our method from "Image of a" to "Image of text that says." Results include impressive understanding. e.g., "The president Kennedy's death" from an image of a paper declaring it or generating "The University of Stanford" from a sign depicting its name.

External knowledge The generated captions comprise a wealth of real-world knowledge on a variety of topics. In Fig. 5 we show samples of famous people (e.g., Trump), animated shows (e.g., Simpsons), cities (e.g., Manhattan), movies (e.g., Avengers), games (e.g., Mario driving), and places (e.g., Stanford).

5.2. Visual-Semantic Arithmetic Study

We demonstrate how our method can generate text for subtraction to explain semantic directions. Next, we demonstrate that the summation operator allows guidance of the generated text through visual cues. One can then apply the above insights to solve visual analogy puzzles.

Subtraction Subtracting vectors intuitively represents a direction between the vectors. In Fig. 6 we demonstrate our method's ability to express relations through several examples. "A caricature illustration" is the result of subtracting a real photo of an airplane from a caricature. To put it another way, adding the concept "A caricature illustration" to the right image of a real plane will match the image on the left of a caricature plane. Concepts of quantity and color can also be seen, for example, a comparison of a green apple versus a red apple yields 'Red,' and vice-versa, subtracting one basketball from many basketballs results in "a bunch." Furthermore, we find directions related to a geographical area, e.g., 'Snow,' and 'Desert.' Further, a concept directly tied to day and night, and a concept of prison (i.e., 'Jailed'). It should be noted that the operator is not symmetric, and cannot always be derived textually. For instance, on the right images, the concept direction from a skateboard to a skateboard tournament can be generated as "The event." However, the direction from a skateboard tournament to a skateboard generated "schematic fossil view," which is irrelevant.

Summation Through the addition operation, the generated text can be guided through visual semantics. In Fig. 7 we show examples of guidance. On the left side, with the addition of a police officer's hat, the caption describes a man running as "A police officer...," if we add a hammer to a man, we get "The judge." On the right side, we show that a concept can be abstract. For example, the Apple company can be represented by an apple. Thus, adding an apple to a phone, results in the text "Apple's iPhone released." Additionally, a country's concept can be represented visually with flags. If Canada's flag is added to a tree, "Toronto Maple" results.

Guidance with Visual-Relations In the field of natural language processing, semantic relations have long been studied [50]. Previous efforts studied visual relations with expensive annotated language-priors [45]. With the introduction of CLIP, richer visual concepts from large-scale data became available [23]. Through visual arithmetic, we are able to exploit this richer embedding space.

In Fig. 8, we show our proposed strategy. Using subtraction, we first determine the direction. For example, the concept of leadership is represented by an image of Obama minus the American flag. With this direction in hand, we can now manipulate the case of other nations. By adding the direction to the German flag, we obtain "Angela Merkel." A different example is to examine the concept direction of CEO-to-company. With different images (e.g., Bill Gates

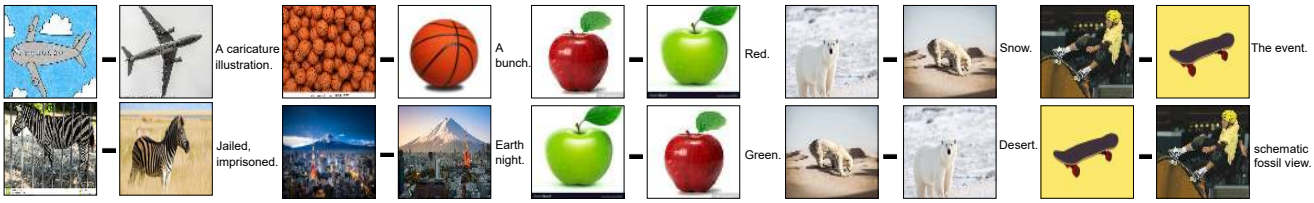


Figure 6. Examples of vector directions derived by subtracting representations from CLIP's embedding space. By generating text for the given direction, the concept is revealed.

Figure 7. Examples of caption guidance with an image through the addition operator.

and Microsoft, Jeff Bezos and Amazon), the direction can be summed to Mark Zuckerberg and Steve Jobs generating 'Facebook' and 'Apple,' respectively. On the right side, we study various interactions with country-related representations. We guide the image of a baguette to generate 'France' by taking photos of pizza and Italy and deriving the country-to-food direction.

The Visual Relations benchmark To further study the relation capabilities of our technique quantitatively, we introduce a new benchmark of visual relations, VR for short. This benchmark comprises 320 relations of the following templates: buildings→country, countries→capital, foods→country, leaders→country, and CEO→company. These were chosen because they are roughly many to one, i.e., a country has many buildings, but a building only relates to one country. The benchmark is designed to measure both the ability to model relations visually and to apply real-world knowledge to perform the task.

We constructed the benchmark through the following steps: (i) we created semantic directions by subtracting visual pairs and (ii) we then used each direction and added it to a visual element in another pair to create its corresponding text companion. As an example, we used images of ('japan,' 'sushi') to convey the direction of food→country, and then we added this direction to an image of a pizza and examined the appearance of Italy in the generated text.

We focused on single-word answers. The three evaluation metrics we find relevant to this setting are (1) BLEU-1, which measures unigram precision; (2) Recall@5, which indicates a word's appearance within the first five words generated; and (3) CLIP-score, which indicates semantic relatedness. To calculate the CLIP-score, we first add "Image of" as a prefix to the ground truth. Using CLIP's textual encoder, we then use a cosine distance. More details are provided in the supplementary material.

In Tab. 2, we show performance for each relation. While this task is challenging, our approach resulted in a significant success rate of 30% at R@5 in most relations. Note that, since the benchmark lacks multiple references, it is still limited, e.g., we mark a miss if the generated word is 'US,' while the ground truth is 'USA'. Observing the returned answers reveals that some mistakes are understandable, e.g., answering Sydney instead of Canberra or the Sinai province instead of the country Egypt. However, other cases return truncated sentences, e.g., returning 'flag' instead of a country name or returning general concepts such as "flickr image". See supplementary for a discussion. When employing the softer CLIP-Score metric, which is based on a semantic distance, a correlation of 70% is observed.

We compared our results with ClipCap [51] that encodes the image with CLIP's image encoder and uses it as an initial token for GPT-2. The method is fine-tuned based on COCO dataset. As can be seen, this method fails to retrieve the correct response, despite employing the same large-scale models as we do and performing arithmetic in the same CLIP embedding space. CLIP-VL [64] and the supervised captioning methods cannot be tested on this benchmark since it uses spatial grid features as embedding.

Multi-modal Arithmetic Our method enables multi-modal reasoning, which involves manipulating images and text simultaneously in the same embedding space. Using CLIP's textual encoder, E_{Text}. In Fig. 9, we show that a day-to-night direction can be obtained with text inputs, i.e., "image of a night," and "image of a day." The direction steers an image of breakfast to "Nighttime dinner."

6. Discussion and Limitations

The zero-shot capabilities presented by CLIP [57] pave a new path for computer vision. However, these are limited to multiclass classification. DALL-E [59] presents an impressive ability to generate images that are very different from its training images in what is termed zero-shot generation ability. However, this ability is exactly the generative task DALL-E was trained to do, only in new domains. No previous computer vision work, as far as we can ascertain, has presented a generative semantic zero-shot capability of the sort that is revolutionizing the NLP world with transformers, such as GPT-3 [6]. Our work is the first to present a generative visual-semantic work.

While the ability to rely on pre-trained models such as

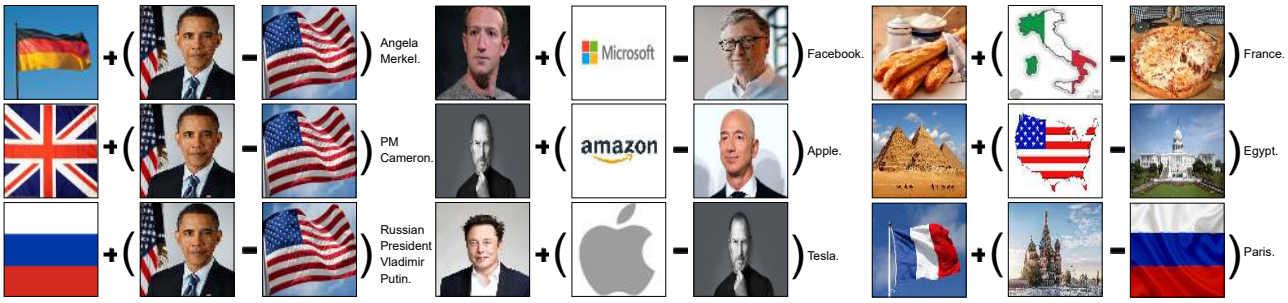


Figure 8. Image arithmetic with both summation and subtraction. For example, on the left side, by removing the American flag from Obama, a leadership direction results. The presidents of different countries are generated when the derived vector is added to their flags.

	Building → Country			**Country → Capital**			**CEO → Company**			**Food → Country**			**Leader → Country**		
Method	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s
ClipCap [51]	0.003	0.035	0.24	0.0	0.0	0.22	0.004	0.05	0.18	0.0	0.0	0.24	0.008	0.24	0.26
Ours	**0.1**	**0.32**	**0.7**	**0.14**	**0.32**	**0.68**	**0.1**	**0.3**	**0.64**	**0.03**	**0.33**	**0.66**	**0.1**	**0.28**	**0.68**

Table 2. Comparison of our method and ClipCap baseline on our novel benchmark for visual relations. B@1 = BLEU-1, R@5 = Recall@5, C-s = CLIP-score.

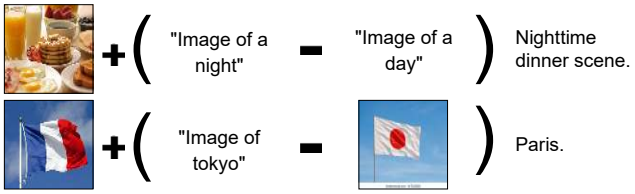


Figure 9. Our method is not only devoted to visual relations, but it also allows arithmetic between image and language.

GPT-2 [58] and CLIP allows us to achieve such new capabilities, they also highlight the uneven playing field AI has become. GPT-2 is far inferior to GPT-3 and other recent LMs in which resources far beyond the reach of most research labs are invested.

On a similar note, it is likely that combining zero-shot with supervised training would lead to a method that outperforms the baselines in all captioning metrics. However, the amount of resources currently used to train supervised captioning methods is becoming a deterring factor from pursuing this direction. For instance, UNITER uses 3645 hours of a V100 GPU [11].

The use of an LM and an image-language matching model trained on large corpora of collected data inevitably leads to biases. For example, the models we employ are clearly oriented towards Western knowledge and can recognize people, places, objects and concepts that are popular in Western media, while being much less knowledgeable about other cultures. For example, our model fails to form relations with the president of China, Xi Jinping.

7. Conclusions

The marriage between a language model and a visual-semantic matching model is a powerful union, with the potential to provide zero-shot captioning that brings together real-world variability in text, recognition abilities that are unrestricted by categories, and real-world knowledge that is embedded in the models through web-scale datasets.

We propose a zero-shot method for combining the two models, which does not involve optimizing over the weights of the models. Instead, we modify, for all layers and attention heads, the key-value pairs of the tokens generated by the language model up to each inference step.

As a captioning model, our method produces results that are less restrictive than those provided by the human annotators on the datasets used by supervised captioning methods. While this lowers the word-to-word metrics, the captions generated seem to be a good match to the image at the semantic level and exhibit real-world information. Moreover, the flexibility of using an embedding-space zero-shot method enables us to perform visual-semantic arithmetic.

We show how we can describe in words the difference between two images and how we can combine concepts from multiple images. Both are novel high-level recognition tasks. Combining these two capabilities, a powerful image analogy machine is obtained, which answers, by providing a text string, questions of the form "A is to B as C is to X" ($X \sim C + B - A$), in which A, B, and C can each be either textual or visual.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research, innovation programme (grant ERC CoG 725974). The contribution of the first author is part of a PhD thesis at Tel Aviv University.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *EMNLP*, 2017. 2

[4] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *CVPR*, 2018. 2

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005. 5

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 7

[7] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *ECCV*, 2018. 2

[8] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. *arXiv preprint arXiv:2110.12427*, 2021. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[10] Yun Chen, Victor OK Li, Kyunghyun Cho, and Samuel R Bowman. A stable and effective learning strategy for trainable greedy decoding. *EMNLP*, 2018. 3

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 8

[12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 2

[13] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *ICLR*, 2020. 3

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019. 2

[16] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*, 2015. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[18] 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. 11

[19] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *TCSVT*, 2020. 2

[20] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, 2019. 2

[21] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *NeurIPS*, 34, 2021. 2

[22] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *NeurIPS*, 2020. 2

[23] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. 4, 6

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 1

[25] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*, 2018. 2

[26] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *EACL*, 2017. 3

[27] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[28] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *CVPR*, 2020. 2

[29] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2

[30] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *NeurIPS*, 2019. 2

[31] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021. 5

[32] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. *AAAI*, 2021. 2

[33] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *ICML*, 2017. 3

[34] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI*, 2021. 2

[35] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019. 3

[36] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *EMNLP*, 2019. 2

[37] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017. 2

[38] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *ICCV*, 2017. 1

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[41] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 2019. 2

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[43] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *CVPR*, 2019. 2

[44] 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. 1, 5, 17

[45] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*. Springer, 2016. 6

[46] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning, corr abs/1612.01887 (2016). *CVPR*, 2017. 2

[47] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 2

[48] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. *AAAI*, 2021. 2

[49] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *CoRR*, abs/1412.6632, 2014. 2

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *NIPS*, 2013. 6

[51] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5, 7, 8, 12, 17

[52] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020. 2

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2001. 5

[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[55] 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. 11

[56] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. 1, 2, 6, 7

[58] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 3, 4, 8

[59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1, 7

[60] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021. 2

[61] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. High-order attention models for visual question answering. *NIPS*, 2017. 2

[62] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019. 2

[63] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *CVPR*, 2019. 2

[64] 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. 1, 2, 5, 7, 12, 17

[65] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *NIPS*, 2017. 3

[66] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 3

[68] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2014. 5

[69] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017. 2

[70] Sahil Verma, Michael Ernst, and Rene Just. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054*, 2021. 2

[71] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. corr abs/1411.4555 (2014). *ICML*, 2015. 2

[72] Liwei Wang, Alexander G Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *NIPS*, 2017. 2

[73] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[74] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 2

[75] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 2

[76] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014. 1

[77] L Yu, W Zhang, J Wang, and Y Yu. Seqgan: Sequence generative adversarial nets with policy gradient. arxiv e-prints, page. *AAAI*, 2017. 3

[78] 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. 1, 2, 5, 12, 17

[79] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3

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

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

B. 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 [55].

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 [18].

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.

C. Ablation Study

Effect of CLIP-based optimization: A further ablation was performed, in which CLIP's score is used directly to

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

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

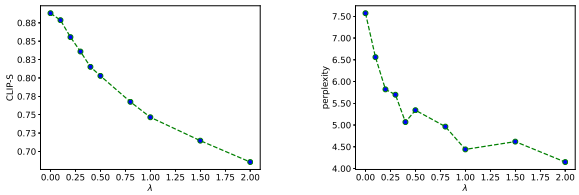
Method	CLIP-S	Perplexity
A1	**0.98**	8.61
A2	0.91	6.04
Ours	0.87	**5.50**

Table 3. 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.

optimize the LM. In Fig. 10, 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. 3), 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.

D. Additional Qualitative Results

Image Captioning: In Fig. 15 (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 [51], CLIP-VL [64], and VinVL [78]. Our method generates original captions that are completely different both in vocabulary and pattern from the baselines' captions.

E. 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. 11; 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. 12, 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. 13, 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.

F. Visual Relations Benchmark Study

Our benchmark combines real-world knowledge with the ability to represent visual relationships. In Appendix F. 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 univer-

sity. 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.

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

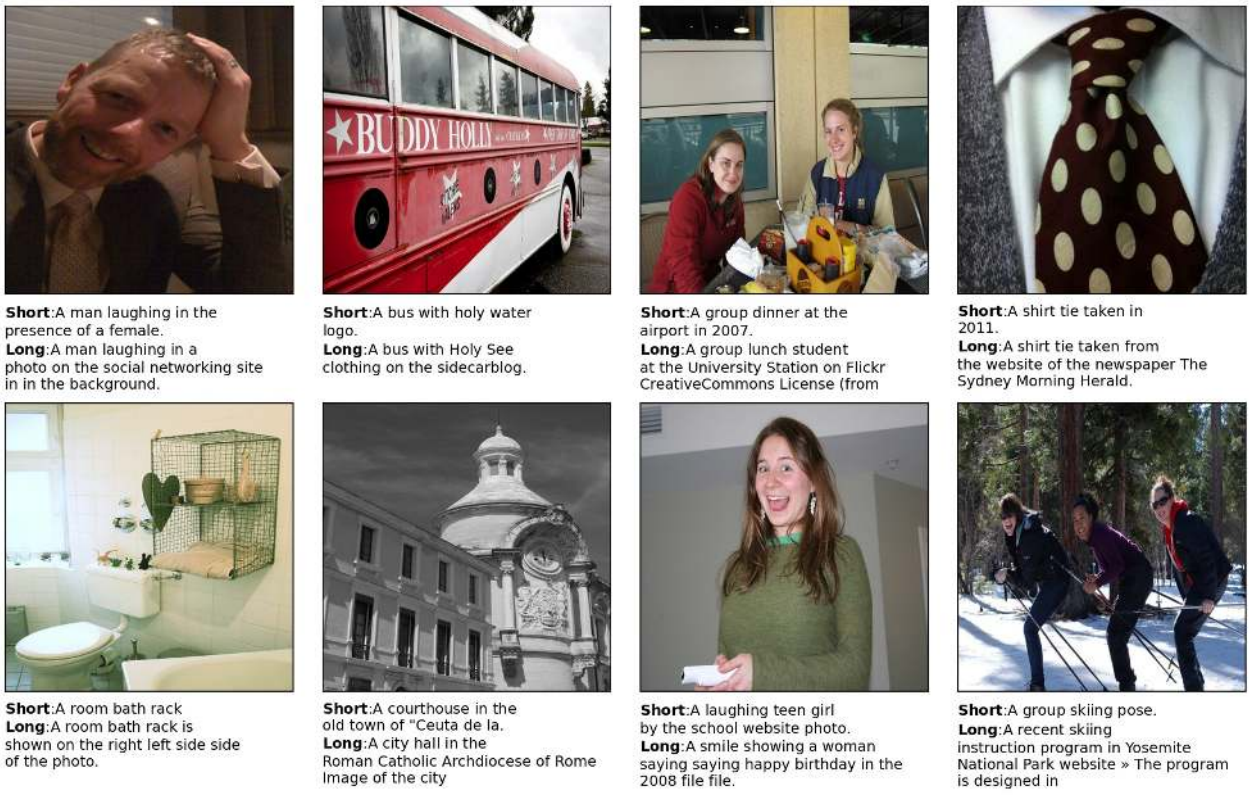
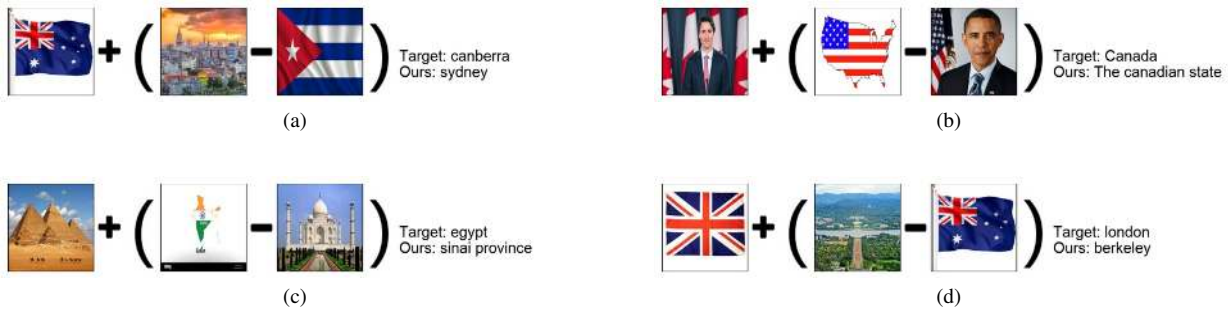


Figure 12. The effect of our length-control mechanism. With the mechanism (Short) and without the mechanism (Long).

Figure 13. Bias cases against distinct groups.



(e) (f)

Figure 14. Error analysis of the visual relations benchmark.

Figure 15. Generated captions by our method and by the baseline methods for images from the MS-COCO [44] test-set. CP=ClipCap [51], CVL=CLIP-VL [64], VVL=VinVL [78].

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.