Teaching CLIP to Count to Ten

Roni Paiss[1,2] Ariel Ephrat[1] Omer Tov[1] Shiran Zada[1]
Inbar Mosseri[1] Michal Irani[1,3] Tali Dekel[1,3]
[1]Google Research [2]Tel Aviv University [3]Weizmann Institute of Science

Our supplementary materials contain the following:

1. Our full CountBench dataset (image urls and captions) is provided as a json file in the supplementary zip file. Sample examples are included in this document.

2. Implementation details and hyperparameter setting of our method, described in Sec. 3 in the paper.

3. Extended results of our count-based image retrieval experiment, reported in Sec. 5.2 in the paper.

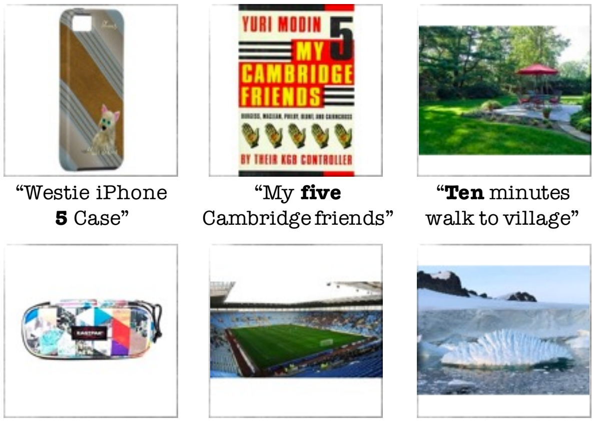
4. General (non-counting) image retrieval results, extending our analysis in Sec. 5.3 in the paper.

5. Additional details about our count-aware image generation, described in Sec. 5.5 in the paper, including the implementation details, evaluation protocol and additional results.

1. CountBench benchmark

Our full CountBench dataset (image urls and captions) is provided as a json file in the supplementary zip file. Sample examples for each number (“two'' – “ten'') are shown in Figs. 6 to 14. As seen, the images vary in resolution and aspect ratios, and the captions vary in length. In order to evaluate the diversity of the in-the-wild captions in CountBench we (a) manually extracted the words describing the counted object in each caption, then (b) automatically associated these words with categories from MS-COCO dataset, according to the cosine distance in CLIP text embedding space. Table 1 reports counting accuracy for the 20 most prevalent classes (81% of the benchmark). Our model consistently outperforms the base models by a large margin, across all categories.

In addition, Fig. 1 presents examples of samples filtered out by our filtering pipeline. While the captions indeed specify numbers, the numbers relate to dates, addresses etc. rather than counts of objects and are therefore not suitable as counting data.



“Westie iPhone **5** Case”

“My **five** Cambridge friends”

“**Ten** minutes walk to village”

“Eastpak Estuche ovalitri, **5** x 22 x **9** cm”

“Coventry have been deducted **ten** points”

“Antarctica is now melting **three** times faster than ever"

Figure 1. **Examples of image captions where the numbers are NOT related to object counts. These are automatically filtered-out by our method.** In all above examples the numbers indicated in the caption do not refer to an actual object count. Numbers often specify measures, versions, dates, time, written numbers in the image, or numbers that refer to things not visible in the image.

2. Implementation details

Models. We tested our method with two classes of state-of-the-art VLMs: BASIC [5] and CLIP [6], in order to verify its robustness to different architectures (Sec. 5.1 in the paper). For CLIP, we experiment with both CLIP-B/32 and CLIP-L/14 configurations, as they are both widely used in recent work. For BASIC, we experiment with BASIC-S.

Training. We finetune all models for $20K$ steps using a cosine schedule with an initial learning rate of $5e^{-6}$ and a batch size of 32,768. Our method introduces two additional hyperparameters: the portion $p \in [0, 1]$ of the batch size dedicated to the counting subset, and the weight λ of our counting loss L_{count}. We empirically chose $p = \frac{1}{32}$ and $\lambda = 1$, based on our ablation below. We deploy a linear warm-up on λ for the first 10,000 steps.

Class	person	book	train	clock	bird	chair	pizza	cap	wine glass	dog	bottle	cake	knife	sheep	bowl	apple	cat	sports ball	bear	car
# samples	117	70	28	27	26	24	16	16	14	13	13	12	9	9	8	8	8	7	7	7
Baseline CLIP	0.38	0.17	0.25	0.41	0.42	0.25	0.12	0.44	0.29	0.23	0.31	0.25	0.11	0.44	0.12	0.12	0.12	0.57	0.57	0.29
Ours	**0.70**	**0.73**	**0.82**	**0.77**	**0.85**	**0.87**	**0.75**	**0.56**	**0.86**	**0.86**	**0.69**	**0.75**	**0.77**	**0.88**	**0.75**	**0.62**	**0.62**	**0.86**	**0.86**	**0.57**

Table 1. **Counting accuracy on CountBench for prominent categories** *The samples in CountBench are divided into the categories from MS-COCO. As can be seen, our model performs well across all categories.*

Dataset	$p = \frac{1}{32}$	$p = \frac{1}{8}$	$p = \frac{1}{4}$
CountBench	**75.93**	70.19	69.81
ImageNet	64.06	64.11	63.96
CIFAR10	60.65	61.69	63.04
CIFAR100	33.56	33.74	34.01
Caltech101	82.36	83.58	83.51
EuroSAT	37.69	39.07	41.56
Food101	80.53	80.59	80.80
ImageNetA	29.81	30.84	30.60
ImageNetR	70.30	70.15	69.98
ImageNetV2	56.62	56.54	56.37
Oxford Pets	87.41	87.14	86.64
Oxford Flowers	67.39	67.21	67.91

Table 2. **Ablation of hyperparameter p.** *p denotes the fraction of the batch size dedicated to samples from the counting subset. As the subset is significantly smaller than the entire curated dataset we found that large values for p lead to overfitting.*

Hyperparameters. We empirically set our hyperparameters $p \in [0, 1]$ and λ used in our method, by comparing the performance of models trained with different weightings on CountBench. As shown in Tab. 2, large portions p result in lower accuracy due to overfitting to the counting training set, which is relatively small compared to the the general image-text training set. Therefore, we set $p = \frac{1}{32}$. Table 3 reports accuracy on CountBench for different choices of the weight λ of the counting-loss. As can be seen, setting $\lambda = 1$ results in the highest counting accuracy. A lower value of $\lambda = 0.1$ results in lower counting accuracy, implying that it does not sufficiently incentivize the model to regard the object counts. Larger values (i.e $\lambda = 5$, $\lambda = 10$) seem to cause overfitting to the objects in the training data.

3. Additional count-based retrieval results

We provide extended qualitative results for the count-based image retrieval task, described in Sec 5.2 in the paper. Figure 2 presents the top-5 retrieved images using the original CLIP model and our counting-aware CLIP model for prompts that specify the number of objects. As can be seen, as the numbers grow larger, the baseline often retrieves images with arbitrary numbers of objects, and tends to retrieve the same images for several different requested numbers. This further demonstrates that the baseline model mostly focuses on the existence of the requested object in the image, rather than their count. In contrast, our counting-aware model retrieves images that depict accurate object counts in most cases.

Dataset	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$
CountBench	69.44	**75.93**	73.15	72.59
ImageNet	64.50	64.06	63.84	63.53
CIFAR10	63.20	60.65	63.79	63.82
CIFAR100	34.51	33.56	35.35	34.15
Caltech101	84.39	82.36	81.82	81.76
EuroSAT	39.48	37.69	39.93	42.20
Food101	80.73	80.53	80.33	79.98
ImageNetA	31.67	29.81	29.55	29.45
ImageNetR	70.92	70.30	69.87	69.77
ImageNetV2	56.70	56.62	56.30	56.09
Oxford Pets	87.65	87.41	87.79	86.97
Oxford Flowers	67.00	67.39	65.33	65.90

Table 3. **Ablation of the auxilary loss weight λ** *We ablate different weights for the auxilary loss. We found $\lambda = 1$ to work best, as lower values lead to suboptimal results and higher values cause overfitting.*

	$Image \rightarrow Text$			$Text \rightarrow Image$		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	53.0	78.4	85.6	38.6	65.0	74.9
Ours	52.3	77.1	85.1	37.3	64.5	73.4

Table 4. **General (non-counting) text-based image retrieval on *COCO* dataset.** *"$R@N$" denotes Top-N retrieval accuracy (whether the retrieved image/caption corresponds to its query caption/image, respectively). Retrieval results of our counting-aware CLIP on general (non-counting) tasks are on par with the CLIP baseline.*

4. Zero-shot non-counting image retrieval

A key property of our method is that it preserves the original non-counting capabilities of the model, as demonstrated through our performance in a various zero-shot classification tasks (Sec 5.3 in the paper). To further validate our performance on non-counting tasks, we evaluate our method on general text-based image retrieval on the COCO dataset [4]. Table 4 reports the results for the baseline CLIP

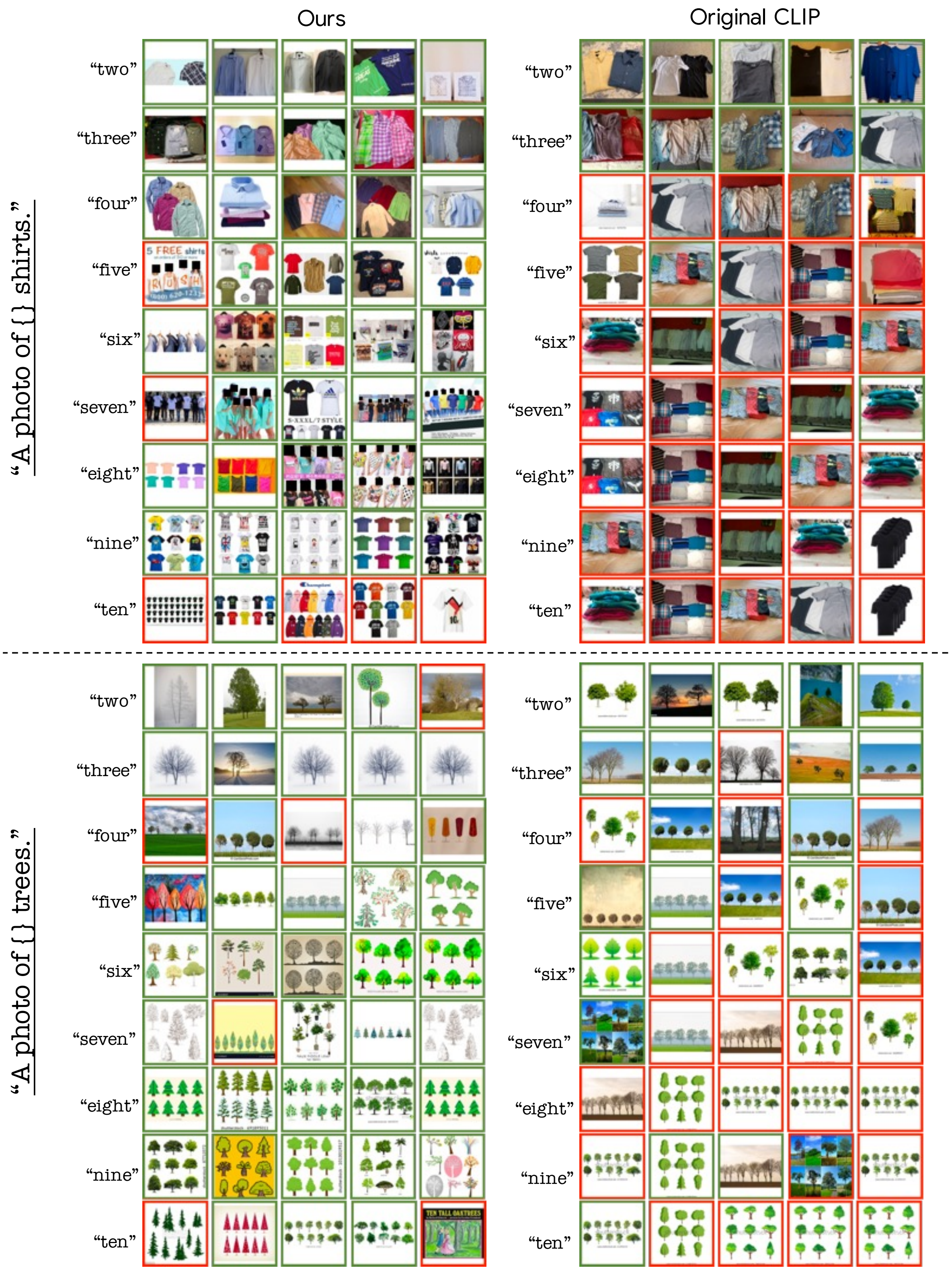


Figure 2. **Top-5 count-based retrieval.** *Text-to-image retrieval results for different counts of objects; retrieved images with correct count are marked with a green border, while images with incorrect counts are marked in red. The images in each row are ordered according to their similarity scores (descending scores from left to right). The images retrieved by our model are significantly more accurate than the original CLIP model, especially for counts higher than three.*

and our counting-aware CLIP. As can be seen, our counting-aware CLIP performs on par with the CLIP baseline, which further confirms the preservation of CLIP's non-numeric capabilities.

5. Count-Aware Image Generation

We next provide further details and additional evaluation for the task of text-conditioned image generation (Sec. 5.5 in the paper).

Training. We train Imagen [7] models from scratch for $500K$ steps with a batch size of 512 on 64 TPUv4 chips. We employ the Adam [3] optimizer with a cosine learning rate schedule where the peak learning rate is $1e{-}4$, as in [7]. We replace the central-cropping used in [7] with padding to prevent mismatch between the image content and the count indicated in the caption. We set 3% of each batch to contain samples from our counting set. As the number of objects in the image determined by the 64×64 model, we do not train the 256×256 and 1024×1024 super-resolution models, as they do not affect the number of objects in the generated image.

Evaluation. We extend the evaluation reported in Sec. 5.5 in the paper to evaluate the model on captions with larger numbers of objects. We construct an additional set of prompts by creating all possible combinations of "$\{number\}$ $\{label\}$" where $number \in \{$"two", .., "ten"$\}$ and $label$ is one of the class labels of CIFAR-10 dataset [1]. This process, which is illustrated in Fig. 4, results in 90 distinct text prompts.

For each text prompt, we generate 12 images using a DDPM sampler [2] with different random seeds, resulting in a total of 1296 images. We manually count the number of instances of the requested object contained in each generated image, and compare it to the number specified in the prompt.

Results. Table 5 reports results for the full set of prompts (including the one form DrawBench [7] used in the paper). In addition to counting accuracy, we also report the mean deviation of the predicted number from the correct number: $MAE = \frac{1}{N}\sum_{i=1}^{N}|gt_i - pred_i|$, where N is the number of prompts we use for evaluation, and for each sample i, gt_i is the number specified in the caption and $pred_i$ is the number of requested objects in the generated image. While accuracy tells how often the models are wrong in the number of objects they generate, the MAE metric quantify *how wrong* they are.

	prompts from DrawBench		CIFAR-10 class labels	
	Accuracy ↑	**MAE ↓**	**Accuracy ↑**	**MAE ↓**
Baseline CLIP	24.12	0.94	20.00	3.32
Ours	**40.35**	**0.81**	**50.18**	**1.09**

Table 5. **Text-conditioned image generation evaluation.** *We compare an Imagen model conditioned on the baseline CLIP against a model trained with our counting-aware CLIP. We evaluate the models on prompts from the DrawBench counting category and prompts created from CIFAR-10, as described in Sec. 5. We report both accuracy and MAE. As can be seen, our model is significantly superior to the baseline in both metrics.*

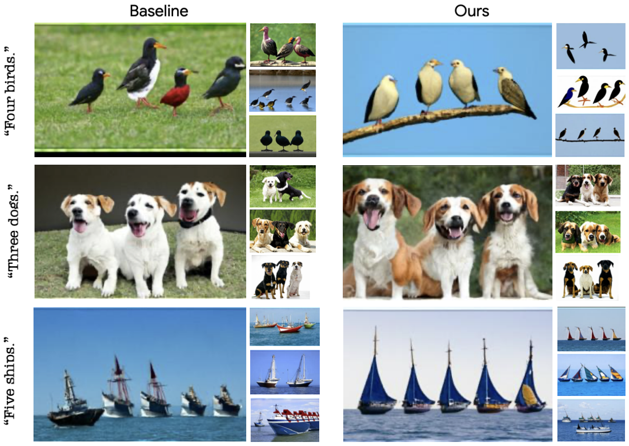


Figure 3. **Qualitative comparison of generated images.** *We compare an Imagen model trained with the original CLIP model against an Imagen model trained with our counting-aware CLIP. The images are generated using textual prompts created from CIFAR-10 class names (as described in Sec. 5).*

As can be seen in Tab. 5, the counting accuracy of the Imagen model trained with our counting-aware CLIP is around 2× better than the results of the Imagen model trained with the baseline CLIP, indicating that is betters follows the number requested in the caption. Our method also achieves lower MAE, indicating that when the model generates the wrong number of objects it still comes much closer to the desired number than the baseline.

We also provide additional qualitative results. Figure 3 shows a qualitative comparison between images generated with our method and the baseline. As can be observed, while the baseline model occasionally generates the correct number of objects, our method produces specific counts of objects more reliably. Fig. 5 presents additional images generated with the Imagen model trained with our counting-aware CLIP for prompts that specify the number of objects.

References

[1] Raveen Doon, Tarun Kumar Rawat, and Shweta Gautam. Cifar-10 classification using deep convolutional neural network. *2018 IEEE Punecon*, pages 1–5, 2018. 4

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information*

Processing Systems, 33:6840–6851, 2020. 4

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 4

[4] Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014. 2

[5] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined Scaling for Open-Vocabulary Image Classification. *arXiv preprint arXiv:2111.10050*, Nov. 2021. 1

[6] 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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 4

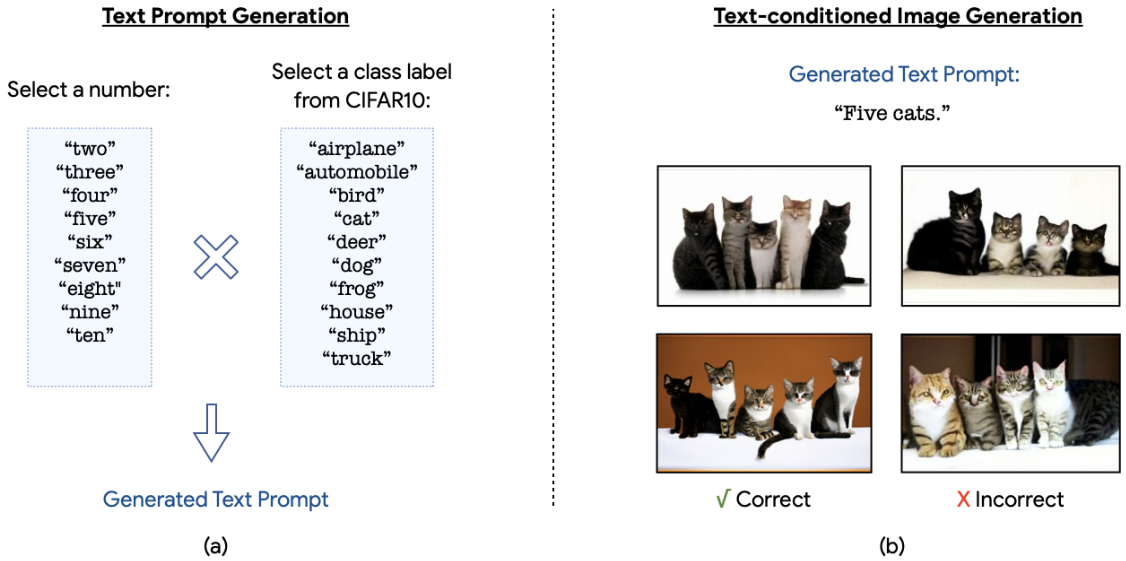


Figure 4. **An overview of the prompt generation pipeline.** *As detailed in Sec. 5, we create a set of captions containing the numbers "two", .., "ten" and the class labels from CIFAR-10. (a) Each combination of number and class label is used to create a text prompt (b) We use the Imagen models to generate images based on the text prompt and measure accuracy and MAE.*

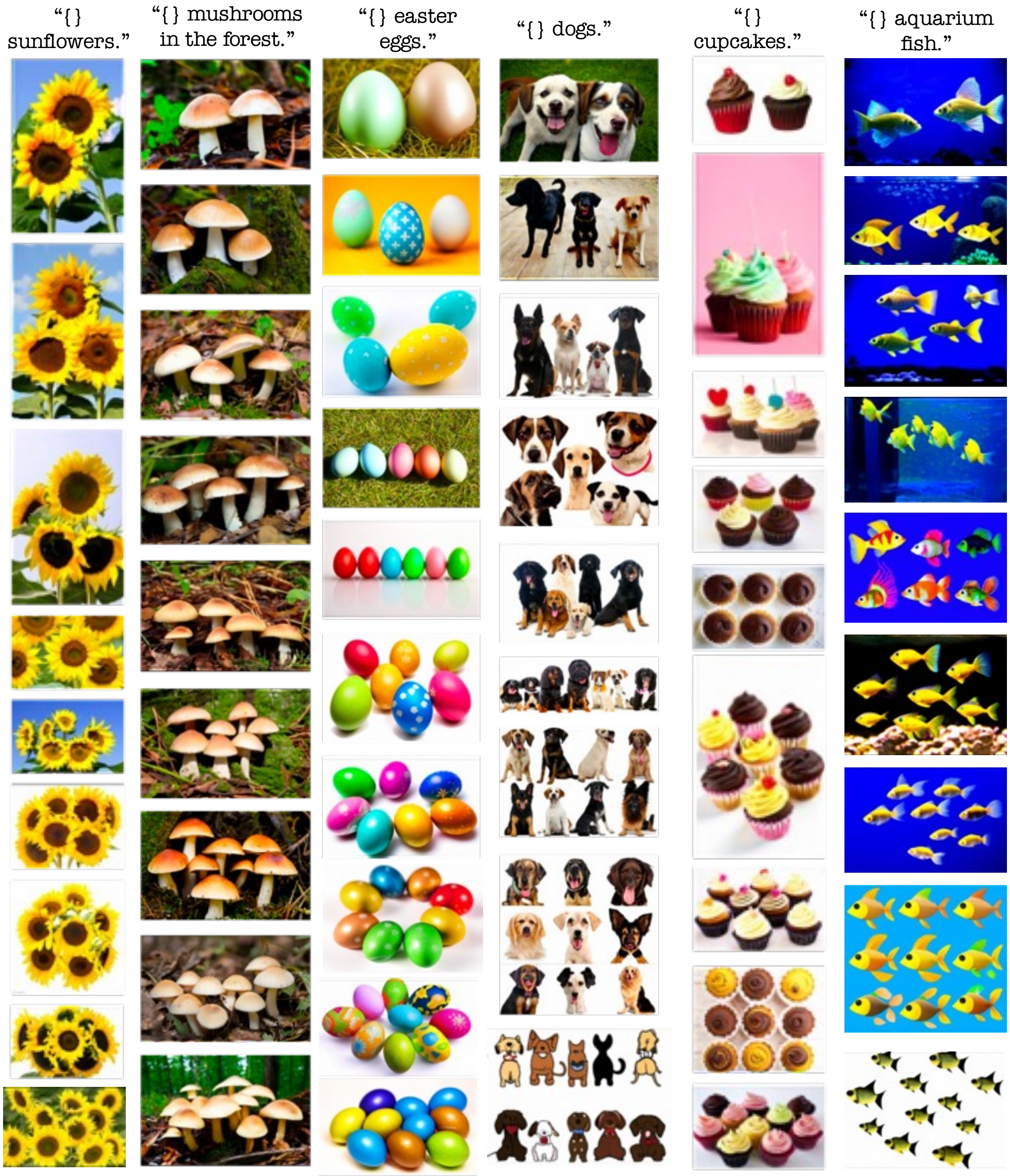


Figure 5. **Images generated with the Imagen model trained with our counting-aware CLIP**. *For each of the caption templates at the top we inject numbers between “two” and “ten”. The images generated conditioned on these prompts are ordered according to the injected number, such that the top-most images contain two objects and the bottom images contain ten objects.*

“A custom cabinet between two pedestal sinks means you can have the style of a pedestal along with the function of a vanity”

“A well furnished bedroom with two double beds a television and balcony.”

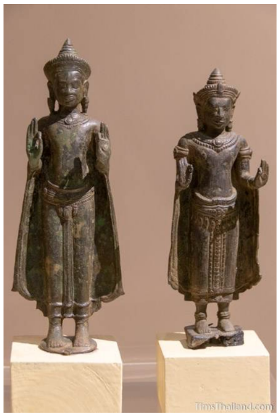
“Still life with bottle of red wine, two wineglasses and grape in”

“two red pingpong rackets on white surface table tennis zoom background”

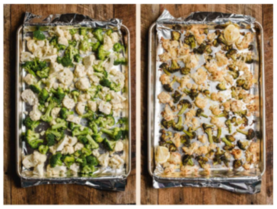
“set of two eames rar chairs black. Black Bedroom Furniture Sets. Home Design Ideas”

“set of two glass star christmas tree decorations amazoncouk kitchen home”

“Dog Leash Coupler - Walk two dogs with a single leash”

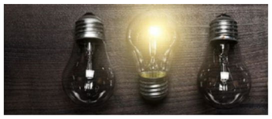


“two brass crowned Buddhas”



“two baking sheets of broccoli and cauliflower florets: one raw, one baked”

Figure 6. **Sampled images from CountBench labeled as “two”.**



“background photo of three light bulbs”

LONDON
PARIS
New York

“City prints: Set of three big prints - $150.00 USD”

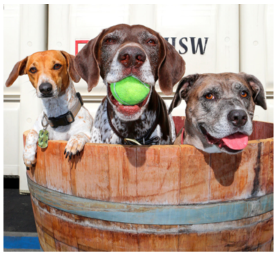
“Three little pigs - cute pig - three pigs paper plate”

“three men new orleans by jules pascin”

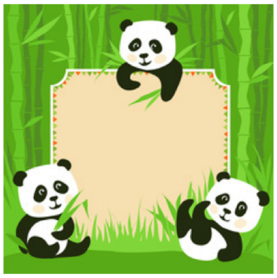
“Mother Orsa (L) and her three young bears discover the open-air enclosure at the wildlife park Tripsdrill, near Cleebronn, Germany, April 10.”

“three candles”

“Choice of three doors opening to possible vacation or getaway destinations Imagens”

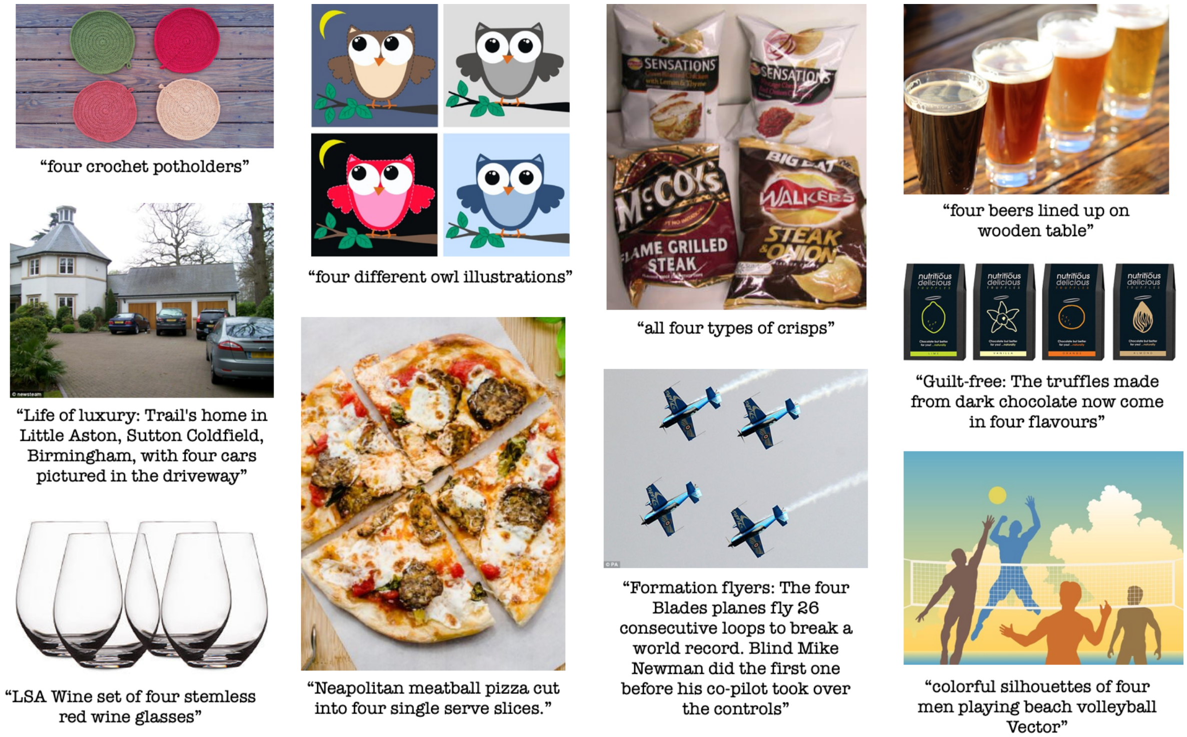


“three dogs in a wine barrel”



“Cartoon frame - bamboo & three little pandas illustration”

Figure 7. **Sampled images from CountBench labeled as “three”.**



"four crochet potholders"

"Life of luxury: Trail's home in Little Aston, Sutton Coldfield, Birmingham, with four cars pictured in the driveway"

"LSA Wine set of four stemless red wine glasses"

"four different owl illustrations"

"Neapolitan meatball pizza cut into four single serve slices."

"all four types of crisps"

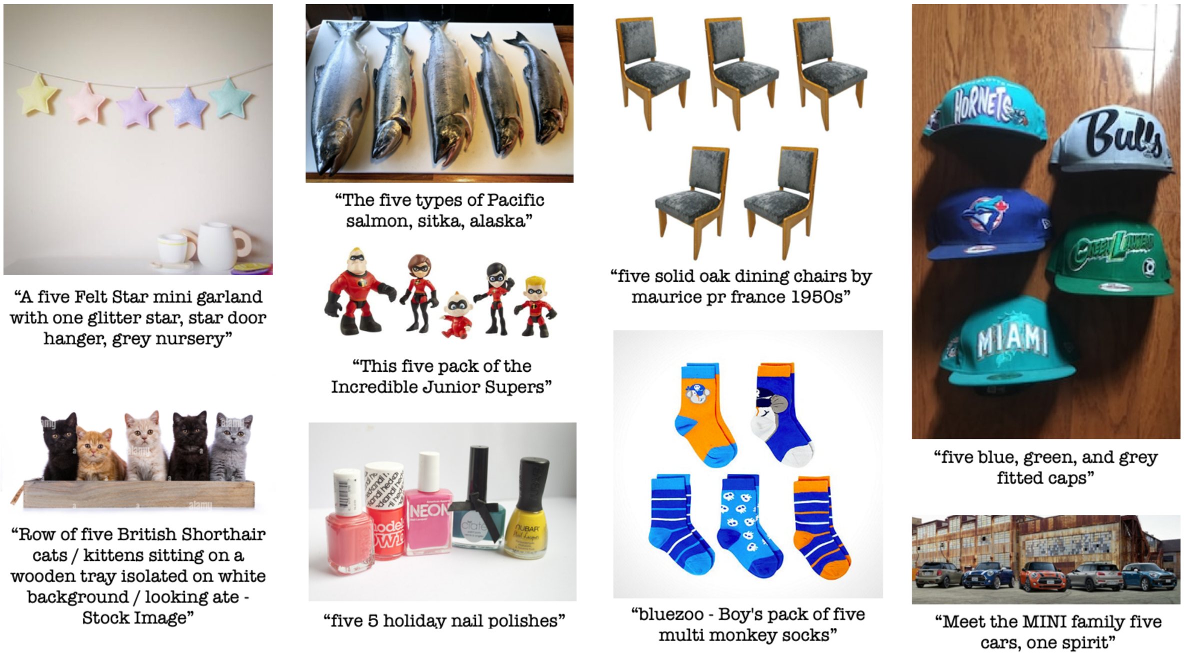
"Formation flyers: The four Blades planes fly 26 consecutive loops to break a world record. Blind Mike Newman did the first one before his co-pilot took over the controls"

"four beers lined up on wooden table"

"Guilt-free: The truffles made from dark chocolate now come in four flavours"

"colorful silhouettes of four men playing beach volleyball Vector"

Figure 8. **Sampled images from CountBench labeled as "four".**



"A five Felt Star mini garland with one glitter star, star door hanger, grey nursery"

"Row of five British Shorthair cats / kittens sitting on a wooden tray isolated on white background / looking ate - Stock Image"

"The five types of Pacific salmon, sitka, alaska"

"This five pack of the Incredible Junior Supers"

"five 5 holiday nail polishes"

"five solid oak dining chairs by maurice pr france 1950s"

"bluezoo - Boy's pack of five multi monkey socks"

"five blue, green, and grey fitted caps"

"Meet the MINI family five cars, one spirit"

Figure 9. **Sampled images from CountBench labeled as "five".**

“vintage silver plate tablespoons, serving spoon set of six 1847 Rogers Ambassador pattern”

“Moller #71 Chair. Set of six dining chairs in Rosewood.”

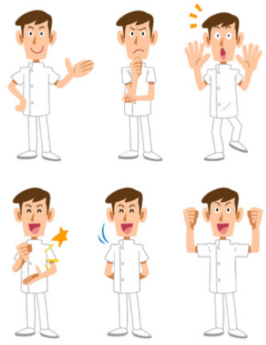
“six cute kittens sitting inside”

“Set of six flowers in pots blue background card vector”

“slam dunk collection with cool six pose”

“six solid yellow salad plates, vintage ceramic Harlequin Homer Laughlin china.”

“Pack of six LED battery operated tea lights/candles”



“Nurse male six types of poses and facial expressions of the white coat”

“Gallery wall idea with six framed pictures arranged on a wall depicting Nature, Animals, and Country Life”

Figure 10. **Sampled images from CountBench labeled as “six”.**

“The essential oil blends for all seven chakras from the chakra alignment therapy workshop, The 49 Professions of Joy,"" by personal trainer Jack Kirven.”

“Photo of seven pubs in Paddington”

“Line up of seven different styles of Bean Boots.”

“Lot of seven Brazil (Rio mint) copper coins of Joao VI: XX reis, 1821-R (two), 1822-R (two); X reis,”

“The seven bags, filled with the Frankfurt artists artworks”

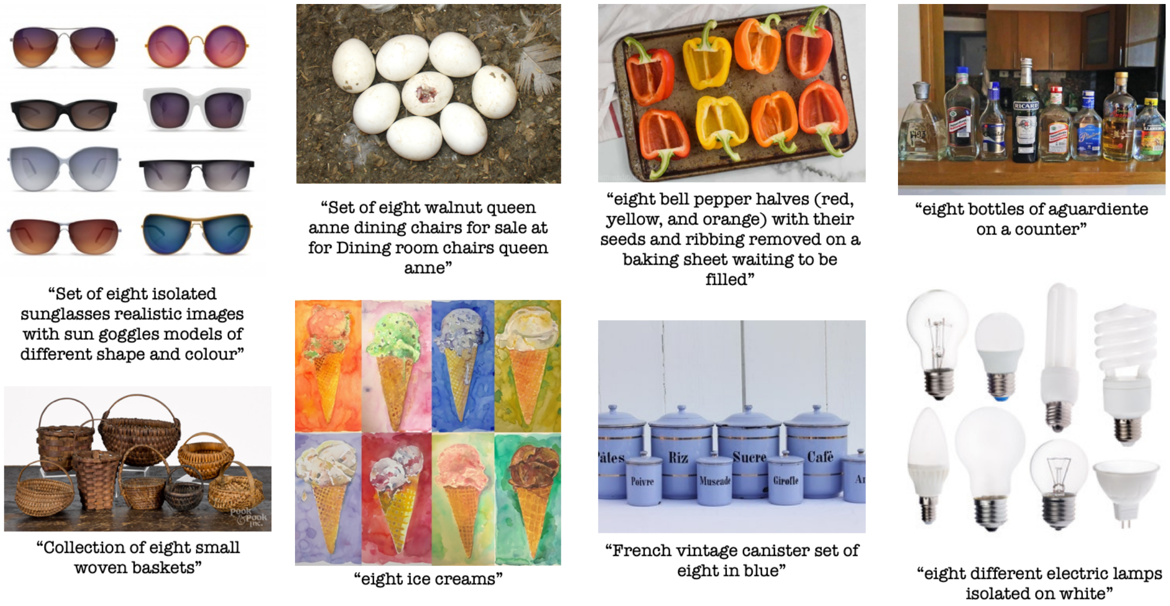
“Overhead shot of seven Chewy Peanut Butter Cookies with Chocolate M&M's on a white plate.”

“A set of seven red plastic apples on a white background”

“Artist's paintbrushes holding all seven colors of the rainbow - red, orange, yellow, green, blue, indigo and violet - stock photo”

“Set of seven vintage retro beer labels with sample text vectorkunst illustratie”

Figure 11. **Sampled images from CountBench labeled as “seven”.**



"Set of eight isolated sunglasses realistic images with sun goggles models of different shape and colour"

"Collection of eight small woven baskets"

"Set of eight walnut queen anne dining chairs for sale at for Dining room chairs queen anne"

"eight ice creams"

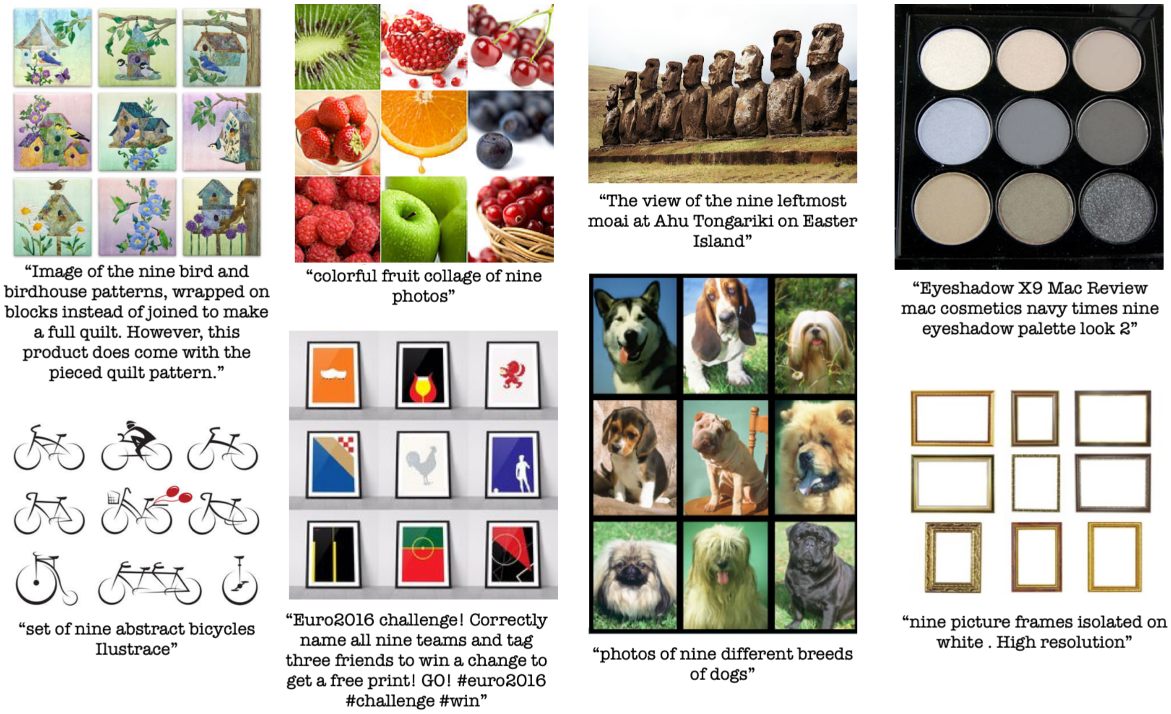
"eight bell pepper halves (red, yellow, and orange) with their seeds and ribbing removed on a baking sheet waiting to be filled"

"French vintage canister set of eight in blue"

"eight bottles of aguardiente on a counter"

"eight different electric lamps isolated on white"

Figure 12. **Sampled images from CountBench labeled as "eight".**



"Image of the nine bird and birdhouse patterns, wrapped on blocks instead of joined to make a full quilt. However, this product does come with the pieced quilt pattern."

"set of nine abstract bicycles Ilustrace"

"colorful fruit collage of nine photos"

"Euro2016 challenge! Correctly name all nine teams and tag three friends to win a change to get a free print! GO! #euro2016 #challenge #win"

"The view of the nine leftmost moai at Ahu Tongariki on Easter Island"

"photos of nine different breeds of dogs"

"Eyeshadow X9 Mac Review mac cosmetics navy times nine eyeshadow palette look 2"

"nine picture frames isolated on white . High resolution"

Figure 13. **Sampled images from CountBench labeled as "nine".**

"We review the ten best gaming headsets in the market"

"Photo Set of ten giraffe portraits, isolated on white background"

"All ten Christmas Poinsettia Kanzashi"

"Ten science fiction paperbacks for ten bucks-small"

"Top ten best fall boots"

"d10 set of ten - Black"

"A group of ten dollhouse needlepoint firescreens"

"ten white and brown chicken eggs in a carton box"

"d10 set of ten - Black"

"Photograph by Jane Mucklow of ten greetings cards of Otford, Kent"

Figure 14. **Sampled images from CountBench labeled as "ten".**