Appendix

A. Additional details regarding our approach

A.1. Pseudocode for classifier-free guidance

Below, we provide pseudocode for greedy decoding with classifier-free guidance. Note that, in practice, we perform decoding in batches.

```
# captioner: Captioning model (returns token log probs)
# img_embed: Image embedding
# gamma: Classifier-free guidance scale
# max_length: Maximum number of tokens in caption
# BOS: Beginning of sequence token
# EOS: End of sequence token

tokens = [BOS]
for i in range(0, max_length):
  # Eq. 3 (without the softmax, since it does not affect the argmax).
  cond_log_probs = captioner(tokens, img_embed)
  uncond_log_probs = captioner(tokens, zeros_like(img_embed))
  scores = uncond_log_probs + gamma * (cond_log_probs - uncond_log_probs)

  # Greedily take the next token.
  next_token = argmax(scores)
  tokens.append(next_token)
  if next_token == EOS: break
```

A.2. Derivation of language model guidance

Assume that we have two joint distributions of captions x and images y, $p(\mathrm{x},\mathrm{y})$ and $q(\mathrm{x},\mathrm{y})$, and these distributions have the same pointwise mutual information between any image-caption pair, i.e. $\log \frac{q(\mathrm{x},\mathrm{y})}{q(\mathrm{x})q(\mathrm{y})} = \log \frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})}$, and thus $\frac{q(\mathrm{x},\mathrm{y})}{q(\mathrm{x})q(\mathrm{y})} = \frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})}$. Starting with the leftmost expression from Eq. 2, there exists an expression that uses the joint distribution from p but only marginals of captions from q,

$$q(\mathrm{x}) \left(\frac{q(\mathrm{x}|\mathrm{y})}{q(\mathrm{x})}\right)^{\gamma} = q(\mathrm{x}) \left(\frac{q(\mathrm{x},\mathrm{y})}{q(\mathrm{x})q(\mathrm{y})}\right)^{\gamma} \tag{5}$$

$$= q(\mathrm{x}) \left(\frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})}\right)^{\gamma}. \tag{6}$$

In Eq. 4, we further decouple the exponents for the numerator and denominator of the above equation. As we note, this decoupling is reminiscent of pmi^k. To see this relationship, first note that $\frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})}$ is the exponential of $\mathrm{pmi}(\mathrm{x},\mathrm{y}) = \log \frac{p(\mathrm{x},\mathrm{y})}{p(\mathrm{x})p(\mathrm{y})}$. Replacing $\mathrm{pmi}(\mathrm{x},\mathrm{y})$ with $\mathrm{pmi}^k(\mathrm{x},\mathrm{y}) = \log \frac{p(x,y)^k}{p(x)p(y)}$, Eq. 6 becomes $q(\mathrm{x}) \left(\frac{p(\mathrm{x},\mathrm{y})^k}{p(\mathrm{x})p(\mathrm{y})}\right)^{\gamma}$. Setting $\alpha = k\gamma$ and $\beta = \gamma$ gives

$$q(\mathrm{x}) \left(\frac{p(\mathrm{x},\mathrm{y})^{\alpha}}{p(\mathrm{x})^{\beta}p(\mathrm{y})^{\beta}}\right) = q(\mathrm{x}) \left(\frac{p(\mathrm{x}|\mathrm{y})^{\alpha}}{p(\mathrm{x})^{\beta}p(\mathrm{y})^{\beta-\alpha}}\right) \propto q(\mathrm{x}) \left(\frac{p(\mathrm{x}|\mathrm{y})^{\alpha}}{p(\mathrm{x})^{\beta}}\right), \tag{7}$$

where the proportionality holds because $p(\mathrm{y})$ is fixed.

A.3. Pseudocode for language model guidance

```
# captioner: Captioning model (returns token log probs)
# lm: Language model (returns token log probs)
# prompt_tokens: Tokenized prompt for language model
# img_embed: Image embedding
# alpha, beta: Cond/uncond exponents from Eq. 4
# max_length: Maximum number of tokens in caption
# BOS: Beginning of sequence token
```

```
# EOS: End of sequence token
# NEWLINE: Newline token

tokens = [BOS]
for i in range(0, max_length):
  # Log of Eq. 4.
  lm_log_probs = lm(concat(prompt_tokens, tokens))
  cond_log_probs = captioner(tokens, img_embed)
  uncond_log_probs = captioner(tokens, zeros_like(img_embed))
  scores = lm_log_probs + alpha * cond_log_probs - beta * uncond_log_probs

  # Transfer probability mass from NEWLINE to EOS.
  scores[EOS] = logsumexp([scores[EOS], scores[NEWLINE]])
  scores[NEWLINE] = -inf

  # Greedily take the next token.
  next_token = argmax(scores)
  tokens.append(next_token)
  if next_token == EOS: break
```

A.4. Manually written prompts

Below, we include the manually written prompts that we use in our language model guidance experiments. Each caption is separated by two newlines.

A.4.1 Descriptive caption prompt

```
a bathroom with goldenrod circular patterned tiles contains a toilet bidet sink mirror
    tissue dispenser and hairdryer\n
donuts being sorted on the conveyor belt of a device labeled donut robot in an industrial
    kitchen\n
a green glass mug containing 3 toothbrushes and 1 tube of toothpaste sitting on a windowsill
    \n
a man wearing sunglasses and a gray shirt poses with a woman wearing a white shirt next to a
     giraffe with a fence behind them\n
a snow covered wooden bench in front of a fence with snow covered evergreen plants behind it
    \n
two white horses pull a plow with a man in a white shirt and cyan cap and a man in a red
    shirt with sunglasses behind them next to a fence under a sky with cumulus clouds\n
a man in a blue shirt and a small child in a red striped shirt play frisbee next to trees in
     a park\n
a black clock tower with a lit up white clock face with roman numerals in front of a
    dilapidated five story warehouse after dusk\n
a decorative pool flanked by palm trees in front of a stone clock tower next to a large ten
    story building with a bright advertisement on top in a city at night\n
cows with gray bodies and white heads eating grass on a hill with a foggy mountain in the
    background\n
```

A.4.2 Counting prompt

```
a photo of four clouds\n
a photo of one cat\n
a photo of three horses\n
a photo of seven candles\n
a photo of sixteen keys\n
a photo of one rat\n
a photo of five carrot sticks\n
```

```
a photo of one turtle\n
a photo of two boats\n
a photo of one orange\n
a photo of nine books\n
a photo of ten fingers\n
a photo of twelve eggs\n
a photo of one microwave\n
a photo of two children\n
a photo of six leaves\n
a photo of two monitors\n
a photo of one toilet\n
a photo of one house\n
a photo of five pairs of pants\n
a photo of eight apples\n
a photo of eleven stars\n
a photo of one hat\n
a photo of two chairs\n
a photo of seven coins\n
a photo of three birds\n
```

A.5. Difference between attention pooling and bottleneck CoCa architecture

Yu et al. [47] perform attentional pooling over the token representations of the image encoder and pass the resulting tokens into the multimodal text decoder (Figure A.1 left). By contrast, our bottleneck architecture uses the same embedding for the contrastive loss and multimodal text decoder (Figure A.1 right). We create this bottleneck because a goal of our work is to invert contrastive embeddings, producing a caption that lies close to the contrastive image embedding when it is embedded by the text encoder. As we show below in Appendix B.1, this bottleneck is not necessary for CFG to yield improvements. The attention pooling architecture is equally compatible with our approach and yields slightly better performance.

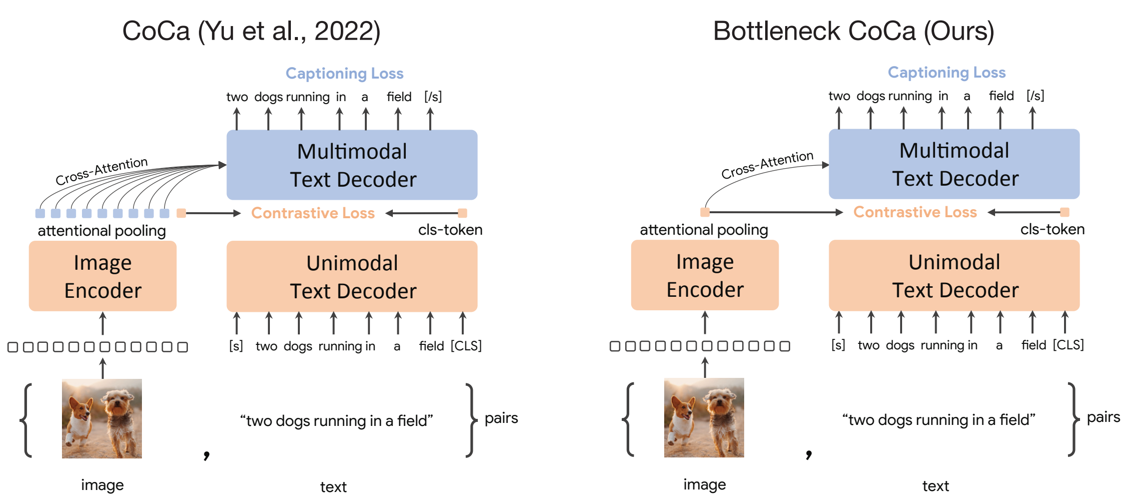


Figure A.1. Comparison of CoCa architecture introduced by Yu et al. [47] (left) with our bottleneck CoCa architecture (right).

B. Additional experimental results

B.1. Attention pooling CoCa architecture

Classifier-free guidance yields similar qualitative results (and slightly better quantiative results) when using the standard CoCa architecture with attention pooling (Figure A.1 left) rather than the bottleneck architecture used in the main text (Figure A.1 right). We fine-tune CoCa-Base for 20,000 steps with a max learning rate of 1×10^{-5} and a conditioning masking proportion of 0.5, following the same procedure that gave the near-optimal bottleneck model described in Section 3.3. Figure B.1 plots reference-based metrics on the x-axis and reference-free metrics on the y-axis, showing a similar trade-off to Figure 2. Table B.1 provides quantitative results demonstrating that the attention pooling architecture performs slightly better across both reference-based and reference-free evaluations. Nonetheless, we adopt the bottleneck architecture for our main experiments for the reasons described in Appendix A.5 above.

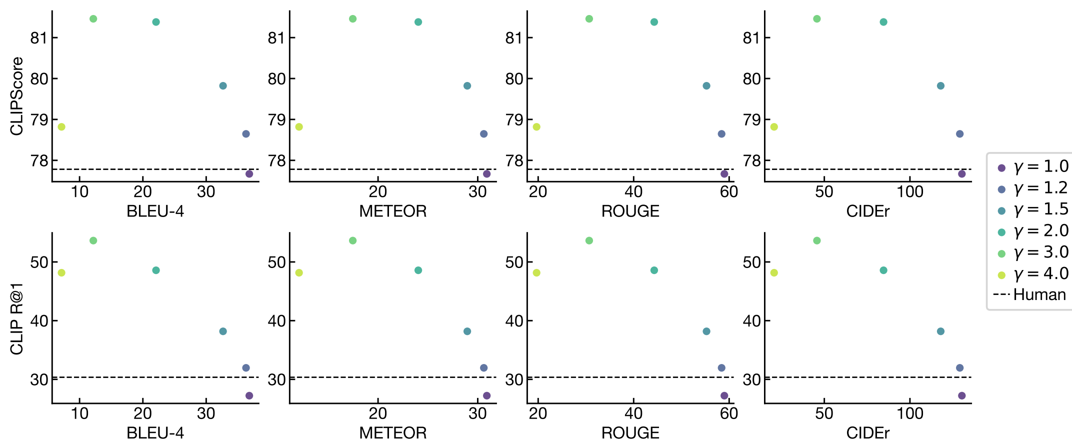


Figure B.1. Effect of classifier-free guidance on captioning metrics with the attention pooling CoCa model. All points reflect the same fine-tuned model; each color represents a γ value used to decode. Models are evaluated with different guidance scales γ, using reference-free captioning metrics based on CLIP ViT-B/32 (y-axes; top: CLIPScore, bottom: recall@1) and reference-based captioning metrics (x-axes). The dashed line reflects the value of the reference-free captioning metric for the ground-truth captions obtained from MS-COCO. See Figure 2 for results with the bottleneck model.

Model	Reference-Based Metrics					Reference-Free Metrics				
	BLEU-4	METEOR	ROUGE	CIDEr	RefOnlyCLIP-S	CLIP-S	R@1	R@5	R@10	RefCLIP-S
Bottleneck ($\gamma = 1.0$)	**36.1**	**30.5**	**58.2**	**126.1**	**0.900**	0.775	26.5	51.9	64.1	0.830
Bottleneck ($\gamma = 1.2$)	35.1	30.0	57.5	124.1	0.899	0.785	31.3	57.4	69.3	0.835
Bottleneck ($\gamma = 1.5$)	31.5	28.4	54.4	113.2	0.891	0.796	36.6	64.0	75.0	**0.838**
Bottleneck ($\gamma = 2.0$)	20.9	23.3	43.0	78.6	0.862	**0.808**	44.6	71.7	81.7	0.831
Bottleneck ($\gamma = 3.0$)	11.5	17.1	29.4	41.7	0.820	**0.808**	**49.4**	**75.7**	**84.7**	0.811
Bottleneck ($\gamma = 4.0$)	6.5	12.3	18.4	17.3	0.766	0.782	44.7	71.3	80.9	0.771
Att. Pooling ($\gamma = 1.0$)	**36.8**	**30.9**	**59.0**	**130.3**	**0.901**	0.777	27.2	52.7	64.6	0.832
Att. Pooling ($\gamma = 1.2$)	36.3	30.6	58.4	129.1	**0.901**	0.786	32.0	58.0	69.4	0.837
Att. Pooling ($\gamma = 1.5$)	32.7	29.0	55.3	118.0	0.892	0.798	38.2	64.9	75.6	**0.840**
Att. Pooling ($\gamma = 2.0$)	22.1	24.0	44.3	84.6	0.861	0.814	48.6	73.7	83.5	0.833
Att. Pooling ($\gamma = 3.0$)	12.2	17.5	30.7	45.7	0.816	**0.815**	**53.6**	**78.2**	**86.0**	0.812
Att. Pooling ($\gamma = 4.0$)	7.2	12.1	19.7	20.7	0.767	0.788	48.2	72.1	80.1	0.773

Table B.1. Quantitative comparison of results obtained with bottleneck and attention pooling architectures.

B.2. Quantitative assessment of specificity

B.2.1 Evaluation on Stanford Dogs

γ	% Containing Breed	% Breeds Correct
1.0	1.9	61.7
1.2	6.2	69.0
1.5	15.9	69.7
2.0	42.4	58.5
3.0	67.0	53.3

Table B.2. We generate captions for the 8,580 captions in the Stanford Dogs test set and measure the percentage of the captions that contain the name of one of the 120 dog classes ("% Containing Breed") and the percentage of those captions where that name is correct ("% Breeds Correct").

B.2.2 Human evaluation

We performed a human evaluation in which we presented crowdsourcing workers with each image and the two possible captions. We experimented with asking subjects to pick the better caption and the more descriptive caption either on different forms or the same form. When asking subjects to pick only the better caption, we provided the following instructions:

> Please answer a survey about comparing the quality of two captions for each image.
>
> We will present to you an image and ask which caption is better.

When asking subjects to pick the more descriptive caption, we instead provided the following instructions:

> Please answer a survey about comparing the descriptiveness of two captions for each image.
>
> We will present to you an image and ask which caption is a more detailed description of the image. Please ignore grammatical errors that do not affect readability.

When asking both questions simultaneously, we instructed the subjects as follows:

> Please answer a survey about comparing two captions for each image.
>
> We will present to you an image and ask a couple questions about:
>
> 1) descriptiveness: "Which caption is a more detailed description of the image?"
>
> 2) quality: "Which caption is better?"

In each case, subjects saw the image along with the two captions (in random order) as well as the option "I'm indifferent." Subjects clicked the radio button next to their preferred choice. We excluded 55 images for which the captions generated without guidance and at $\gamma = 2.0$ were identical, resulting in a total of 4,945 images. We obtained a single rating for each image in each condition.

Results are shown in Table B.3. When we asked which caption was "better" and which was "more descriptive" in separate surveys, we found that subjects preferred each caption at a statistically indistinguishable rate. When we asked subjects to pick the "better" and "more descriptive" captions in the same survey, we found that $\gamma = 1.0$ was more likely to be chosen as "better" whereas $\gamma = 2.0$ was more likely to be chosen as "more specific." Comparing the odds ratios obtained with the two ways of posing the questions using Fisher's exact test, we find that the difference between them is statistically significant ("better": $p = 0.004$; "more descriptive": $p = 0.01$) indicating that human judgments are significantly affected by whether the questions are posed on the same form or separately.

Question	$\gamma = 1.0$	$\gamma = 2.0$	Indifferent	p-value
Separate forms:				
Better	**48.0%** (2375)	**49.8%** (2461)	2.2% (109)	$p = 0.22$
More descriptive	**47.7%** (2359)	**49.5%** (2446)	2.8% (140)	$p = 0.21$
Same form:				
Better	**50.5%** (2497)	46.6% (2306)	2.9% (142)	$p = 0.006$
More descriptive	45.8% (2265)	**52.7%** (2606)	1.5% (74)	$p = 10^{-6}$

Table B.3. Human evaluation results. We report the percentage and overall number of the 5,000 MS-COCO Karpathy test set images where subjects preferred captions generated at $\gamma = 1.0$ or $\gamma = 2.0$ or were indifferent, as well as the p-value for the null hypothesis that users are equally likely to select the captions generated at $\gamma = 1.0$ and $\gamma = 2.0$, computed by a binomial test. When $p < 0.05$, we bold-face the best result in each row. Otherwise, we bold-face both results.

B.3. Reference-free metrics with retrieval models

In Table B.4, we show cosine similarity between generated captions and image embeddings and caption→image retrieval accuracy for the CoCa 2B model and the CoCa-Base model fine-tuned on MS-COCO that was used to generate the captions. In both cases, we find that $\gamma > 1$ yields much better metrics than no guidance. Retrieval accuracies (but not cosine similarities) are directly comparable across models; both models offer better retrieval accuracy than CLIP ViT-B/32.

γ	CoCa 2B				Captioning Model (CoCa Base)			
	Cos.	R@1	R@5	R@10	Cos.	R@1	R@5	R@10
1.0	0.125	40.1	65.3	75.1	0.843	49.4	75.0	84.1
1.2	0.128	46.5	72.0	80.3	0.859	56.2	80.1	88.1
1.5	0.131	55.5	78.9	86.4	0.877	64.6	85.9	91.5
2.0	**0.135**	64.9	86.4	91.3	0.887	73.0	91.6	95.3
3.0	0.134	**66.5**	**87.0**	**91.4**	**0.890**	**77.7**	**92.4**	**95.8**
4.0	0.126	60.3	81.8	87.5	0.875	74.7	90.1	94.0

Table B.4. CFG improves caption$\rightarrow$image retrieval in the embedding spaces of CoCa models on MS-COCO. “Cos.” = mean cosine similarity between the image and text embeddings.

C. Additional examples

γ=1.0 a vase filled with red and yellow flowers
γ=1.5 tulips in a clear vase on a table
γ=2.0 tulips in a clear glass vase on a tablecloth
γ=3.0 tulips in a clear punchov glass setting on doily
GT Fresh red and yellow tulips in a vase.

γ=1.0 a group of birds standing on top of a wooden post
γ=1.5 seagulls lined up on posts in a lake
γ=2.0 seagulls lined up along a pond line
γ=3.0 seagulls lined up along posts in shallow water
GT Wood post lined up in the water with birds perched on them.

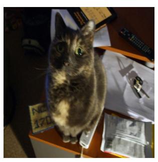
γ=1.0 a living room with a couch and a table
γ=1.5 a living room with a couch and a window
γ=2.0 living room with large window overlooking woods
γ=3.0 livingroom with view out the window
GT A living room in a remotely located home.

γ=1.0 a herd of sheep grazing in a field
γ=1.5 a herd of sheep grazing in a field
γ=2.0 sheep are gathered in a field near piles of hay
γ=3.0 bales of sheep are gathered in formation near rocks
GT A herd of sheep standing on top of a grass covered field.

γ=1.0 a pair of skis sitting on a tiled floor
γ=1.5 pair of skis and ski boots on tiled floor
γ=2.0 skis and pair of skis on linoleum floor
γ=3.0 skis and pair of bottle opener sit on vct floor
GT Skis and ski boots sit together on a tiled floor.

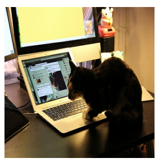
γ=1.0 a cat sitting on a blue chair with a white wall behind it
γ=1.5 a cat sitting on a blue chair outside
γ=2.0 calico colored cat sitting on blue chair outside
γ=3.0 calico colored cat sitting on blue metal chair
GT A furry cat sits on a blue chair.

γ=1.0 a bathroom with a toilet sink and bathtub
γ=1.5 a bathroom with blue and white tiles and a blue and white towel
γ=2.0 bathroom with blue accents and blue and white towels
γ=3.0 spotless uncroom bathroom with blue accents fisheye fisheye fisheye fisheye fisheyemmangles viewersquallly fisheye and fisheye lens
GT Bathroom with white pedestal sink, bathtub and shower, and commode.

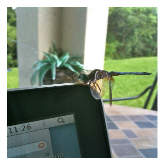


γ=1.0 a cat sitting on top of a desk next to a box
γ=1.5 a cat sitting on top of a desk
γ=2.0 a cat sitting on top of files on a cabinet
γ=3.0 tortoiseshell mittedtabkat sitting inquisitive on papers
GT Cat sitting next to remote control on small counter.

γ=1.0 a box of assorted donuts with a variety of toppings
γ=1.5 a box of assorted donuts with different toppings
γ=2.0 six glazed and chocolate sprinkled doughnuts in a box
γ=3.0 krispy box of dozen glazed and decorated doughnuts
GT Half a dozen donuts from Krispy Kreme of various different flavors.

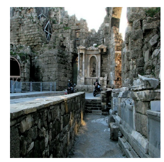


γ=1.0 a cat sitting on a desk next to a laptop
γ=1.5 a cat sitting on a desk next to a laptop
γ=2.0 cat sitting on desk looking at lap top screen
γ=3.0 calico laptop sitting on computer desk with calico cat sitting on top of screen
GT A cat standing on top of a laptop computer.



γ=1.0 a large brown and black insect on top of a laptop
γ=1.5 a bug sitting on the edge of a laptop
γ=2.0 dragonfly perched on television outside on patio
γ=3.0 dragonfly perched on television outside on cantilever table
GT A bug sitting on the side of a laptop computer.

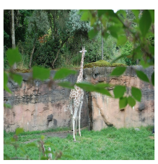
γ=1.0 a red traffic light sitting on the side of a road
γ=1.5 a traffic light with a red pedestrian crossing sign on it
γ=2.0 red traffic light sitting on the side of a street
γ=3.0 pedestrian signal red on a black light pole
GT A red traffic light with a sad face drawn over it.



γ=1.0 a stone wall with a clock tower and a stone wall
γ=1.5 ruins of a building with people walking around
γ=2.0 ruins at a castle in turkey
γ=3.0 ruins at diocletianopolis roman ruins
GT A city made out of stone brick with large arches.

γ=1.0 a pizza that is sitting on a pan
γ=1.5 pepperoni pizza on metal pan with cutter
γ=2.0 pepperoni pizza on metal pan with cutter
γ=3.0 pepperoni steel traybake pepperoni steel tray pizza cutter pepperoni steel tray
GT A pan with three pieces of pepperoni pizza.

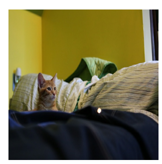
γ=1.0 a basket of bananas and coconuts on a table
γ=1.5 coconut and bananas in a basket with a banana inside
γ=2.0 coconut basket with bananas and nuts in it
γ=3.0 coconut basket bananas coconut husknus and husk laid out
GT a basket with a few things of fruit in it



γ=1.0 a giraffe standing in a grassy area next to a rock wall
γ=1.5 a giraffe standing in a grassy area next to a rock wall
γ=2.0 giraffe standing in enclosure near trees and rock wall
γ=3.0 girafe confined motionless zoo confined wild confined into captivity
GT A giraffe walking through a lush green field.

γ=1.0 a group of teddy bears sitting on a bed
γ=1.5 three teddy bears sitting on a bed together
γ=2.0 four teddy bears sitting on a bed together
γ=3.0 cuddling teddy bears lay piled on a sofa
GT Three different teddy bear on a blanket on a chair.

γ=1.0 two black suitcases are sitting next to each other
γ=1.5 two suitcases with wheels on white background
γ=2.0 two suitcases facing each other 3d illustration
γ=3.0 cgi suitcases rendered cgi cgi cgi looks like luggages cgi cgie cgih cgih cgih cgih cgih cgih cgih cgih cgih cgih cgih cgih cgih travelshpinky like nexushxm gif 3dding
GT A couple of pieces of very nice looking luggage.

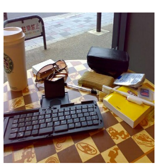


γ=1.0 a cat sitting on a bed next to a blanket
γ=1.5 a cat sitting on a bed under a blanket
γ=2.0 a tabby kitten sitting on top of a comforter on a bed
γ=3.0 tabby kitten sitting on uncovered rumple drapes on unmade unmade unmade bed
GT A brown and white cat lying on a bed

γ=1.0 a bird is standing on the ground in the grass
γ=1.5 weeds and rocks in a grassy area with dirt
γ=2.0 weeds and rocks litter a gravel path in a grassy area
γ=3.0 weeds and gravel strewn away along gravel trail strewn with bird rocks
GT A bird walking through some gravel as its baby chicks follow.

γ=1.0 a cake with a dog and horse on it
γ=1.5 a cake with dogs and horses on it
γ=2.0 cake decorated dog horse and dog motif with three horses
γ=3.0 cake puppy horse dog dog and cats decorated for a first birthday
GT A cake that has paw prints and miniatures dogs on it.

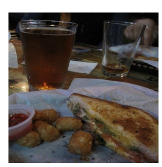
γ=1.0 a city street with a clock tower and cars
γ=1.5 a city street at night with cars and buildings
γ=2.0 cars are driving down a busy city street at night
γ=3.0 ginza at night with cars lights and edifice in asia
GT The traffic and people on a commercial street corner at night



γ=1.0 a table with a keyboard a cup of coffee and a keyboard
γ=1.5 a keyboard coffee cup and glasses on a table
γ=2.0 keyboard coffee sunglasses pen and cup on outdoor table
γ=3.0 keyboard coffee sunglasses pen wallet keyboard starbucks cup on outdoor table
GT a keyboard on a table with a toothbrush a book some sunglasses and coffee

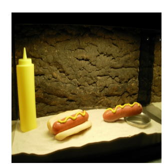
γ=1.0 two cats sitting on a rug in a room
γ=1.5 two cats sitting on a rug in a room
γ=2.0 two cats sitting on rug in room with orange carpet
γ=3.0 cats sitting next to each other on patterned carpet
GT A black cat and an orange cat are sitting on the floor.

γ=1.0 a sandwich and a drink in a basket on a table
γ=1.5 a sandwich and a drink in a basket on a table
γ=2.0 sandwich basket with drink and pickle relish
γ=3.0 sandwich basket drink relish relish pickle hot dog and drink
GT A hotdog with toppings served in a red basket



γ=1.0 a plate of food with a sandwich and a drink
γ=1.5 tater tots and a sandwich and tater tots are on a paper plate
γ=2.0 tater tots toast and a beer on a restaurant table
γ=3.0 tater tots toast club sandwich tater tots and beer on a restaurant table
GT A tray with a cheese and meat sandwich with tater tots.

γ=1.0 a wooden bench sitting in the middle of a forest
γ=1.5 a bench sitting in the middle of a hedge
γ=2.0 hedges and bench in a forested area
γ=3.0 hedges hedge bench hedges bush hedges
GT A bench out by a hedge by the woods



γ=1.0 a hot dog and a mustard bottle on a table
γ=1.5 a hotdog and mustard are on wax paper next to a counter
γ=2.0 hot dog and mustard candles on wax paper
γ=3.0 dug hot dog and mustard candles on wax paper under counter
GT Two hot dogs sitting on top of tissue paper.

Figure C.1. Additional examples of captions generated with classifier-free guidance at different strengths.