Supplementary for
Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval

A. Matching table between noun phrases and class names.

When synthesizing the text by removing noun phrase chunks, we should match the noun phrase with the class names of the object to be removed. While this (class name, noun phrase) pair is annotated in the Flickr30k dataset, we manually list the matching pair in the MS-COCO dataset. If the noun phrase contains a word related to the given class name, we regard that noun phrase as matching the given class name. The matching table is given in Table 1. These matching pairs are based on the implementation done in the previous literature [1], but we added and removed some pairs to make the pairs more relevant. For brevity, we did not list the word that is identical to the class name on the right-hand side of the table.

B. Analysis of distribution shift between the synthetic (D') and the original (D) datasets.

CLIP	ODmAP@1	i2t R@1
zero-shot	58.6	50.6
D_s	61.5	60.5
D'	66.4	58.1
D+D'	70.1	65.6

As $|D'| < |D|$ (one-third smaller), we made a new dataset $D_s \subset D$ where $|D_s| = |D'|$ for comparison. Fine-tuning CLIP with D' and D_s, respectively, resulted in pretty similar results (differing by 2.4% i2t R@1). Considering the 9.9% improvement from zero-shot to D_s, the data distribution of D' seems not much shifted from the data distribution of D even with somewhat broken visual and linguistic coherence in D'. Also, compared to $D+D'$, D' lowers ODmAP@1 by 3.7%. We think this is because information on de-correlated objects in D is not learned by the model trained only with D'.

C. Pseudo-code.

We present the pseudo-code for the implementation of our proposed data synthesis in Listing 1.

Class name in MS-COCO	Word in noun phrase chunk
person	man, woman, player, child, girl, boy, boys, people, lady, guy, kid, kids, surfer, cowboy, cowboys, adult, adults, cop, soldier, police, catcher, pitcher, jockey, baby, men, women, biker, spectator, rider, batter, gay, anyone, someone, reporter, somebody, anybody, everyone, worker, workers
airplane	plane, jet, aircraft
bicycle	bike, biking, cycling
motorcycle	motor
bus	trolley
car	van, taxi, trunk, truck, suv
train	tram, subway
traffic light	traffic
stop sign	sign
parking meter	meter
fire hydrant	hydrant, hydrate, hydra
bird	beak, duck, goose, gull, pigeon, chicken, penguin
cat	kitty, kitten
dog	puppy, puppies
sheep	lamb
horse	pony, foal
cow	cattle, oxen, ox, herd, calves, bull, calf
handbag	bag
suitcase	bag, luggage, case
frisbee	disc, disk, frisby
sports ball	ball
baseball bat	bat
baseball glove	glove
skateboard	board, skate
surfboard	board
snowboard	board
skis	ski
tennis racket	racket, racquet
wine glass	glass, wine, beverage
bottle	thermos, flask, beer, beverage
cup	glass, mug, beverage, coffee, tea
spoon	siverware
donut	doughnut, dough
cake	dessert, frosting
dining table	desk, table, tables
chair	stool
potted plant	plant, flower
vase	pot, vase
tv	television, screen
laptop	computer, monitor, screen
cell phone	phone
refrigerator	fridge
book	novel
scissors	scissor
toothbrush	brush
hair drier	drier
teddy bear	teddy, toy, bear, doll

Table 1. **Matching table between class names and noun phrases.** We regard the noun phrase as matching the given class name if the word related to the class name is contained in the noun phrase.

```
# Threshold for data synthesis (Section 3.1)
alpha1 = 0.4
alpha2 = 0.8
alpha3 = 0.7

# get synthetic data
for image_idx in ranger(n_images):
    image, caption, bboxes, bbox_classnames = dataset.__getitem__(image_idx)

    # extract nounphrases from caption using NLTK tool
    nounphrases = get_nounphrases(caption)

    # get masks for each classname in the image
    classname_set = list(set(bbox_classnames))
    mask_list = []
    for classname_to_remove in classname_set:
        bbox_idxs_to_remove = [i for i, _cat in enumerate(bbox_classnames)
                               if _cat == classname_to_remove]
        bboxes_to_remove = bboxes[bbox_idxs_to_remove]
        mask = union_bboxes(bboxes_to_remove) # union all the bboxes
        mask_list.append(mask)

    if len(classnames_set) >= 2:
        for i, classname_to_remove in enumerate(classname_set):
            # classname_to_remove to be removed
            mask_q = mask_list[i]
            mask_gs = [mask for j, mask in enumerate(mask_list) if j != i]
            classname_gs = [_c for j, _c in enumerate(classname_set) if j != i]

            # check size of removed region
            if mask_q.sum() / (mask_q.size(2) * mask_q.size(3)) > alpha3:
                continue

            # check overlap between bbox from selected class and others
            overlaps = torch.tensor(
                [torch.logical_and(mask_q, mask_g).sum() / mask_g.sum()
                 for mask_g in mask_gs])

            # when removing a single class
            if all(overlaps < alpha1):
                # synthetic image using inpainting GAN
                synth_image = synthesize_image(image, mask_q)
                # synthetic caption using matching table in Appendix
                synth_caption = synthesize_caption(caption, classname_to_remove)

            # when removing multiple classes
            elif any(overlaps > alpha2):
                bool_overlaps = overlaps > alpha2
                mask_qs = \
                    [mask_q] + [_m for j, _m in enumerate(mask_gs) if bool_overlaps[j]]
                classnames_to_remove = \
                    [classname] + [_c for j, _c in enumerate(classname_gs) if bool_overlaps[j]]
                # synthetic image using inpainting GAN
                synth_image = synthesize_image(image, mask_qs)
                # synthetic caption using matching table in Appendix
                synth_caption = synthesize_caption(caption, classnames_to_remove)
```

Listing 1. Pseudo-code for the proposed data synthesis method to reduce spuriousness.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020. 1