Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding

Supplementary Material

6. Analysis of learned representations

In this section, we examine the image and text representations learned by our model. In particular, we investigate whether our method learns more distinct representations for positive and hard negative examples compared to those learned by CLIP and NegCLIP. For each of CLIP, NegCLIP and CE-CLIP, we measure the intra-modal similarity between positive and hard-negative captions, as well as, the cross-modal similarity gap between positive and hard-negative image-caption pairs. We expect our method to reduce the intra-modal similarity and enlarge the cross-modal similarity gap compared to CLIP and NegCLIP. We report the results in Fig. 8, which shows that CE-CLIP achieves statistically significantly better intra-modal similarity (lower is better) and cross-modal similarity gap (higher is better) compared to CLIP and NegCLIP. To compute statistical significance, we used bootstrapping with 50,000 samples with confidence interval of 99%.

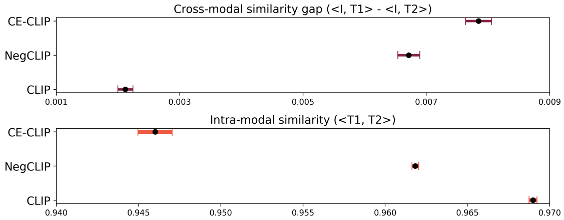

Figure 8. Analyzing the intra-modal similarities and cross-modal similarity gaps yielded by different methods on the ARO benchmark. T1 refers to positive caption while T2 refers to hard-negative caption. The red lines denote the standard errors obtained with bootstrapping 50,000 samples with confidence interval of 99%.

7. Qualitative Examples

Fig. 9 and Fig. 10 illustrate some side-by-side comparisons of hits and misses by CE-CLIP versus NegCLIP.

8. Benchmark

The statistics of benchmarks we use are shown in Tab. 6.

Benchmark	Task	# image-text pairs
ARO-Relation	Relation	24k
ARO-Attribution	Attribution	28.7k
VALSE	Linguistic Phenomena	6.8k
VL-CheckList	Objects, Attributes and Relations	410k
SugarCrepe	Objects, Attributes and Relations	7.5k

Table 6. Overview of vl-compositional benchmarks.

9. Acknowledgement

We thank the Mila IDT team and their technical support for maintaining the Mila compute cluster. We also acknowledge the material support of NVIDIA in the form of computational resources. Throughout this project, Aishwarya Agrawal received support from the Canada CIFAR AI Chair award.

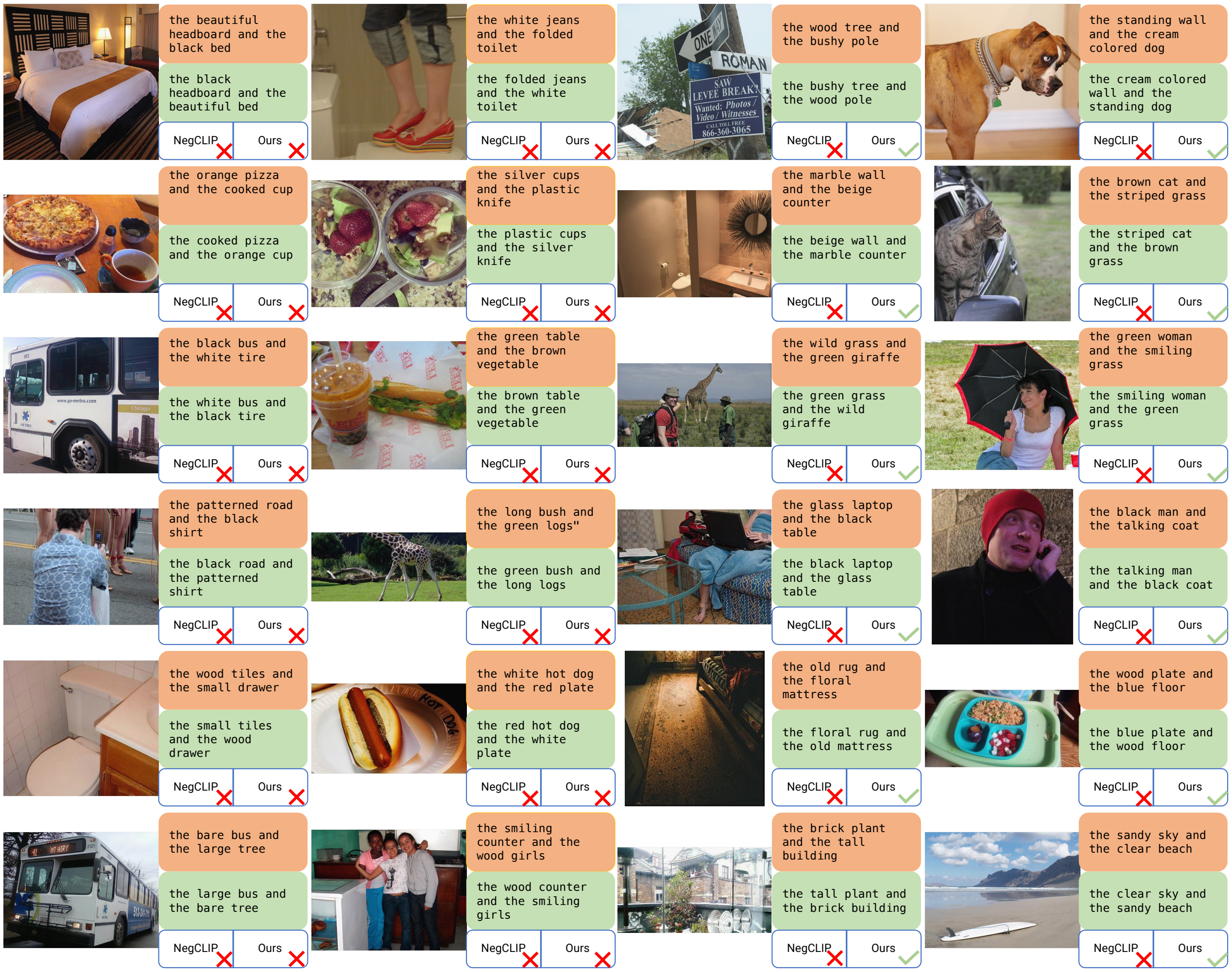
the beautiful headboard and the black bed
the black headboard and the beautiful bed
NegCLIP ✗
Ours ✗
the white jeans and the folded toilet
the folded jeans and the white toilet
NegCLIP ✗
Ours ✗
ONE WAY
ROMAN
SAW LEVEE BREAK?
Wanted: Photos / Video / Witnesses
the wood tree and the bushy pole
the bushy tree and the wood pole
NegCLIP ✗
Ours ✓
the standing wall and the cream colored dog
the cream colored wall and the standing dog
NegCLIP ✗
Ours ✓
the orange pizza and the cooked cup
the cooked pizza and the orange cup
NegCLIP ✗
Ours ✗
the silver cups and the plastic knife
the plastic cups and the silver knife
NegCLIP ✗
Ours ✗
the marble wall and the beige counter
the beige wall and the marble counter
NegCLIP ✗
Ours ✓
the brown cat and the striped grass
the striped cat and the brown grass
NegCLIP ✗
Ours ✓
the black bus and the white tire
the white bus and the black tire
NegCLIP ✗
Ours ✗
the green table and the brown vegetable
the brown table and the green vegetable
NegCLIP ✗
Ours ✗
the wild grass and the green giraffe
the green grass and the wild giraffe
NegCLIP ✗
Ours ✓
the green woman and the smiling grass
the smiling woman and the green grass
NegCLIP ✗
Ours ✓
the patterned road and the black shirt
the black road and the patterned shirt
NegCLIP ✗
Ours ✗
the long bush and the green logs"
the green bush and the long logs
NegCLIP ✗
Ours ✗
the glass laptop and the black table
the black laptop and the glass table
NegCLIP ✗
Ours ✓
the black man and the talking coat
the talking man and the black coat
NegCLIP ✗
Ours ✓
the wood tiles and the small drawer
the small tiles and the wood drawer
NegCLIP ✗
Ours ✗
Hot Dog
the white hot dog and the red plate
the red hot dog and the white plate
NegCLIP ✗
Ours ✗
the old rug and the floral mattress
the floral rug and the old mattress
NegCLIP ✗
Ours ✓
the wood plate and the blue floor
the blue plate and the wood floor
NegCLIP ✗
Ours ✓
the bare bus and the large tree
the large bus and the bare tree
NegCLIP ✗
Ours ✗
the smiling counter and the wood girls
the wood counter and the smiling girls
NegCLIP ✗
Ours ✗
the brick plant and the tall building
the tall plant and the brick building
NegCLIP ✗
Ours ✓
the sandy sky and the clear beach
the clear sky and the sandy beach
NegCLIP ✗
Ours ✓

Figure 9. Some qualitative examples from ARO-Attribute. Caption in red box is unmatched and in green box is matched. ✓ represents model predicates correctly and ✗ means wrong.

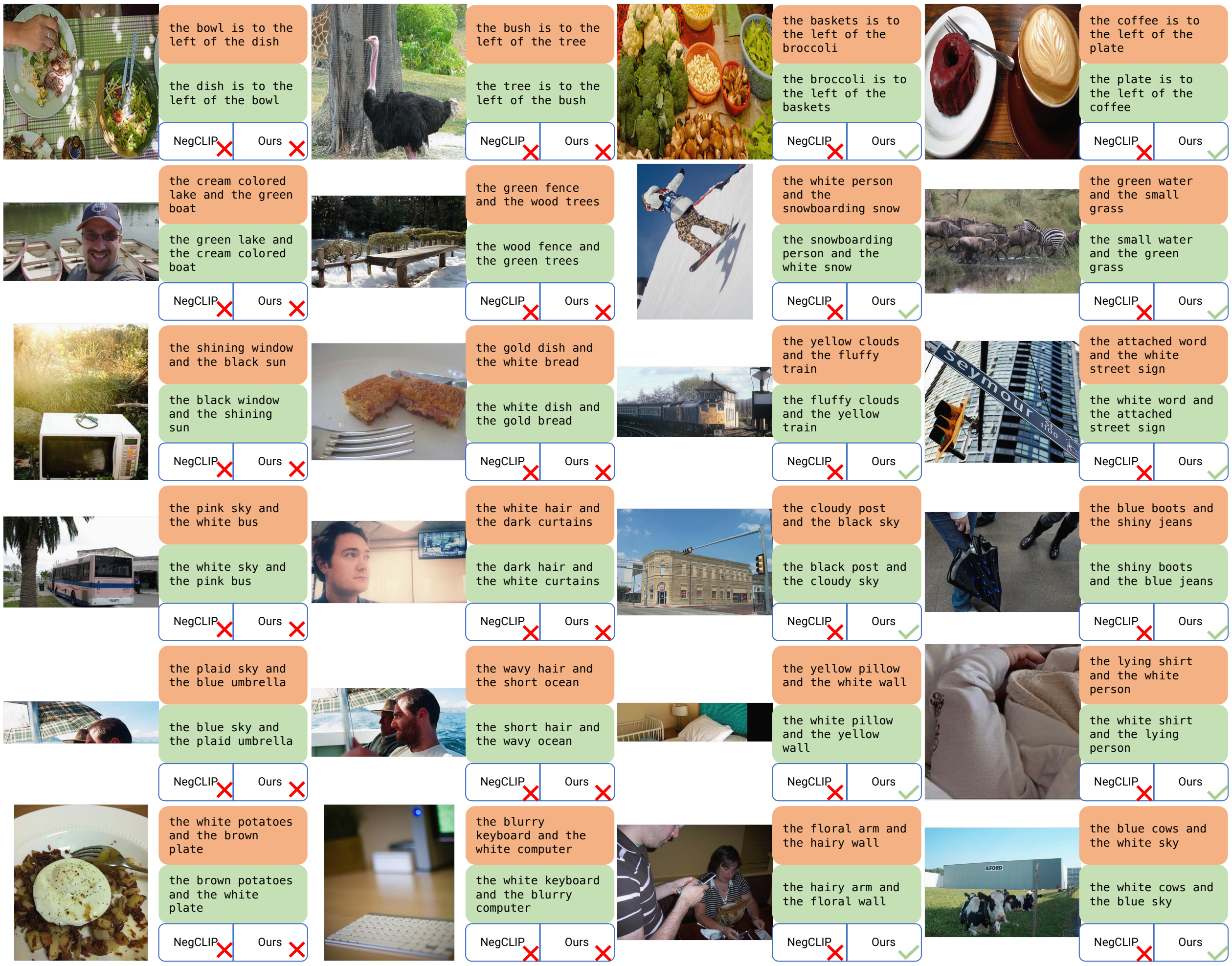
the bowl is to the left of the dish
the dish is to the left of the bowl
NegCLIP ✗ Ours ✗
the bush is to the left of the tree
the tree is to the left of the bush
NegCLIP ✗ Ours ✗
the baskets is to the left of the broccoli
the broccoli is to the left of the baskets
NegCLIP ✗ Ours ✓
the coffee is to the left of the plate
the plate is to the left of the coffee
NegCLIP ✗ Ours ✓
the cream colored lake and the green boat
the green lake and the cream colored boat
NegCLIP ✗ Ours ✗
the green fence and the wood trees
the wood fence and the green trees
NegCLIP ✗ Ours ✗
the white person and the snowboarding snow
the snowboarding person and the white snow
NegCLIP ✗ Ours ✓
the green water and the small grass
the small water and the green grass
NegCLIP ✗ Ours ✓
the shining window and the black sun
the black window and the shining sun
NegCLIP ✗ Ours ✗
the gold dish and the white bread
the white dish and the gold bread
NegCLIP ✗ Ours ✗
the yellow clouds and the fluffy train
the fluffy clouds and the yellow train
NegCLIP ✗ Ours ✓
the attached word and the white street sign
the white word and the attached street sign
NegCLIP ✗ Ours ✓
the pink sky and the white bus
the white sky and the pink bus
NegCLIP ✗ Ours ✗
the white hair and the dark curtains
the dark hair and the white curtains
NegCLIP ✗ Ours ✗
the cloudy post and the black sky
the black post and the cloudy sky
NegCLIP ✗ Ours ✓
the blue boots and the shiny jeans
the shiny boots and the blue jeans
NegCLIP ✗ Ours ✓
the plaid sky and the blue umbrella
the blue sky and the plaid umbrella
NegCLIP ✗ Ours ✗
the wavy hair and the short ocean
the short hair and the wavy ocean
NegCLIP ✗ Ours ✗
the yellow pillow and the white wall
the white pillow and the yellow wall
NegCLIP ✗ Ours ✓
the lying shirt and the white person
the white shirt and the lying person
NegCLIP ✗ Ours ✓
the white potatoes and the brown plate
the brown potatoes and the white plate
NegCLIP ✗ Ours ✗
the blurry keyboard and the white computer
the white keyboard and the blurry computer
NegCLIP ✗ Ours ✗
the floral arm and the hairy wall
the hairy arm and the floral wall
NegCLIP ✗ Ours ✓
the blue cows and the white sky
the white cows and the blue sky
NegCLIP ✗ Ours ✓

Figure 10. Some qualitative examples from ARO-Relation. Caption in red box is unmatched and in green box is matched. ✓ represents model predicates correctly and ✗ means wrong.