Probing Conceptual Understanding of Large Visual-Language Models

Supplementary Material

The supplementary will provide additional details about our proposed datasets, finetuning CLIP and the models evaluated on in this benchmark. Additional details and results for Probe-R, Probe-C and Probe-B are in Section 7. We provide more details about finetuning CLIP and additional results in Section 8. In Section 9 we provide additional details about the models we evaluated in this benchmark.

7. Datasets Details

In this section we will provide additional results for the different dataset benchmarks.

7.1. Probe-R: Relational Understanding

This dataset was created using Visual Genome (VG) [21]. To collect unlikely "<subject, predicate, object>" triplets, we first cleaned the relationship aliases. This was done by mapping repeated aliases that meant the same thing into one, for example "are standing next to" would become "standing next to". This was done to reduce the space to map all objects to aliases they have been associated with as well as to confirm they have not been associated with one similar. We then collect all the objects each cleaned alias was associated with using regex and NLTK part-of-speech (POS) tagging [3]. Using these object collections, we iterated through $100,000$ VG annotations of $R_1 = \langle s_1, r_1, o_1 \rangle$ to (1) replace the existing alias with an alias that the current subject and object are not associated with as swap ($R_2 = \langle s_1, \overline{r}_1, o_1 \rangle$) and (2) replace the existing subject with an object that is not associated with the current alias ($R_3 = \langle \overline{s}_1, r_1, o_1 \rangle$). To better collect images with specific objects in them, we iterated through VG and generated a mapping of each image ID to all objects present in the image according to the relationships annotations. We extract positive images X_{O_1} that do not have the relation but have the subject and no other objects present in the anchor image X_{R_1}.

The results for all models for the Probe-R benchmark are shown in Table 4. We include CLIP models we finetuned on RelComp, training either the text encoder (T), visual encoder (V) or both encoders (VT). Training only the text encoder seems to have the highest improvement, but as mentioned in the paper, the largest occurrence of "catastrophic forgetting" when evaluated on ImageNet. A TSNE plot of model features that includes CLIP Patched (VT) is shown in Figure 9. In black we have the image features, in red we have the predicate swapped text features (P_{R_2}), and in green we have the ground truth relation text features (P_{R_1}). This finetuned and patched version appears to have tighter clusters compared to the original CLIP model.

7.2. Probe-C: Compositional Understanding

This dataset was generated using MSCOCO [24]. To guarantee that the images had no similarity or overlap, we focused on using antonyms of select attributes. We started by using NLTK POS [3] to find adjective-noun pairs. We then manually cleaned and extracted the adjectives to guarantee the attribute is a visual one such as "red" or "young" as opposed to a subjective one such as "hungry" or "thirsty". While these are useful attributes, we are primarily interested in visual perception as opposed to subjective inference. We then iterated through all images and mapped each attribute to their corresponding image IDs, and we did the same with objects. Using this collection, we were able to create groups of pairs based on either swapping the attribute to one of its antonyms or swapping the object with one that has the same attribute.

The overall results for Probe-C for all models is in Table 5. The mappings we used to categorize different attributes is shown in Table 6, these were manually generated. A visual break down of different model performances for each attribute is shown in Figure 10. From there, you can see the changes in score based on whether it is matching the caption given the image versus given text. We also see that most models struggle with "visibility" and often "texture".

7.3. Probe-B: Context Understanding

In set 1, for each image we remove the background using segmentation masks from original annotations. We replace the background with 1 of four fillers: black, gray, Gaussian noise, or a random scene. Random scenery was collected from the Indoor Scenes Dataset [30] and the Kaggle Landscape dataset [34]. These images were manually filtered to ensure none of the 80 MSCOCO classes were present. The total collection is 31,745 images with 4 fillings each for a total of 126,980 images. We filtered images based on a threshold for how much background can be removed to ensure that some context was actually removed. In set 2, for each image we remove all other objects and the background using segmentation masks. In this case, x_0 is the image with all objects with just the background removed while $\tilde{x_1}$ is the image with just one object remaining and all other objects and the background removed. This allows us to isolate whether it is the other objects compared to background removal. Like in set 1, we replace them with the different possible fillers. Images are chosen if they do not have overlapping bounding boxes and if their object area is over a threshold to allow for better visibility. Prompts for

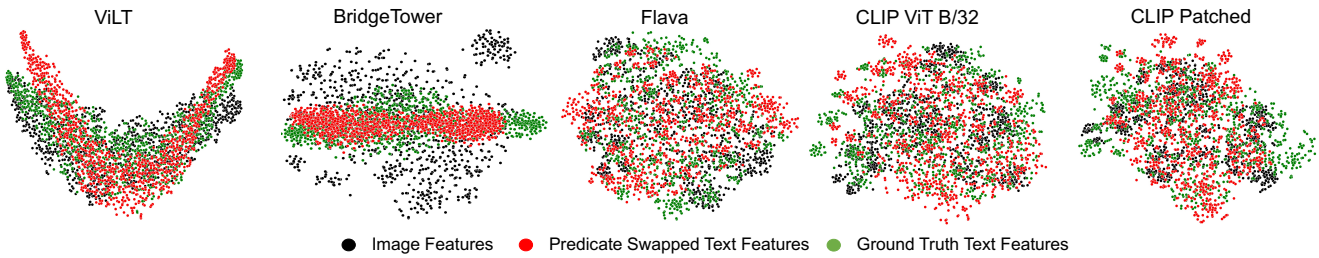

Figure 9. TSNE plots of image features (black) and text features from the Probe-R. Text features are prompts generated from either the ground truth relation R_1 (green) or the relation with the predicate swapped to an unrealistic one R_2 (red). Both ViLT and BridgeTower rely on cross-attention heavily, showing the impact on the feature space. While the features for the other models are more visibly in the same space, ViLT and BridgeTower generally show higher performance. CLIP patched is finetuning both visual and text encoders using RelComp and patching with an alpha of 0.2 [16]. CLIP ViT is ViT/L-14@336px while CLIP CNN is RN50x4.

Table 4. **Overall results for relation evaluation.** The anchor image X_{R_1} contains the relation $R_1 = \langle s, r, o \rangle$, image X_{O_1} contains $O_1 = \langle s \rangle$. Prompts contain either the relation P_{R_1}, $P_{R_2} = \langle s, \overline{r}, o \rangle$, $P_{R_3} = \langle \overline{s}, r, o \rangle$, $P_{O_1} = \langle s \rangle$, or $P_{O_3} = \langle \overline{s} \rangle$. The mean confidence $\mu(c)$ is for the correct prompt to image. Models with CLIP Patched are those we finetuned on our training dataset RelComp. We finetuned either the text encoder (T), the visual encoder (V) or both (VT). Models show higher performance for when objects are switched but lower performance when the relation is switched, showing the models are confused.

Model	X_{R_1}						X_{O_1}	
	P_{R_1} vs. P_{R_3}		P_{R_1} vs. P_{R_2}		P_{R_1} vs. P_{O_1}		P_{R_1} vs. P_{R_2}	
	$\mu(c)$	Acc	$\mu(c)$	Acc	$\mu(c)$	Acc	$\mu(c)$	Acc
CLIP RN50	69.77	72.14	51.33	51.13	61.19	61.69	78.44	89.10
CLIP ViT L/14	71.59	73.68	52.44	52.59	59.09	58.67	84.17	93.23
CLIP ViT-B/16	71.08	73.40	52.84	53.37	61.69	62.07	79.62	89.96
CLIP ViT/B-32	69.00	71.21	53.02	53.53	58.83	58.56	82.21	92.32
CLIP ViT	72.09	74.27	53.52	53.97	59.53	59.14	83.97	93.50
CLIP RN101	70.62	73.28	54.01	55.11	60.66	60.83	79.08	91.17
CLIP RN50x64	72.79	74.79	56.66	58.03	64.10	64.88	78.10	87.15
CLIP CNN	72.71	75.59	56.35	58.14	62.31	62.81	78.29	90.77
CLIP RN50x16	73.91	76.57	58.08	60.52	59.80	59.79	83.03	94.05
CLIP Patched (V)	78.58	81.41	59.36	62.27	66.56	68.07	81.32	90.79
FLAVA	76.79	79.09	64.65	68.29	64.40	65.56	84.19	90.12
ViLT	76.41	78.45	64.77	69.00	54.84	54.78	**94.23**	**99.10**
CLIP Patched (VT)	80.56	84.46	64.53	71.12	67.63	70.07	81.74	92.40
CLIP Patched (T)	82.37	86.25	66.28	72.55	67.93	70.51	79.76	90.70
BridgeTower	**83.03**	**89.01**	**72.93**	**82.04**	**71.73**	**78.90**	76.58	94.38
BLIP	62.38	69.2	56.8	65.02	48.31	46.68	76.65	97.15
BLIP2	70.82	81.57	59.31	68.39	47.26	41.9	78.06	96.51
OTTER	49.87	42.18	50.02	52.33	49.6	24.48	50.49	84.28
ALIGN	75.68	79.81	56.88	60.34	65.35	66.24	73.42	90.51
MetaCLIP	72.66	74.53	52.72	53.42	54.93	54.16	88.68	96.14
SigLIP	73.88	75.78	54.14	55.31	63.86	63.92	82.77	91.86

set 2 only include objects not present in the original image and the target object.

To better compare CLIP backbones, Figure 11 shows a comparison between the change in confidence from a patched image $\tilde{x_0}$ to the image where all other objects and background $\tilde{x_1}$ is removed aggregated over CLIP backbones. Table 8 shows what objects are assigned to which category and how many samples are present in the annotations. The main differences are in objects they struggle with by how much and in which order.

Table 5. **Overall results for the compositional evaluation** on select models with highest scores in **bold** and second highest underlined. Mean confidence for the correct prompt-to-image is $\mu(c)$. CLIP ViT is ViT/L-14@336px while CLIP CNN is RN50x4.

	Composition Switch				Object Switch			
Model	$\mu(c)\uparrow$	Image $\uparrow$	Text $\uparrow$	Group $\uparrow$	$\mu(c)\uparrow$	Image $\uparrow$	Text $\uparrow$	Group $\uparrow$
CLIP ViT	69.69	33.06	52.82	26.64	88.15	61.96	81.89	58.05
CLIP RN50	69.47	33.41	54.60	26.92	87.00	61.40	80.17	56.81
CLIP ViT-B/16	69.23	34.29	52.23	26.94	88.02	63.53	81.44	59.12
CLIP ViT L/14	69.41	33.73	52.36	27.01	87.89	61.93	81.31	57.86
CLIP RN101	69.24	34.95	51.82	27.42	86.99	61.75	80.58	57.46
CLIP RN50x64	70.44	35.21	52.89	27.95	87.75	63.09	80.55	58.27
CLIP ViT/B-32	69.79	34.71	53.85	27.96	87.75	62.01	80.92	57.65
CLIP RN50x16	69.77	35.51	53.24	28.07	87.91	63.12	82.23	59.38
CLIP CNN	69.75	36.07	54.56	28.79	87.24	61.29	81.24	57.06
FLAVA	67.45	60.93	39.65	33.09	83.85	82.66	70.08	65.37
CLIP Patched (T)	71.94	40.96	58.79	33.83	89.58	68.81	84.36	65.19
CLIP Patched (V)	73.65	42.30	59.10	34.48	89.79	66.17	84.00	62.45
CLIP Patched (VT)	73.65	44.53	61.92	37.18	90.30	70.01	85.41	66.83
ViLT	79.02	53.74	66.84	46.65	**90.78**	73.82	85.88	70.26
BridgeTower	**81.88**	**65.95**	**75.02**	**59.28**	90.05	77.44	**87.63**	**74.54**
BLIP	73.1	65.64	70.91	56.74	81.59	74.24	81.37	47.26
BLIP2	70.98	62.55	72.03	54.69	81.8	74.01	81.31	67.5
OTTER	50.05	12.71	22.24	7.14	50.21	31.17	24.5	14.62
ALIGN	71.85	61.48	39.16	33.13	87.9	**83.6**	68.79	65.08
MetaCLIP	71.56	36.55	56.01	29.63	87.31	66.02	79.41	60.53
SigLIP	74.5	40.59	60.82	33.59	90.1	70.55	83.65	66.64

Table 6. **The attributes that belong to each category for the compositional analysis** on specific attributes in Probe-C.

Attribute	Category	Groups
age	[young, old, new]	2,051
color	[greyscale, coloured, sepia, reddish, bronze, greenish, green, turquoise, blue, tan, red, white, silver, purple, gold, pink, navy, brown, teal, gray, black, yellow, grey, golden, camo, pinkish, beige, orange, blonde]	39,971
expression	[happy, unhappy, smiling, laughing, smiley, sad]	2,088
gender	[male, female]	2,346
material	[tin, aluminum, cloth, gravel, unpaved, wooden, stainless, marble, metallic, metal, grassy, porcelain, wooded, pebbled]	3,875
pattern	[checkered, patterned, striped, spotted, plaid, stripped, checkerboard]	3,08
shape	[triangular, flat, circular, triangle, oval, round, dotted, rectangular, square]	1,164
size	[bulky, long, thin, large, big, tall, short, small, huge, tiny, giant, little, chubby, pudgy]	16,575
texture	[smooth, fluffy, fuzzy, dry, wet, rusty, bald, hairy, stony]	1,090
visibility	[shiny, unclear, sun, nightime, blurry, shadowy, lit, shady, light, darkened, hazy, dark, barren, cloudy, clear, sunlit, bright, foggy, rainy, sparkling]	10,454

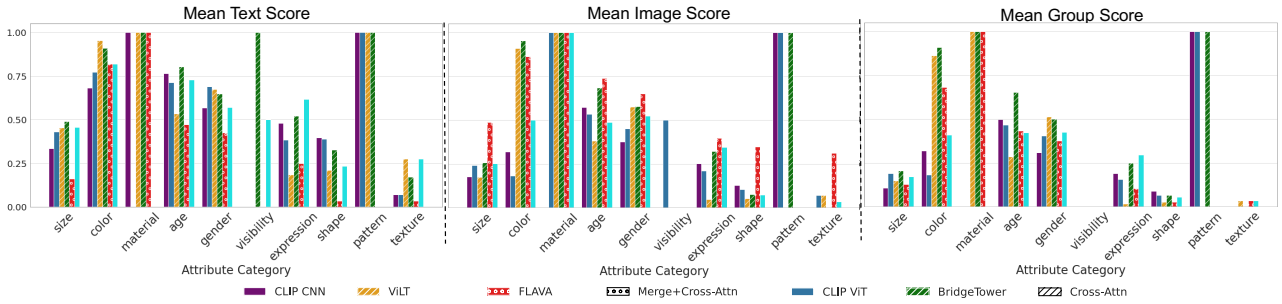

Figure 10. **Performance on compositional understanding**Mean image, text and group scores for a subset of models. Models are typically better matching a caption given an image rather than the reverse.

Table 7. **Mean image, text and group scores** for each category of attributes for each model.

Model	age			color			expression			gender			material		
	Image	Text	Group	Image	Text	Group	Image	Text	Group	Image	Text	Group	Image	Text	Group
CLIP RN50	47.74	68.84	39.20	18.18	13.64	0.00	21.07	35.45	13.14	34.71	60.75	30.18	0.00	0.00	0.00
CLIP RN50x64	53.77	82.41	49.25	4.55	13.64	0.00	20.08	36.28	13.97	39.64	**70.61**	36.29	100.00	100.00	100.00
CLIP RN101	48.74	70.35	39.20	18.18	63.64	18.18	37.60	43.97	24.30	32.35	55.23	27.02	0.00	100.00	0.00
CLIP ViT	53.27	71.36	46.73	18.18	77.27	18.18	20.91	38.26	15.62	45.17	69.03	40.43	100.00	0.00	0.00
CLIP ViT-B/16	45.73	76.88	40.70	27.27	77.27	22.73	28.84	**66.94**	24.05	41.03	62.92	35.90	0.00	0.00	0.00
CLIP ViT L/14	53.77	75.88	47.74	31.82	72.73	27.27	18.43	16.28	11.82	41.62	69.43	37.67	100.00	0.00	0.00
CLIP CNN	57.29	76.38	49.75	31.82	68.18	31.82	25.21	47.93	19.01	37.67	56.80	30.77	0.00	100.00	0.00
CLIP Patched (T)	41.71	79.40	37.69	31.82	68.18	31.82	37.27	61.98	**31.74**	46.94	54.83	36.09	100.00	100.00	100.00
CLIP ViT/B-32	40.20	74.37	35.68	18.18	13.64	4.55	23.22	58.02	20.08	39.05	60.75	33.93	100.00	100.00	100.00
CLIP Patched (V)	44.72	76.38	39.70	40.91	90.91	40.91	31.65	60.74	27.19	53.06	60.16	43.20	100.00	100.00	100.00
CLIP Patched (VT)	48.74	72.86	42.21	50.00	81.82	40.91	34.55	61.57	29.67	52.47	57.00	42.60	100.00	0.00	0.00
FLAVA	**73.87**	47.24	43.22	86.36	81.82	68.18	**39.67**	24.88	10.33	**65.09**	42.60	37.67	100.00	100.00	100.00
CLIP RN50x16	52.76	**83.92**	49.75	9.09	13.64	9.09	21.98	42.98	14.71	40.83	62.72	35.11	100.00	100.00	100.00
BridgeTower	68.34	80.40	**65.33**	**95.45**	90.91	**90.91**	32.23	52.07	24.79	57.79	64.89	49.90	100.00	100.00	100.00
ViLT	38.19	53.27	28.64	90.91	**95.45**	86.36	4.79	18.43	1.57	57.59	67.46	**51.28**	**100.00**	**100.00**	**100.00**

Model	pattern			shape			size			texture			visibility		
	Image	Text	Group	Image	Text	Group	Image	Text	Group	Image	Text	Group	Image	Text	Group
CLIP RN50	100.00	100.00	100.00	3.69	23.88	2.56	21.21	47.95	17.11	0.00	20.69	0.00	0.00	0.00	0.00
CLIP RN50x64	100.00	100.00	100.00	9.46	22.92	6.25	18.83	33.40	12.37	3.45	20.69	0.00	0.00	0.00	0.00
CLIP RN101	100.00	100.00	100.00	9.13	27.88	5.45	20.96	37.63	12.72	3.45	**31.03**	3.45	50.00	0.00	0.00
CLIP ViT	100.00	100.00	100.00	10.26	38.94	6.41	24.13	42.98	18.85	6.90	6.90	0.00	50.00	0.00	0.00
CLIP ViT-B/16	0.00	100.00	0.00	8.01	21.63	3.21	15.69	28.33	7.73	6.90	17.24	0.00	**100.00**	0.00	0.00
CLIP ViT L/14	0.00	100.00	0.00	13.46	37.02	8.49	24.84	45.11	19.34	3.45	6.90	0.00	50.00	0.00	0.00
CLIP CNN	100.00	100.00	100.00	12.82	**39.58**	**8.81**	17.51	33.43	10.62	0.00	6.90	0.00	0.00	0.00	0.00
CLIP Patched (T)	0.00	0.00	0.00	6.25	22.92	3.04	20.40	43.08	16.09	3.45	10.34	0.00	50.00	50.00	0.00
CLIP ViT/B-32	0.00	0.00	0.00	7.85	26.60	4.97	13.63	34.47	7.05	0.00	3.45	0.00	50.00	0.00	0.00
CLIP Patched (V)	0.00	100.00	0.00	10.10	28.04	6.41	16.29	44.60	10.42	0.00	6.90	0.00	50.00	50.00	**50.00**
CLIP Patched (VT)	0.00	0.00	0.00	7.37	23.24	5.45	25.14	45.67	17.03	3.45	27.59	3.45	0.00	50.00	0.00
FLAVA	0.00	0.00	0.00	**34.62**	3.37	2.72	**48.66**	16.09	12.60	**31.03**	3.45	3.45	0.00	0.00	0.00
CLIP RN50x16	**100.00**	100.00	**100.00**	7.85	19.87	4.49	22.93	**54.13**	18.88	3.45	20.69	0.00	0.00	0.00	0.00
BridgeTower	100.00	100.00	100.00	7.53	32.53	6.41	25.54	49.06	**20.40**	0.00	17.24	0.00	0.00	**100.00**	0.00
ViLT	0.00	**100.00**	0.00	5.13	20.99	2.56	17.44	45.36	14.75	6.90	27.59	**3.45**	0.00	0.00	0.00

Table 8. **The objects that belong to each category for the object-context analysis** on specific objects in Probe-B.

Object	Category	Groups
accessories	[backpack, umbrella, handbag, tie, suitcase]	174
animals	[bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe]	627
appliances	[microwave, oven, toaster, refrigerator]	591
decor	[clock, vase]	138
electronics	[tv, laptop, mouse, remote, keyboard, cell phone]	1095
fixtures	[toilet, sink]	387
foods	[sandwich, hot dog, pizza, donut, cake]	258
fruits	[banana, orange]	120
furniture	[chair, couch, bed, dining table]	546
kitchenware	[bottle, wine glass, cup, fork, knife, spoon, bowl]	399
people	[person]	720
plants	[potted plant]	108
recreation	[frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket]	117
roadway	[traffic light, fire hydrant, stop sign, parking meter]	144
street furniture	[bench]	42
tools	[scissors, hair drier, toothbrush]	15
toys	[book, teddy bear]	144
vegetables	[broccoli, carrot]	111
vehicles	[bicycle, car, motorcycle, airplane, bus, train, truck, boat]	603

Overall results for Probe-B are in Table 9 and 10. In both cases, replacing with scene and noise produces worse results compared to black and gray fillers. For aggregating across filler, we only include CLIP ViT-L/14@336px, CLIP RN50x4, FLAVA, ViLT, BridgeTower, BLIP, BLIP2, OTTER, ALIGN, MetaCLIP and SigLIP. When comparing individual model results in Table 10, performance tends to increase when only the other object remains, meaning that

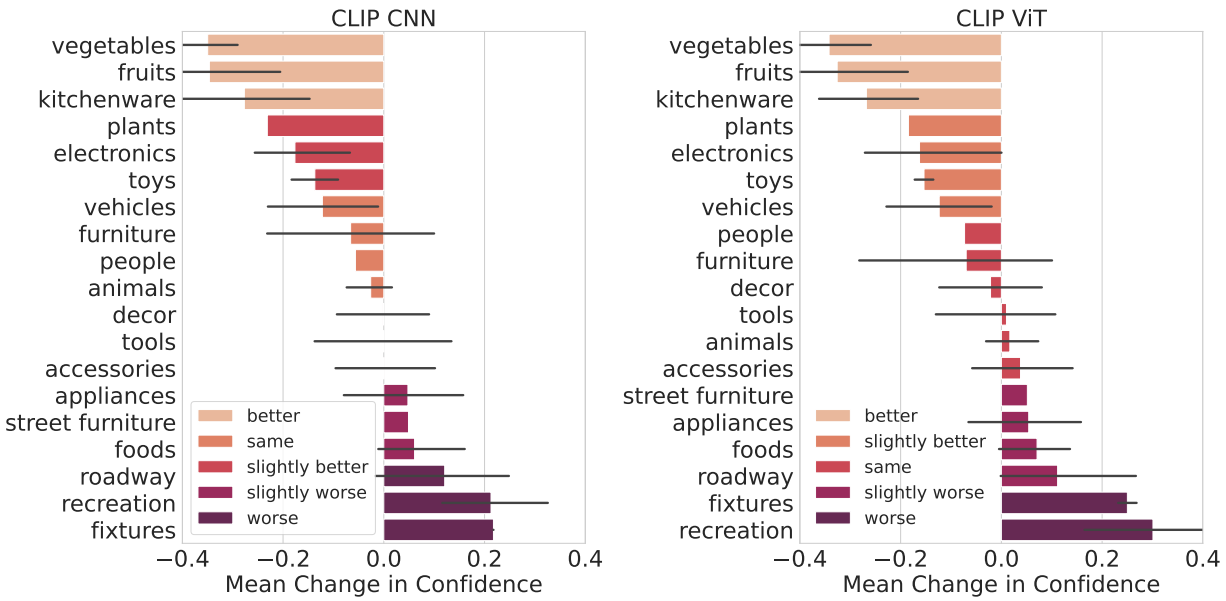

Figure 11. **Comparing the change in confidence** from a patched image $\tilde{x_0}$ to the image where all other objects and background $\tilde{x_1}$ is removed aggregated over CLIP backbones.

Table 9. **Mean results for Probe-B_{MR}** when the background of an image is replaced with each filler (top) and for each model averaged over fillers (bottom). Comparisons are between the original image x_0, original image with a random patch $\tilde{x_0}$ and the modified image $\tilde{x_1}$ where the background is removed. The metrics are mean average precision (mAP), relative robustness (γ_r) measuring the relative drop/increase in performance, and mean change in softmax confidence $\mu(\nabla(c))$ for the objects.

	Average Precision (mAP)			Relative Robustness (γ_r)			Mean Change Confidence ($\mu(\nabla(c))$		
Filler	x_0	$\tilde{x_0}$	$\tilde{x_1}$	$(x_0, \tilde{x_1})$	$(x_0, \tilde{x_0})$	$(\tilde{x_0}, \tilde{x_1})$	$(c_0 - \tilde{c_0})$	$(\tilde{c_0} - \tilde{c_1})$	$(c_0 - \tilde{c_1})$
black	69.76	69.62	70.6	1.09	**1.05**	1.18	0.95	−0.91	**0.04**
noise	69.8	67.75	68.14	1.04	1.02	1.2	1.97	−0.91	1.06
gray	**69.84**	**69.68**	**71.02**	**1.1**	**1.05**	1.2	**0.92**	**−0.86**	0.05
scene	69.8	66.5	67.85	1.04	1	**1.29**	2.18	−1.31	0.87
Model	x_0	$\tilde{x_0}$	$\tilde{x_1}$	$(x_0, \tilde{x_1})$	$(x_0, \tilde{x_0})$	$(\tilde{x_0}, \tilde{x_1})$	$(c_0 - \tilde{c_0})$	$(\tilde{c_0} - \tilde{c_1})$	$(c_0 - \tilde{c_1})$
CLIP RN50	65.05	65.47	60.30	1.00	1.02	0.99	−0.05	4.86	4.82
CLIP ViT/B-32	68.77	67.49	61.10	0.95	0.99	0.98	0.40	5.12	5.52
CLIP CNN	63.23	63.56	63.46	1.12	1.02	1.10	0.46	2.64	3.11
CLIP RN101	64.56	65.00	63.80	1.09	1.02	1.08	0.21	2.79	3.00
CLIP ViT-B/16	69.97	68.54	65.15	0.99	0.98	1.02	0.36	2.92	3.28
FLAVA	72.05	74.47	66.75	0.98	1.05	0.94	0.00	0.01	0.01
CLIP ViT L/14	70.98	69.38	68.99	1.04	0.98	1.08	0.94	1.17	2.12
CLIP ViT	71.05	70.94	71.50	1.08	1.01	1.08	0.70	0.52	1.22
ViLT	**83.49**	71.38	83.26	1.00	0.87	**1.62**	6.68	**−6.61**	0.08
BridgeTower	81.85	**81.88**	**83.40**	1.05	1.06	1.23	**−0.79**	−0.42	**−1.22**
BLIP	68.67	73.43	71.76	1.07	**1.16**	1.08	−0.78	0.16	−0.62
BLIP2	77.37	76.19	77.1	1.03	1.05	1.17	3.47	−1.96	1.51
OTTER	30.75	29.23	30.8	1.05	1.08	1.3	0	0	0
ALIGN	50.53	50.52	48.89	**1.48**	**1.15**	1.46	0	0.04	0.04
MetaCLIP	67.1	61.79	66.76	1	0.99	1.43	5.22	−4.86	0.36
SigLIP	69.96	70.46	69.77	1	1.08	1.16	2.19	−2.07	0.13

Table 10. **Results for when the background and all other objects are replaced** with a filler $\tilde{x_1}$, compared to the original image x_0, and an image with a random patch of the same filler type $\tilde{x_0}$. Metrics used are the accuracy of detecting the object compared to other objects that are not present in the image and the relative robustness γ_r, which is the relative change in confidence.

	Accuracy			Relative Robustness γ^r		
Filler	x_0	$\tilde{x_0}$	$\tilde{x_1}$	$(x_0, \tilde{x_1})$	$(x_0, \tilde{x_0})$	$(\tilde{x_0}, \tilde{x_1})$
noise	54.48	54.95	60.09	1.1	1.01	1.09
scene	49.08	52	56.1	**1.14**	**1.06**	1.08
black	56.62	**57.13**	63.79	1.13	1.01	1.12
gray	**57.08**	56.83	**63.98**	1.12	1	**1.13**
Model	x_0	$\tilde{x_0}$	$\tilde{x_1}$	$(x_0, \tilde{x_1})$	$(x_0, \tilde{x_0})$	$(\tilde{x_0}, \tilde{x_1})$
BridgeTower	**77.36**	**76.14**	**77.70**	1.01	0.98	1.02
FLAVA	56.33	59.16	58.32	1.03	1.06	0.99
CLIP ViT/B-32	46.19	50.50	54.20	1.17	1.10	1.07
CLIP ViT-B/16	51.12	51.31	60.30	1.18	1.01	1.17
CLIP ViT L/14	56.27	54.18	67.16	1.19	0.96	1.24
CLIP RN50	45.79	48.64	55.68	1.21	1.07	1.14
CLIP ViT	59.31	56.08	71.76	1.21	0.95	1.28
CLIP RN101	46.48	47.92	57.32	1.23	1.03	1.20
CLIP RN50x16	53.88	50.92	66.18	1.23	0.95	**1.30**
CLIP RN50x64	56.86	53.70	70.04	1.23	0.95	**1.30**
CLIP CNN	50.08	49.66	62.18	1.24	1.00	1.25
ViLT	54.73	55.82	72.09	**1.33**	1.02	**1.30**
BLIP	68.41	73.95	64.62	0.94	1.084	0.87
BLIP2	68.02	65.13	65.67	0.97	0.964	1.01
OTTER	12.16	10.45	11.89	0.98	0.864	1.14
ALIGN	63.12	61.66	69.17	1.1	0.984	1.12
MetaCLIP	48.01	55.03	51.37	1.07	1.154	0.93
SigLIP	63.51	73.82	62.22	0.98	**1.164**	0.84

other objects may actually distract models. BridgeTower is the highest performer and has the lowest robustness from x_0 to x_1 meaning that it may be using some level of object relationship understandings to help recognize objects. However, this difference is minor and therefore inconclusive. Other models' robustness though is higher indicating they perform better when objects are in isolation, indicating they are not using object relationship understanding to help object detection of particular objects. In Table 9, when only background is removed, we see little change. However, in ViLT, which is one transformer that takes both text and visual tokens, adding a patch reduced performance noticeably worse when compared to other models. This may indicate a weakness in a single-stream, transformer based approach.

8. Exploring Improving Dual-Stream Only Conceptual Models

Based on our evaluation of these models, we see that cross-attention between modalities improves the learning of conceptual models about objects and actions in a system and the relationships between them. However, a limitation of this approach is its use for downstream tasks. Both ViLT and BridgeTower require image-text pairs of input, making other tasks like image classification *computationally expensive and difficult*. Meanwhile, dual-stream encoders like CLIP and FLAVA allow uni-modal feature representations that can be extracted and used for a variety of downstream tasks. Improving models that do not require paired input would provide greater value and stronger representations. To explore this idea, we fine-tune CLIP on a new dataset inspired by this benchmark called RelComp.

Table 11. **The results for varying the alpha values** for patching [16] finetuned CLIP models on either text encoder (T), visual encoder (V), or both (VT). There is a clear trade-off with downstream ImageNet classification and finetuning on a smaller, compositional and relational focused dataset.

		RelComp			ImageNet	
Stream	alpha	Group Score	Image Score	Text Score	Top1	Top5
v	0.2	31.52	54.67	53.17	**61.45**	**87.73**
v	0.3	32.58	55.61	53.97	58.25	85.69
v	0.4	33.29	56.40	54.89	54.19	82.72
v	0.5	34.15	57.31	55.60	49.36	78.87
v	0.6	34.43	57.62	55.94	44.04	73.96
vt	0.2	42.19	64.30	62.45	54.62	83.05
t	0.2	47.18	67.86	66.62	57.42	85.14
vt	0.3	49.58	69.83	68.00	44.92	73.93
vt	0.4	50.11	70.14	68.78	34.95	63.07
vt	0.5	51.57	71.12	70.35	27.00	52.54
vt	0.6	53.38	72.62	71.77	20.98	43.88
t	0.3	56.31	74.56	73.55	50.69	79.35
t	0.4	62.51	78.57	78.29	44.39	72.39
t	0.5	70.36	83.55	83.21	38.66	66.23
t	0.6	**74.63**	**86.26**	**85.62**	33.54	60.21

Table 12. **The pre-training datasets** include MSCOCO [24], SBU Captions, Localized Narratives (LN), Visual Genome (VG) [21], Wikipedia Image Text (WIT) [40], Conceptual Captions (CC) [37], Conceptual Captions 12M (CC12) [6], Red Caps (RC) [9], YFCC100M [41], and LAION-400M [36].

Model	Params	Datasets	Images	Captions	Arch.	Attn
CLIP RN50 [31]	102M	LAION-400M	400M	400M	dual-stream	modality-specific
CLIP RN101 [31]	121M	LAION-400M	400M	400M	dual-stream	modality-specific
CLIP ViT B16/32 [31]	150M	LAION-400M	400M	400M	dual-stream	modality-specific
CLIP ViT L14 [31]	428M	LAION-400M	400M	400M	dual-stream	modality-specific
FLAVA [38]	358M	MSCOCO, SBU, LN, CC, CC12, VG, WIT, RC, YFCC100M	70M	70M	dual-stream	modality-specific, merged
ViLT [20]	112M	MSCOCO,VG,SBU,CC	4.20M	9.58M	single-stream	modality-specific, merged
Bridgetower [47]	865M	MSCOCO,VG,SBU,CC	4.20M	9.58M	dual-stream	modality-specific, co-attn, merged

8.1. Method

In order to improve CLIP for compositional and relational understanding, we propose using selective negative and positive pairing based on compositional and predicate swaps. We propose using two losses, an image-text matching (ITM) loss and a contrastive loss (C) similar to CLIP [31] and FLAVA [38]. The ITM loss is a triplet loss with two instances [7], maximizing the distance between an anchor and a negative sample while minimizing the distance between an anchor and a positive sample. We use this in order to focus model learning on compositions and relations. The first is where the anchor is the image x, the positive is the caption p, and the negative $\overline{p}$ is the same caption but with either the predicate or the composition swapped. The second uses a real-world caption y as an anchor and the corresponding image x as a positive. The final ITM loss is the average of the two. For the contrastive loss, we maximize the cosine similarities between image and text pairs and minimize those for the image and negative text pairs. We use two versions, the first uses the real-world captions y and their corresponding images, and the second uses the positive text prompts p and their images. The final contrastive loss is the average of the two. A summary of this approach is shown in Figure 7.

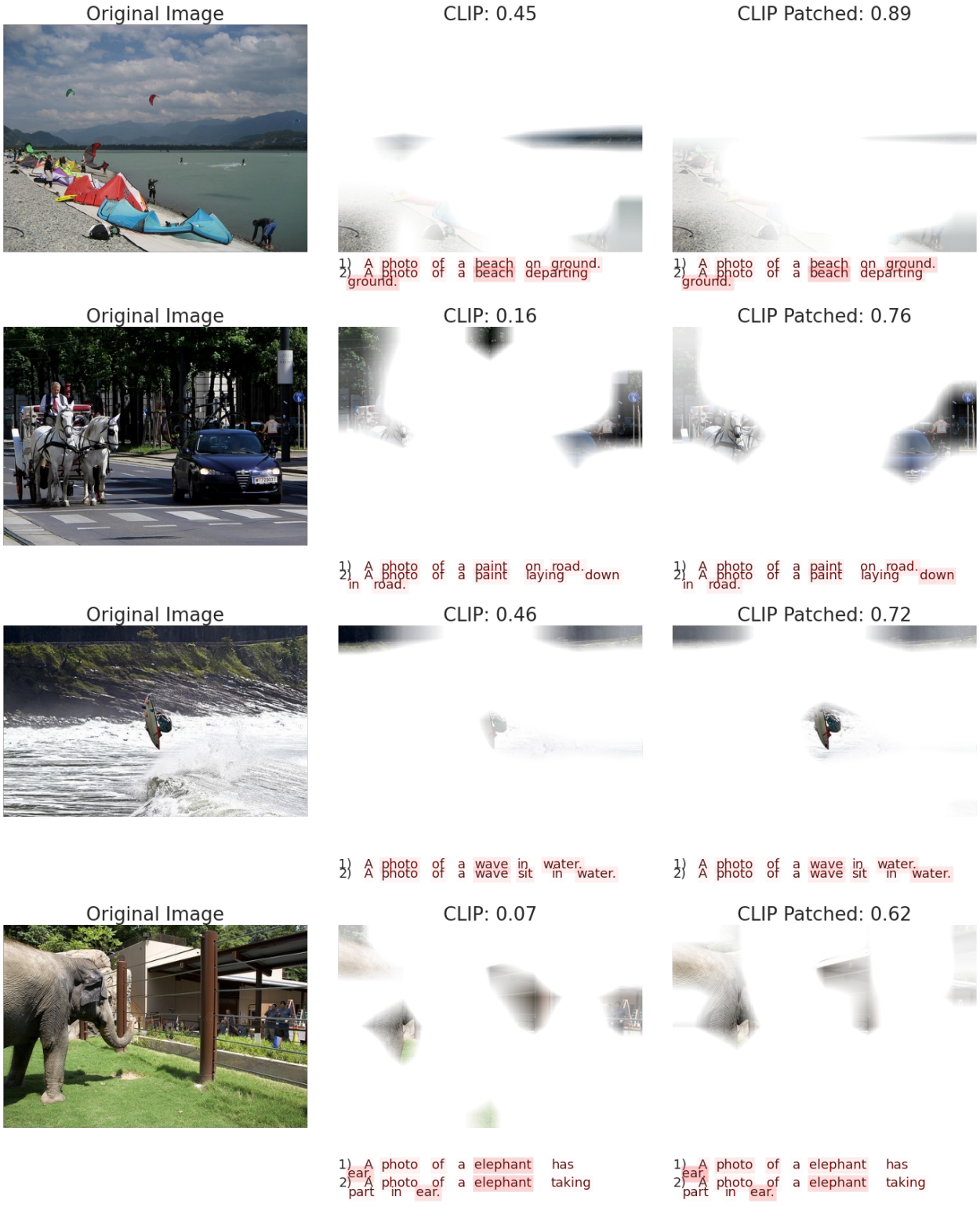

Figure 12. **Examples from Probe-R** comparing CLIP ViT-B/32 to the same model finetuned on RelComp for both the visual and text encoder, then patched [16]. The values are the softmax confidence for the correct prompt P_{R_1} shown as 1) vs the incorrect prompt 2), where the predicate is swapped, or P_{R_2}.

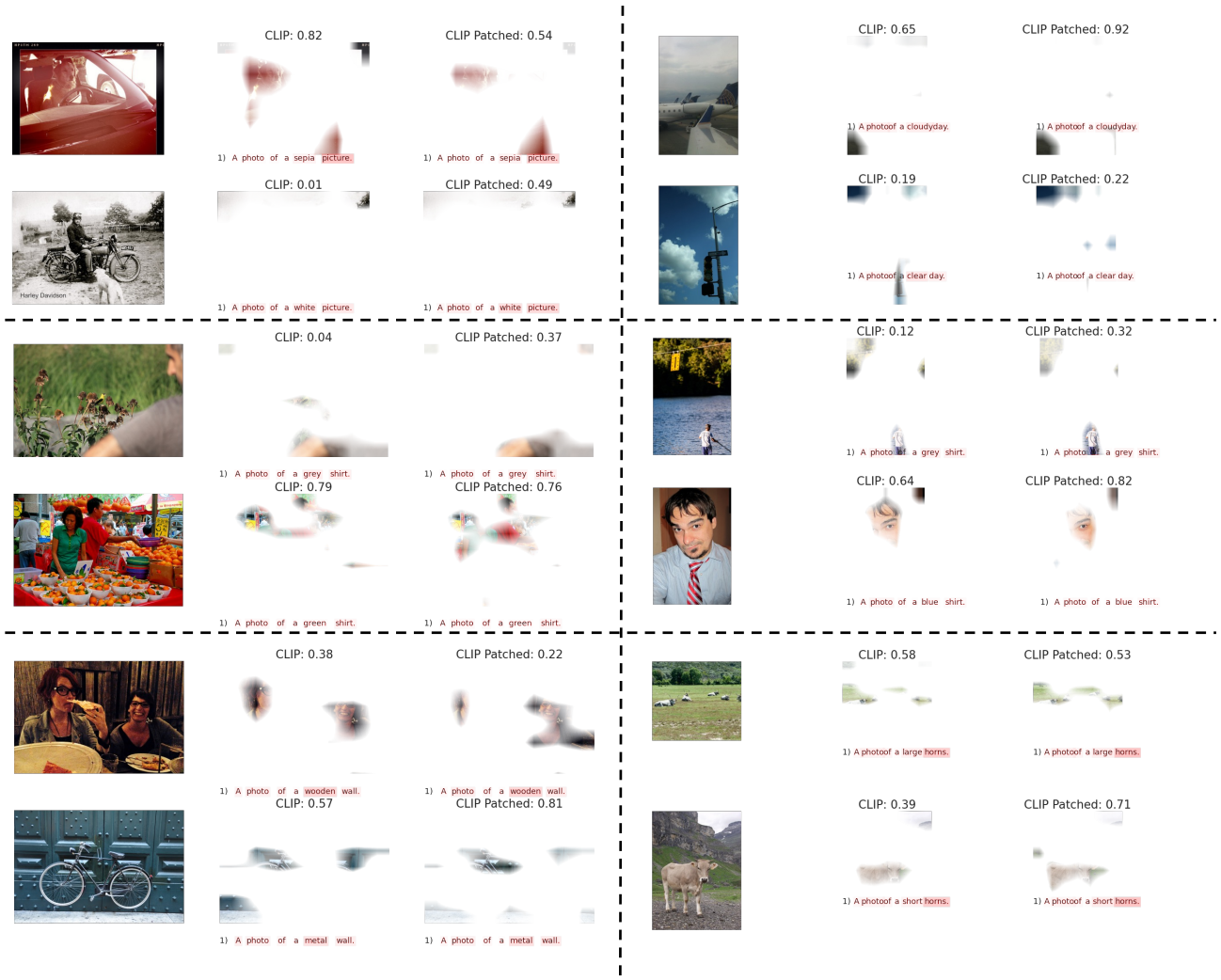

Figure 13. **Examples from Probe-C** comparing CLIP ViT-B/32 to the same model finetuned on RelComp for both the visual and text encoder, then patched [16]. For each group, the first image and its corresponding prompt are on top, and the second image and prompt are on the bottom. The values are the softmax confidence for the corresponding prompt when compared to the alternative prompt.

8.2. Dataset: RelComp

We used our existing knowledge of the benchmark to generate a new training and testing dataset. For compositions, we use images and captions from the MSCOCO dataset [24]. For anchor text we use the real-world caption, for positive we replace all compositions with synonyms, and for negatives we replace all compositions with antonyms. No captions seen in this dataset are also seen in Probe-C. For relations, we use images, region descriptions and relationships from the VisualGenome dataset [21]. For each image, we find the region description that has the most overlap with prompts generated in the same way as Probe-R and use this as our anchor caption. For negative, we use the same template but use prompt with the predicate swapped to an unlikely one, as in Probe-R. To prevent exact prompts from the benchmark being included, we filtered for images that are not present in Probe-R. This results in 149,166 groups with 78,155 of those swapping compositions and 71,011 swapping predicates for training. The test set has 15,836 groups and of those, 8,734 are swap compositions and 7,102 swap predicates.

8.3. Implementation

We finetune the CLIP ViT-B/32 model using our proposed ITM and contrastive loss on the proposed dataset RelComp. We use stochastic gradient descent with a cosine learning rate scheduler with a minumum learning rate of .001, momentum 0.9, weight decay of .0001. We train for 40 epochs using an 11GB GPU and a batch size of 128. We use these smaller configurations to show the benefits with just light

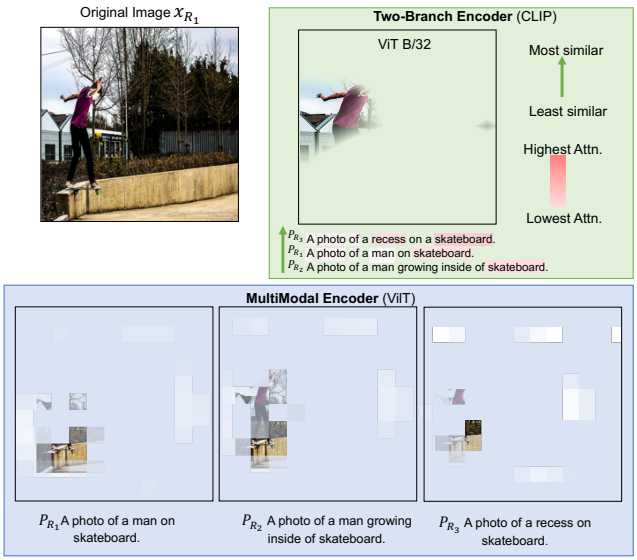

Figure 14. We design probes that measure **relational understanding** in V+L models, in this case we compare ViLT [20] that uses cross-attention (top) and CLIP [31] which does not (bottom). With cross-attention, the model can change its focus based on the prompt and performs better when compositions and relations are swapped for unrealistic/non-present ones. Meanwhile, CLIP does not adapt and focuses more highly on objects, like "man" and "skateboard".

tuning. One of the many challenges of fine-tuning a large model, is that the distribution shift may lead to a loss of the original feature space. In order to prevent this "catastrophic forgetting" of the original feature space, we linearly interpolate the original CLIP weights with our finetuned weights using an alpha= 0.2, leaning more towards the original weights, in order to reduce this shift [16, 44]. This is referred to as patching and therefore we call the finetuned and patched version "CLIP Patched". We finetune three configurations based on which encoders we finetune: visual only (V), text only (T) or both (VT).

8.4. Results

Overall results for our experiment are shown in Table 13. When finetuning on the new dataset, there is an issue of drift from the original CLIP performance as measured by ImageNet accuracy, even when patching. When finetuning using the visual-encoder only, the drift is less pronounced, but so is the improvement on RelComp. The largest increase in RelComp is seen when just training the text encoder. (1) This may indicate that for non-cross-attention models text is more important for conceptual mapping. Overall, (2) our findings indicate that it is possible by using selective negative sampling to enforce compositional and relational learning without extensive co-attention and computational complexity. Limitations of this experiment is our training data

Table 13. **Overall results for finetuning and patching the CLIP ViT-B/32 on the proposed RelComp dataset**. ImageNet accuracy is shown to measure the drift from the original CLIP space. RelComp is the image score for the correct image-to-prompt matching. Probe-C/R are the mean accuracy for the correct image-prompt match. Top scores are in **bold** while second are underlined.

Model	ImageNet	RelComp	Probe-C	Probe-R
ViLT	–	76.00	90.78	69.00
BridgeTower	–	85.00	90.06	82.20
FLAVA	56.83	47.12	83.85	68.29
CLIP ViT B32	**63.60**	51.93	88.15	53.52
CLIP Patched (T)	57.85	**67.85**	89.49	71.14
CLIP Patched (V)	61.45	54.66	89.81	61.40
CLIP Patched (VT)	54.61	64.27	**90.30**	**71.20**

is very small in comparison to recent works, further work should investigate this relationship with a larger dataset with more variation. Table 11 shows the results based on different alphas for RelComp and ImageNet. There is a definite trade-off between original performance and performance on the new task. We also see that training only the text encoder yields the greatest improvement in these tasks but also the largest "forgetting". Some examples of where CLIP patched improved over CLIP in Probe-R is shown in Figure 12. The first column are the original images, the second the attention maps of visual and text features for CLIP ViT-B/32 and the third are the attention maps for CLIP Patched (VT). The values are the softmax confidence for the correct prompt P_{R_1} shown as 1) versus the incorrect prompt P_{R_2} where the predicate is switched 2). Similar examples for Probe-C are shown in Figure 13. For each group, the first image and its corresponding prompt are on top, and the second image and prompt are on the bottom. The values are the softmax confidence for the corresponding prompt when compared to the alternative prompt.

9. Model Details

A summary of the model details can be found in Table 12. The highest performing model is BridgeTower but it also had the largest number of parameters and the slowest. Additionally, BridgeTower utilizes a pre-trained CLIP visual encoder, improving upon CLIPs performance. All models require image-text pairs, making a greater number of comparisons difficult, especially for downstream tasks like image classification on ImageNet where there are 1000 classes. However, because FLAVA merges dual-stream encoder output prior to cross-encoding, it is easier to extract feature embeddings prior to the cross-encoding for a greater number of comparisons. This however does not utilize its full potential for performance. Figure 14 shows examples of how this image-text pair input is a strength for performance in these kinds of tasks. The bottom shows ViLT and how its visual attention changes based on its input while the

top shows CLIP which has consistent attention no matter the text, visual input. Table **??** shows the reported results for the selected models and some CLIP models on the MSCOCO [24] and Flicker [48] datasets. We do see correlation between performance on these datasets and performance on the proposed datasets in this benchmark. This indicates that retrieval tasks on datasets like MSCOCO may be a good indicator of "understanding" at a high-level. Code to run these models is

10. Dataset Labelling/Preprocessing/Cleaning

Probe-R This dataset is built off of Visual Genome [21]. This dataset was created by cleaning the annotations/relationship aliases with relations that are specifically an interaction rather than an attribute which was often an erroneous annotation and grouping relations that are the same despite spelling errors. Objects and predicates are additionally cleaned based on spelling errors. Using the extracted (subject, predicate, object) triplets, unlikely relationships are determined if there is no existing combination of an object-predicate pair or subject-object pair. For each image, the ground truth relation is compared to a highly unlikely swap of subject and predicate. There are set of "positive" images that are images with the subject being swapped and no other objects from the original image. There are also a set of "negative" images that are images with the swapped subject and no other objects from the original image. The predicate swapped is based on the predicates that have not been found in the original dataset to be associated with the original subject and therefore are highly unlikely.

Probe-C This dataset is built of the COCO Validation 2014 [24] dataset. Using the NLP library NLTK [4] and the COCO caption annotations, words are tagged and pairs of adjective and nouns are extracted. These pairs are then manually cleaned to ensure the attribute is indeed an adjective and the object is indeed an object. Instead of using unlikely combinations, antonyms were manually mapped to each attribute in order to ensure that the attribute is not present in the image. For example, if there is a "a small dog", the comparison prompt is "a large dog". There are two splits for this dataset. The first is where the composition is swapped with an antonym and the other is where an object is switched. Each dataset has two images and two captions and comparisons are based on how well the model can match the captions to the correct images.

Probe-B This dataset is built off COCO Validation 2014 [24] dataset. Segmentation annotations from the original COCO dataset were used for removing background and/or objects. For set 1, for each image, it removes all other objects using either the segmentation(retaining shape cues). These are replaced with either "black", "gray", "scene" or "noise". Images are chosen if they do not have overlapping bounding boxes and if their object area is over a threshold to allow for better visibility, making the task easier. "Scene" fillers are extracted from landscape scenery from a subset of the Kaggle Landscape dataset [33] and Indoor Scenes Dataset [30]. The subset of 290 scene filler images were selected based on whether there were any objects in the image that are also in the annotations.