# The Hard Positive Truth about Vision-Language Compositionality: Supplementary

Amita Kamath[1,2], Cheng-Yu Hsieh[1], Kai-Wei Chang[2], and Ranjay Krishna[1,3]

[1] University of Washington
[2] University of California, Los Angeles
[3] Allen Institute for AI
https://github.com/amitakamath/hard_positives

## A Related work

We contextualize our study within research aiming to improve the compositionality of vision-language models.

**Benchmarks for vision-language compositionality.** There has been a surge of benchmarks to assess how well vision-language models represent compositional concepts [15,20,29,36,47,55,61]. These tools often reveal that, despite achieving impressive results in various applications [1, 27, 35, 43, 50, 51, 60], these models struggle with basic compositional tasks. Issues include difficulty in processing sentences with the same words in a different order [47], and in recognizing relationships between objects or associating objects with their attributes [2, 15, 36, 55, 61]. Benchmarks also reveal that many models struggle with spatial reasoning [12, 21, 34, 56]. Our evaluation dataset complements these benchmarks by introducing the notion of hard positives which allows us to uncover that hard negative finetuning induces behaviors that bring into question their semantic understanding of concepts.

**Hard negative finetuning for compositionality.** Efforts to bolster the compositional capabilities of vision-language models have introduced strategies that incorporate new data, methodologies, and loss functions [3, 6, 36, 44, 55]. A key strategy involves training models to differentiate between correct captions and procedurally-generated hard negatives [5, 6, 55]. However, it remains uncertain whether these approaches genuinely foster a deeper understanding of compositionality or merely enable models to perform well on dataset biases [15]. Our study explores this question to provide evidence that models do in fact *appear* to perform better on existing benchmarks, but produce the undesirable side effect of being overly sensitive even to semantic-preserving perturbations.

**Mitigating biases in datasets.** The challenge of biased datasets, which can artificially inflate the perceived effectiveness of models, has been well-documented [10].

Several studies propose methods for de-biasing these datasets to ensure evaluations more accurately reflect model capabilities [24, 37, 39, 56]. Techniques like adversarial filtering [56] use a set of classifiers to eliminate easily guessable instances, creating a tougher benchmark. AFLite builds on this by offering a simplified approach to filtering without needing iterative model retraining, leading to benchmarks that more closely align with the intended tasks [24, 39]. In the context of vision-language compositionality evaluation, SugarCrepe identifies and fixes several textual biases exhibiting in procedurally-generated hard negatives in prior benchmarks, yet it only uses hard negatives as in prior benchmarks [15]. We complement these benchmarks by introducing hard positives to allow a comprehensive evaluation of vision-language models' compositionality.

**Augmenting model training with rewritten captions.** In addition to hard negative mining, several recent works have explored augmenting data with caption-rewriting methods to improve vision-language models' performance [5–7]. These works typically utilize large language models [32, 52] to rewrite a given caption into a very different, new caption describing the same scene, in the hope that the generated captions enrich language supervision for model learning. In this work, we show that even by augmenting model training with the rewritten *positive* captions, the oversensitivity introduced by hard negative finetuning [5,6] is so dire that models still fail to correctly identify hard positives from negatives. However, we show that by training with *hard* positives, we are able to better mitigate models' oversensitivity issue.

## B Additional Benchmark Details

This section contains further details about the creation of the `REPLACE` benchmark, as well as a random sample of both benchmarks.

### B.1 Further details about `REPLACE`

This dataset consists of hard negatives selected from VL-Checklist [61] where one word or phrase in the caption is replaced with another in a way that changes the meaning of the caption, and hard positives we create where we replace one word or phrase in the caption with another in a way that does *not* change the meaning of the caption. As discussed in Section 2.4, we focus on the VL-Checklist hard negatives that target relations and attributes, as they are more challenging for models to understand. Additionally, we ignore objects because their replacements in VL-Checklist are not very targeted to be similar to the original object (e.g., positive: "train has wheels", negative: "stir fry"), as the object class from which the hard negatives are created (all objects) is much broader than the relation or attribute classes (e.g., spatial relations, colors). We thus focus on relations and attributes, which have much harder hard negatives. We select the Visual Genome [23] subset of VL-Checklist to stay consistent with the `SWAP` benchmark, which is sourced from the same dataset.

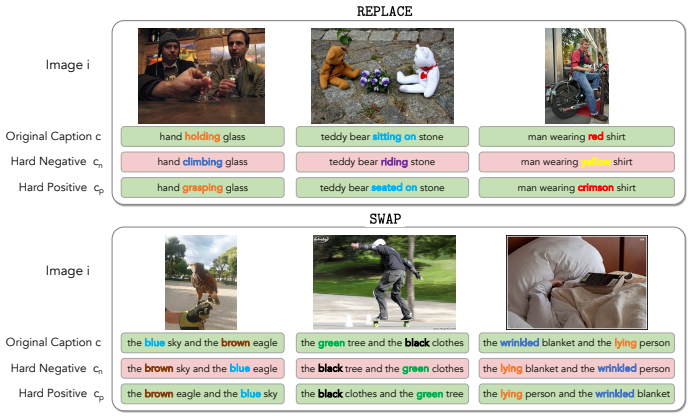


**Fig. 1:** Random samples of REPLACE and SWAP. The first two REPLACE samples are from Relations, and the third from Attributes.

The VL-Checklist Relations benchmark has two types of relations: actions and spatial. The VL-Checklist Attributes benchmark has five types of relations: action, color, material, size, and state. As discussed in Section 2.4, for each of these types, we collect the ten most common relations/attributes, and hand-write a fixed replacement that holds for the various word senses of each original word. If no replacement can be found, we discard the sample. Finally, we replace 14 relations and 24 attributes, resulting in a benchmark of 16,868 hard positives targeting relations, and 10,575 hard positives targeting attributes, for a total of 27,443 examples.

The replaced relations and attributes, their replacements, their frequency in the benchmark, and an example caption containing each is provided in Tables 1, 2 and 3.

### B.2 Random samples of REPLACE and SWAP

Figure 1 contains random samples of REPLACE-Relations, REPLACE-Attributes and SWAP. As the benchmarks are created from Visual Genome region annotations, they occasionally only discuss a part of the image; however, the hard negative captions are created such that they are always a mismatch for the corresponding image — i.e., they do not satisfy any part of the image [55, 61].

## C Additional Results

This section contains additional results, splitting the REPLACE results in the main paper into the separate Relations and Attributes subsets (Table 4), as well as the results of various other models on our benchmarks: varying model size, architecture, pretraining data, and training objective (Table 5). We also explain

the Random Chance and Human Performance numbers in the main paper.

**Random Chance Performance.** For Original Test Accuracy, random chance is 50%, as there are only two possible rankings for the two captions (original and hard negative). For Augmented Test Accuracy, random chance is 33.3%, as two of six possible rankings for the three captions (original, hard negative and hard positive) satisfy the condition: $s(c|i) > s(c_n|i)$ and $s(c_p|i) > s(c_n|i)$. For Brittleness, random chance is again 33.3%, as two of six possible rankings for the three captions satisfy the condition: $s(c|i) > s(c_n|i) > s(c_p|i)$ or $s(c_p|i) > s(c_n|i) > s(c|i)$.

**Human Performance.** The errors in human performance on `REPLACE` arise from noise caused by errors in the underlying hard negative annotation (e.g., VL-Checklist containing a hard negative caption that is still a match for the image) or Visual Genome annotation (e.g., an incorrect region caption).

**Replacing relations vs replacing attributes.** Table 4 contains the results for the models in the main paper, split across `REPLACE` Relations and `REPLACE` Attributes. It is clear that model performance is worse on Relations, likely because relations are more challenging than attributes for models to understand — following simple combinatorial logic, it is more likely that within one training batch, the same *object* appears twice with different attributes, than that the same *pair of objects* appears twice with different relations between them. This contributes towards why contrastively trained models are more likely to understand attributes than relations.

Following a similar trend, our model finetuned on both hard positives and hard negatives performs extremely well on `REPLACE`-Attributes (more so than on `REPLACE`-Relations), achieving high Augmented Test Accuracy and low Brittleness — in fact, the drop from Original Accuracy is only 6.7 points, almost four times lower than the average drop of 24.7 points across models from existing work.

**Changing CLIP model size.** From Table 5(b), it is clear that increasing the model size of CLIP does not necessarily improve its performance on our benchmarks — there is no clear pattern in the results of various models.

**Changing CLIP text encoder.** From Table 5(c), we see the effect of using pretrained RoBERTa weights in the CLIP text encoder. The model performance is fair for `REPLACE`, but very poor for `SWAP`— likely due to the fact that only the word order changes across all three captions, and masked language models have been shown to struggle with word order.

**Changing CLIP pretraining data.** From Table 5(d), DataComp [8] seems to hurt model performance, more so on `REPLACE` than on `SWAP`.

**Changing CLIP vision encoder.** From Table 5(e), we see that replacing the ViT vision encoder with a ResNet-based vision encoder seems to improve performance slightly, in the case of the RN50 models.

**Comparing CLIP to XVLM [58].** Table 5(f) shows the performance of XVLM-16M (pretrained) on our benchmarks, as it has been shown to perform well on hard negative-focused benchmarks [2]. At first glance, the performance is shockingly high compared to CLIP — however, it is important to note that XVLM is trained on Visual Genome region captions, from which all of our benchmarks are sourced. It is possible that there is data leakage, as the XVLM training data was curated to prevent leakage with popular test sets *at the time*, and predates ARO [55] and VL-Checklist [61], from which our benchmarks are sourced. This may also explain the results of [2].

## D Hard Positive Training Data Generation Details

In this section, we discuss the details of generating hard positive training data. First, we discuss the prompts used to generate data from the LLM LLAMA2 [48]. Then, we discuss the implementation details of the generation. Finally, we provide a random sample of the data generated using the prompts.

### D.1 Prompts

The prompt for `REPLACE` is:

```
Replace one word in this sentence with a synonym, without changing the
meaning of the sentence. Only output the changed sentence.

{example}
```

The prompt for `SWAP` is:

```
Swap the words around the word "and" in a sentence without changing the
meaning. Only respond with the changed sentence.

Input: three giraffes and two antelope
Output: two antelopes and three giraffes

Input: a blue and white stained glass clock shows the time
Output: a white and blue stained glass clock shows the time

Input: a mixture of rice and broccoli are put together
Output: a mixture of broccoli and rice are put together

Input: a bathroom with a sink, toilet and shower
Output: a bathroom with a sink, shower and toilet

Input: there is a man wearing glasses and holding a wine bottle
Output: there is a man holding a wine bottle and wearing glasses

Input: {example}

Output:
```

We arrived at the examples in the SWAP prompt by looking at patterns of common mistakes in the LLM outputs. No such examples were needed for REPLACE, as it appears to be an easier task, e.g., not requiring correct dependency parsing of text inputs, which can be potentially ungrammatical captions.

### D.2 Implementation details

We generate hard positive training data by feeding the above prompt to the LLAMA2 70B-Chat model [48]. The examples are sourced from COCO train (note: Hard negatives are generated from COCO train as well, following the CREPE [29] procedure). SWAP hard positives are created for COCO train captions containing the word "and" and less than 15 words, which amounts to 119,071 captions, and REPLACE hard positives are created for all 591,753 COCO train captions. In total, we generate 710,824 hard positives — although we subsample these during finetuning, as discussed in Section E.1.

We run inference on LLAMA2 with Flash Attention on a batch size of 32, on 4xA100s, which takes 36 hours to generate all hard positives (we parallelize this across 8 similar machines). For SWAP we set the maximum number of generated tokens to 20 (as we filter out captions of greater than 15 words), and for REPLACE we set it to 30 (as we do no such filtering).

Note: We considered using Spacy to get dependency parses of the sentences and write code to perform the swapping, but Spacy fails often on COCO image captions, which are often only noun phrases (e.g., "a person on a brown horse") or ungrammatical. Thus, we used an LLM instead, which had almost perfect performance in swapping sentences from a random sample of 100 inputs we went through manually.

### D.3 Random sample of generated data

Below is a random sample of the generated data for SWAP:

```
A cabinet setting with green vases and a wooden backboard →
A cabinet setting with a wooden backboard and green vases

A couch and a television in a room →
A television and a couch in a room

An older gentleman in a white shirt and black bow tie →
An older gentleman in a black bow tie and white shirt

Two giraffes standing next to one another with trees and bushes near them →
Two giraffes standing next to one another with bushes and trees near them

a lady wearing snow skis and a man holding snow skis →
a man holding snow skis and a lady wearing snow skis

An adorable little girl wearing sunglasses and holding a stack of frisbee →
An adorable little girl holding a stack of frisbee and wearing sunglass
```

Below is a random sample of the generated data for REPLACE:

```
a person holding an piece of an eaten sandhwich next to a lap top computer →
```

```
a person holding a morsel of a devoured sandwich next to a portable computer

Two baby goats stand together on worn stones →
Two baby kids stand together on worn rocks

a field that ha a bunch of sheep in it →
a meadow that has a flock of sheep in it

A side view mirror on the handle bars of a motorcycle →
A side view mirror on the handle bars of a motorbike

A variety of vegetables sits in a pile on a stand →
A collection of vegetables sits in a pile on a stand

a man going down a handle on some stairs on a skate board →
a man going down a rail on some stairs on a skate board
```

We notice that the LLM frequently changes grammatical errors if present in the original caption when generating the hard positive caption, e.g., "a field that *ha* ..." → "a meadow that *has* ...".

We also notice that, while generating `REPLACE` hard positives, the LLM tends to replace the objects ("field" → "meadow"), more than the attributes ("eaten" → "devoured"), more than the relations (none in this sample) — which we hypothesized may be the reason our finetuned model performs better on `REPLACE` Attributes than Relations (c.f. Table 4). We separately generate more relation-targeted hard positives (with separate prompts to replace verbs and spatial prepositions), then sampling an equal number for relations and attributes, but the results when finetuning a model on this data did not differ significantly from those of our earlier finetuned model. Further study is required to improve model performance on `REPLACE` Relations.

## E Finetuning on Hard Positives and Hard Negatives

### E.1 Implementation details

The finetuning follows the procedure outlined in SVLC [6]. For each training sample, one hard positive and one hard negative is retrieved and added to the batch. The loss consists of: a contrastive loss across the batch, as in CLIP; a hard negative loss on each image with its original and negative captions; and a hard positive loss (called an analogy loss in SVLC) on each image with its original and positive captions. We finetune the model for 5 epochs on 4xA100 GPUs, which takes approximately 3 hours.

### E.2 Finetuning on both hard positives and hard negatives prevents reduction in model score of original caption

As discussed in Section 3.2 and 4.4, hard negative finetuning causes the model to award a lower score to *all* captions, not the hard negative caption alone. This has negative implications for various use cases where the image-text matching score of the model is used directly, rather than as a ranking mechanism.

Table 6 shows the mean score awarded to the original caption $c$ by CLIP as well as various hard-negative finetuned models, showing that they all reduce the score of $c$ across both `REPLACE` and `SWAP` (by 0.031 on average). In comparison, our model, finetuned on both hard positives and hard negatives, reduces the score of the original caption much less (by 0.006 on average) than all models except CREPE-Swap. CREPE-Swap assigns a higher score to $c$, but also an incorrectly higher score to $c_N$, resulting in much worse performance than our model on `SWAP` and `REPLACE` (c.f. Table 1). Our model strikes the best balance of high benchmark performance without significantly reducing the image-text matching score of the original caption.

## F Standard Evaluations

We conduct standard evaluations of our model on vision and vision-language tasks to ensure that our model did not experience catastrophic forgetting during finetuning. Table 7 contains the results of our models evaluated on a wide range of zero-shot tasks. Specifically, we include zero-shot classification results on ImageNet-1K and 20 different VTAB tasks [59], as well as zero-shot retrieval performances on COCO and Flickr30k. We include a CLIP model without finetuning, and a CLIP model finetuned on COCO alone (without hard positives or hard negatives) to serve as controlled baselines.

**Zero-shot classification performance drops.** From Table 7, we see that the models finetuned on the COCO training set show significant performance gains on COCO and Flickr30k retrieval, while losing performance on ImageNet-1K and VTAB classification tasks. This observation agrees with prior work [54], which shows that finetuning can decrease the robustness of CLIP models, particularly on different domains. Various methods have been proposed to effectively tackle the problem [53, 54], and are orthogonal to this work.

**Adding hard positives improves compositionality while maintaining robustness, compared to training only with hard negatives.** Comparing finetuning with hard positives and hard negatives to finetuning with hard negatives alone (as well as the COCO finetuning baseline with neither hard positives nor hard negatives), we see that adding hard positives to finetuning largely maintains the model's robustness on standard tasks while achieving significant improvements on compositionality.

| Orig. Rel. | Replaced Rel. | Freq. | Example |
|---|---|---|---|
| in | within | 6173 | O: white horse in field<br>HP: white horse within field<br>HN: white horse out of field |
| behind | to the rear of | 1057 | O: van behind truck<br>HP: van to the rear of truck<br>HN: van in front of truck |
| on top of | on | 683 | O: dishes on top of table<br>HP: dishes on table<br>HN: dishes below table |
| near | next to | 657 | O: deck near water<br>HP: deck next to water<br>HN: deck far from water |
| next to | near | 621 | O: person next to train<br>HP: person near train<br>HN: person far from train |
| under | beneath | 467 | O: street under animals<br>HP: street beneath animals<br>HN: street above animals |
| by | near | 394 | O: road by building<br>HP: road near building<br>HN: road far from building |
| above | on top of | 298 | O: cloud above hill<br>HP: cloud on top of hill<br>HN: cloud below hill |
| wearing, wears | in | 3976 | O: man wearing shirt<br>HP: man in shirt<br>HN: man hugging shirt |
| holding | grasping | 950 | O: woman holding fork<br>HP: woman grasping fork<br>HN: woman helping fork |
| sitting | seated | 639 | O: cow sitting next to man<br>HP: cow seated next to man<br>HN: cow chasing man |
| hanging | dangling | 382 | O: banner hanging from building<br>HP: banner dangling from building<br>HN: banner driving building |
| walking | strolling | 288 | O: man walking on beach<br>HP: man strolling on beach<br>HN: man enclosing beach |
| riding on | traveling on | 283 | O: person riding motorcycle<br>HP: person traveling on motorcycle<br>HN: person herding motorcycle |

**Table 1:** Benchmark details of `REPLACE` Relations, which consist of spatial relations and transitive actions. `O`, `HP` and `HN` denote the Original, Hard Positive and Hard Negative captions respectively, randomly sampled from each relation.

| Orig. Att. | Replaced Att. | Freq. | Example |
|---|---|---|---|
| standing | upright | 153 | O: turned head of a standing person<br>HP: turned head of a upright person<br>HN: turned head of a sitting person |
| sitting | seated | 88 | O: sitting man<br>HP: seated man<br>HN: crouching man |
| walking | strolling | 64 | O: foot of walking man<br>HP: foot of strolling man<br>HN: foot of lying man |
| eating | ingesting | 41 | O: eating woman<br>HP: ingesting woman<br>HN: driving woman |
| hanging | dangling | 29 | O: hanging branch<br>HP: dangling branch<br>HN: looking up branch |
| looking | gazing | 27 | O: looking elephant<br>HP: gazing elephant<br>HN: playing elephant |
| white | ivory | 2742 | O: white toilet<br>HP: ivory toilet<br>HN: orange toilet |
| black | ebony | 1790 | O: black socks<br>HP: ebony socks<br>HN: dark brown socks |
| blue | sapphire | 1253 | O: lady wearing blue shirt<br>HP: lady wearing sapphire shirt<br>HN: lady wearing yellow shirt |
| brown | chestnut | 947 | O: edge of brown beach<br>HP: edge of chestnut beach<br>HN: edge of purple beach |
| red | crimson | 827 | O: red glove<br>HP: crimson glove<br>HN: blue glove |
| green | emerald | 755 | O: cooler has green lid<br>HP: cooler has emerald lid<br>HN: cooler has dark blue lid |
| silver | metallic | 242 | O: silver fork<br>HP: metallic fork<br>HN: light brown fork |

**Table 2:** Benchmark details of `REPLACE` Attributes (Part I, split due to space constraints), which consist of intransitive actions and colors. `O`, `HP` and `HN` denote the Original, Hard Positive and Hard Negative captions respectively, randomly sampled from each attribute.

| Orig. Att. | Replaced Att. | Freq. | Example |
|---|---|---|---|
| large | big | 571 | O: tire on large truck<br>HP: tire on big truck<br>HN: tire on tiny truck |
| small | tiny | 358 | O: toilet inside small bathroom<br>HP: toilet inside tiny bathroom<br>HN: toilet inside huge bathroom |
| long | lengthy | 271 | O: person carrying a long skateboard<br>HP: person carrying a lengthy skateboard<br>HN: person carrying a short skateboard |
| big | large | 146 | O: big elephant<br>HP: large elephant<br>HN: tiny elephant |
| huge | big | 31 | O: kites under huge sky<br>HP: kites under big sky<br>HN: kites under tiny sky |
| wet | damp | 62 | O: wet road<br>HP: damp road<br>HN: cloudless road |
| smiling | happy | 50 | O: snowboard with smiling man<br>HP: snowboard with happy man<br>HN: snowboard with sad man |
| old | aged | 46 | O: old train<br>HP: aged train<br>HN: young train |
| clear | unclouded | 43 | O: clear sky<br>HP: unclouded sky<br>HN: partly cloudy sky |
| young | youthful | 36 | O: shoes on young man<br>HP: shoes on youthful man<br>HN: shoes on unhappy man |

**Table 3:** Benchmark details of REPLACE Attributes (Part II, split due to space constraints), which consist of sizes and states. The fifth attribute, material, had no synonyms for each word (e.g., "brick"), so we discard it. O, HP and HN denote the Original, Hard Positive and Hard Negative captions respectively, randomly sampled from each attribute.

| | Model | REPLACE-Rel Orig. Test Acc. | REPLACE-Rel Aug. Test Acc. | REPLACE-Att Orig. Test Acc. | REPLACE-Att Aug. Test Acc. | REPLACE-Rel Brittleness (↓) | REPLACE-Att Brittleness(↓) |
|---|---|---|---|---|---|---|---|
| (a) | CLIP ViT-B/32 | 57.6 | 45.3 (-12.3) | 68.1 | 49.0 (-19.1) | 21.7 | 25.5 |
| (b) | NegCLIP | 65.6 | 48.2 (-17.4) | 73.4 | 58.2 (-15.2) | 22.3 | 20.3 |
| | CREPE-Swap | 56.6 | 43.0 (-13.7) | 74.4 | **62.2** (-12.1) | **21.2** | 17.6 |
| | CREPE-Replace | 70.5 | 49.4 (-21.1) | 78.8 | 61.1 (-17.7) | 25.3 | 21.6 |
| | SVLC | 72.0 | 42.1 (-29.9) | 83.8 | 48.2 (-35.6) | 41.6 | **37.3** |
| | SVLC+Pos | 62.1 | 44.7 (-17.4) | 68.0 | 45.6 (-22.4) | 30.3 | 29.0 |
| | DAC-LLM | 88.1 | 51.5 (-36.6) | 86.8 | 44.9 (-41.9) | 38.4 | 42.7 |
| | DAC-SAM | 89.2 | **59.6** (-29.5) | 86.9 | 55.9 (-31.0) | 31.2 | 32.5 |
| (c) | Our HN | 71.6 | **52.6** (-19.0) | 77.5 | 60.8 (-16.8) | 23.5 | 21.0 |
| | Our HP+HN | 65.5 | 51.9 (-13.6) | 74.5 | **67.7** (-6.7) | **19.9** | **12.2** |
| (d) | Our HP+HN (Swap-only) | 57.0 | 44.4 (-12.6) | 75.1 | 63.1 (-11.9) | **19.4** | 17.2 |
| | Our HP+HN (Replace-only) | 68.8 | **53.7** (-15.1) | 74.2 | **67.3** (-6.8) | 21.0 | **12.7** |
| | Random Chance | 50.0 | 33.3 | 50.0 | 33.3 | 33.3 | 33.3 |
| | Human Estimate | 97 | 97 | 100 | 100 | 0 | 0 |

**Table 4:** Detailed results of various ITM models on our REPLACE benchmark: (a) CLIP, (b) Hard-Negative finetuned versions of CLIP from previous work (Section 3.2), (c) Our improved model (Section 4.2). The purple cells indicate the models have seen perturbations of the type we are testing for during finetuning, blue cells indicate otherwise. We report performance on the Relations and Attributes subsets of REPLACE separately here; they are averaged in the main paper for brevity.

| | Model | REPLACE Orig. Test Acc. | REPLACE Aug. Test Acc. | SWAP Orig. Test Acc. | SWAP Aug. Test Acc. | REPLACE Brittleness (↓) | SWAP Brittleness(↓) |
|---|---|---|---|---|---|---|---|
| (a) | CLIP ViT-B/32 | 61.6 | 46.8 (-14.9) | 60.5 | 49.6 (-10.9) | 23.2 | 21.7 |
| (b) | CLIP ViT-B/16 | 61.8 | 45.0 (-16.8) | 61.1 | 51.1 (-10.0) | 24.8 | 19.8 |
| | CLIP ViT-L/14 | 64.2 | 48.4 (-15.8) | 61.1 | 49.9 (-11.2) | 24.0 | 21.9 |
| | OpenCLIP ViT-H/14 | 56.5 | 43.7 (-12.8) | 62.9 | 51.7 (-11.2) | 20.5 | 21.7 |
| | OpenCLIP ViT-g/14 | 59.5 | 45.8 (-13.7) | 63.5 | 52.1 (-11.4) | 22.2 | 22.4 |
| | OpenCLIP ViT-G/14 | 58.6 | 44.4 (-14.2) | 61.9 | 50.5 (-11.3) | 22.9 | 22.4 |
| (c) | RoBERTa-CLIP ViT-B/32 | 57.5 | 44.3 (-13.3) | 48.7 | 29.4 (-19.3) | 28.7 | 40.3 |
| (d) | DataComp-CLIP ViT-B/32 | 53.0 | 42.4 (-10.6) | 58.5 | 44.8 (-13.7) | 21.2 | 27.1 |
| | DataComp-CLIP ViT-B/16 | 51.7 | 40.8 (-10.9) | 56.8 | 43.6 (-13.2) | 21.5 | 26.5 |
| | DataComp-CLIP ViT-L/14 | 55.7 | 42.7 (-13.1) | 60.0 | 47.6 (-12.4) | 22.0 | 24.2 |
| (e) | CLIP-RN50x16 | 63.2 | 45.8 (-17.5) | 62.2 | 51.9 (-10.3) | 24.9 | 20.0 |
| | CLIP-RN50x64 | 66.3 | 49.2 (-17.1) | 62.2 | 51.3 (-10.9) | 25.4 | 21.1 |
| | CLIP-RN101 | 58.3 | 43.9 (-14.4) | 61.9 | 52.0 (-9.9) | 23.2 | 19.3 |
| (f) | XVLM-16M* | 72.9 | 63.8 (-9.1) | 89.3 | 84.8 (-4.5) | 16.3 | 8.1 |
| | Random Chance | 50.0 | 33.3 | 50.0 | 33.3 | 33.3 | 33.3 |
| | Human Estimate | 97 | 97 | 100 | 100 | 0 | 0 |

**Table 5:** Results of additional ITM models on our benchmark: (a) CLIP, (b) Different model sizes of CLIP, (c) CLIP where the text encoder is initialized with RoBERTa-pretrained weights, (d) CLIP trained on DataComp [8] rather than WIT [35] or LAION [40], (e) CLIP with different vision encoders, (f) XVLM*. The * on XVLM depicts that it is not a fair comparison with the other models, as XVLM is trained specifically on VG region captions, from which our benchmarks are sourced. REPLACE averages performance on Attributes and Relations.

| Mean $c$ Score | CLIP | Neg-CLIP | CREPE -Swap | CREPE -Repl. | SVLC | SVLC +Pos | DAC -LLM | DAC -SAM | Ours |
|---|---|---|---|---|---|---|---|---|---|
| `REPL.` | 0.234 | 0.225 | 0.233 | 0.214 | 0.202 | 0.223 | 0.157 | 0.228 | 0.231 |
| `SWAP` | 0.255 | 0.239 | 0.250 | 0.228 | 0.211 | 0.228 | 0.132 | 0.224 | 0.247 |

**Table 6:** Mean image-text matching score of original caption $c$ per benchmark of all evaluated models. All hard negative-finetuned models reduce the image-text matching score of $c$, nearly all more so than our model finetuned on both hard negatives and hard positives.

| | | ImageNet1k | | COCO | | Flickr30k | | VTAB | |
|---|---|---|---|---|---|---|---|---|---|
| | Model | Acc@1 | Acc@5 | Image Recall@1 | Text Recall@1 | Image Recall@1 | Text Recall@1 | Acc@1 | Acc@5 |
| (a) | CLIP ViT-B/32 | 63.33 | 88.83 | 30.46 | 50.14 | 58.82 | 77.40 | 39.00 | 70.90 |
| (b) | CLIP-COCO | 53.18 | 81.98 | 50.34 | 66.76 | 68.48 | 83.40 | 34.67 | 68.55 |
| (c) | Our HN | 50.40 | 79.58 | 49.61 | 63.98 | 67.80 | 80.10 | 32.40 | 67.53 |
| | Our HP+HN | 49.85 | 79.70 | 49.67 | 65.02 | 67.52 | 80.60 | 33.24 | 67.75 |

**Table 7:** Evaluation results on standard zero-shot tasks of (a) CLIP ViT-B/32, (b) CLIP ViT-B/32 finetuned on COCO train captions with neither hard positives nor hard negatives, (c) Our models. We report Acc@1 and Acc@5 for zero-shot classification on ImageNet1k and VTAB. For VTAB, we report the average over 20 zero-shot classification tasks [18, 59]. For COCO and Flicker30k, we report Recall@1 for both image and text retrieval. Comparing training with both hard positives and hard negatives ("Our HP + HN") to training with hard negatives alone ("Our HN"), we see that we maintain — or even improve — performance on standard evaluation tasks, while improving model compositionality (c.f. Table 1).