The Hard Positive Truth about Vision-Language Compositionality

Amita Kamath[1,2] **Cheng-Yu Hsieh**[1] **Kai-Wei Chang**[2] **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

Image i

Existing work

		Captions	CLIP	Hard Negative Finetuned	Ours
Original Caption	c	brown grass	0.236	0.152	0.240
Hard Negative	c_n	blue grass	0.240	0.143	0.231
Hard Positive	c_p	chestnut grass	0.249	0.134	0.241

Our work

Figure 1: Prior work shows that CLIP is insensitive to minor changes to the input caption, incorrectly assigning a higher score to a hard negative caption c_n than to the original caption c. While hard negative finetuning (here, Doveh et al. (2023a)) fixes the ordering between the original caption and the hard negative, we reveal that the resulting model becomes oversensitive and incorrectly assigns a lower score to a hard *positive* caption c_p. We mitigate this by finetuning with both hard negatives and hard positives, leading to an overall correct understanding of the different captions, and achieving a more well-rounded sense of compositionality (real example shown).

Abstract

Several benchmarks[1] have concluded that our best vision-language models (*e.g.*, CLIP) are lacking in compositionality. Given an image, these benchmarks probe a model's ability to identify its associated caption amongst a set of compositional distractors. In response, a surge of recent proposals show improvements by finetuning CLIP with distractors as **hard negatives**. Our investigations reveal that these improvements have, in fact, been significantly overstated — because existing benchmarks do not probe whether finetuned vision-language models remain invariant to **hard positives**. By curating an evaluation dataset with $112,382$ hard negatives and hard positives, we uncover that including hard positives decreases CLIP's performance by 12.9%, while humans perform effortlessly at 99%. CLIP finetuned with hard negatives results in an even larger decrease, up to 38.7%. With this finding, we then produce a 1,775,259 image-text training set with both hard negative and hard positive captions. By training with both, we see improvements on existing benchmarks while simultaneously improving performance on hard positives, indicating a more robust improvement in compositionality. Our work suggests the need for future research to rigorously test and improve CLIP's understanding of semantic relationships between related "positive" concepts.

1 Introduction

Compositionality is a fundamental characteristic of both human vision as well as natural language. It suggests that "the meaning of the whole is a function of the meaning of its parts"(Cresswell, 1973). For instance, compositionality allows people to differentiate between a photo of "a brown dog holding a white frisbee" and "a white dog running after a brown frisbee". For a while now, research on vision-language models has sought to inject such compositional structure as inductive priors so that models can comprehend scenes and express them using compositional language (Krishna et al., 2017; Ji et al., 2020; Lu et al., 2016; Grunde-McLaughlin et al., 2021). However, with the rise of large-scale pretraining, vision-language models today are trained from image-text pairs scraped

[1]Correspondence to: kamatha@cs.washington.edu

from the internet (Thomee et al., 2016; Schuhmann et al., 2022a; Sharma et al., 2018), and thus, are not explicitly given structural priors.

To probe whether large-scale pretrained vision-language models, such as CLIP (Radford et al., 2021), are capable of compositional reasoning, a number of contemporary benchmarks have been released (Thrush et al., 2022; Zhao et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023; Ray et al., 2023; Hsieh et al., 2023; Kamath et al., 2023a). Evaluation is primarily conducted through an image-to-text retrieval task formulation (Zhao et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023): by measuring how often models pick the description, "a brown dog holding a white frisbee" when presented with an image of it, and avoid choosing the incorrect **hard negative** description, "a white dog running after a brown frisbee". This second sentence is considered a hard negative because the colors are swapped and the verb is replaced. Surprisingly, these benchmarks unanimously find that state-of-the-art models demonstrate little to no compositionality (Hsieh et al., 2023).

As a natural follow up, many approaches have been proposed to remedy this lack of compositionality (Zheng et al., 2024). The most common method finetunes the CLIP model with similar hard negatives. Intuition suggests that by exposing CLIP to hard negatives, it will learn when such perturbations change the semantic meaning of the caption, and therefore should be sensitive to them (Yuksekgonul et al., 2023; Doveh et al., 2023b). With hard negative finetuning, results on benchmarks appear to suggest that CLIP models become more compositional (Hsieh et al., 2023). However, our results indicate otherwise.

We create a new evaluation dataset of $56,191$ images with $28,748$ swap and $27,443$ replace **hard positives**. Hard positives, in contrast to their negative counterparts, make semantic-*preserving* changes to concepts in an original caption. For example, "a brown dog holding ..." and "a brown dog grasping ..." are replaced hard positives. Ideally, models should be invariant to semantics-preserving perturbations. We validate this evaluation set with a human evaluation, where our participants effortlessly achieved 99%.

Our experiments reveal that the default CLIP model (Radford et al., 2021) performs 14.9% worse on our data versus on existing benchmarks. Worse, we test 7 CLIP finetuning approaches (Yuksekgonul et al., 2023; Ma et al., 2023; Hsieh et al., 2023; Doveh et al., 2023b,a) to find even sharper decreases in performance, up to 38.7%. We find that hard negative-finetuned models are "oversensitive", *i.e.*, they more often rank hard negatives higher than one but not both the original caption and the hard positive. We summarize these ideas in Figure 1.

To mitigate oversentitivity and this general degradation of performance, we curate a larger training set of $591,753$ hard positives and explore a simple data-augmentation training technique wherein CLIP models are finetuned simultaneously with both hard negatives and positives, in addition to the original caption. Compared to the original CLIP model, exposure to both improves performance in existing benchmarks and our evaluation data. When compared to models finetuned only on hard negatives, our model retains most of the performance improvements on existing benchmarks while improving on our evaluation set. We also find that exposure to only swap positives mitigates oversensitivity on the swap evaluation set and not on replace evaluation set, and vice versa.

Taken together, our investigations expose another dimension of compositionality which was previously unexplored by existing benchmarks. We lay out a number of implications of our findings in our discussion. We release our code, datasets and models at `https://github.com/amitakamath/hard_positives`.

2 Related work

We contextualize our study within research aiming to evaluate and 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 (Yuksekgonul et al., 2023; Thrush et al., 2022; Zhao et al., 2022; Ma et al., 2023; Ray et al., 2023; Hsieh et al., 2023; Kamath et al., 2023a). These tools often reveal that, despite achieving impressive results in various applications (Radford et al., 2021; Li et al., 2022b; Singh et al., 2022; Alayrac et al., 2022; Wang et al., 2022, 2023; Zhai et al., 2022), these models struggle with basic compositional tasks. Issues include difficulty in processing sentences with the same words in a different order (Thrush et al., 2022), and in recognizing relationships between objects or associating objects with their attributes (Zhao et al., 2022; Yuksekgonul et al., 2023; Ray et al., 2023; Hsieh et al., 2023; Bugliarello et al., 2023). Benchmarks also reveal that many models struggle with spatial reasoning (Zellers et al., 2018; Parcalabescu et al., 2022; Hendricks and Nematzadeh, 2021; Kamath et al., 2023b). 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 (Yuksekgonul et al., 2023; Cascante-Bonilla et al., 2023; Ray et al., 2023; Doveh et al., 2023b; Singh et al., 2023). A key strategy involves training models to differentiate between correct captions and procedurally-generated hard negatives (Yuksekgonul et al., 2023; Doveh et al., 2023b,a). However, it remains uncertain whether these approaches genuinely foster a deeper understanding of compositionality or merely enable models to perform well on dataset biases (Hsieh et al., 2023). 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 (Gururangan et al., 2018). Several studies propose methods for de-biasing these datasets to ensure evaluations more accurately reflect model capabilities (Reif and Schwartz, 2023; Zellers et al., 2018; Sakaguchi et al., 2021; Le Bras et al., 2020). Techniques like adversarial filtering (Zellers et al., 2018) 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 (Sakaguchi et al., 2021; Le Bras et al., 2020). 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 (Hsieh et al., 2023). We complement these benchmarks by introducing hard positives to allow a comprehensive evaluation of vision-language 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 (Doveh et al., 2023a,b; Fan et al., 2023). These works typically utilize large language models (OpenAI, 2022; Workshop et al., 2022) 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 (Doveh et al., 2023a,b) 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.

3 Evaluating for compositionality

This section formalizes the principle of compositionality to a well-defined evaluation scheme (Hupkes et al., 2020). First, we establish how vision-language compositionality is defined (Section 3.1). Then, we explain how existing benchmarks evaluate compositionality (Section 3.2) and their limitations under this definition (Section 3.3). Finally, we explain how we overcome this limitation by developing a new evaluation dataset (Section 3.4).

3.1 Definition of compositionality

To evaluate the compositionality of vision-language models, most existing benchmarks define a compositional language consisting of *scene graph* visual concepts (Ma et al., 2023) or a subset of scene graphs (*e.g.* some focus only on spatial relationships (Parcalabescu et al., 2022; Kamath et al., 2023b)). Within this language, an *atom* a is defined as a singular visual concept, corresponding to a single scene graph node. A *compound* c is defined as a primitive composition of multiple atoms, which corresponds to connections between scene graph nodes. Scene graphs admit two compound types: the attachment of attribute to objects ("brown dog"), and the attachment of two objects via a relationship ("dog runs after frisbee").

In most cases, we use entire captions to represent compounds c found in existing vision-language datasets. Conversely, captions can be parsed to become scene graphs. It has been shown that scene graphs, through this compositional language, are capable of capturing a number of linguistic phenomena (Suhr et al., 2019; Parcalabescu et al., 2022), including the existence of concepts ("a photo with *dog*"), spatial relationships ("a grill *on the left of* a staircase"), action relationships ("a dog *holding* a frisbee"), prepositional attachment ("A *brown* dog"), and negation ("There are *no* cats").

3.2 Evaluation protocol

A majority of existing compositionality benchmarks for vision-language models formulate the evaluation task as image-to-text retrieval (Zhao et al., 2022; Yuksekgonul et al., 2023; Ma et al., 2023). Given an image, the model is probed to select text that correctly describes the image from a pool of candidates. Unlike standard retrieval tasks where the negative (incorrect) candidates differ significantly from the positive (correct) text, compositionality benchmarks intentionally design **hard negative** texts that differ minimally from the positive text, in order to test whether the model understands the fine-grained atomic concepts that compose the scene. Under the definition above, hard negatives are defined as compounds with an atom either swapped or replaced. Both operations modify the compound such that their semantic interpretation violates the visual concepts in their corresponding image.

Re-using the example from the introduction, we have an image of "a brown dog holding a white frisbee". In comparison, "a white dog running after a brown frisbee" is a compound with multiple negative operations. The attributes white and brown are swapped and the relationship holding is replaced by running after. Most benchmarks curate evaluation sets with multiple hard negatives per image-text pair.

Using such a benchmark, they define the compositionality evaluation protocol as follows: Given a query image i, the model is tasked with retrieving its corresponding compound caption c amongst a set of distractors. Without loss of generality, assume there is one distractor c_n per image. The protocol first estimates a matching score between the image and each of the captions (image-text matching score): $s(c, i)$, $s(c_n, i)$. If a model is compositional, $s(c, i) > s(c_n, i)$, resulting in retrieving the correct caption over the hard negative.

3.3 Limitations with existing evaluations

The assumption made by existing benchmarks is that all atomic swaps or replacements necessarily cause a change in semantics. However, this is not the case with language. For example, "a brown dog holding ..." and "a brown dog grasping ..." are replaced hard positives since the replacement of holding to grasping does not alter the caption's grounding with respect to the image.

As such, we posit that existing benchmarks are incomplete. They have left out a vital component of compositionality: **hard positives**. Compositional models should be able to reason about two kinds of operations: (1) when a modification to c produces a hard negative c_n, the $s(c_n, i)$ should reduce when compared to $s(c, i)$; and (2) when a modification to c produces a hard positive c_p, then $s(c_p, i)$ should remain relatively similar to $s(c, i)$. In summary, hard positives should not alter the score $s(c, i) \approx s(c_p, i)$.

3.4 Curating a hard positive evaluation dataset

We respond to this incomplete evaluation by curating an evaluation dataset with hard positives. We focus on the two main types of perturbations in existing work: replacing one word or phrase in the caption; or swapping two words or phrases within the caption. Although other forms of perturbations exist, we choose these two as they are the most well-represented in prior benchmarks.

Therefore, we can consider each image in our dataset to be associated with three captions: the original caption c, a hard negative c_n (sourced from an existing hard negative benchmark) and a hard positive c_p (generated by us). Figure 2 shows examples from our benchmarks.

Generating replacements. The most popular type of hard negative considered by existing work is `REPLACE`, where one word or phrase in the caption is replaced with another in a way that changes the meaning of the caption (Zhao et al., 2022; Parcalabescu et al., 2022; Ma et al., 2023; Doveh et al., 2023a,b; Kamath et al., 2023a,b; Hendricks and Nematzadeh, 2021). To create hard positives, we replace one word or phrase in a way that does *not* change the meaning of the caption.

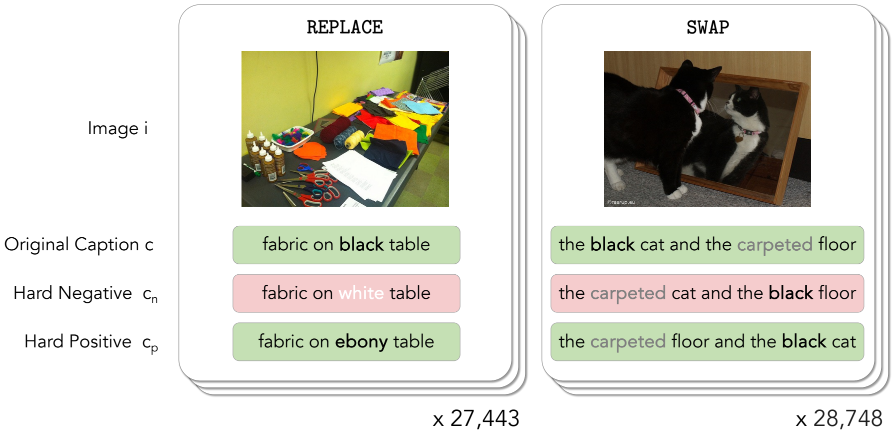


Figure 2: Our REPLACE and SWAP evaluation sets. REPLACE replaces either an attribute or a relation in the original caption c to obtain c_n and c_p. SWAP swaps object-attribute associations in the original caption c to obtain c_n and c_p.

We begin with examples from VL-Checklist (Zhao et al., 2022). This benchmark contains REPLACE hard negatives targeting either objects, attributes or relations. We focus on attributes and relations, as they have been shown to be more challenging for vision-language models to understand (Doveh et al., 2023a,b; Hsieh et al., 2023), and select the subset of VL-Checklist based on Visual Genome (Krishna et al., 2017) to stay consistent with our SWAP benchmark. The VL-Checklist Relations benchmark has two types of relations: actions and spatial. The VL-Checklist Attributes benchmark has five types of attributes: action[1], color, material, size, and state.

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 (details in Appendix A).

E.g., for the Visual Genome caption "cutting board next to pan", VL-Checklist constructs a hard negative by replacing the relation with an antonym: "cutting board *far from* pan". We construct a hard positive by replacing the relation with a synonym: "cutting board *near* pan". While there may be minor differences between the original and hard positive captions (e.g., "next to" may imply a closer spatial relation than "near"), they are both a match for the image, while the hard negative caption is not.

[1]The action *relation* is a transitive verb, e.g., "a person wearing a shirt", whereas the action *attribute* is an intransitive verb, e.g., "a person standing".

Generating swaps. The other popular type of hard negative considered by existing work is SWAP, where two words or phrases in a caption are swapped with each other in a way that changes the meaning of the caption (Yuksekgonul et al., 2023; Parcalabescu et al., 2022; Ma et al., 2023; Thrush et al., 2022). To create hard positives, we swap two phrases in a way that does *not* change the meaning of the caption.

We begin with the Visual Genome Attribution (VGA) set from the Attribute-Relation-Order benchmark (Yuksekgonul et al., 2023), which switches object-attribute associations in a Visual Genome caption to create a hard negative, e.g., "the crouched cat and the open door" → "the open cat and the crouched door". To create a hard positive, we switch the word order while retaining the object-attribute associations, thus retaining the meaning of the caption, e.g., "the open door and the crouched cat". While there are small linguistic differences between the original and hard positive captions (e.g., people tend to describe the most salient object first), they are both a match for the image, where the hard negative caption is not. We create a hard positive for each example in the VGA dataset, resulting in a benchmark of 28,748 examples.

4 Hard negative finetuning induces brittleness

In this section we investigate existing models' performance, utilizing our more complete evaluation. We especially focus on evaluating whether recently introduced methods that train models with hard negatives indeed improve compositionality.

The goal of hard negative finetuning is to encourage CLIP models to understand how structural changes in language can affect the semantic interpretation of the caption. For example, finetuning on hard negatives targeting swaps should, in intuition, teach models that the directionality of a relationship between objects matters; finetuning on hard negatives targeting replacement should teach models to be sensitive to changes to any single word in the caption. Ideally, we want the model to understand that perturbations to the caption (e.g., swaps, replacements) are important, and to recognize when a perturbed sentence has the same meaning as the original sentence, and when it does not. However, we posit that solely emphasizing on hard negatives does not teach the model *when* perturbations to the caption change meaning, they teach the model that perturbations *do* change meaning, *always*.

To validate our hypothesis, we benchmark a suite of CLIP models, trained regularly or with different hard negative augmentation strategies in Section 4.1. We uncover that hard negative finetuning improves performance on hard negative evaluations at the cost of performance degradation on hard positives in Section 4.2. We finally discuss why this happens in Section 4.3.

4.1 Evaluation

Task. To evaluate model understanding of hard positives in addition to hard negatives, we use the image-text matching (ITM) task, consistent with existing benchmarks discussed in Section 3.2. In our benchmark, the input is an image paired with three captions: two captions match the image (the original caption and the hard positive), and the third does not match the image (the hard negative). The model must return a high image-text matching score s for the correct matching captions, and a low score for the incorrect one.

Metrics. The first metric we use is the percentage of images in the benchmark for which the model-assigned score of the correct captions is higher than that of the incorrect caption.

For an image i, let the original caption be c, the hard negative from the existing benchmark (VGA for SWAP and VL-Checklist for REPLACE) be c_n, and the hard positive that we construct (per Section 3.4) be c_p. The vision-language model returns an image-text matching score $s(C|I)$ for some caption C and image I. We measure *Augmented Test Accuracy*: the fraction of instances in the benchmark where:

$$s(c|i) > s(c_n|i) \text{ and } s(c_p|i) > s(c_n|i) \quad (1)$$

We do not require $s(c|i)$ to be equal to $s(c_p|i)$, as there are minor linguistic differences between the original caption and hard positive (c.f. Section 3.4), and it is reasonable to predict that one of these captions matches the image slightly better than the other. However, as these two captions are both correct matches for the image and the hard negative is not, their model-assigned score should be higher than that of the hard negative caption.

The second metric we use is the percentage of images in the benchmark where the model treats c and c_p *differently* when ranking with respect to c_n: ranking one of them above c_n and one below. We measure this oversensitivity as *Brittleness* (↓): the fraction of instances in the benchmark where:

$$\begin{aligned} s(c|i) > s(c_n|i) > s(c_p|i) \quad \text{or} \\ s(c_p|i) > s(c_n|i) > s(c|i) \end{aligned} \quad (2)$$

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 Condition (1). For Brittleness, random chance is again 33.3%, as two of six possible rankings for the three captions satisfy Condition (2).

Human-estimated performance. We estimate human performance on our benchmark. We sample 100 data points each from the SWAP and REPLACE benchmarks and solicit two expert annotations per data point. Each data point contains the image, the original caption, the hard negative and the hard positive. We ask the annotators to rank the captions based on the match for the image, allowing them to give multiple captions the same rank. The annotators have all taken at least one graduate-level course in NLP or Machine Learning. A point is awarded to the example if both annotators agree on the correct rank[2].

[2]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).

	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)	NegCLIP	68.6	52.1 (-16.6)	70.9	56.7 (-14.2)	21.5	26.4
	CREPE-Swap	63.5	50.4 (-13.1)	70.6	56.7 (-13.9)	**19.8**	26.0
	CREPE-Replace	73.7	53.9 (-19.8)	71.1	57.7 (-13.4)	23.9	25.4
	SVLC	76.6	44.5 (-32.1)	72.4	**61.6** (-10.9)	39.9	**20.8**
	SVLC+Pos	64.3	45.0 (-19.3)	56.5	45.4 (-11.1)	29.8	22.8
	DAC-LLM	87.6	48.9 (-38.7)	72.0	61.1 (-10.9)	40.1	21.6
	DAC-SAM	86.9	**55.9** (-31.0)	69.5	56.5 (-13.0)	32.5	25.6
(c)	Our HN	73.9	55.7 (-18.2)	74.3	60.5 (-13.8)	21.0	25.1
	Our HP+HN	69.0	**58.0** (-11.0)	73.2	**61.1** (-12.1)	**16.9**	**22.9**
(d)	Our HP+HN (Swap-only)	63.9	51.6 (-12.3)	73.0	**61.9** (-11.2)	18.6	**21.2**
	Our HP+HN (Replace-only)	70.9	**59.0** (-11.9)	69.7	55.6 (-14.1)	**17.8**	26.5
	Random Chance	50.0	33.3	50.0	33.3	33.3	33.3
	Human Estimate	97	97	100	100	0	0

Table 1: Results of various ITM models on our benchmark: (a) CLIP; (b) Hard-Negative finetuned versions of CLIP from previous work (Section 4.2); (c,d) Our improved model (Section 5.2). The purple cells indicate the models have seen perturbations of the type we are testing for during finetuning, blue cells indicate otherwise. REPLACE averages performance on Attributes and Relations; refer to Appendix B for detailed results.

Models evaluated. Without loss of generality, we adopt the ViT-B/32 architecture for all our experiments. So, CLIP ViT-B/32 is our baseline CLIP model (Radford et al., 2021). We then evaluate several training interventions that finetune CLIP ViT-B/32 using different types of hard negatives: NegCLIP (Yuksekgonul et al., 2023) is finetuned on hard negatives targeting word order shuffling; CREPE-Swap (Ma et al., 2023; Hsieh et al., 2023) is finetuned on hard negatives targeting single-phrase swaps; CREPE-Replace (Ma et al., 2023; Hsieh et al., 2023) is finetuned on hard negatives targeting single-phrase replacements; SVLC (Doveh et al., 2023b) is finetuned on hard negatives targeting single-phrase replacements generated by LLMs and rule-based methods; SVLC+Pos (Doveh et al., 2023b) is finetuned on the aforementioned hard negatives as well as paraphrases of the caption; DAC-LLM (Doveh et al., 2023a) is finetuned on several LLM-generated captions of the image as well as hard negatives generated by the SVLC method; and DAC-SAM (Doveh et al., 2023a) is finetuned on SAM-generated captions of the image as well as hard negatives generated by the SVLC method.

It is worth noting that SVLC+Pos, DAC-LLM and DAC-SAM contain "positives" in their finetuning, *i.e.*, alternate captions that also match the image. However, these are not *hard* positives, as in our work. Our alternate captions are *minimal* perturbations to the original caption, swapping or replacing only single phrases while retaining the caption's meaning.

4.2 Results

Hard negative finetuning doesn't help models understand *when* perturbations matter. In Table 1, we first compare ITM model scores on only the original caption c and the hard negative c_n, given an image i — as is done in existing work (Original Test Score). We then introduce the hard positive c_p central to our work, and check: is the model score for the hard positive caption greater than that of the hard negative caption? Per Section 4.1, we evaluate the cases when $s(c|i) > s(c_n|i)$ and $s(c_p|i) > s(c_n|i)$ (Augmented Test Score).

We find that, when including hard positives, the performance of models finetuned on hard negatives drops (Aug. Test Score < Orig. Test Score, difference depicted in red) by an average of 24.4 points for REPLACE and 12.5 points for SWAP— greater than the base model CLIP's 14.9 point and 10.9 point drops respectively. In fact, we see that as much as 39 points of model performance on hard negative benchmarks is misleading, as the model did not understand the underlying concept (e.g., word order) enough to recognize when the perturbation retained caption semantics.

Hard negative finetuned models are oversensitive. Per Section 4.1, to evaluate model brittleness, we calculate the percentage of instances in the benchmark where $s(c|i) > s(c_n|i) > s(c_p|i)$ or $s(c_p|i) > s(c_n|i) > s(c|i)$. In these instances, it is clear that the model does not understand that c and c_p have the same meaning and c_n has a different meaning from both of them, *i.e.*, it is oversensitive to the perturbation. In Table 1, we see that in almost all cases, Brittleness increases after finetuning (rows (a) vs (b)) — *i.e.*, that hard negative finetuning makes the models more oversensitive to perturbations.

Oversensitivity transfers across pertubation types. We see that, for each type of hard positive (`SWAP`, `REPLACE`), the most oversensitive models are those finetuned on the corresponding hard negative (the purple cells in Table 1), e.g., NegCLIP and CREPE-SWAP are finetuned on `SWAP` hard negatives, and are the most oversensitive models under the `SWAP` hard positives, and similarly for the other models on `REPLACE`. This is unsurprising, as the finetuning has taught the model to be sensitive to that specific type of perturbation.

However, we see that models trained on `REPLACE` hard negatives are still brittle to `SWAP` hard positives (with an average score of 23.2), more so than the original CLIP baseline. We also see that models trained on `SWAP` hard negatives are brittle to `REPLACE` hard positives (with an average score of 20.7), although less so than the original CLIP baseline — potentially because a swap can be seen as two replacements. In essence, we see that the oversensitivity introduced by finetuning on hard negatives of one type of perturbation transfer to the other type of perturbation (blue cells in Table 1).

"Non-hard" positive finetuning increases oversensitivity. Three of the models we evaluate include finetuning on multiple correct captions ("positives") for the image. For SVLC+Pos and DAC-LLM, these are generated by LLMs that see the caption alone, and for DAC-SAM, these are generated by BLIP2 (Li et al., 2023) which sees segments of the image extracted by SAM (Kirillov et al., 2023).

However, c.f. Table 1, this addition of positives to training does not improve model understanding of *hard* positives compared to models finetuned on hard negatives alone; in fact, these models usually perform much worse. Comparing SVLC with SVLC+Pos, where the only difference is the addition of positives to training, it is clear that positive finetuning significantly increases oversensitivity.

Why? The alternate captions tend to be structurally very different from the original caption, and in the case of SAM-generated captions, contain different focuses entirely, as they only describe a segment of the image. Thus, they may give the model a more holistic understanding of the overall image (Doveh et al., 2023a), but not the fine-grained understanding we evaluate with our hard positives.

Hard Negative finetuning lowers scores of the original captions too. Image-text matching scores are used to filter out data during web-scale corpora curation (Schuhmann et al., 2022b; Gadre et al., 2024), to evaluate captions for images (Hessel et al., 2021), to evaluate text-to-image generation (Saharia et al., 2022; Hu et al., 2023), and to evaluate text-to-video generation (Ho et al., 2022). Thus, while our evaluations focus on ranking, it is worth paying attention to the absolute value of the image text matching score itself.

Across all benchmarks, models with hard negative finetuning lower the image-text matching score of the *original* caption with the image as well — not just the negative caption (c.f. Table 2 and Appendix D.2). In fact, the model that achieves one of the the highest performance on VL-Checklist, DAC-LLM, reduces the original caption scores on `REPLACE` from 0.23 to 0.16, a very large drop. This could cause significant errors in the aforementioned downstream applications. Examples are shown in Section 5.4.

Different variants of CLIP all perform poorly. In Appendix B, we study the performance of CLIP with different model sizes, text encoders, pretraining data, and vision encoders. However, none of these variants significantly improve CLIP's poor compositionality on our benchmarks.

4.3 Why does hard negative finetuning induce brittleness?

From these results, it is clear that hard negative finetuning does not improve vision-language models' compositionality holistically. Performance on hard negatives is necessary but insufficient for compositionality, and by focusing on hard negatives alone, hard negative finetuning exacerbates poor performance on hard positives. We now discuss why the hard negative finetuning setup leads to worse performance on hard positives, as shown by our evaluation.

Let there be a set $\mathbb{P}$ of all possible small perturbations to the caption. During training on original captions and hard negatives alone, all perturbations $\mathcal{P} \in \mathbb{P}$ to the caption c seen by the model $\mathcal{M}$ change the label of the caption. The loss always penalizes $\mathcal{M}$ if $\mathcal{P}(c)$ matches the image under $\mathcal{M}$, *i.e.*, the model is taught to reduce $s(\mathcal{P}(c)|i)$ for all seen $\mathcal{P}$. Thus, it is consistent with the training data to identify whether a text input c somewhat matches the image and comes from the original caption distribution $\mathcal{C}$, and award it a high score if so, and a low score if not, *i.e.*, if the caption appears to have been perturbed. Essentially, it is sufficient for $\mathcal{M}$ to learn perturbation detection.

We see empirical proof of this in two ways (c.f. Section 4.2): firstly, we see that $\mathcal{M}$ awards low scores to all perturbed captions, whether the meaning of the caption has changed or not; secondly, we see that this behavior transfers across *types* of perturbations — a model trained with SWAP hard negatives awards low scores to REPLACE hard negatives and hard positives, and vice versa. Thus, by only showing models that perturbations *do* change the input, not *when* they change the input, we fail to attain improved compositionality.

5 Exploring hard positive finetuning

After establishing that finetuning on hard negatives alone teaches models that perturbations always change meaning, which causes poor compositionality, we explore a more well-rounded finetuning technique, incorporating hard positives into finetuning to determine whether that improves compositionality.

5.1 Method

We first generate hard positives using LLAMA-2 70B-Chat (Touvron et al., 2023). We prompt this text-only model to modify a given caption without changing the meaning, either with word replacements, or swaps (if the caption contains the word "and"). The inputs we provide the model are COCO-train captions. Prompting and generation details are provided in Appendix C.

We then add these hard positives to model finetuning. We finetune CLIP ViT-B/32 on COCO-train with hard positives, generated as discussed above, and hard negatives, generated by the CREPE (Ma et al., 2023) process, as in SugarCrepe (Hsieh et al., 2023). One hard positive and one hard negative is generated for each of the 591,753 COCO-train captions, resulting in an overall train set of 1,775,259 examples. We release this data to support further research in compositionality.

	Mean score		
Model	c (↑)	c_n (↓)	c_p (↑)
CLIP ViT-B/32	0.234	0.226	0.229
DAC-LLM	0.160	0.134	0.131
Ours	0.232	0.220	0.231

Table 2: Mean score for c, c_n, and c_p in REPLACE produced by CLIP, a hard negative finetuned model (DAC-LLM) and Our model. Our model exhibits better compositionality than CLIP and DAC-LLM by correctly lowering the score of c_n but not c or c_p. Refer to Appendix D.2 for results across all models.

The finetuning follows the procedure outlined in SVLC (Doveh et al., 2023b). We separately finetune CLIP ViT-B/32 on COCO-train with hard negatives only, to serve as a direct comparison for how the inclusion of hard positives in finetuning impacts model performance. We also finetune CLIP ViT-B/32 on COCO-train alone to serve as a control. Refer to Appendix D.1 for implementation details.

5.2 Results

Adding hard positives to finetuning improves model performance. On REPLACE and SWAP, our model finetuned on hard positives and hard negatives achieves the highest augmented test accuracy and lowest brittleness, compared to our model finetuned on hard negatives alone (Table 1(c)).

On REPLACE, our model also outperforms all hard negative finetuned models in Table 1(b) in augmented test accuracy and brittleness. On SWAP, our model outperforms NegCLIP, the CREPE-finetuned models, and DAC-SAM, but has slightly worse brittleness than the other models and slightly worse augmented test accuracy than SVLC. This could be due to the inherent difficulty of the SWAP task — not only could it be considered two replacements, but the word identities are unchanged, which causes added difficulty (Thrush et al., 2022; Yuksekgonul et al., 2023).

Table 2 shows the mean image-text matching scores of CLIP, DAC-LLM, and our finetuned model for the original, hard negative, and hard positive captions in REPLACE. CLIP awards similar scores to all, seeming to ignore the replacement for both hard negatives and hard positives. For DAC-

	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)	0 HN	58.5	49.8 (-8.6)	64.1	51.2 (-12.9)	**15.8**	25.0
	0.25 HN	66.0	55.5 (-10.5)	71.6	59.8 (-11.8)	16.6	22.8
	0.50 HN	67.3	56.9 (-10.5)	72.5	60.5 (-12.0)	16.4	22.8
	0.75 HN	68.2	**57.6** (-10.6)	72.9	**61.0** (-11.9)	16.6	**22.7**
(c)	Our HN	73.9	55.7 (-18.2)	74.3	60.5 (-13.8)	21.0	25.1
	Our HP+HN	69.0	**58.0** (-11.0)	73.2	**61.1** (-12.1)	**16.9**	**22.9**
	Random Chance	50.0	33.3	50.0	33.3	33.3	33.3
	Human Estimate	97	97	100	100	0	0

Table 3: Results of ITM models on our benchmark while varying the ratio of hard negatives to hard positives during finetuning: (a) CLIP, (b) Ablated versions of our improved model, (c) Our improved model (Section 5.2). REPLACE averages performance on Attributes and Relations.

LLM, the model recognizes the replacement for hard negatives and lowers the score significantly — however, it incorrectly lowers the score of the hard positives by an even greater amount, although the meaning of the caption has not changed. Our finetuned model exhibits the correct behavior — it reduces the score of the hard negative but maintains the score of the hard positive compared to the original caption. Moreover, unlike DAC-LLM, it does not lower the score of all captions, which could otherwise have repercussions downstream (c.f. Section 4.2).

Oversensitivity transfers across perturbations, but improved invariance does not. We additionally finetuned two CLIP ViT-B/32 models on hard positives and hard negatives targeting only SWAP and only REPLACE respectively (c.f. Table 1(d)). While neither of these models perform significantly better than the multi-task version on their respective evaluations (purple cells), we see that the Swap-Only finetuned model performs poorly on REPLACE, and likewise for the Replace-only finetuned model on SWAP (blue cells). As such, while we saw that oversensitivity transferred across types of perturbations (Section 4.2), it appears that improved invariance to a certain type of perturbation does not.

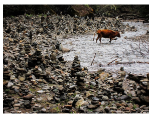
Performance on standard benchmarks. In order to ensure that models do not experience catastrophic forgetting while finetuning on our data, we evaluate our finetuned models on standard benchmarks. As in (Yuksekgonul et al., 2023), we evaluate on ImageNet-1K, CIFAR-10, CIFAR-100, COCO Retrieval and Flickr30K Retrieval. Our models improve at hard positives and hard negatives while not losing overall performance. Refer to Appendix E for further details.

5.3 Changing the ratio between hard positives and hard negatives

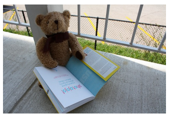
In this section, we study the impact of changing the ratio between hard positive and hard negative losses during model finetuning. Table 3 contains results of models trained on differing weights of hard negative loss while keeping the weight of hard positives loss fixed. We vary the weight of hard negative loss from 0 (which equates to a model trained only on hard positives) to 1 (which equates to our default proposed model, c.f. Table 1) in increments of 0.25.

Hard negatives are needed. Rather unsurprisingly, the hard positive-only trained model performs poorly on our evaluation — it has no sense of the existence of hard negatives, and learns from finetuning the *opposite* of what hard negative-only finetuned models learn in existing work: rather than that perturbations *always* change the label, this model learns that perturbations *never* change the label. It is clear from these results that hard negatives are needed in addition to hard positives to improve model compositionality.

As the ratio of hard negatives to hard positives increases, test accuracy increases, but so may brittleness. As the hard negative loss weight increases from 0 to 1, we see the Original and Augmented Test Accuracies both increasing. However, so too does the brittleness, for REPLACE. This trend

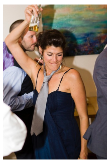


Captions	CLIP	DAC-LLM	Ours
c: standing cow	0.203	0.164	0.249
c_n: lying cow	0.210	0.155	0.242
c_p: upright cow	0.217	0.140	0.246



Captions	CLIP	DAC-LLM	Ours
c: the open book and the concrete floor	0.247	0.146	0.293
c_n: the concrete book and the open floor	0.254	0.142	0.283
c_p: the concrete floor and the open book	0.24	0.139	0.286

Captions	CLIP	DAC-LLM	Ours
c: plane flying in white sky	0.25	0.166	0.272
c_n: plane flying in yellow sky	0.245	0.146	0.234
c_p: plane flying in ivory sky	0.248	0.136	0.275



Captions	CLIP	DAC-LLM	Ours
c: the brown hair and the gray tie	0.248	0.103	0.269
c_n: the gray hair and the brown tie	0.244	0.102	0.257
c_p: the gray tie and the brown hair	0.245	0.095	0.267

Figure 3: Sample predictions of CLIP, a hard negative finetuned model (Doveh et al., 2023a), and our model. Top: Considering hard negatives alone provides an incomplete picture of compositionality. Bottom: Hard negative finetuning can harm model performance. Both: Hard negative finetuning incorrectly lowers scores of the *original* caption, unlike our model.

continues: when the hard positives are dropped (i.e. a ratio of ∞), we see in Table 3(c) that the hard negative-only finetuned model achieves the highest Original Test Accuracy, but also has the highest brittleness for both REPLACE and SWAP. This trade-off suggests the need for careful tuning to achieve the best understanding of both hard positives and hard negatives.

5.4 Qualitative Analysis

Figure 3 depicts examples of outputs of the original CLIP ViT-B/32 model, the hard-negative finetuned DAC-LLM, and our model finetuned on both hard positives and hard negatives.

The top part shows similar behavior as depicted in Figure 1: the hard negative finetuned model appears to have achieved high compositionality when its performance on c and c_n is compared to CLIP — however, this is an incomplete picture. The hard negative finetuned model actually awards a lower score to c_p than to c_n, showing that its understanding of compositionality is still lacking. In contrast, our model correctly awards higher scores to c and c_p than to c_n.

The lower part shows instances of interesting behavior: where CLIP ranked the three captions correctly, and hard negative finetuning causes the model to now rank the captions incorrectly (awarding a low score to c_P). Clearly, hard negative finetuning can hurt the original model's performance.

In all shown examples, the hard negative finetuned model awards a lower score to *all* captions than CLIP (including the *original* caption), as discussed in Section 4.2. Our model does not exhibit this behavior (c.f. Table 2 and Appendix D.2).

6 Discussion

Our investigations explore a component of compositionality that has, until now, been largely underexplored. While a few efforts have studied the effects of training with positive rewritings (Fan et al., 2023), the use of *hard* positives has been absent from the literature. We uncovered not just that CLIP models finetuned with hard negatives become oversensitive to changes, but that the de facto CLIP model itself performs poorly on our augmented set. This calls into question whether CLIP models have a grounded sense of relational semantics (Hsieh et al., 2023): for example, even basic text encoders such as word2vec (Mikolov et al., 2013a,b) understand that "white" and "ivory" have closer meanings to each other than either does to "blue" — so why should CLIP models fail to understand this, given *additional* signal from the image, and millions of image-text pairs of supervision?

Although training with hard positives mitigates the oversensitivity of CLIP models, models' performance is still far behind human performance. There is a need for further designs that incentivize compositionality by exploring alternative architecture designs and training objectives (Bugliarello et al., 2023; Zeng et al., 2022a; Tschannen et al., 2024). Our work calls for further research investigating more rigorously how finetuning methods targeting specific behaviors can cause adverse effects to overall model behavior, compared to the current status quo of simply evaluating on standard downstream evaluations. More research is also required to arrive at finetuning techniques that do not cause such adverse effects, and achieve the goal of improved robust vision-language compositionality.

Limitations

While we have further analysis in the Appendix, our work, like most work in vision-language compositionality today, is limited to CLIP-style models. There is a need to evaluate vision-language generation models, including Flamingo (Alayrac et al., 2022), BLIP (Li et al., 2022a, 2023), and GPT-4V (OpenAI, 2023), to isolate the effects of architecture and training objective. Additionally, while our models achieve higher performance on hard positives, more research is required to further improve performance and generalize to types of hard positives unseen during finetuning.

Acknowledgements

We thank Jieyu Zhang, Zixian Ma, the rest of the members of the RAIVN lab, William Merrill, as well as the anonymous reviewers, for helpful discussion and feedback. This work was partially supported by the Allen Institute for AI and ONR award N00014-23-1-2780.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada. Association for Computational Linguistics.

Paola Cascante-Bonilla, Khaled Shehada, James Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogério Schmidt Feris, and Leonid Karlinsky. 2023. Going beyond nouns with vision & language models using synthetic data. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20098–20108.

MJ Cresswell. 1973. Logics and languages.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023a. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith.

2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023a. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4v(ision) system card.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: How to adapt vision-language models to compose objects localized with attributes?

Yuval Reif and Roy Schwartz. 2023. Fighting bias with bias: Promoting model robustness by amplifying dataset biases. In *Findings of the Association for*

Computational Linguistics: ACL 2023, pages 13169–13189, Toronto, Canada. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022a. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022b. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 869–893, Singapore. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. 2024. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022a. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022b. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022a. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022b. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE. Association for Computational Linguistics.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

A Additional Benchmark Details

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

A.1 Further details about REPLACE

This dataset consists of hard negatives selected from VL-Checklist (Zhao et al., 2022) 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 3.4, 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 (Krishna et al., 2017) subset of VL-Checklist to stay consistent with the SWAP benchmark, which is sourced from the same dataset.

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 3.4, 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 4, 5 and 6.

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

Table 4: 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.

A.2 Random samples of REPLACE and SWAP

Figure 4 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 (Zhao et al., 2022; Yuksekgonul et al., 2023).

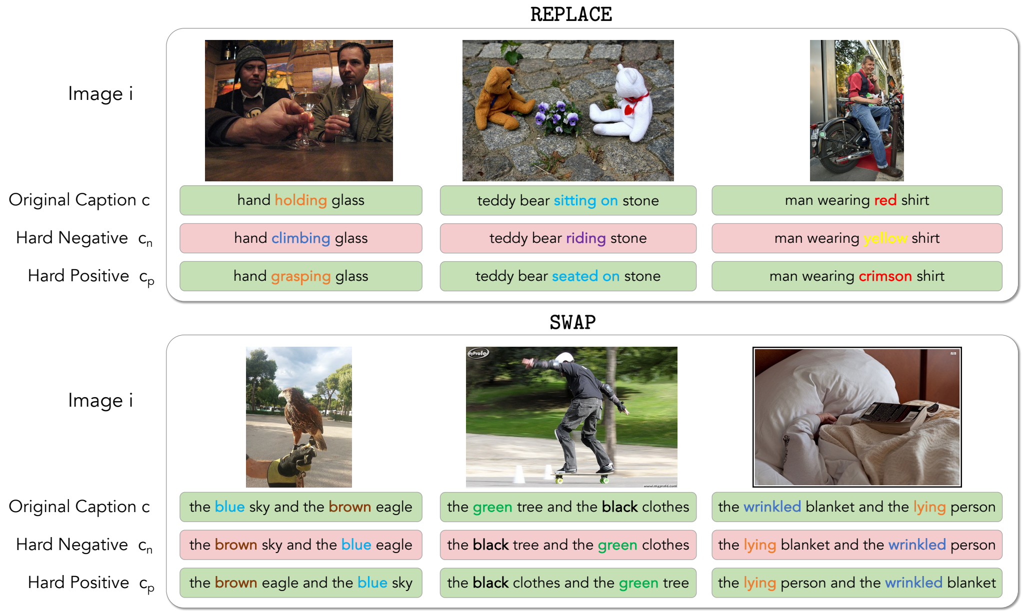


Figure 4: Random samples of REPLACE and SWAP. The first two REPLACE samples are from Relations, and the third from Attributes.

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

Table 5: 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 HP: tire on big truck HN: tire on tiny truck
small	tiny	358	O: toilet inside small bathroom HP: toilet inside tiny bathroom HN: toilet inside huge bathroom
long	lengthy	271	O: person carrying a long skateboard HP: person carrying a lengthy skateboard HN: person carrying a short skateboard
big	large	146	O: big elephant HP: large elephant HN: tiny elephant
huge	big	31	O: kites under huge sky HP: kites under big sky HN: kites under tiny sky
wet	damp	62	O: wet road HP: damp road HN: cloudless road
smiling	happy	50	O: snowboard with smiling man HP: snowboard with happy man HN: snowboard with sad man
old	aged	46	O: old train HP: aged train HN: young train
clear	unclouded	43	O: clear sky HP: unclouded sky HN: partly cloudy sky
young	youthful	36	O: shoes on young man HP: shoes on youthful man HN: shoes on unhappy man

Table 6: 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.

B Additional Results

This section contains additional results, splitting the REPLACE results in the main paper into the separate Relations and Attributes subsets (Table 7), as well as the results of various other models on our benchmarks: varying model size, architecture, pretraining data, and training objective (Table 8).

Replacing relations vs replacing attributes. Table 7 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 better 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 8(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 8(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 8(d), using DataComp (Gadre et al., 2024) as the pretraining data for CLIP seems to hurt model performance, more so on REPLACE than on SWAP.

Changing CLIP vision encoder. From Table 8(e), replacing the ViT vision encoder with a ResNet-based vision encoder seems to improve performance slightly, in the case of RN50 models.

Comparing CLIP to XVLM (Zeng et al., 2022b). Table 8(f) shows the performance of XVLM-16M (pretrained) on our benchmarks, as it has been shown to perform well on hard negative-focused benchmarks (Bugliarello et al., 2023). 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 pre-dates ARO (Yuksekgonul et al., 2023) and VL-Checklist (Zhao et al., 2022), from which our benchmarks are sourced. This may also explain the results of (Bugliarello et al., 2023).

C 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 (Touvron et al., 2023). Then, we discuss the implementation details of the generation. Finally, we provide a random sample of the data generated using the prompts.

C.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:
```

	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 7: Detailed results of various ITM models on our REPLACE benchmark: (a) CLIP, (b) Hard-Negative finetuned versions of CLIP from previous work (Section 4.2), (c) Our improved model (Section 5.2). 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 8: 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 (Gadre et al., 2024) rather than WIT (Radford et al., 2021) or LAION (Schuhmann et al., 2022a), (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.

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.

C.2 Implementation details

We generate hard positive training data by feeding the above prompt to the LLAMA2 70B-Chat model (Touvron et al., 2023). The examples are sourced from COCO train (note: Hard negatives are generated from COCO train as well, following the CREPE (Ma et al., 2023) 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 D.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.

C.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 7). 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.

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 9: 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.

D Finetuning on Hard Positives and Hard Negatives

D.1 Implementation details

The finetuning follows the procedure outlined in SVLC (Doveh et al., 2023b). 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.

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

As discussed in Section 4.2 and 5.4, hard negative finetuning causes the model to award a lower image-text matching score to *all* captions, not the hard negative caption alone. This has negative implications for various use cases where the score of the model is used directly, rather than as a ranking mechanism.

Table 9 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 score of the original caption.

E 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 10 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 (Zhai et al., 2019), 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 10, 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 (Wortsman et al., 2022b), 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 (Wortsman et al., 2022b,a), 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.

	Model	ImageNet1k		COCO		Flickr30k		VTAB	
		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 10: 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 (Zhai et al., 2019; Ilharco et al., 2021). 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).