The Hard Positive Truth about Vision-Language Compositionality

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

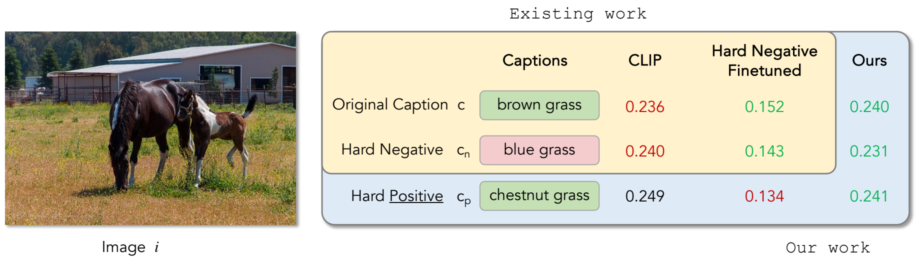
Abstract. Several benchmarks have concluded that our best vision-language models (*e.g.*, CLIP) are lacking in compositionality. Given an image, these benchmarks probe a model's ability to identify its associated caption amongst a set of compositional distractors. In response, a surge of recent proposals show improvements by finetuning CLIP with distractors as **hard negatives**. Our investigations reveal that these improvements have, in fact, been overstated — because existing benchmarks do not probe whether finetuned 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.

Keywords: Vision-Language · Compositionality

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" [4]. 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 [9, 19, 23, 28]. However, with the rise of large-scale pretraining,

* Correspondence to: kamatha@cs.washington.edu



	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

Fig. 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, [5]) 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 (real example shown).

vision-language models today are trained from image-text pairs scraped from the internet [40, 42, 46], and thus, are not explicitly given structural priors.

To probe whether large-scale pretrained vision-language models, such as CLIP [35], are capable of compositional reasoning, a number of contemporary benchmarks have been released [15, 20, 29, 36, 47, 55, 61]. Evaluation is primarily conducted through an image-to-text retrieval task formulation [29, 55, 61]: 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 [15].

As a natural follow up, many approaches have been proposed to remedy this lack of compositionality [62]. 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 [6, 55]. With hard negative finetuning, results on benchmarks appear to suggest that CLIP models become more compositional [15]. 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 [35] performs 14.9% worse on our data versus on existing benchmarks. Worse, we test 7 CLIP finetuning approaches [5,6,15,29,55] 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 Evaluating for compositionality

This section formalizes the principle of compositionality to a well-defined evaluation scheme [17]. First, we establish how vision-language compositionality is defined (§2.1). Then, we explain how existing benchmarks evaluate compositionality (§2.2) and their limitations under this definition (§2.3). Finally, we explain how we overcome this limitation by developing a new evaluation dataset (§2.4).

2.1 Definition of compositionality

To evaluate the compositionality of vision-language models, most existing benchmarks define a compositional language consisting of *scene graph* visual concepts [29] or a subset of scene graphs (*e.g.* some focus only on spatial relationships [21,34]). 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 [34,45], 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").

2.2 Evaluation protocol

A majority of existing compositionality benchmarks for vision-language models formulate the evaluation task as image-to-text retrieval [29, 55, 61]. 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.

2.3 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 $s(c, i) \approx s(c_p, i)$.

2.4 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:

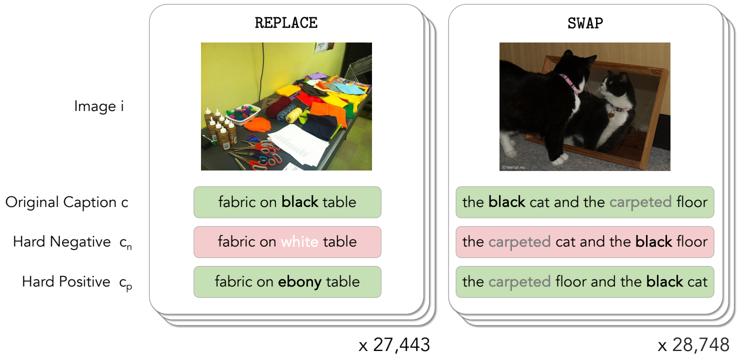


Fig. 2: Our REPLACE and SWAP evaluation sets. REPLACE replaces either an attribute or a relation in the original caption c to obtain c_n and c_p. SWAP swaps the object-attribute associations in the original caption c to obtain c_n and c_p.

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 [5, 6, 11, 20, 21, 29, 34, 61]. To create hard positives, we replace one word or phrase in a way that does *not* change the meaning of the caption.

We begin with examples from VL-Checklist [61]. 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 [5, 6, 15], and select the subset of VL-Checklist based on Visual Genome [23] 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. Refer to the Supplementary for further details.

[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".

For example, 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.

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 [29, 34, 47, 55]. 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 [55], 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.

3 Hard negative finetuning induces brittleness

In this section we investigate existing models' performance, utilizing the more complete evaluation we created. We especially focus on evaluating whether recently introduced methods that train models with hard negatives indeed improve models' 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 3.1. We uncover that hard negative finetuning improves performance on hard negative evaluations at the cost of performance degradation on hard positives in Section 3.2. We finally discuss why this happens in Section 3.3.

3.1 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 2.2. 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 image-text matching score for the incorrect one.
Metrics. The first metric we use is the percentage of images in the benchmark for which score of the correct captions is higher than that of incorrect captions.

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 2.4) 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 the Augmented Test Accuracy: the fraction of instances in the benchmark where:

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

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 2.4), 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:

$$s(c|i) > s(c_n|i) > s(c_p|i) \text{ or } s(c_p|i) > s(c_n|i) > s(c|i)$$

Human-estimated performance. We estimate human performance on our benchmark. We sample 100 data points each from SWAP and REPLACE benchmarks and solicit two expert annotations per data point. Each data point contains 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 if both annotators agree on the correct rank.
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 [35]. We then evaluate several training interventions that finetune CLIP ViT-B/32

	Model	REPLACE Orig. Test Acc.	 Aug. Test Acc.	SWAP Orig. Test Acc.	 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 (§3.2), (c,d) Our improved model (§4.2). 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 Supplementary for details.

using different types of hard negatives: NegCLIP [55] is finetuned on hard negatives targeting word order shuffling; CREPE-Swap [15,29] is finetuned on hard negatives targeting single-phrase swaps; CREPE-Replace [15, 29] is finetuned on hard negatives targeting single-phrase replacements; SVLC [6] is finetuned on hard negatives targeting single-phrase replacements generated by LLMs and rule-based methods; SVLC+Pos [6] is finetuned on the aforementioned hard negatives as well as paraphrases of the caption; DAC-LLM [5] is finetuned on several LLM-generated captions of the image as well as hard negatives generated by the SVLC method; and DAC-SAM [5] 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.

3.2 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 3.1, 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) 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 3.1, 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 (the 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 [25] which sees segments of the image extracted by SAM [22].

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 [5], 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 [8, 41], to evaluate captions for images [13], to evaluate text-to-image generation [16,38], and to evaluate text-to-video generation [14]. 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 Supplementary). 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 4.4.

3.3 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 3.2): 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.

4 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.

	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 Supplementary for results across all models.

4.1 Method

We first generate hard positives using LLAMA-2 70B-Chat [48]. 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 the Supplementary.

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 [29] process, as in SugarCrepe [15]. 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.

The finetuning follows the procedure outlined in SVLC [6]. 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 the Supplementary for implementation details.

4.2 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 [47, 55].

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		SWAP		REPLACE	SWAP
	Model	Orig. Test Acc.	Aug. Test Acc.	Orig. Test Acc.	Aug. Test Acc.	Brittleness (↓)	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 4.2). REPLACE averages performance on Attributes and Relations.

REPLACE. CLIP awards similar scores to all, seeming to ignore the replacement for both hard negatives and hard positives. For DAC-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 3.2).

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 3.2), 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 [55], 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 the Supplementary for further details.

4.3 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 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 tradeoff suggests the need for careful tuning to achieve the best understanding of both hard positives and hard negatives.

4.4 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.

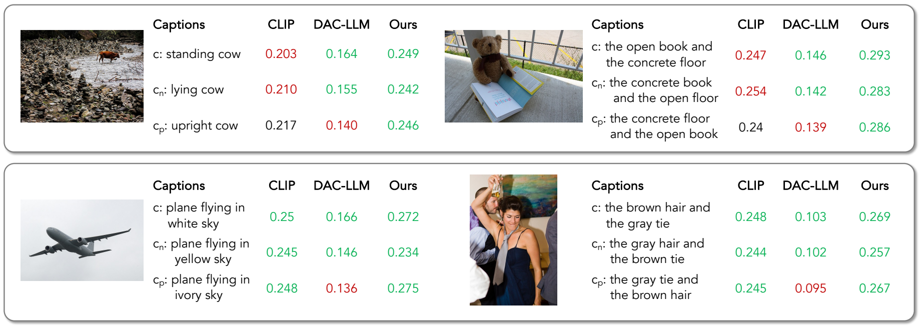
The top part shows similar behavior as depicted in Figure 1: the hard negative finetuned model appears to have achieved high compositionality when its performance on c and c_n is compared to CLIP — however, this is an incomplete picture. The hard negative finetuned model actually awards a lower score 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 §3.2. Our model does not exhibit this behavior (c.f. Table 2 and Supplementary).

5 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 [7], the use of *hard* positives has been absent from



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

Fig. 3: Sample predictions of CLIP, a hard negative finetuned model, 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.

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 [15]: for example, even basic text encoders such as word2vec [30, 31] 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?

Related Work. We contextualize our study within research aiming to improve the compositionality of vision-language models in the Supplementary. Our work complements benchmarks that assess vision-language models' compositionality [15, 20, 29, 36, 47, 55, 61] by introducing the notion of hard positives.

Limitations. While we have further analysis in the Supplementary, 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 [1], BLIP [25, 26], and GPT-4V [33], 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.

Conclusion. 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 [2, 49, 57]. 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.

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

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Bugliarello, E., Sartran, L., Agrawal, A., Hendricks, L.A., Nematzadeh, A.: Measuring progress in fine-grained vision-and-language understanding. arXiv preprint arXiv:2305.07558 (2023)
3. Cascante-Bonilla, P., Shehada, K., Smith, J.S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., Karlinsky, L.: Going beyond nouns with vision & language models using synthetic data (2023)
4. Cresswell, M.: Logics and languages (1973)
5. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., Karlinsky, L.: Dense and aligned captions (DAC) promote compositional reasoning in VL models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
6. Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L.: Teaching structured vision & language concepts to vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2657–2668 (2023)
7. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving CLIP training with language rewrites. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), `https://openreview.net/forum?id=SVjDiiVySh`
8. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
9. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
10. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324 (2018)
11. Hendricks, L.A., Nematzadeh, A.: Probing image-language transformers for verb understanding. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021)
12. Hendricks, L.A., Nematzadeh, A.: Probing image-language transformers for verb understanding. arXiv preprint arXiv:2106.09141 (2021)
13. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021)

14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A.A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. ArXiv **abs/2210.02303** (2022), `https://api.semanticscholar.org/CorpusID:252715883`
15. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
16. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
17. Hupkes, D., Dankers, V., Mul, M., Bruni, E.: Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research **67**, 757–795 (2020)
18. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`
19. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10236–10247 (2020)
20. Kamath, A., Hessel, J., Chang, K.W.: Text encoders bottleneck compositionality in contrastive vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
21. Kamath, A., Hessel, J., Chang, K.W.: 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 (2023)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
24. Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., Choi, Y.: Adversarial filters of dataset biases. In: International Conference on Machine Learning. pp. 1078–1088. PMLR (2020)
25. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
26. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
27. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022), `https://proceedings.mlr.press/v162/li22n.html`
28. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: 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. pp. 852–869. Springer (2016)

29. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? arXiv preprint arXiv:2212.07796 (2022)
30. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
31. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
32. OpenAI: Chatgpt (2022)
33. OpenAI: Gpt-4v(ision) system card (2023)
34. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., Gatt, A.: 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) (2022)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), `http://proceedings.mlr.press/v139/radford21a.html`
36. Ray, A., Radenovic, F., Dubey, A., Plummer, B.A., Krishna, R., Saenko, K.: Cola: How to adapt vision-language models to compose objects localized with attributes? (2023)
37. Reif, Y., Schwartz, R.: Fighting bias with bias: Promoting model robustness by amplifying dataset biases. arXiv preprint arXiv:2305.18917 (2023)
38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
39. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM **64**(9), 99–106 (2021)
40. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: 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 (2022), `https://openreview.net/forum?id=M3Y74vmsMcY`
42. Sharma, P., Ding, N., Goodman, S., Soricut, R.: 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). pp. 2556–2565 (2018)
43. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)

44. Singh, H., Zhang, P., Wang, Q., Wang, M., Xiong, W., Du, J., Chen, Y.: Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality (2023)
45. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6418–6428. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1644, https://aclanthology.org/P19-1644
46. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
47. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
48. Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D.M., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A.S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I.M., Korenev, A.V., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. ArXiv (2023)
49. Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. Advances in Neural Information Processing Systems **36** (2024)
50. Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.G., Yuan, L.: Omnivl: One foundation model for image-language and video-language tasks. arXiv preprint arXiv:2209.07526 (2022)
51. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)
52. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
53. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. pp. 23965–23998. PMLR (2022)
54. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7959–7971 (2022)

55. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), `https://openreview.net/forum?id=KRLUvxh8uaX`
56. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326 (2018)
57. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276 (2021)
58. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 25994–26009. PMLR (2022)
59. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., et al.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867 (2019)
60. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022)
61. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221 (2022)
62. Zheng, C., Zhang, J., Kembhavi, A., Krishna, R.: Iterated learning improves compositionality in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)