NatuRel: Advancing Relational Understanding in Vision-Language Models with Natural Language Variations

Stanford CS224N Custom Project

Laya Iyer
Stanford University
laya@stanford.edu

Ashna Khetan
Stanford University
ashnak@stanford.edu

Isabel Sieh
Stanford University
isabelrs@stanford.edu

Abstract

Vision-Language Models (VLMs) face challenges in accurately encoding compositional relationships, as highlighted by their modest performance on benchmarks like Winoground and the Attribution, Relation, and Order (ARO) tests. These benchmarks are designed to evaluate visio-linguistic compositional reasoning capabilities through tasks such as image-text retrieval. Despite notable advancements achieved through fine-tuning strategies on datasets like CoCo-Order by utilizing contrastive learning and hard negative mining, VLMs continue to struggle with out-of-domain reasoning. This is particularly evident in cases of unconventional captioning (e.g., "the water rests below the sail"), as demonstrated in the Winoground benchmark. To address these challenges, we introduce NatuRel, an expansive dataset comprising over 100,000 images and 660,000 image-caption-pair examples, enriched with 2,500 uniquely crafted captions ranging from easy to hard, positive to negative, and varied in sentence structure and natural language expression. This dataset is designed to encapsulate both simplistic and intricate captions, aiming to bridge the existing gaps in visio-linguistic compositional understanding. We fine-tune two models, NegCLIP and SigLIP, on the NatuRel dataset. Notably, SigLIP is adapted to fine-tune using multiple positive and negative captions for each image, diverging from the traditional single-caption fine-tuning method. This comprehensive fine-tuning process on visio-linguistic compositional reasoning enables us to achieve state-of-the-art accuracies on our NR-Simple, NR-Complex, and Winoground-Object benchmarks, as well as achieve competitive performance on ARO's VG-Relations and Winoground-All benchmarks.

Mentor: Yuhui Zhang
Contributions:
Laya Iyer: SigLIP evaluation and fine-tune; related works, figures, future work, poster
Ashna Khetan: dataset generation, fine-tune NegCLIP; approach, experiments, VG-R analysis
Isabel Sieh: prompting, CLIP-related eval; Winoground analysis, intro, related work, results write-up

1 Introduction

Vision-Language Models (VLMs) have demonstrated remarkable capabilities across various tasks including text-image generation and image-text retrieval (Radford et al. (2021); Singh et al. (2022); Li et al. (2022)). However, the recent Winoground benchmark (Thrush et al. (2022)) testing compositional understanding, indicates that embedding-based VLMs like CLIP, Flava, VisualBERT perform near chance. The task asks to match two images with two captions of the same text but with two words swapped (e.g."there is a mug in some grass" vs. "there is some grass in a mug").

Findings suggest that VLMs perform like a bag of words; Yuksekgonul et al. (2023) observed that models trained on contrastive loss (e.g. CLIP (Radford et al. (2021))) can achieve high performance

on traditional benchmarks even without order information unless datasets are carefully designed. Embedding-based VLMs compresses compositional information from an image or text into a single vector representation which usually encodes object importation but loses information on relationships between the objects Rösch and Libovický (2022a). Improving datasets and our model architecture, will help avoid VLMs taking shortcuts that avoids understanding nuanced compositional relationships.

Yuksekgonul et al. (2023) proposes the Attribution, Relation, and Order (ARO) benchmark that assess composition and order understanding on a larger scale. Their accompanying fine-tuned model NegCLIP saw significant improvements on their ARO benchmark but marginal improvements on Winoground. This points to the challenges of Winoground (Diwan et al. (2022)) not addressed in current data; for instance, ARO uses simple language (e.g. "cat on table") but avoids awkward captioning (e.g. "the water rests below the sail"). Thus, our first contribution is NatuRel, a 100,000+ image and 660,000+ text dataset where each image contains multiple positive (correct) and negative (incorrect) relationship captions with natural language variation .

Further, NegCLIP (Yuksekgonul et al. (2023)) and other standard contrastive models are only exposed to one positive and if any, one hard negative example (Zhai et al. (2023)). This limits the possible captions. Our second contribution is the exploration of novel architectural strategies by modifying NegCLIP and SigLIP (Zhai et al. (2023)) to allow for fine-tuning on multiple positive and negative examples per image. SigLIP, trained on sigmoid loss, also allows us to explore if there is a variance in performance between sigmoid loss and traditional contrastive loss.

Our contribution – a rich dataset with natural language variation and expanding the fine-tuning process to encompass multiple examples – advances VLMs towards a more nuanced understanding of language for visio-linguistic compositional reasoning.

2 Related Work

Language-Image Pretraining. VLMs like CLIP (Radford et al. (2021)) often use contrastive learning to further their compositional understanding. Contrastive learning is the method by which a model is exposed to positive – similar or related – and negative – dissimilar or unrelated – examples from a dataset (Chen et al. (2020)) thus refining its understanding by distinguishing between aligned and misaligned pairs. CLIP's approach to image-to-text training treats pairs of images and their matching text descriptions as positive examples, and all other pairs as the negative examples (Radford et al. (2021)). NegCLIP doubles the input matrix that CLIP takes for a training example, increasing computation cost, to allow a hard negative caption and image (Yuksekgonul et al. (2023)). A hard negative is an example that is mismatched but closely resembles the correct example (Robinson et al. (2021)). Hard negatives work well for encoding compositional relationships because they focus on fine-grained differences (Yuksekgonul et al. (2023)); for instance, this can distinguish between "cat on table" and "table on cat". Other models like BLIP (Li et al. (2022)), and FLAVA (Singh et al. (2022)) also use contrastive loss. However, it often requires a global view of losses which is computationally expensive. Instead, SigLIP (Zhai et al. (2023)) uses sigmoid loss which is a single image-text loss and allows for an increase in batch size without affecting the task; SigLip can achieve CLIP's accuracy in less than a fifth of the time. Simultaneously, SigLIP allows for multiple negative and positive captions for each image without incurring the large cost of having to compare against the rest of the dataset.

VLM Compositionality Benchmarks. Winoground (Thrush et al. (2022)) tested viso-linguistic reasoning with 400 test cases based on a wide variety of linguistic tags, revealing that VLMs perform near chance. Winoground requires not just compositional understanding, but also common sense and the ability to untangle complex natural-sounding sentences (Diwan et al. (2022)). The ARO benchmark (Yuksekgonul et al. (2023)) tackles composition and order on a larger scale (50,000 test cases) by introducing four tasks: Visual Genome Attribution, testing the understanding of object's properties; Visual Genome Relations, testing relational understanding; and COCO Order and Flickr30k Order, testing order sensitivity. ARO works by having an image with a true and false (swapped words) caption. ARO prompted the creation of other compositionality benchmarks including ColorSwap (Burapacheep et al. (2024)) and SugarCrepe (Hsieh et al. (2023)). SugarCrepe found significant biases in ARO, claiming ARO is hackable; SugarCrepe attempts to "faithfully" measure the compositional understanding of vision-language models by using ChatGPT to generate

plausible and fluent hard negatives. Compared to SugarCrepe, we add captions with more natural language variety.

Improving VLM Compositional Understanding. Aside from NegCLIP, other approaches have attempted to help VLMs form strong compositional understanding. Rösch and Libovický (2022b) proposed a spatial relations classifier based on the coordinates of objects in an image and their bounding boxes and performed contrastive learning to encode positional information in VLMs. CoCoT (Zhang et al. (2024)) uses prompting for large language models like GPT-4V and Gemini to guide their understanding of images, and then tests its performance on benchmarks like Winoground. While performance is significantly improved, more information from prompting is provided to the model, whereas we aim to focus on improving the encoding of compositional information. We focus on a language and model architecture approach.

3 Approach

This section contains the initial context of our NatuRel dataset. It also explains the architecture of two models, NegCLIP and SigLIP, which we fine-tune, and describes our baselines.

NatuRel Dataset. NatuRel contains over 100,000 images and 330,000 examples. For each image, we have 8 total captions: 1 simple positive and negative caption (e.g. "woman plays guitar" v.s. "guitar plays woman"), and 3 complex positive and negative (e.g. "a musical performance given by a woman with a guitar" "A woman is far away from a guitar") captions. The images and annotations (e.g. class: "woman", relationship: "plays") come from OpenImages V4 (Kuznetsova et al. (2020)) but the $\sim$ 2,500 complex captions are an original contribution.

NegCLIP. NegCLIP (Yuksekgonul et al. (2023)) takes in negative captions and images for a given sample. The negative image is sampled from strong alternatives in the dataset, and the negative caption is sampled from a list generated by perturbing the order of the positive caption.

The architecture of CLIP's (Radford et al. (2021)) loss calculation is altered as seen in Figure 2. Given a batch of images I_N and captions T_N, NegCLIP concatenates each I_N's negative caption to T_N to obtain T_{2N} and computes the similarity matrix $S \in R^{N2N}$. Both row-wise and column-wise cross-entropy losses are computed like in CLIP, but NegCLIP does not compute loss column-wise for the negative captions since there is no corresponding image.

To focus on the *language* encoder and decoder, we utilize NegCLIP, holding the negative image constant as a sentinel blank image. For each sample image, we train it with one positive caption and one negative caption at a time, training it on the same relation multiple times for each caption pair, as discussed in *Experiments*.

We also made minor modifications to NegCLIP for it to better handle data samples with empty caption lists, GPU space constraints (emptied torch cache after every batch, chunking CSV reads), and to save checkpoints more frequently.

SigLIP. SigLIP (Zhai et al. (2023)) is a vision-language model that uses a sigmoid loss approach rather than the traditional contrastive learning approach used in CLIP (Radford et al. (2021)). Rather than requiring a global view of losses in contrastive loss, sigmoid loss is calculated using a single image-text loss, making it more computationally efficient. Figure 3details how sigmoid loss is performed for the model, resulting in a lower training time than contrastive learning.

We fine-tuned SigLIP using two positive and two negative captions but were only able to run a single epoch due to computational limitations. Details about possible approaches to fine-tuning are discussed in the future work section below.

Baselines. We have three fine-tuned models: NegCLIP-SP, NegCLIP-AP, and SigLIP-TP (see: *5.3 Experimental details*). We compare these three fine-tuned models against three baseline models: CLIP, NegCLIP, and SigLIP. Note that NegCLIP is a ViT-B/32 variant of CLIP on the COCO dataset with hard negatives. We compare the models on the performance on Visual Genome Relation (VG-Relations), Winoground, SugarCrepe, and NatuRel test data (see: *5.2 Evaluation method*). NegCLIP-NR2 includes the first 20% of Winoground, so our Winoground evaluation on NegCIP-NR2 uses the second 80% of Winoground.

4 Experiments

4.1 Data

NatuRel Data Overview. NatuRel contains ∼100,000 images, ∼330,000 examples, with 330 unique (class, relationship, class) triplets, and ∼2,500 captions. The final form of our data consists of images paired with three complex positive captions, formed directly from the relationship triplet. It also contains a list of six negative captions, from which one will be randomly sampled for use with NegCLIP.

(Image) Filepath	List of 3 Complex Positive Captions	List of 3 Complex Negative Captions	1 Simple Positive Caption	1 Simple Negative Caption

NatuRel Data Generation. We detail how we generated the NatuRel dataset.

1. First, we filter the OpenImages V4 dataset (Kuznetsova et al. (2020)) for the 100,000 images (330,000 examples, as there may be multiple relations per image) that contain annotations in the form of relationship triplets (class, relationship, class). We also filtered out relationships with relationship triplet (class, relationship, adjective) and the relationship "interacts with", as that can be interpreted as a bidirectional relationship.
2. We then use the relationship triplets to generate 330 positive and negative simple and complex captions. *Simple* positive and negative captions are made directly from the form (class1, relationship, class2) and reversed (class2, relationship, class1), respectively (e.g. "cat on table", "table on cat"). Then, we use GPT-3.5-Turbo (Brown et al. (2020)) to generate additional *complex* captions that vary in natural language and sentence structure. For "cat on table", additional positive and negative captions are:

Positive Complex	**Negative Complex**
A cat rests atop the table.	A table rests atop the cat.
On top of the table, a cat can be found.	On top of the cat, a table can be found.
Above the table is a cat.	A cat is next to a table.
The cat finds itself upon a table.	A table is to the right of the cat.

Prompting details. Negative captions either switch the order of the classes (e.g. "A table rests atop the cat." or the type of relationship (e.g. "A cat is next to a table.) We filtered the negative complex captions that make commonsense. VLMs are more likely to confuse captions that both make commonsense, having negative captions with incorrect semantics may provide a shortcut for VLMs to choose the positive caption (Hsieh et al. (2023)). Negative examples such as "The beer grips the man." may not contribute to meaningful learning. Our multi-step generation prompt to GPT can be found in the appendix. We then cleaned excessive whitespace, commas, bullet points, and header words, and regenerated missing captions, to maintain data consistency. Our exploration on prompting is visualized in Figure 4, and our multi-step generation prompt to GPT can be found in Figure 5.

Training Data. We test our fine-tuned models on NatuRel, and additionally on the VG-Relations benchmark proposed by Yuksekgonul et al. (2023) and the Winoground benchmark (Thrush et al. (2022)). Since our images vary in style, source, and caption language from these datasets, we added images from COCO and Winoground to our training data, to avoid an out-of-distribution problem. We added COCO since NegCLIP was fine-tuned on a part of it, which was in-distribution for VG-Relations. Specifically, our training data consists of the first 20% of COCO Validation 2014, and the first 20% of Winoground, where we consider each image its own example.

Test and Validation Data. We used a 70-10-20 train-valid-test split for NatuRel.

4.2 Evaluation method

We test on 3 benchmarks: Visual Genome Relation (VG-Relations), Winoground, and NatuRel test data. For all benchmarks we calculate accuracy. Figure 6provides more detail about what each benchmark tests for as well as the number of images in each benchmark.

VG-Relations is one of the four tasks in ARO Yuksekgonul et al. (2023). Given an image and two captions X *relation* Y and Y *relation* X, we test if the model can pick the correct caption. The original VG-Relations test seen in Yuksekgonul et al. (2023) crops images based on the smallest bounding box containing both classes. By evaluating NegCLIP on VG-Relations we found that removing this feature made a marginal difference in performance (0.80 cropped, 0.79 not cropped). Since the other tests do not crop the images, we took this feature out to show the whole image.

Winoground (Thrush et al. (2022)) comes in entries of two images I_0 and I_1 and captions C_0 and C_1. Each entry is assessed on text, image, and overall score. An entry has a text score of 1 if a model can select the correct caption given an image (selecting C_0 over C_1 given I_0 *and* vice versa given I_1). An entry has an image score of 1 if a model can select the correct image given a text (selecting I_0 over I_0 given C_0 *and* vice versa given C_1) The overall score is 1 if both text and image score are 1. We focus on the text score since it best emulates our other benchmarks. Since Winogrand assesses all kinds of word swaps (noun, adjective, adverb, etc.), we also look to observe the text score for Object word swaps (e.g. "[a person] holding up [books]").

NatuRel (NR) has NR-Simple and NR-Complex. NR-Simple contains an image with a true and false simple caption. NR-Complex contains an image with a randomly-selected true and false complex caption. For both tests, we test if the model can pick the correct caption.

SugarCrepe Hsieh et al. (2023) tests if, given an image and a positive and hard negative caption, the model can pick the positive caption. SugarCrepe has 3 forms of hard negatives: REPLACE form, which replaces an atomic object, attribute, or relation with a new concept; SWAP form, which swaps two atomic object or attribute concepts in the text; ADD form, which adds a new atomic object or attribute concept. Each test (REPLACE-Object, REPLACE-Attribute, REPLACE-Relation, SWAP-Object, SWAP-Attribute, ADD-Object, ADD-Attribute) is assessed separately. On SugarCrepe, given an image, a model is to select the positive caption that correctly describes the image against another hard negative text distractor that differs from the positive only by small compositional changes.

For SigLIP, given its novelty in HuggingFace Transformers API, more work was involved to provide images in the form expected by the model for evaluation. Images with alpha layers were reduced to RGB-formatted images and Black and White Images were modified to include a third channel.

4.3 Experimental details

We ran two fine-tuning experiments on NegCLIP and another on SigLIP. Each of these fine-tuned models was then evaluated, as described above, on our three benchmarks.

NegCLIP Single-Pair (NegCLIP-SP). Our training data consisted of 1 randomly selected complex positive and 1 randomly selected complex negative caption from NatuRel. We limited to 1 to fit NegCLIP's architecture and to be mindful of computation. This resulted in 165,501 examples. We used the pre-trained OpenAI ViT-B-32 weights, and our images were preprocessed using CLIP's ViT-B-32 preprocessing technique.

We used a batch size of 32 to fit our GPU constraints and trained on 5 epochs. We used an initial learning rate of 1e-6 with an AdamW Optimizer and cosine learning rate scheduling, which adjusts the learning rate on every batch update and allows the model to converge to a different local minimum on each restart. Additionally, we were able to use 2 workers on 1 GPU and a warmup of 50, so that the model could view some samples at initialization without expending much time.

The model took 7 hours to train ($\sim$ 1.4 per epoch), and our loss dropped from $\sim$ 7.8 to 2.6.

NegCLIP All-Pairs (NegCLIP-AP). In this experiment, we paired up each positive complex caption with a negative complex and paired up the two simple captions, generating four caption pairs per image. We included each as a separate example in our training data, bringing the size to 662,004 examples. We also included the 20% of COCO and Winoground subsets in this training data, bringing the grand total to 850k examples.

For this reason, we decided to run 2 epochs on a random 50% of this data. All other parameters remained the same. Our experiment took 12 hours, and the loss dropped from $\sim$4.8 to 2.3.

SigLIP Two-Pairs (SigLIP-TP). We attempted to fine-tune SigLIP using NatuRel but due to a lack of computational resources, we were only able to run one epoch (with a batch size of 8 and learning rate 1e-5) and didn't adjust hyperparameters. We believe that a more thorough finetuning of SigLIP would

lead to substantial improvements in performance. Future research could focus on the fine-tuning and evaluation of a fine-tuned SigLIP.

4.4 Results

Model	VG-Relations	NatuRel		Winoground (text)	
		Simple	Complex	All	Object
CLIP	0.59	0.47	0.52	0.31	0.36
SigLIP	0.46	0.38	0.49	0.14	0.12
SigLIP-TP	0.53	0.45	0.31	0.14	0.12
NegCLIP	0.81	0.87	0.62	0.31	0.31
NegCLIP-SP	0.65	0.70	0.84	0.19	0.23
NegCLIP-AP	0.80	0.99	0.84	0.28	0.33

Table 1: Results of Various Models on Various Benchmarks

Note that NegCLIP indicates CLIP fine-tuned on hard-negatives from CoCo order, and NegCLIP-SP and NegCLIP-AP similarly indicate CLIP fine-tuned on their respective data. NegCLIP-SP and NegCLIP-AP are not fine-tuned on top of NegCLIP but rather use the same architecture. SigLIP-TP represents the version of SigLIP that is finetuned on two positive and two negative captions.

In the NatuRel (NR) task, we find that for our first two baselines, CLIP and SigLIP, performance on NR-Complex is slightly better than NR-Simple, whereas, in NegCLIP, performance in Simple is much higher than complex. This could be because NegCLIP is finetuned on CoCO Order data similar to this format. The disparity between NR-Simple and NR-Complex reveals that NegCLIP may be taking a shortcut to relation tasks that take the form of NR-Simple, rather than a more comprehensive understanding that NR-Complex may assess.

NegCLIP-SP, which was only trained on complex data performs well on Complex, and worse on Simple. However, the disparity between these two tasks is less and may indicate a more holistic understanding. Before including COCO in our training data, NegCLIP-SP also performed worse on VG-Relations, a problem fixed with NegCLIP-AP. Finally, NegCLIP-AP performs well on both NR-Simple and NR-Complex. NegCLIP-AP and NegCLIP-SP have nearly the same performance on NR-Complex, which may be because NegCLIP-SP also had exposure to complex data.

In the VG-Relations and Winoground task, NegCLIP-AP performed similarly to NegCLIP, which points to further discussion on the strengths of NegCLIP-AP explored in *Analysis*.

Model	REPLACE			SWAP		ADD	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
CLIP	0.91	0.80	0.69	0.61	0.63	0.77	0.68
NegCLIP	0.93	0.86	0.76	0.76	0.75	0.89	0.83
NegCLIP-SP	0.89	0.74	0.68	0.60	0.62	0.81	0.66
NegCLIP-AP	0.91	0.83	0.73	0.69	0.72	0.84	0.76

Table 2: Results on SugarCrepe

In the SugarCrepe Task, NegCLIP-AP performs slightly lower than NegCLIP on all sub-tasks. Moreover, aside from the ADD-Object, NegCLIP-SP performs slightly worse than CLIP. This could be due to SugarCrepe's benchmark being built on CoCo's 2017 validation dataset; NegCLIP is finetuned on CoCo-Order, whereas NegCLIP-SP is only fine-tuned on OpenImages V4 and NegCLIP-AP is finetuned on a mix of image-text pairs.

5 Analysis

Examples. We look at the images and true captions where NegCLIP classifies the image properly but NegCLIP-AP does not, and vice-versa.

Starting with *VG-Relations*, the models demonstrate differing levels of accuracy when tasked with classifying captions that describe the same image scene in varying linguistic structures (Table 3). For example, in Figure 1, NegCLIP correctly classifies the captions: ['the cows is to the left of the grass', 'the cow is to the left of the grass', 'the cow is to the left of the pasture', 'the dog is to the left of the pasture', 'the dog is to the left of the pasture'], while NegCLIP-AP correctly classifies: ['the pasture is to the right of the dog', 'the pasture is to the right of the cow', 'the pasture is to the right of the cows', 'the pasture is to the right of the dog', 'the grass is to the right of the cow']. We observe that NegCLIP-AP performs well on captions that are considered 'reverse' descriptions which, while linguistically valid, are less intuitive from a human perspective, as they reverse the more natural subject-object spatial ordering ('the dog is to the left of the pasture' vs 'the pasture is to the right of the dog'). This is consistent with the way we trained NegCLIP-AP with natural language versions of these 'reverse' relations, making them appear as more viable answers.

Figure 1: Example from VG-Relations

Captions NegCLIP gets correct, while NegCLIP-AP does not	**Captions NegCLIP-AP gets correct, while NegCLIP does not**
['the door is to the left of the shirt', 'the man is to the right of the door', 'the plate is on top of the table', 'the girl is to the left of the shirt', 'the legs is of the cat', 'the cows is to the left of the grass', 'the cow is to the left of the grass', 'the cow is to the left of the pasture']	['the door is to the left of the man', 'the shirt is to the right of the door', 'the table is under the plate', 'the plate is of the sandwiches', 'the player is playing with the racket', 'the building is to the left of the sky', 'the pasture is to the right of the dog', 'the pasture is to the right of the cow']

Table 3: VG-Relations: examples that NegCLIP and NegCLIP-AP disagree on.

For *Winoground*, we also explored examples where NegCLIP and NegCLIP-AP performed differently (Table 4. Despite the similar text accuracies, there were 34 examples where NegCLIP correctly matched captions to a given image and NegCLIP-AP did not, and 34 vice versa. However, upon observation, we found no strong language or visual pattern of errors. Examples NegCLIP-AP got correct over NegCLIP ranged from "[sail] [boat]" to "the orange on the [left] is moldy while the orange on the [right] is fresh". The range of errors could be from Winoground's extreme diversity in test cases; this makes finding failure patterns across a small (400) benchmark difficult.

By-Relation Performance We noticed a few interesting stats when looking at the by-relationship accuracies. Initially, while most relations ("at", "inside of", "on") had high accuracies above 95%, a few ("holds", "hits") seemed lower, around 66%, with "hits" receiving 3% accuracy. We note that "hits" was only in a singular relationship triple in the test set, so if this triple was largely misclassified, the low accuracy makes sense.

Data Quality We also noticed that for a lone few examples like "girl holds guitar", one of the complex "negative" captions was a positive caption, due to GPT misgeneration or mis-parsing of the response.

Generalizability Compared to the hard-negative mining approach taken from Yuksekgonul et al. (2023), we offer the model a breadth of negative examples, encouraging it to learn patterns on its own rather than guiding it directly to the image-text pairs it pairs incorrectly. Our primary purpose for this is to support development of a generalizable model rather than optimizing for accuracy. This might provide insight into why our model does not perform as well as NegCLIP on some benchmarks.

Natural Language Variation and Negative Hard Images NegCLIP-AP which is one of our fine-tuned versions of Neg-CLIP incorporating some data from CoCo and Winoground in order to avoid an out-of-distribution issue, performs close to the accuracy of negCLIP on VisualGenome. This tells us that even without providing a negative image example with each caption (we provided a white box), we are able to achieve similar results to vanilla NegCLIP. Natural Language variation achieves similar results to providing a negative image example.

6 Future Work

Hyperparameter tuning. Adjusting the learning rate and batch size for CLIP-finetune would find the most efficient hyperparameters to improve accuracy. This would require multiple iterations to compare different combinations of batch size and learning rate. Also, we must find the optimal number of epochs to run to make sure that the model does not overfit to the data.

SigLIP Fine Tuning. Although we attempted fine-tuning, we were unable to achieve significant changes in evaluation. Reasons for this that could be examined by future studies are: the number of epochs was not enough to see a significant difference and the learning rate is either too high or low to see a significant difference.

VG-Attributions. Currently, our NatuRel Dataset only contains caption variations focused on the "[class] [relationship] [class]" format but in the future this dataset could be expanded to include attributional information about the classes. As an example, if we have an image with a caption "a black cat and white sofa", we can generate variations in captions such as "a white sofa and black cat." This would allow us to evaluate our fine-tuned model on VG-Attributions in addition to VG-Relations.

Annotator's Review of the Dataset. Captions generated by GPT may not be accurate 100% of the time so having human annotators go through the dataset and remove malformed examples would help improve the accuracy of our fine-tuning and evaluation. This is a costly process for a large dataset, so a possible extension is taking a subsection of our dataset and cleaning up the captions for that subsection. In order to provide varied data, we could consider optimizing for distinct relationship triplets.

Negative Image Mining. To strengthen our contrastive approach, we could leverage NegCLIP's hard image for each sample, as well. Currently, we pass a white rectangular image as our negative image with each negative caption and instead, we could use an image of the negative caption to give our model more useful information when fine-tuning.

Experiment with Natural Language. Though we experimented with the number of natural language captions, we didn't vary the way we generated them. For example, we could analyze whether it is better to filter out non-common sense captions. We could also test whether extra irrelevant details in a caption e.g. "On a sunny day, a cat sits atop the table." leads to overfit as hypothesized, or adds valuable variation. Like SugarCrepe, we could compose a taxonomy for data generation that includes specific types of 'natural language variations'.

7 Conclusion

Overall, our contribution is threefold: a larger image dataset with positional information, natural language variation in captions, and the effect of sigmoid loss on vision-language model's understanding of positionality. Current datasets focused on probing or informing vision language models with positional information are very limited in the number of unique images they provide. In fact, what is considered the most robust compositional benchmark, Winoground, is a dataset of only 800 images. Using a large image dataset such as openimages as a baseline allows us to be more thorough with the positional examples that we provide the vision-language models during finetuning. As previously mentioned, awkward captioning is another issue with negative caption generation for these datasets, resulting in a bias for the vision language model to pick out the caption that sounds less like natural language rather than understanding the compositionality of the image; vision language models tend to find shortcuts. We generate our captions with natural language variation to prevent this issue from occurring. Finally, we evaluate the effect of the state-of-the-art sigmoid loss function on a SigLIP's understanding of image compositionality.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. 2024. Colorswap: A color and word order dataset for multimodal evaluation.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

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

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.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples.

Philipp J. Rösch and Jindřich Libovický. 2022a. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Philipp J. Rösch and Jindřich Libovický. 2022b. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model.

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.

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?

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs.

A Appendix: Figures and Tables

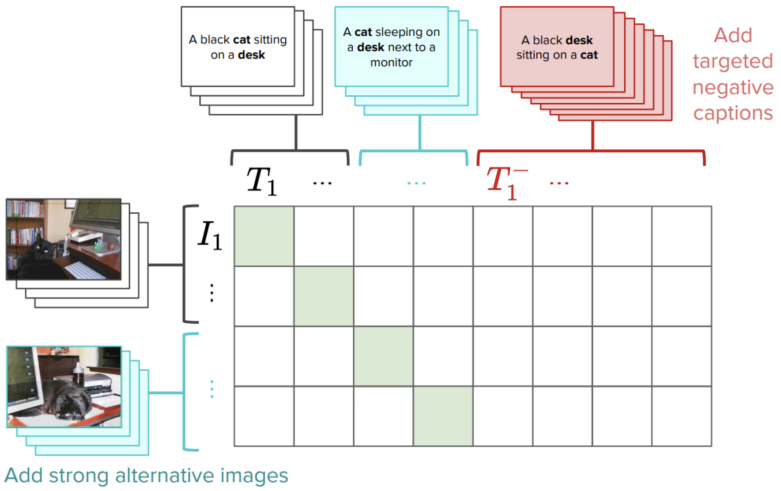
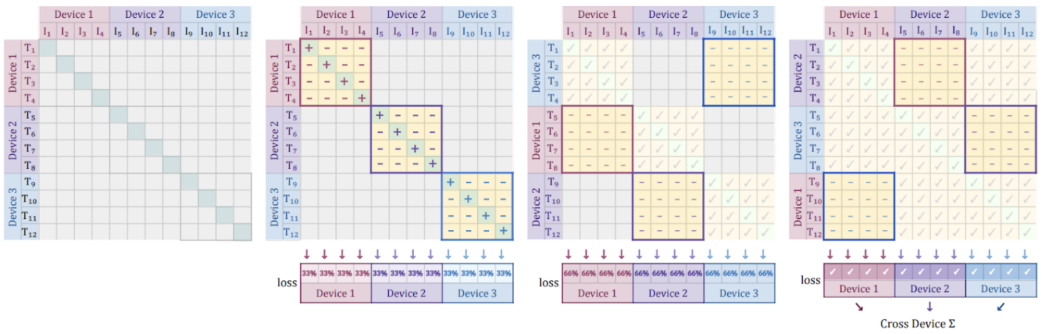


Figure 2: Diagram taken from Yuksekgonul et al. (2023) visualizing the NegCLIP input matrix



(a) Initially each device holds 4 image and 4 text representations. Each device needs to see the representations from other devices to calculate the full loss.

(b) They each compute the component of the loss (highlighted) for their representations, which includes the positives.

(c) Texts are swapped across the devices, so device 1 now has $I_{1:4}$ and $T_{5:8}$ etc. The new loss is computed and accumulated with the previous.

(d) This repeats till every image & text pair have interacted, e.g. device 1 has the loss of $I_{1:4}$ and $T_{1:12}$. A final cross-device sum brings everything together.

Figure 3: Diagram taken from Zhai et al. (2023) describing how sigmoid loss operates

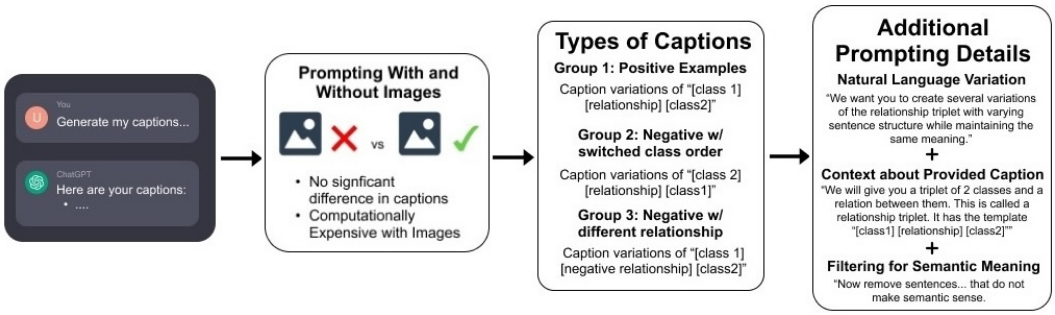


Figure 4: Our exploratory steps taken with prompting to generate captions

PROMPT 1

We will give you a triplet of 2 classes and a relation between them. This is called a relationship triplet. It has the template "[class1] [relationship] [class2]"

For example if the relationship triplet is "cat on table", then [class1] is "cat", [class2] is "table", and [relationship] is "on".

From this relationship triplet, we want you to produce 3 groups each with 6 phrases.

The first group has several variations of the relationship triplet with varying sentence structure while maintaining the same meaning. With the "cat on table" example we expect a response of:

A cat rests atop the table.
Atop the table sits the cat.
On top of the table, a cat can be found.
A table is under the cat.
Above the table is a cat.
The cat finds itself upon a table.

The second group has the 2 classes switched to create a negative relationship triplet template "[class2] [relationship] [class1]". Take the literal output of the first group and directly switch the classes. With the "cat on table" example the negative relationship triplet becomes "table on cat", and we expect a response of:

A table rests atop the cat.
Atop the cat sits the table.
On top of the cat, a table can be found.
A cat is under the table.
Above the cat is a table.
The table finds itself upon a cat.

The third group maintains the same class order but with a negative relationship of a different meaning to create a negative2 relationship triplet template "[class1] [negative-relationship] [class2]", and has several variations of the negative2 relationship triplet with varying sentence structure while maintaining an opposite meaning to the original relationship. Some possibilities for relationships are: "plays", "on", "holds", "at", "wears", "inside of", "under", "hits", "to the right of", "to the left of" . With the "cat on table" example the negative2 relationship triplet becomes "cat [negative-relationship] table", and we expect a response of:

A cat is next to a table.
Table is beside a cat.
The cat is in front of the table.
A table is to the right of the cat.
The cat is to the left of a table.
A cat inside a table.

Now do the same for "<s>". The output should be a list of these 3 groups in order.

PROMPT 2

Now remove sentences from the second and third group (negative relationships) that do not make semantic sense.
Create a list of these 3 groups in order, with 3 examples from the first group, and 3 examples chosen randomly from the second and third group.

Output just this unnumbered list of 6 total examples.

Figure 5: This image shows our final prompt used to generate all our positive and negative captions.

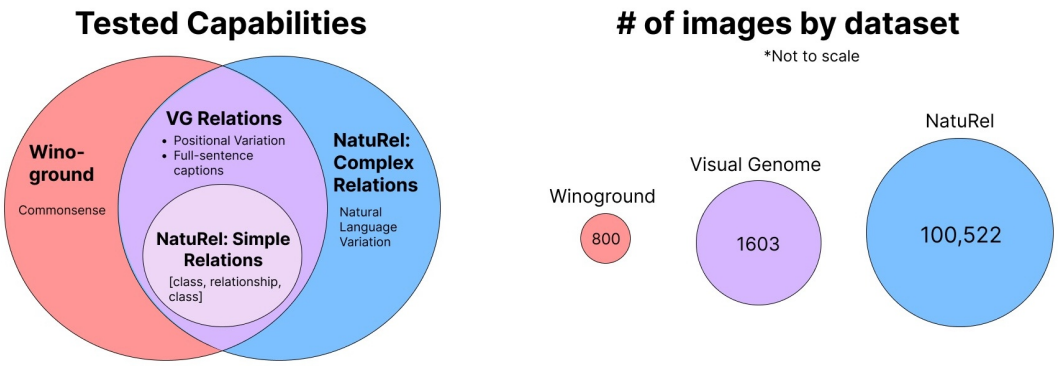


Figure 6: This figure shows what each benchmark tests for as well as the sizes of each of the image datasets for each of them.

Winoground: Captions NegCLIP gets correct, while NegCLIP-AP does not	**Winoground: Captions NegCLIP-AP gets correct, while NegCLIP does not**
the flip flops are too big for these feet	a dog sitting on a couch with a person lying on the floor
the feet are too big for these flip flops	a person lying on a couch with a dog sitting on the floor
handing a hammer	meat with potatoes
hammering a hand	potatoes with meat
there's one blue and many yellow balls	The dog rides without a visible tongue
there's one yellow and many blue balls	The dog rides with a visible tongue out
there are more silver coins than gold coins	three white and two brown eggs
there are more gold coins than silver coins	two white and three brown eggs
milk cow	boat house
cow milk	house boat

Table 4: Winoground Analysis: examples that NegCLIP and NegCLIP-AP disagree on.