CREPE: Can Vision-Language Foundation Models Reason Compositionally?

Zixian Ma\(^1\)*, Jerry Hong\(^1\)*, Mustafa Omer Gul\(^2\)*, Mona Gandhi\(^3\), Irena Gao\(^1\), Ranjay Krishna\(^4\)
Stanford University\(^1\), Cornell University\(^2\), University of Pennsylvania\(^3\), University of Washington\(^4\)
{zixianma, jerryhong, irena}@cs.stanford.edu mog29@cornell.edu mona09@seas.upenn.edu ranjay@cs.washington.edu

Abstract

A fundamental characteristic common to both human vision and natural language is their compositional nature. Yet, despite the performance gains contributed by large vision and language pretraining, we find that—across 7 architectures trained with 4 algorithms on massive datasets—they struggle at compositionality. To arrive at this conclusion, we introduce a new compositionality evaluation benchmark, CREPE, which measures two important aspects of compositionality identified by cognitive science literature: systematicity and productivity. To measure systematicity, CREPE consists of a test dataset containing over 370K image-text pairs and three different seen-unseen splits. The three splits are designed to test models trained on three popular training datasets: CC-12M, YFCC-15M, and LAION-400M. We also generate 325K, 316K, and 309K hard negative captions for a subset of the pairs. To test productivity, CREPE contains a test dataset containing over 17K image-text pairs with nine different complexities plus 278K hard negative captions with atomic, swapping and negation foils. The datasets are generated by repurposing the Visual Genome scene graphs and region descriptions and applying handcrafted templates and GPT-3. For systematicity, we find that model performance decreases consistently when novel compositions dominate the retrieval set, with Recall@1 dropping by up to 9%. For productivity, models’ retrieval success decays as complexity increases, frequently nearing random chance at high complexity. These results hold regardless of model and training dataset size.

1. Introduction

Compositionality, the understanding that “the meaning of the whole is a function of the meanings of its parts” [11], is held to be a key characteristic of human intelligence. In language, the whole is a sentence, made up of words. In vision, the whole is a scene, made up of parts like objects, their attributes, and their relationships [31, 35].

*Equal contribution

Figure 1. We introduce CREPE, a benchmark to evaluate whether vision-language foundation models demonstrate two fundamental aspects of compositionality: systematicity and productivity. To evaluate systematicity, CREPE utilizes Visual Genome and introduces three new test datasets for the three popular pretraining datasets: CC-12M, YFCC-15M, and LAION-400M. These enable evaluating models’ abilities to systematically generalize their understanding to seen compounds, unseen compounds, and even unseen atoms. To evaluate productivity, CREPE introduces examples of nine complexities, with three types of hard negatives for each.

Through compositional reasoning, humans can understand new scenes and generate complex sentences by combining known parts [6, 27, 30]. Despite compositionality’s importance, there are no large-scale benchmarks directly evaluating whether vision-language models can reason compositionally. These models are pretrained using large-scale image-caption datasets [62, 64, 74], and are already widely applied for tasks that benefit from compositional reasoning, including retrieval, text-to-image generation, and open-vocabulary classification [10, 57, 60]. Especially as such models become ubiquitous “foundations” for other models [5], it is critical to understand their compositional abilities.

Previous work has evaluated these models using image-text retrieval [32, 56, 82]. However, the retrieval datasets used either do not provide controlled sets of negatives [45, 74] or study narrow negatives which vary along a single axis
Figure 2. An overview of the systematicity retrieval set generation process. First, a model’s image-caption training set is parsed to identify what atoms and compounds the model has seen. Then, an evaluation set is divided into three compositional splits according to whether the model has seen all the compounds (Seen Compounds), only all the atoms of the caption (Unseen Compounds), or neither (Unseen Atoms). Finally, hard negative captions HN-ATOM and HN-COMP are generated for the hard negatives retrieval set $D_{HN}^{test}$.

(e.g., permuted word orders or single word substitutions as negative captions) [21, 51, 65, 75]. Further, these analyses have also not studied how retrieval performance varies when generalizing to unseen compositional combinations, or to combinations of increased complexity.

We introduce CREPE (Compositional REPresentation Evaluation): a new large-scale benchmark to evaluate two aspects of compositionality: systematicity and productivity (Figure 1). Systematicity measures how well a model is able to represent seen versus unseen atoms and their compositions. Productivity studies how well a model can comprehend an unbounded set of increasingly complex expressions. CREPE uses Visual Genome’s scene graph representation as the compositionality language [35] and constructs evaluation datasets using its annotations. To test systematicity, we parse the captions in three popular training datasets, CC-12M [8], YFCC-15M [74], and LAION-400M [62], to identify atoms (objects, relations, or attributes) and compounds (combinations of atoms) present in each dataset. For each training set, we curate corresponding test sets containing $385K$, $385K$ and $373K$ image-text pairs respectively, with splits checking generalization to seen compounds, unseen compounds, and unseen atoms. To test productivity, CREPE contains $17K$ image-text pairs split across nine levels of complexity, as defined by the number of atoms present in the text. Examples across all datasets are paired with various hard negative types to ensure the legitimacy of our conclusions.

Our experiments—across 7 architectures trained with 4 training algorithms on massive datasets—find that vision-language models struggle at compositionality, with both systematicity and productivity. We present six key findings: first, our systematicity experiments find that models’ performance consistently drops between seen and unseen compositions; second, we observe larger drops for models trained on LAION-400M (up to a 9% decrease in Recall@1); third, our productivity experiments indicate that retrieval performance degrades with increased caption complexity; fourth, we find no clear trend relating training dataset size to models’ compositional reasoning; fifth, model size also has no impact; finally, models’ zero-shot ImageNet classification accuracy correlates only with their absolute retrieval performance on the systematicity dataset but not systematic generalization to unseen compounds or to productivity. \(1\)

2. Related Work

Our work lies within the field of evaluating foundation models. Specifically, we measure visio-linguistic compositionality. To do so, we create a retrieval benchmark with hard negatives.

Contrastive Image-Text Pretraining. The recently released contrastively trained CLIP model [56] has catalyzed a wide array of work at the intersection of Computer Vision and Natural Language Processing. Since its release, CLIP has enabled several tasks, ranging from semantic segmentation to image captioning, many of which have remarkable zero-shot capability [12, 16, 38, 56, 71, 73]. CLIP has been used as a loss function within image synthesis applications [29, 44, 46, 54, 79, 83], acted as an automated evaluation metric [22, 52], used successfully as a feature extractor for various vision and language tasks [66], and incorporated into architectures for various tasks including dense prediction and video summarization [43, 50, 55, 58, 67, 68]. This success has also encouraged the design of other contrastive vision and language pretraining algorithms for image [15, 18, 40–42, 48, 69, 80, 81] and video domains [39, 76, 78]. Our work evaluates how well

\(1\) We release our datasets, and code to generate and evaluate on our test sets at https://github.com/RAIVNLab/CREPE.
such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).

In our work, we focus on two aspects of compositionality: systematicity and productivity. While there is a plethora of benchmarks for systematic generalization within Computer Vision [3, 4, 19, 33] and Machine Learning [34, 36, 59], the subject has been almost unexplored for vision-language models, largely due to lack of benchmarks complementary to the different large-scale training datasets. To address this, CREPE provides a benchmark with three different datasets to evaluate the compositional generalization of vision-language models. Productivity, on the other hand, has been studied only for specialized tasks [19] or toy domains [27, 36, 59]. CREPE evaluates productivity by using an image-text retrieval task featuring captions of varying compositional complexity.

**Evaluation with hard negatives.** Like us, past work evaluating models has commonly designed tasks featuring hard negatives to isolate particular model capabilities while overcoming the limitations of prior evaluation tasks. Using atomic foils that replace an atom in the image or text with a distractor has been the most common strategy [4,9,21,24,51,53,65]. Notably, Park et al. [53] targets verbs and person entities in videos; COVR [4] studies question answering with distractor images; VALESE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).

In our work, we focus on two aspects of compositionality: systematicity and productivity. While there is a plethora of benchmarks for systematic generalization within Computer Vision [3, 4, 19, 33] and Machine Learning [34, 36, 59], the subject has been almost unexplored for vision-language models, largely due to lack of benchmarks complementary to the different large-scale training datasets. To address this, CREPE provides a benchmark with three different datasets to evaluate the compositional generalization of vision-language models. Productivity, on the other hand, has been studied only for specialized tasks [19] or toy domains [27, 36, 59]. CREPE evaluates productivity by using an image-text retrieval task featuring captions of varying compositional complexity.

**Evaluation with hard negatives.** Like us, past work evaluating models has commonly designed tasks featuring hard negatives to isolate particular model capabilities while overcoming the limitations of prior evaluation tasks. Using atomic foils that replace an atom in the image or text with a distractor has been the most common strategy [4,9,21,24,51,53,65]. Notably, Park et al. [53] targets verbs and person entities in videos; COVR [4] studies question answering with distractor images; VALESE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).

In our work, we focus on two aspects of compositionality: systematicity and productivity. While there is a plethora of benchmarks for systematic generalization within Computer Vision [3, 4, 19, 33] and Machine Learning [34, 36, 59], the subject has been almost unexplored for vision-language models, largely due to lack of benchmarks complementary to the different large-scale training datasets. To address this, CREPE provides a benchmark with three different datasets to evaluate the compositional generalization of vision-language models. Productivity, on the other hand, has been studied only for specialized tasks [19] or toy domains [27, 36, 59]. CREPE evaluates productivity by using an image-text retrieval task featuring captions of varying compositional complexity.

**Evaluation with hard negatives.** Like us, past work evaluating models has commonly designed tasks featuring hard negatives to isolate particular model capabilities while overcoming the limitations of prior evaluation tasks. Using atomic foils that replace an atom in the image or text with a distractor has been the most common strategy [4,9,21,24,51,53,65]. Notably, Park et al. [53] targets verbs and person entities in videos; COVR [4] studies question answering with distractor images; VALESE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).

In our work, we focus on two aspects of compositionality: systematicity and productivity. While there is a plethora of benchmarks for systematic generalization within Computer Vision [3, 4, 19, 33] and Machine Learning [34, 36, 59], the subject has been almost unexplored for vision-language models, largely due to lack of benchmarks complementary to the different large-scale training datasets. To address this, CREPE provides a benchmark with three different datasets to evaluate the compositional generalization of vision-language models. Productivity, on the other hand, has been studied only for specialized tasks [19] or toy domains [27, 36, 59]. CREPE evaluates productivity by using an image-text retrieval task featuring captions of varying compositional complexity.

**Evaluation with hard negatives.** Like us, past work evaluating models has commonly designed tasks featuring hard negatives to isolate particular model capabilities while overcoming the limitations of prior evaluation tasks. Using atomic foils that replace an atom in the image or text with a distractor has been the most common strategy [4,9,21,24,51,53,65]. Notably, Park et al. [53] targets verbs and person entities in videos; COVR [4] studies question answering with distractor images; VALESE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).

In our work, we focus on two aspects of compositionality: systematicity and productivity. While there is a plethora of benchmarks for systematic generalization within Computer Vision [3, 4, 19, 33] and Machine Learning [34, 36, 59], the subject has been almost unexplored for vision-language models, largely due to lack of benchmarks complementary to the different large-scale training datasets. To address this, CREPE provides a benchmark with three different datasets to evaluate the compositional generalization of vision-language models. Productivity, on the other hand, has been studied only for specialized tasks [19] or toy domains [27, 36, 59]. CREPE evaluates productivity by using an image-text retrieval task featuring captions of varying compositional complexity.

**Evaluation with hard negatives.** Like us, past work evaluating models has commonly designed tasks featuring hard negatives to isolate particular model capabilities while overcoming the limitations of prior evaluation tasks. Using atomic foils that replace an atom in the image or text with a distractor has been the most common strategy [4,9,21,24,51,53,65]. Notably, Park et al. [53] targets verbs and person entities in videos; COVR [4] studies question answering with distractor images; VALESE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap such contrastively trained models capture a fundamental property present in human vision and language: compositionality.

**Compositionality.** Compositionality allows us to comprehend an infinite number of scenes and utterances [37]. For an AI model, compositionality would not only allow for systematic, combinatorial generalization, but would also confer benefits such as controllability [5]. This promise prompted a wealth of work on both designing [2, 23, 25] and evaluating to empirically measure those behaviors via retrieval (3.4).
Figure 3. An overview of the productivity retrieval set generation process. By performing random walks on the scene graphs of an evaluation dataset, we generate subgraphs of various complexities. Then, for complexities $n \in \{4, 5, \ldots, 12\}$ and three hard negative types, we populate the retrieval set $D_{\text{HN, test}}^n$ by generating a ground truth caption for each $n$-subgraph and hard negatives for each caption.

Table 1. We summarize the sizes of the eight evaluation datasets we create for systematicity and productivity evaluation.

<table>
<thead>
<tr>
<th>Systematicity</th>
<th>Productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{\text{RAW, test}}$ (number of image-text pairs)</td>
<td>$D_{\text{HN, test}}^n$ (number of texts)</td>
</tr>
<tr>
<td>Training data</td>
<td>Dataset size</td>
</tr>
<tr>
<td>CC-12M</td>
<td>385,777</td>
</tr>
<tr>
<td>YFCC-15M</td>
<td>385,777</td>
</tr>
<tr>
<td>LAION-400M</td>
<td>373,703</td>
</tr>
<tr>
<td>Any</td>
<td>17,553</td>
</tr>
</tbody>
</table>

Comprehension is testing comprehension over increasingly complex scenes. Now, an image $I$ does not have a notion of complexity, since it is theoretically infinitely describable; on the other hand, we can define a notion of complexity for a caption $T$: the number of atoms in its corresponding scene graph $|S_T|$. Therefore, a productive vision-language model should be able to match a given image to the correct corresponding caption, regardless of that caption’s complexity. To evaluate productivity, we define a range of productivity complexity (in our case, $n = 4, 5, \ldots, 12$). We need splits of the evaluation dataset based on these complexities, where image-text pairs in a given split have a fixed complexity $n$, and evaluate a model’s performance over each split.

3.4. Compositional evaluation via retrieval

We evaluate compositional reasoning using zero-shot image-to-text and text-to-image retrieval. This formulation probes the representation space as directly as possible and is already the most common evaluation method for vision-language foundation models [56]. Theoretically, any existing image-text dataset can be used as retrieval sets for our evaluation. However, one challenging limitation in existing datasets renders the metrics evaluated on them inaccurate. Consider using an image query of a “plant inside a yellow vase on top of a black television.” Retrieving unintended alternative positives (e.g., “a black television”) is not necessarily incorrect. Similarly, if no other texts in the retrieval set contain a “plant” and a “television”, retrieving the correct text doesn’t suggest that the model comprehends the image. Ideally, to properly evaluate a model, the retrieval dataset should contain hard negatives for every query. A hard negative is a caption that does not faithfully represent the corresponding image, and differs from the ground truth caption by some minimal atomic shift. An example hard negative for the query above is “man inside a yellow vase on top of a black television.” By erring in a single, granular syntactic or semantic fashion, hard negatives allow for variations in retrieval performance to be attributable to a specific failure mode of a model’s compositional comprehension (see Appendix). We address this need for a new benchmark dataset to evaluate the systematicity and productivity of vision-language models.

4. CREPE: a large-scale benchmark for vision-language compositionality

There are several challenges to creating image-text retrieval datasets that evaluate compositional systematicity and productivity. For systematicity, the primary challenge lies in parsing the training dataset for seen atoms and compounds in order to split the data into the three compositional splits. For productivity, the major challenge is generating image-text pairs across different text complexities for the retrieval sets. For both datasets, it is crucial to enumerate different types of hard negatives, and to design an automated hard negative generator which ensures the incorrectness of the
negatives it generates. We detail our methods for tackling these challenges for future efforts that attempt to create similar benchmarks for other training datasets.

4.1. Creating systematicity datasets

To create the three systematicity splits—SC, SA, UA—we parse a given training dataset \( D \) into its constituent atoms and compounds, filter low-quality data, and generate hard negatives (Figure 2).

**Parsing a dataset into atoms and compounds** Since we utilize the scene graph representation as our compositional language, we use the Stanford Scene Graph Parser [63,77] to parse texts in \( D_{\text{train}} \) into their corresponding scene graphs with objects, attributes and relationships. Since the parser only parses for objects and relationships, we further extract the attributes from the text via spaCy’s natural language processing parser by identifying adjective part-of-speech tags. These connected objects, attributes, and relationships constitute our seen atoms and compounds. Similarly, we parse a given \( D_{\text{test}} \) and divide all the image-text pairs into the three splits based on the presence of unseen atoms and/or compounds in the parsed training set. Details on the quality of the scene graph parser can be found in the Appendix.

**Filtering low-quality data** We perform the following filtering steps on the image-text pairs in all splits: we only keep region crops which have an area greater than or equal to \( 40K \) pixels, occupy at least 10\% of the whole image, and whose width-to-height ratio is between 0.5-2.0. We only include text which have at least 2 atoms and 1 compound and de-duplicate text using their corresponding scene graphs.

**Generating hard negatives** We introduce two types of hard negatives: HN-ATOM and HN-COMP. HN-ATOM replaces \( A_a, A_o, \text{ or } A_r \) in the text with an atomic foil. For example, for the caption “a grill on top of the porch”, one HN-ATOM can replace “underneath the porch”, where the \( A_r \) “on top of” is replaced by “underneath”. Since captions and scene graphs are not exhaustive, this replacement must be done carefully. For example, if a dog is white and furry, but only “white” is annotated, replacing the atom “white” with “furry” would result in a correct caption. To minimize errors, we employ WordNet [49] to pick replacement atoms that are either antonyms (“black dog”) or share the same grand-hypernym (“pink dog”) with respect to the original atom. Furthermore, we use BERT to select the most sensical negatives for each ground truth caption [13,51]. HN-COMP concatenates two compound foils where each contains an atomic foil. For instance, one HN-COMP of the caption “a pink car” can be “a blue car and a pink toy”, where “blue” and “toy” are the atomic foils in the two compounds foils “blue car” and “pink toy”. We only generate negatives for one-compound examples for systematicity evaluation, as productivity covers complex captions with more atoms.

4.2. Creating productivity datasets

We first generate ground truth captions for scene graphs of varying complexity, filter for data quality, and then generate hard negatives for each example (Figure 3).

**Generating captions** We systematically generate captions of different atom counts for each image. Given a scene graph, we perform a random walk of length \( n \) through the graph to generate a subgraph. Each subgraph corresponds to a specific region of the image, determined by the union of the bounding boxes of the subgraph atoms. We filter out low-quality regions using the same process as systematicity with additional deduplication on patches that overlap by \( \geq 75\% \). For simple subgraphs \( (n = 4) \), we produce captions using handcrafted templates. For larger subgraphs \( (n \geq 5) \), we leverage GPT-3 [7] (text-davinci-002) to generate captions based on a text description of the scene graph, which lists all objects and relationships. We prompt GPT-3 using 5 manually written captions per complexity, filtering out captions where GPT-3 errs and omits atoms from the subgraph during generation (see more details in Appendix).

**Generating hard negatives** For productivity, we employ three hard negatives types (HN-AToM from systematicity, HN-SWAP, and HN-NEG) corresponding to three hypothesized model error modes. First, as a caption’s complexity increases, a model may begin to ignore individual atoms. HN-AToM randomly selects an atom from the caption and replaces it with an incorrect atom. Second, as a caption’s complexity increases, a model may treat captions as “bags of words”, ignoring syntactic connections built out of word order. A swap hard negative (HN-SWAP) accordingly permutes atoms of the same subtype in a caption. This hard negative is similar to Winoground [75], but in the context of varying caption complexity. On top of Wordnet, we use entailment with RoBERTa to further filter errant HN-SWAP hard negatives [47]. Finally, as a caption’s complexity increases, a model may begin to lose comprehension of negations. A negation hard negative (HN-NEG) either negates the entire caption or a specific atom. Refer to the Appendix for details on generating HN-SWAP and HN-NEG.

4.3. The final benchmark datasets

For both productivity and systematicity, we generate two test datasets: \( D_{\text{test}}^{HN} \), which contains image-ground truth text pairs along with all generated hard negatives, and \( D_{\text{test}}^{RAW} \), which contains only image-ground truth text pairs. To measure the data quality, we randomly sample 2\% of productivity ground truth captions generated by GPT-3 and 1\% of the queries in the productivity and systematicity \( D_{\text{test}}^{HN} \) sets for manual human verification. We assign 2 annotators to each set and measure both generated quality and intra-annotator agreement. 87.9\% of sampled productivity ground truth captions generated by GPT-3 are rated as faithful to the image, with an average pairwise annotator agreement of
Figure 4. We plot models’ recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types. We observe a consistent drop in models’ performance from the SC to UC split when the hard negative set consists of both HN-ATOM and HN-COMP or HN-ATOM only.

Figure 5. Productivity Analysis. We plot models’ Recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets. We find that models’ ability to retrieve the ground-truth degrades as complexity increases.

88.8%. 83.7% of productivity and 86.0% of systematicity hard negatives were rated as genuine negatives (i.e. made factually incorrect statements about the image), with pairwise annotator agreements of 84.3% and 83.7% respectively.

5. Experiments

We present our experimental setup and results with six takeaways. First, our systematicity experiments show performance decreases consistently on compounds unseen in training. Second, the greatest drop between splits occurs for models trained on LAION-400M. Third, our productivity results reveal models’ retrieval performance decays with increasing complexity. Fourth, we find that dataset size has no impact on compositionality. Fifth, we find no clear trend relating model size to compositionality. Finally, models’ zero-shot ImageNet classification accuracy correlates with retrieval performance on the systematicity dataset but not systematic generalization to the UC split or productivity.

Datasets. We utilize Visual Genome to create our test datasets. For systematicity, image patches and corresponding spelling-corrected region descriptions are used. We provide three different splits for $D_{test}^{HN}$, for three training datasets: CC-12M, YFCC-15M and LAION-400M. For productivity, Visual Genome’s image-scene graph pairs are used to create captions and hard negatives for $D_{test}^{RAW}$ and $D_{test}^{HN}$ (Table 1).

Models. We firstly evaluate seven vision-language models pretrained with contrastive loss [70] across three commonly used image-text datasets: Conceptual Captions 12M (CC-12M) [8], a subset of the YFCC100M dataset (YFCC-15M) [56, 74] and LAION-400M [62]. We limit our evaluation to models openly released in the OpenCLIP repository [28] for systematicity evaluation. These include ResNet (RN) [20] and Vision Transformer (ViT) [14] encoders of different sizes: RN50, RN101, ViT-B-16, ViT-B-16-plus-240, ViT-B-32 and ViT-L-14. Additionally, since productivity evaluation is not restricted to models that were trained on publicly released datasets, we conduct productivity evaluation on other foundation vision-language models as well. Specifically, we consider OpenAI’s CLIP [56] with ResNet and ViT backbones, CyCLIP [18] (a variant of CLIP introducing auxiliary losses that regularize the gap in similarity scores between mismatched pairs, trained on Conceptual Captions 3M [64] with a ResNet-50 [20] backbone), AL-BEF [41] (additionally trained with a masked language modeling and image-text matching loss) and FLAVA [69] (which
We plot models’ Recall@1 on the productivity hard negatives retrieval set against complexity, where OpenAI CLIP’s performance is averaged across five models RN50, RN101, ViT-B-16, ViT-B-32 and ViT-L-14. We find that all models’ retrieval performance decreases as complexity increases in both the HN-ATOM and HN-SWAP retrieval sets. For the HN-NEG set, all models except for CLIP either drop in performance or remain at random chance.

Figure 7. A plot showing the correlation between zero-shot top-1 accuracy on ImageNet and Recall@1 on CREPE’s systematicity hard-negative sets. We observe a strong correlation, with an $R^2$ score of 0.9914 for the SC split and 0.9534 for the UC split.

further adds unimodal losses for image and text domains. **Retrieval.** For $D_{test}^{HN}$, we perform image-to-text retrieval and stratify results by split and hard negative type. For systematicity, the splits are SC, UC, and UA; for productivity, the splits are by caption complexity $n$ (denoted $D_{test}^{HN,n}$). Each retrieval task is between one image and its ground truth caption plus $h$ hard negatives of a single type (see Appendix). We adopt commonly used retrieval metrics Recall@1, 3, 5 and Average Recall@K. For $D_{test}^{RAW}$, retrieval experiments are described in the Appendix.

5.1. Systematicity evaluation

**Model performance on the $D_{test}^{HN}$ dataset for systematicity decreases monotonically when compounds are unseen.** We first observe a monotonic decrease in recall@1 from the Seen Compounds to the Unseen Compounds split on the systematicity $D_{test}^{HN}$ set consisting of both HN-ATOM and HN-Comp (Figure 4 left). This drop is relatively small ($1 - 5\%$) for the CC-12M and YFCC-15M trained models and the most pronounced for models trained on the largest dataset LAION-400M [62], with the decrease reaching 7.9% for the ViT-B-32 model. However, CC-12M and YFCC-15M models also significantly underperform LAION-400M models in general, meaning that small drops between sets may be due to overall poor performance rather than improved systematic generalization. In comparison, human oracle experiments generalize with 100% accuracy to $D_{test}^{HN}$.

Similar to the overall results, there is also a consistent discrepancy between the SC and UC split on the $D_{test}^{HN,n}$ subset consisting of HN-ATOM only (Figure 4 center). This drop is consistently smaller ($3 - 6\%$) for models trained on CC-12M and YFCC-15M, but pronounced ($6\%$ or higher, reaching 9.4% drop for ViT-B-32) for LAION-400M models.

On the HN-Comp subset (Figure 4 right), we find little to no difference in performance between the SC and UC split. We hypothesize that this is due to the lower difficulty of the HN-Comp hard negatives, as they introduce more foils to the caption, are always longer than the ground truth, and thus offer more opportunities for the model to correctly distinguish the ground truth. This hypothesis is supported by the fact that Recall@1 values on HN-Comp are similar or higher than the ones on HN-ATOM even though the HN-Comp retrieval set size is larger than that of HN-ATOM.

Figure 6. Productivity Analysis on Additional Foundation Vision-language Models. We plot models’ Recall@1 on the productivity hard negatives retrieval set against complexity, where OpenAI CLIP’s performance is averaged across five models RN50, RN101, ViT-B-16, ViT-B-32 and ViT-L-14. We find that all models’ retrieval performance decreases as complexity increases in both the HN-ATOM and HN-SWAP retrieval sets. For the HN-NEG set, all models except for CLIP either drop in performance or remain at random chance.

5.2. Productivity evaluation

**Models’ performance decreases with complexity on HN-ATOM and HN-SWAP negatives.** At small complexities such as $n = 4$, we observe that model retrieval quality is well above random chance (Figure 5). However, as caption complexity increases, we observe a steady decrease in performance, nearing random chance for HN-Swap and dipping below it for HN-ATOM negatives. Similarly, we find that the same downward trend persists for other vision-language foundation models (Figure 6). Importantly, the downward trend occurs for FLAVA and ALBEF even though their training set contains Visual Genome images. We note that for HN-NEG negatives, the OpenAI CLIP models do not adhere...
Figure 8. A plot showing the correlation between zero-shot top-1 accuracy on ImageNet and Recall@1 on CREPE’s productivity hard negative sets for complexities of 4, 8 and 12. Overall, we find no correlation between ImageNet accuracy and Recall@1 on our productivity sets. The strongest correlations are $R^2$ scores of 0.284 for HN-NEG negatives on $n = 12$ and of 0.222 for HN-ATOM negatives on $n = 8$.

to the downward trend, achieving their lowest scores for the lowest complexity. Their performances on higher complexities, however, show great variation. In short, our conclusion is that vision-language foundation models struggle with productivity. Our results on models released by OpenCLIP [28] as well as other vision-language foundation models demonstrate the challenge of differentiating between atomic and swapping foils is exacerbated by caption complexity.

We see no effect of dataset size on productivity. We do not observe a clear advantage for larger pretraining datasets in our productivity evaluation. For atomic and swapping foils, we see similar performance for models trained on the three datasets, with slightly worse performance on atomic foils for the CC-12M trained models. However, on negation hard negatives (Figure 5), we see variable performance across training sets, with CC-12M models outperforming larger models trained on larger datasets YFCC and LAION.

5.3. Effect of model size

We find no trends relating compositionality to model size. Overall, we note that the LAION trained models (which are both larger models and trained on larger datasets) achieve significantly better absolute performances than smaller models. However, model’s systematicity and productivity remain indifferent to the size of the model itself (Figures 4 and 5).

5.4. Correlation with ImageNet performance

We find that zero-shot ImageNet top-1 accuracy strongly correlates with models’ Recall@1 on the systematic retrieval set. Specifically, we acquire $R^2$ scores of 0.984 and 0.877 for the $SC$ and $UC$ splits respectively (Figure 7). However, this correlation does not imply that models’ zero-shot ImageNet performance correlates with systematic generalization, which is instead indicated by small or no difference between the $SC$ and $UC$ splits. On our productivity dataset, we do not observe such a strong correlation, where the highest $R^2$ score is 0.284 for HN-NEG negatives on a complexity of $n = 12$ (Figure 8). As such, we can infer that successful zero-shot performance on ImageNet does not necessarily lead to better performance on our productivity sets.

6. Discussion

Limitations. First, although our data validation protocols verified our generated hard negatives for productivity as high-quality, approximately 70% of HN-SWAP and of HN-NEG negatives were rated as correct. While this does not invalidate our key productivity result, this noise is a limitation of CREPE and could hinder future evaluations once foundation models begin performing better. Second, our evaluation only covers a limited set of vision-language foundation models that were trained with contrastive loss. Additionally, given the computational requirements associated with training a foundation model, our experiments centered around model architectures that were already available publicly. We hope that future foundation models are evaluated with our publicly available CREPE benchmark. Third, while we observe text-to-image and image-to-text retrieval to have similar trends for our systematicity experiments, we lack text-to-image datasets with hard negatives. Future work can explore mechanisms to generate counterfactual negative images.

Conclusion. We present CREPE, a collection of text-to-image and image-to-text retrieval datasets for evaluating pretrained vision-language models’ systematicity and productivity. We demonstrate that models struggle with compositionality along both axes, with performance drops across different compositional splits and increasing complexity. We expect that CREPE will provide a more systematic evaluation to benchmark the emergence of compositionality as future models improve. Finally, researchers can leverage our hard-negative generation process to create training batches with hard negatives to incentivize vision-language compositionality.
References


[77] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic


