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 $17K$ 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].

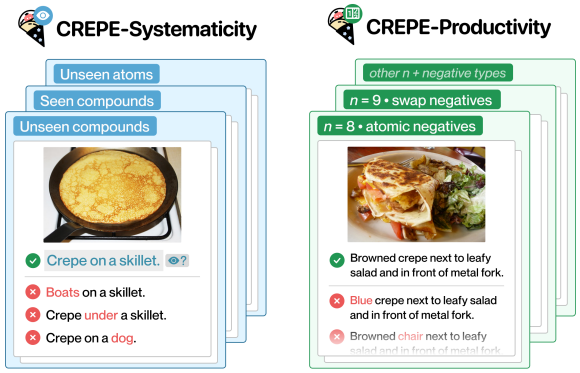

Figure 1. We introduce CREPE, a benchmark to evaluate whether vision-language foundation models demonstrate two fundamental aspects of compositionality: systematicity and productivity. To evaluate systematicity, CREPE utilizes Visual Genome and introduces three new test datasets for the three popular 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

*Equal contribution

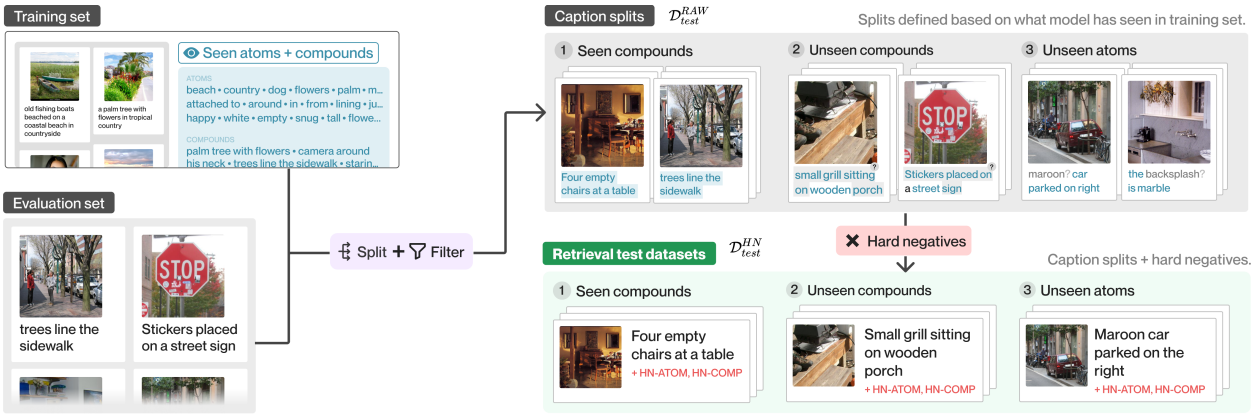

Figure 2. An overview of the **systematicity** retrieval set generation process. First, a model's image-caption training set is parsed to identify what atoms and compounds the model has seen. Then, an evaluation set is divided into three compositional splits according to whether the model has seen all 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 $\mathcal{D}_{test}^{HN}$.

(*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 [17, 19, 26, 36, 72] compositional models. 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; VALSE [51] targets linguistic phenomena such as existence, cardinality and the recognition of actions and spatial relationships. Another strategy has been to swap atoms within a caption to test whether models behave akin to a bag-of-words [1, 51, 75]. In particular, Winoground [75] introduces a set of 800 human edited negatives to evaluate compositionality; it is the closest related work to us. We complement Winoground by scaling it up by three orders of magnitude, by decomposing compositionality into systematicity and productivity, and by studying a variety of different types of hard negatives.

3. Compositional evaluation

The following section builds from the formally vacuous principle of compositionality to a well-defined evaluation scheme [27]. First, we establish the syntax and semantics of the composed language (Section 3.1). Then, we define expected behaviors from a model that achieves comprehension of said language (3.2, 3.3). Finally, we establish how to empirically measure those behaviors via retrieval (3.4).

3.1. Compositional language of visual concepts

To evaluate vision-language models, we find that a compositional language consisting of *scene graph* visual concepts to be an appropriate foundation [35]. Accordingly, an *atom* A is defined as a singular visual concept, corresponding to a single scene graph node. Atoms are subtyped into *objects* A_o, *relationships* A_r, and *attributes* A_a. A *compound* C is defined as a primitive composition of multiple atoms, which corresponds to connections between scene graph nodes. Visual concepts admit two compound types: the attachment of attribute to objects ("black dog") C_{ao}, and the attachment of two objects via a relationship ("man hugs child") C_{oro}.

The composition of these compounds form subgraphs S, which can be translated to natural language captions T. Conversely, captions T derived from image-text datasets $\mathcal{D}$ can be parsed to become scene graphs S. This extensible language is capable of capturing a number of linguistic phenomena identified in existing literature [51, 72], including the existence of concepts ("a photo with *flowers*"), spatial relationships ("a grill *on the left of* a staircase"), action relationships ("a person *throwing* a frisbee"), prepositional attachment ("A bird with *green* wings"), and negation ("There are *no* trucks on the road"). Furthermore, while this study focuses on visual concepts, scene graphs featuring common-sense relationships or other more abstract concepts can be designed; therefore, our methodology is widely applicable [61].

3.2. Systematicity

With our compositional language in place, we now define two dimensions of compositionality—systematicity and productivity—which we adapt to vision-language representations. *Systematicity* evaluates a model's ability to systematically recombine seen atoms in compounds. Concretely, let $\textsc{Seen}(A, D)$ denote if an atom is seen in a training dataset $\mathcal{D}$, namely $\exists (I, S) \in \mathcal{D} : A \in S$, and $\textsc{Seen}(C, D)$ denote if a compound is seen in a dataset $\mathcal{D}$, namely $\exists (I, S) \in \mathcal{D} : C \subseteq S$. To evaluate systematicity, we define three compositional splits: Seen Compounds (SC), Unseen Compounds (UC) and Unseen Atoms (UA). SC is the split where all compounds (and thus all atoms) of every caption have been seen in the training dataset, *i.e.* $\mathcal{D}_{\text{SC}} = \{(I, S) \in \mathcal{D}_{test} \,|\, \forall C \subseteq S : \textsc{Seen}(C, \mathcal{D}_{train})\}$. UC is the split where, for each caption, all atoms have been seen but at least one compound has NOT, *i.e.* $\mathcal{D}_{UC} = \{(I, S) \in \mathcal{D}_{test} \,|\, (\forall A \in S : \textsc{Seen}(A, \mathcal{D}_{train}) \wedge (\exists C \subseteq S : \neg\text{Seen}(C, \mathcal{D}_{train}))\}$. UA is the split where each caption contains at least one atom that has NOT been seen, *i.e.* $\mathcal{D}_{UA} = \{(I, S) \in \mathcal{D}_{test} \,|\, \exists A \in S : \neg\textsc{Seen}(A, \mathcal{D}_{train})\ \}$.

3.3. Productivity

Productivity refers to a capacity to comprehend an unbounded set of expressions. Since the set of atoms in any dataset is finite, a reasonable substitute for testing unbounded

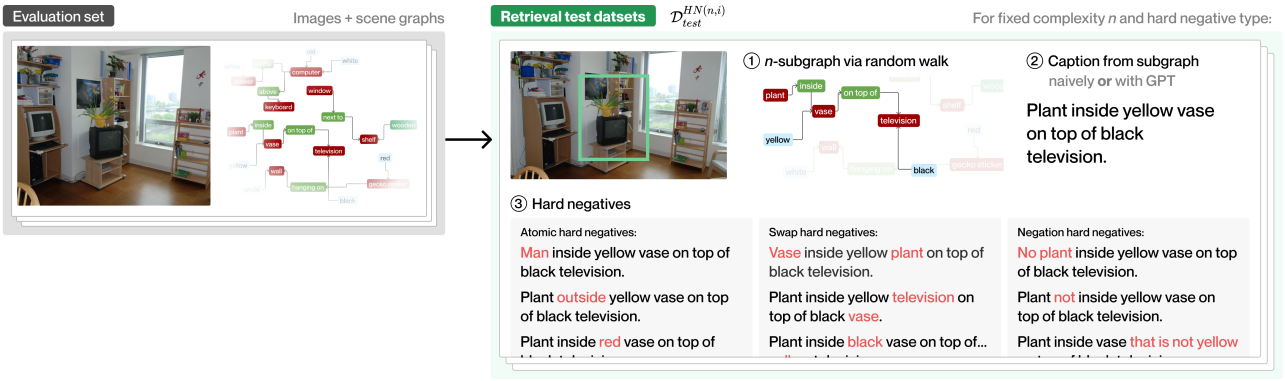

Figure 3. An overview of the **productivity** retrieval set generation process. By performing random walks on the scene graphs of an evaluation dataset, we generate subgraphs of various complexities. Then, for complexities $n \in \{4, , 5 \ldots, 12\}$ and three hard negative types, we populate the retrieval set $\mathcal{D}_{test}^{HN}$ 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.

	Systematicity						Productivity	
	$\mathcal{D}_{test}^{RAW}$ (# of image-text pairs)			$\mathcal{D}_{test}^{HN}$ (# of texts)			$\mathcal{D}_{test}^{RAW}$	$\mathcal{D}_{test}^{HN}$
Training data	CC-12M	YFCC-15M	LAION-400M	CC-12M	YFCC-15M	LAION-400M	Any	Any
Dataset size	385,777	385,777	373,703	325,523	316,668	309,342	17,553	278,730

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 $|S_T|$. [2] 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. A 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

[2] By avoiding captions with redundant objects ("... a lamb and a lamb and...") and abstract modifiers ("there are many lampposts"), we ensure atom count is tightly coupled with caption complexity.

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 $\mathcal{D}$ 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 $\mathcal{D}_{train}$ 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 $\mathcal{D}_{test}$ 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, 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 be "a grill *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: $\mathcal{D}_{test}^{HN}$, which contains image-ground truth text pairs along with all generated hard negatives, and D_{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 $\mathcal{D}_{test}^{HN}$ 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

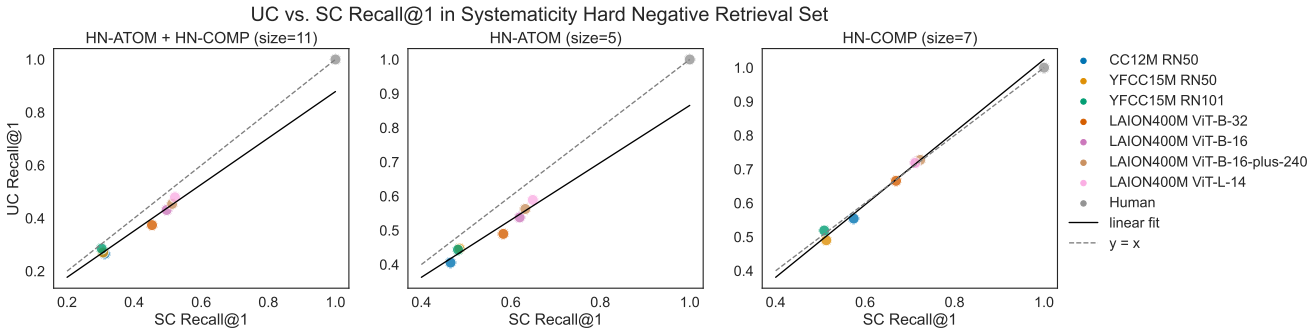

Figure 4. We plot models' recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types. We observe a consistent drop in models' performance from the SC to UC split when the hard negative set consists of both HN-ATOM and HN-COMP or HN-ATOM only.

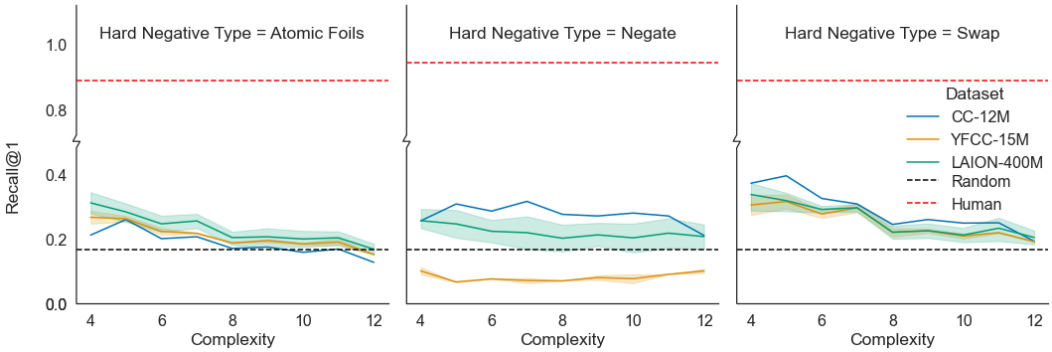

Figure 5. *Productivity Analysis.* We plot models' Recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets. We find that models' ability to retrieve the ground-truth degrades as complexity increases.

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 $\mathcal{D}_{test}^{HN}$, 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 $\mathcal{D}_{test}^{RAW}$ and $\mathcal{D}_{test}^{HN}$ (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), ALBEF [41] (additionally trained with a masked language modeling and image-text matching loss) and FLAVA [69] (which

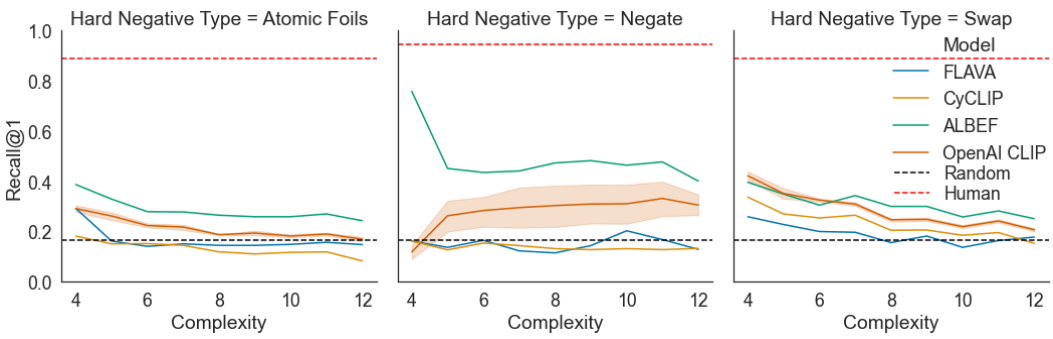

Figure 6. *Productivity Analysis on Additional Foundation Vision-language Models.* We plot models' Recall@1 on the productivity hard negatives retrieval set against complexity, where OpenAI CLIP's performance is averaged across five models RN50, RN101, ViT-B-16, ViT-B-32 and ViT-L-14. We find that all 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.

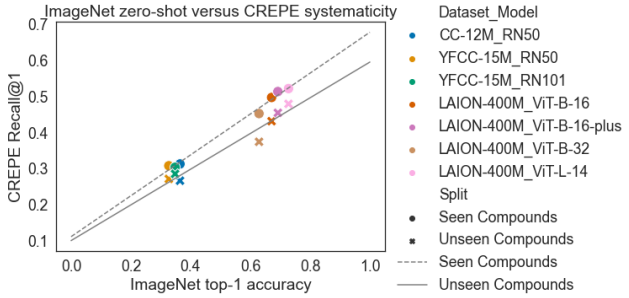

Figure 7. A plot showing the correlation between zero-shot top-1 accuracy on ImageNet and Recall@1 on CREPE's systematicity hard-negative sets. We observe a strong correlation, with an R^2 score of 0.9914 for the SC split and 0.9534 for the UC split.

further adds unimodal losses for image and text domains).

Retrieval. For $\mathcal{D}_{test}^{HN}$, 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 $\mathcal{D}_{test}^{HN,n}$). 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 $\mathcal{D}_{test}^{RAW}$, retrieval experiments are described in the Appendix.

5.1. Systematicity evaluation

Model performance on the $\mathcal{D}_{test}^{HN}$ 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 $\mathcal{D}_{test}^{HN}$ 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 $\mathcal{D}_{test}^{HN}$.

Similar to the overall results, there is also a consistent discrepancy between the SC and UC split on the $\mathcal{D}_{test}^{HN}$ 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.

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

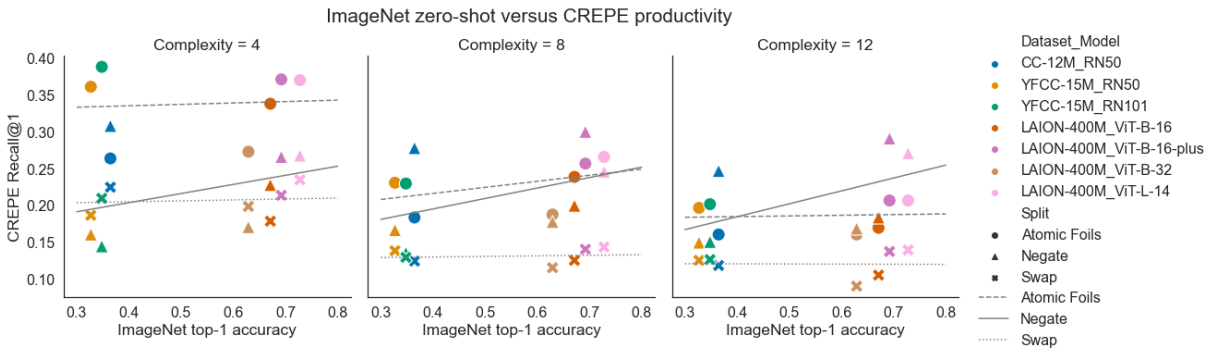

Figure 8. A plot showing the correlation between zero-shot top-1 accuracy on ImageNet and Recall@1 on CREPE's productivity hard negative sets for complexities of $4, 8$ and 12. Overall, we find no correlation between ImageNet accuracy and Recall@1 on our productivity sets. The strongest correlations 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 systematicity 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

[1] Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online, July 2020. Association for Computational Linguistics.

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks, 2015.

[3] Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019.

[4] Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. COVR: A test-bed for visually grounded compositional generalization with real images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9824–9846, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[9] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension, 2020.

[10] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

[11] MJ Cresswell. Logics and languages. 1973.

[12] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In Roya Firoozi, Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel Kochenderfer, editors, *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 893–905. PMLR, 23–24 Jun 2022.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[15] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook, 2022.

[16] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip, 2021.

[17] Mona Gandhi, Mustafa O. Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

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

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding, 2021.

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[23] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.

[24] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Evaluating text-to-image matching using binary image selection (bison), 2019.

[25] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning, 2018.

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[28] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 7 2021.

[29] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021.

[30] Theo MV Janssen and Barbara H Partee. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier, 1997.

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

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[33] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

[34] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020.

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[36] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[37] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[38] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

[39] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4943–4953, 2022.

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

[41] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[42] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO-2: End-to-end unified vision-language grounded learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3187–3201, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[43] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[44] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[46] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance, 2021.

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[48] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval, 2022.

[49] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[50] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. 2021.

[51] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 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)*, pages 8253–8280, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[52] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[53] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, Seattle, United States, July 2022. Association for Computational Linguistics.

[54] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021.

[55] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18071–18081, 2022.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[58] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[59] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[61] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146, 2018.

[62] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[63] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[64] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[65] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[66] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022.

[67] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues, 2022.

[68] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[69] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2021.

[70] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[71] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

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

[73] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021.

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

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

[76] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021.

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

embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019.

[78] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[79] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. *arXiv preprint arXiv:2111.13333*, 2021.

[80] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. 2022.

[81] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[82] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, June 2021.

[83] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.