

CREPE: Can Vision-Language Foundation Models Reason Compositionally?

Zixian Ma^{1*}, Jerry Hong^{1*}, Mustafa Omer Gul^{2*}, Mona Gandhi³, Irena Gao¹, Ranjay Krishna⁴
Stanford University¹, Cornell University², University of Pennsylvania³, University of Washington⁴

{zixianma, jerryhong, irena}@cs.stanford.edu mog29@cornell.edu mona09@seas.upenn.edu
ranjay@cs.washington.edu

A. Additional details on dataset generation

A.1. Hard negative types

In both our productivity and systematicity experiments, we rely on hard negatives to ensure that the retrieval sets we construct meaningfully probe a model’s comprehension. Specifically, to *granularly* probe a model’s comprehension, we identify a set of common failure modes of non-compositional models and design hard negative types that address each of these failure modes. Examples of each failure mode and hard negative type are outlined in Table 1.

A.2. Scene graph parser verification

To generate data splits for our systematicity experiments, we employed a rule-based implementation of the Stanford Scene Graph Parser [15, 19]. To verify its performance, we randomly sample 20 captions from each of CC-12M, YFCC-15M and LAION-400M and manually annotate scene graphs for the captions. We report precision and recall values for object, attribute and relationship atoms and object-relationship-object triplets on Table 2. For CC-12M, YFCC-15M and LAION-400M, the object precision was 88.14, 96.24, 70.00%, attribute precision was 93.00, 94.44, 72.22% and triplet precision was 91.67, 92.31 and 87.00% respectively. For recall values, on the other hand, we found that object recall was 83.06, 93.33, 60.68%, attribute recall was 56.51, 75.56, 36.11% and triplet recall was 64.04, 81.11 and 39.55% respectively. The precision values help determine whether the atoms the parser identifies are valid, while the recall values help determine whether the parser *can* identify the atoms and triplets present in the caption, important for the validity of our seen compounds (SC) and unseen compounds (UC) splits.

We find that the parser’s precision values are high throughout for each dataset. Recall values are lower compared to precision, particularly for the LAION dataset, where captions can be more similar to bags of words rather than well structured sentences. We note, however, that if compounds were incorrectly placed into the UC set due to poor recall,

our systematicity task would become easier. As all models experience drops in performance between SC and UC splits, we do not observe this.

A.3. Productivity caption generation

As discussed in the main text, each instance in the productivity test dataset is a image-text pair of complexity n with a set of hard negative captions. To generate such examples, we begin by sampling a n -node subgraph from a scene graph in Visual Genome [7]. We sample this subgraph using a random walk (see the paragraph titled **Random walk**). This subgraph is then transformed into a caption either using a template or GPT-3 (see the paragraph titled **Caption generation**). Finally, we crop the original image to the union of all object bounding boxes in the subgraph (see main text). We describe these details below.

Random walk Given a scene graph G , we generate an n -atom subgraph ($n \leq |G|$). We initialize a subgraph S with a single random object in G . While this subgraph contains less than n atoms, a compound C consisting of at least one unadded atom is added to S . If C is a relationship compound (C_{oro}), the walk continues from the newly added object; otherwise, the walk is continued from the same object. If the entire connected component of the scene graph is exhausted, another object is selected at random from a different connected component. This process ends when n atoms are added to the subgraph. We discard all walks that result in insufficient number of atom.

Caption generation To generate captions, we either utilize hand crafted templates or use GPT-3. For subgraphs of complexity $n > 4$, we use GPT-3 to generate a coherent caption for each prompt; otherwise, we use the templates. When prompting GPT-3 to produce captions, we populate the the first line of the prompt with a list the objects in the subgraph, prepended with their attributes. If multiple instances of an object type occur (*e.g.*, we have two objects both with name “window” in the graph), we append a numerical suffix to distinguish between them (*e.g.* “window1” from

*Equal contribution

Table 1. A list of the potential failure modes a vision-language model may encounter when parsing increasingly complex scenes, and the corresponding hard negatives generated in our test datasets.

Dataset	Label	Error Mode	Hard Negative	Example
Sys	HN-ATOM	Ignoring incorrect atoms.	Atomic foils. Replace a single atom with a mutually exclusive or antonymic atom, enforced by WordNet.	A grill on top of the porch. →: A grill underneath the porch.
Sys	HN-COMP	Ignoring proper binding of atoms into compounds.	Compound foils. Split the correct atoms of a single compound over two compounds; fill in the partial compounds with atomic foils (see above).	A pink car. →: A blue car and a pink toy . →: A pink flower and a black car.
Prod	HN-ATOM	Ignoring incorrect atoms.	Atomic foils. Replace a single atom with a mutually exclusive or antonymic atom, enforced by WordNet.	Yellow vase on top of television. →: Red vase on top of television. →: Yellow vase underneath television. →: Yellow vase on top of shelf .
Prod	HN-SWAP	Ignoring proper binding of atoms.	Swapping foils. Swap two atoms of the same type – or permute several atoms of the same type.	Yellow vase on top of television. →: Yellow television on top of vase . →: Television on top of yellow vase.
Prod	HN-NEG	Disregarding incorrect negations.	Negation foils. Negate the entire caption or an individual atom with a grammatically correct “not” modifier.	Yellow vase on top of television. →: There is no yellow vase on top of television. →: Vase that is not yellow on top of television.

Table 2. *Scene Graph Parser Validation*: We report precision and recall values the Stanford Scene Graph parser obtains on the CC-12M, YFCC-15M and LAION-400M datasets. For each dataset, we compute values for object, attribute and relationship atoms as well as object-relationship-object triplets. Overall, the scene graph obtains high precision values but lower recall scores. The parser performs the poorest on LAION-400M due its noisier captions.

	CC-12M		YFCC-15M		LAION-400M	
	Precision	Recall	Precision	Recall	Precision	Recall
Object	88.14	83.06	96.24	93.33	69.91	60.68
Attribute	93.00	56.51	94.44	75.56	72.22	36.11
Relationship	92.86	70.18	93.59	83.33	88.33	40.15
Triplet	91.67	64.04	92.31	81.11	87.00	39.55

“window2.”). On the second line of the prompt, we list all the relationships between objects in the graph, in the form `subject relationship object`. Additionally, we manually generate 5 caption examples per complexity from random subgraphs and prepend both the random subgraph and the manually generated caption to the prompt above, as few-shot training examples for GPT-3. We provide examples of graphs, prompts, and their generated captions in Figure 1.

For examples of complexity $n = 4$, we found that stringing together a simple templated prompt was sufficient to produce fluent captions. This was done by prepending attributes in front of objects and stringing together subjects, relations, and objects in the correct order. For example, a subgraph containing `boy=(tall,blue); grass=(green); (boy, on, grass)` would be templated as `tall and blue boy on green grass`. Any disconnected atoms are appended with the prefix “and a.”

Table 3. Productivity ground truth captions’ faithfulness to their paired images, split by caption complexity. Overall, the generated captions’ faithfulness is stable and consistently high across different complexities.

Complexity	Avg faithfulness
$n = 7$	88.7 ± 10.8
$n = 8$	85.7 ± 7.0
$n = 9$	90.0 ± 6.0
$n = 10$	87.7 ± 9.3
$n = 11$	88.1 ± 7.8
$n = 12$	89.1 ± 2.9

Data verification. We manually verify the accuracy of our produced productivity dataset. We provide a breakdown of annotators’ scores for GPT-3 caption faithfulness across complex subgraphs with $n \geq 7$ in Table 3. We see that scores are consistently high for ground-truth captions across

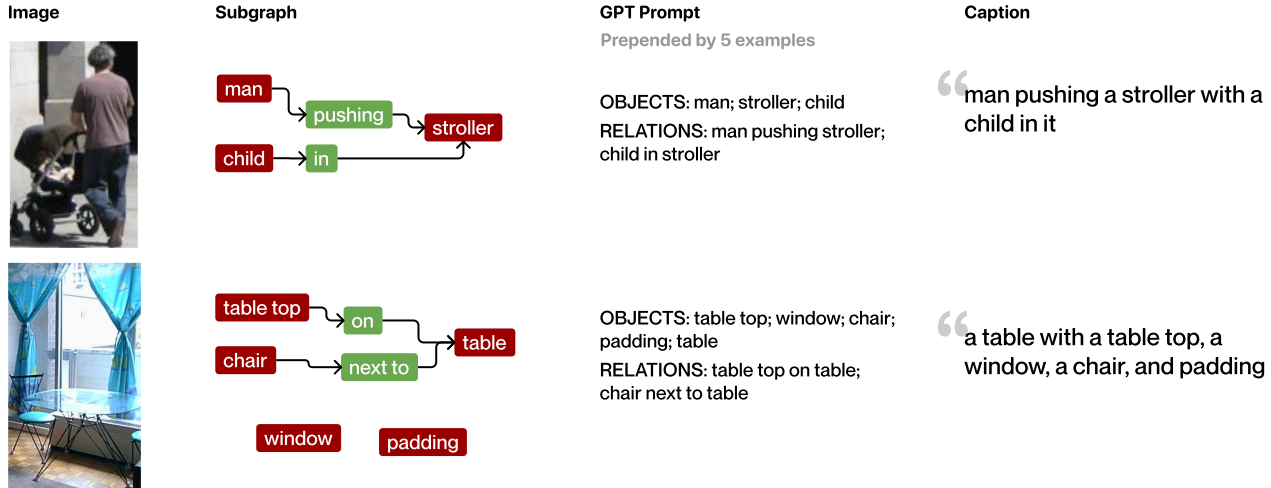


Figure 1. Examples of image-text pairs we generate for our productivity evaluation. The image is the union of the objects bounding boxes in the subgraphs. We also showcase the GPT-3 prompt associated with the subgraph and their corresponding generated ground-truth captions.

complexities.

A.4. Hard negative generation details

We provide additional detail for the procedure of generating a hard negative of types HN-SWAP and HN-NEG. Suppose throughout that for a given image and its annotated scene graph G , we seek to generate a hard negative caption for caption t associated with the subgraph $S \subseteq G$.

HN-SWAP The following pairs of atoms could be swapped to create a hard negative for S :

- The subject A_o and object A'_o of a relationship compound $C_{oro} \in S$.
- Two attributes A_a and A'_a attached to distinct objects A_o , A'_o , such that one attribute is not present for the other object in G and vice versa. $((A_a, A'_o) \notin G$ and $(A'_a, A_o) \notin G$).
- Two objects A_o , A'_o not connected by a relationship such that their swapping within G does not create an identical graph.

Additionally, some swap hard negatives generated are *permutations* rather than a swapped pair:

- One attribute A_a can be transferred from one object A_o to another object A'_o , so long as that attribute doesn't apply to the new object $((A_a, A'_o) \notin G$).
- For low complexities ($n = 4$), any permutation of atoms of the same type are allowed. For example: (“There is a dog on the bed and also a nightstand” \rightarrow “There is a *nightstand* on the *dog* and also a *bed*”)

HN-NEG We verify with G to ensure that negating an atom results in an incorrect caption. If an attribute A_a connected with A_o is negated, we ensure that there does not exist an object of A'_o that doesn't have an attribute A_a but shares all the other attributes of A_o . For example, if we negate “black” in “Black dog on a building”, we ensure there doesn't exist another dog on the building that isn't black. Similar checks are performed for negating relationships and objects. When a relationship A_r connecting A_o and A'_o is negated, there cannot exist another identical subject and object pair connected by a different relationship A'_r . When an object is negated, there cannot exist any other object with the same attributes and relationships.

A.5. Test dataset sizes, examples, and additional verification

Table 4 expands on Table 1 from the main paper to provide a breakdown of the number of image-text pair per hard negative type and, for productivity, for each sentence complexity. We remark that \mathcal{D}_{test}^{RAW} , which contains only image-ground-truth caption pairs, is a *superset* of the ground-truth captions in \mathcal{D}_{test}^{HN} . This is because, for some ground truth captions in \mathcal{D}_{test}^{RAW} , a sufficient number of hard negatives to perform retrieval in \mathcal{D}_{test}^{HN} could not be generated. Additionally, due to the prevalence of rare atoms, we could only generate valid hard negatives for very few captions in the UA split. Therefore, we omit the evaluation on the UA split with hard negatives and focus on the analysis of results between the SC and UC split, which is more interesting as models have seen all the atoms in both splits. Table 5 summarizes the text retrieval set size of each image query for both \mathcal{D}_{test}^{raw} and \mathcal{D}_{test}^{HN} in our systematicity and productivity

Table 4. We report the ground truth caption counts in \mathcal{D}_{test}^{RAW} and hard negative counts in \mathcal{D}_{test}^{HN} s for systematicity and productivity, separated by hard negative type and split.

Systematicity				Productivity				
Split	Ground Truth	HN-ATOM	HN-COMP	Split	Ground Truth	HN-ATOM	HN-SWAP	HN-NEG
CC-12M SC	262,541	104,024	156,036	$n = 4$	1,508	6,290	6,310	2,510
CC-12M UC	113,659	14,348	21,522	$n = 5$	1,734	7,270	6,560	3,425
CC-12M UA	9,577	-	-	$n = 6$	1,905	9,025	8,990	6,565
YFCC SC	194,502	75,948	113,922	$n = 7$	2,171	10,410	10,045	7,845
YFCC UC	172,469	39,204	58,806	$n = 8$	2,247	11,205	11,220	10,210
YFCC UA	18,806	-	-	$n = 9$	1,969	9,485	9,120	8,310
LAION SC	170,253	62,884	94,326	$n = 10$	2,246	11,325	11,185	10,460
LAION UC	201,595	49,604	74,406	$n = 11$	1,895	8,620	8,300	7,925
LAION UA	1,855	-	-	$n = 12$	1,878	10,005	9,950	9,710










Systematicity			
CC-12M	\mathcal{D}_{test}^{RAW}	Ground truth captions	\mathcal{D}_{test}^{HN} Hard negatives
Seen compounds		a purple umbrella	HN-ATOM: a purple awning HN-COMP: a purple marquee and blue umbrella
Unseen compounds		sidewalk next to black train	HN-ATOM: sidewalk next to black sedan HN-COMP: sidewalk next to black bus and brown train
Unseen atoms		green bushes next to pole	No hard negatives for unseen atoms.
LAION-400M			
Seen compounds		dirty fork	HN-ATOM: dirty spoon HN-COMP: dirty china and clean fork
Unseen compounds		hat on the man.	HN-ATOM: swimsuit on the man HN-COMP: hat on lamb and hat off man
Unseen atoms		mauve colored food tray	No hard negatives for unseen atoms.
YFCC-15M			
Seen compounds		the two dogs on the chair	HN-ATOM: the two panda on the chair HN-COMP: wolf on chair and dogs off chair
Unseen compounds		purple couch	HN-ATOM: purple desk HN-COMP: purple booth and green couch
Unseen atoms		green and white stripe wallpaper	No hard negatives for unseen atoms.

Figure 2. A sample of image-caption pairs in the **systematicity** retrieval sets. One ground truth caption is shown for each split of each training dataset, each of which lie in both \mathcal{D}_{test}^{RAW} and \mathcal{D}_{test}^{HN} . Additionally, one example of each hard negative type is shown for each ground truth caption.

evaluation. Figures 2 and 3 present examples of ground truth captions and hard negative captions in our test datasets for systematicity and productivity, respectively.

We provide a breakdown of annotators' scores for the accuracy of productivity hard negatives in Table 6. A hard negative caption is accurate if it contains incorrect facts

about the image. We find that the accuracy and pairwise agreement of the HN-ATOM is the highest and much higher than those of HN-SWAP and HN-NEG.

Productivity	\mathcal{D}_{test}^{RAW} Ground truth captions	\mathcal{D}_{test}^{HN} Hard negatives
n = 4	 speaker beside pope. there is a stand	 HN-ATOM: speaker beside mistress . There is a stand. HN-SWAP: pope besides speaker . there is a stand HN-NEG: there is no speaker beside pope.
n = 5	 tree on a sidewalk next to a meter	 HN-ATOM • tree on a catwalk next to a meter • pineapple on a sidewalk next to a meter HN-SWAP • meter on a tree next to a sidewalk • meter on a sidewalk next to a tree HN-NEG • there is no tree on a sidewalk next to a meter • tree not on a sidewalk next to a meter
n = 6	 a dog behind a surfboard and water with splashes	 HN-ATOM: a dog behind a foil and water with splashes HN-SWAP: a surfboard behind a dog and water with splashes HN-NEG: a dog behind an object which is not a surfboard and water with splashes
n = 7	 chairs in a row with umbrellas above them; there are also tables and chairs	 HN-ATOM: chairs in a row with umbrellas below them; there are also tables and chairs HN-SWAP: chairs in a tables with umbrellas above them; there are also row and chairs HN-NEG: chairs in a row with umbrellas not above them; there are also tables and chairs
n = 8	 a person wearing a cap and another person standing on the sidewalk, and another person on the sidewalk	 HN-ATOM: a person wearing a coverall and another person standing on the sidewalk, and another person on the sidewalk HN-SWAP: a person wearing a sidewalk and another person standing on the cap , and another person on the sidewalk HN-NEG: a person wearing a cap and another person standing not on the sidewalk, and another person on the sidewalk
n = 9	 a laptop and paper on a table. a man is standing by the table with his hands on it.	 HN-ATOM: a laptop and paper on a matrix , a man is standing by the table with his hands on it HN-SWAP: a laptop and paper on a hands . a man is standing by the hands with his table on it. HN-NEG: a laptop and paper on a table. a man is not standing by the table with his hands on it
n = 10	 a black chair with wheels in front of a desk, with a laptop and lamp on it	 HN-ATOM: a black chair with wheels in front of a console , with a laptop and lamp on it HN-SWAP: a black chair with desk in front of a wheels , with a laptop and lamp on it HN-NEG: a black chair with wheels not in front of a desk, with a laptop and lamp on it
n = 11	 a cardboard under a pan, and a deep dish pizza in the pan. the pan is filled with the deep dish pizza and there is a spatula in the deep dish pizza.	 HN-ATOM: A cardboard under a pan, and a deep dish pizza not in the pan. the pan is filled with the deep dish pizza and there is a spatula not in the deep dish pizza. HN-SWAP: a pan under a cardboard , and a deep dish pizza in the pan. the pan is filled with the deep dish pizza and there is a spatula in the deep dish pizza. HN-NEG: a cardboard under a pan, and a deep dish pizza in the pan. the pan is filled with the deep dish pizza and there is an object that is not a spatula in the deep dish pizza.
n = 12	 stand with handles and advertisements, with a tv resting on top of three drawers. the surface of the tv has a reflection, and there is a sign on top of the stand.	 HN-ATOM • stand with handles and advertisements, with a tv resting on top of three drawers. The surface of the tv has a rendering , and there is a sign on top of the stand. • wing with handles and advertisements, with a tv resting on top of three drawers. the surface of the tv has a reflection, and there is a sign on top of the stand HN-SWAP • sign with handles and advertisements, with a tv resting on top of three drawers. the surface of the tv has a reflection, and there is a stand on top of the sign . • stand with handles and reflection , with a tv resting on top of three drawers. the surface of the tv has a advertisements , and there is a sign on top of the stand. HN-NEG • an object which is not a stand with handles and advertisements, with a tv resting on top of three drawers. the surface of the tv has a reflection, and there is a sign on top of the an object which is not a stand. • stand with handles and advertisements, with a tv not resting on top of three drawers. the surface of the tv has a reflection, and there is a sign on top of the stand

Figure 3. A sample of image-caption pairs in the **productivity** retrieval sets. One ground truth (GT) caption is shown for each complexity n . These GT captions lie in both \mathcal{D}_{test}^{RAW} and \mathcal{D}_{test}^{HN} . One example of each hard negative type is shown for each GT caption. For two highlighted example captions ($n = 5, 12$), we show 2 hard negatives per type for comprehensiveness.

Table 5. We summarize the retrieval set sizes for both \mathcal{D}_{test}^{HN} and \mathcal{D}_{test}^{RAW} in our systematicity and productivity evaluation.

Retrieval set size	\mathcal{D}_{test}^{HN}		\mathcal{D}_{test}^{RAW}
	HN-ATOM	HN-COMP	
Systematicity	5	7	1,855
Productivity	6	6	1,508

Table 6. Accuracy of our generated hard negatives for productivity, split by type, in our data verification. While HN-ATOM atoms receive strong human evaluation scores, we find that HN-SWAP and HN-NEG negatives are noisier.

Type	Acc. mean \pm std	Pairwise agreement
HN-ATOM	91.6 \pm 4.2	83.1
HN-SWAP	70.1 \pm 9.1	58.5
HN-NEG	72.4 \pm 0.0	59.5

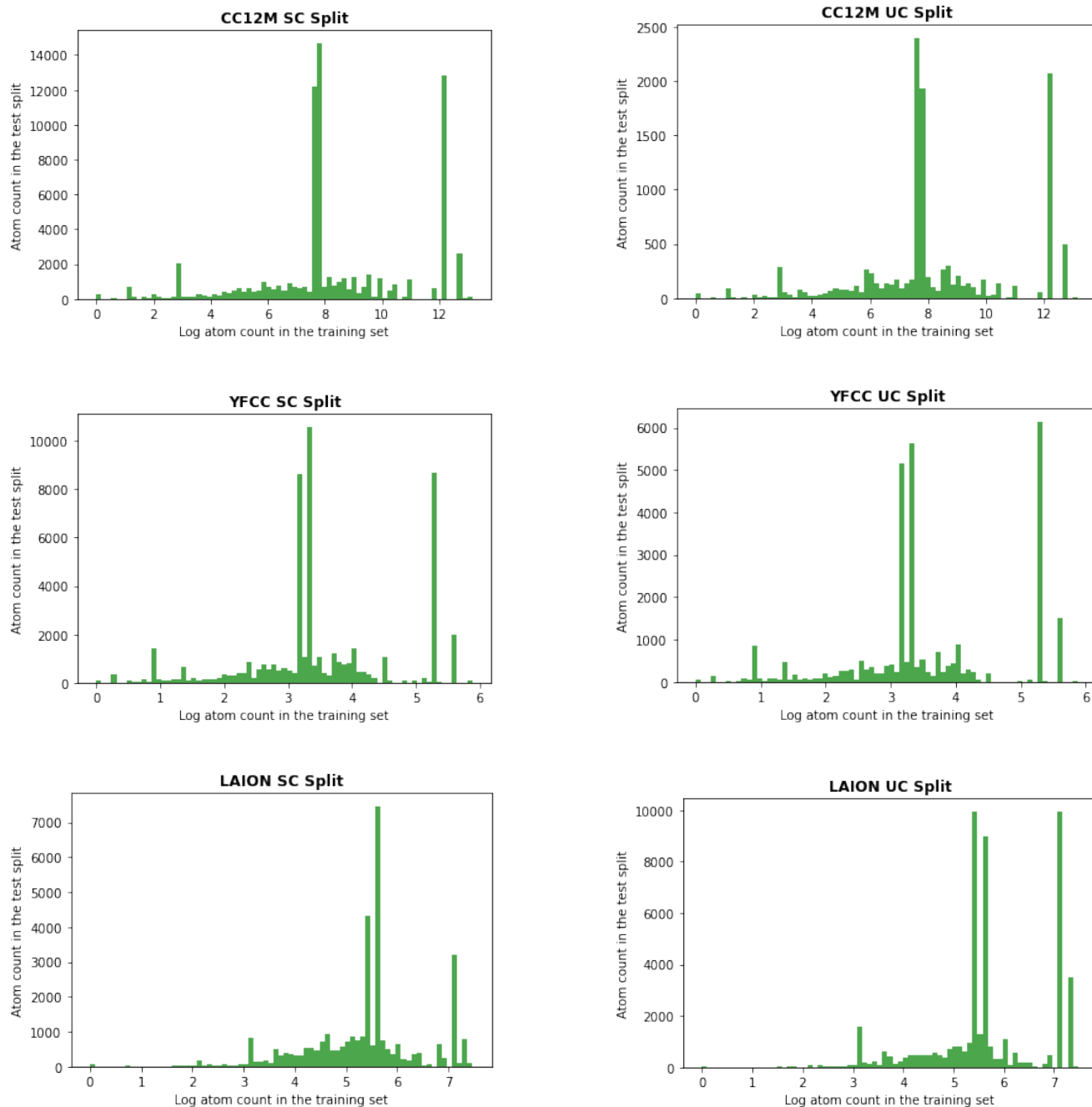


Figure 4. We plot the atom count in training vs. in the systematicity hard negative test set. We observe that the atoms in the SC and UC test splits have similar counts in the training dataset.

A.6. Systematicity hard negative dataset details

Table 7 summarizes the number of unique atoms and compounds in the SC and UC split of the systematicity hard negative set. Additionally, we plot the atom count in the systematicity test set vs. the training set (on a log scale). As shown in Figure 4, we see that the atom count in the training

set is always on the same scale across both splits for the same training dataset (x-axis in each row). We further observe that the atom distributions are similar in the SC and UC splits. These suggest that the atoms appearing in the UC split are not substantially rarer or more difficult than the ones in the SC split.

Table 7. We summarize the unique atom and compound counts in the SC and UC split of the systematicity hard negative set.

Train dataset	SC		UC	
	Atom (seen)	Comp (seen)	Atom (seen)	Comp (unseen)
CC12M	3,348	26,006	946	3,587
YFCC	3,173	18,987	1,405	9,801
LAION	2,968	12,401	1,951	15,721

B. Additional evaluation results

B.1. Full retrieval results on hard negative datasets

Systematicity We additionally include the full retrieval results on \mathcal{D}_{test}^{HN} with both HN-ATOM and HN-COMP, HN-ATOM only and HN-COMP only in Tables 8, 9 and 10. We note that as we relax the metric from R@1 to R@3, the difference between models’ performance in the SC and UC split decreases.

B.2. Retrieval results on raw datasets

In addition to \mathcal{D}_{test}^{HN} retrieval experiments, we perform retrieval experiments with \mathcal{D}_{test}^{RAW} .

We perform *both* image-to-text and text-to-image retrieval within splits of \mathcal{D}_{test}^{RAW} . Each retrieval task is between one image and every caption in the split, or vice versa. We report the mean and standard deviation of Recall@1 across K-fold retrievals (where $K = \min(20, \lfloor \frac{|\mathcal{D}_{test}^{RAW}|}{N} \rfloor$), and $N = \min\{|SC|, |UC|, |UA|\} = 1855$ for systematicity and $N = \min_{n \in \{4 \dots 12\}} |\mathcal{D}_{test}^{RAW, n}| = 1508$ for productivity), as the data size varies across compositional splits and complexities.

Systematicity We present the systematicity retrieval results on \mathcal{D}_{test}^{raw} in Table 11, where each retrieval set for an image consists of the captions of the other images. We continue to observe a monotonic decrease in performance when compounds are unseen. Additionally, we continue to observe a drop in performance for larger training datasets. In particular, we see a similar drop in performance for LAION-trained models across both the image-to-text and text-to-image tasks. We also observe larger drops on LAION-trained models than for \mathcal{D}_{test}^{HN} when moving across the SC \rightarrow UC, and across UC \rightarrow UA splits, with LAION models dropping as much as 13% for ViT-L/14.

Productivity We additionally present the productivity retrieval results on \mathcal{D}_{test}^{raw} in Table 12. We observe that models’ Recall@1 generally increases as the caption complexity increases. We hypothesize that models’ low performance in the low-complexity subset is caused by false negatives in the original dataset: since the captions are simple and likely true for multiple images, there are multiple false negatives in the retrieval set, making these numbers unreliable. As the captions become more complex, however, the chance of

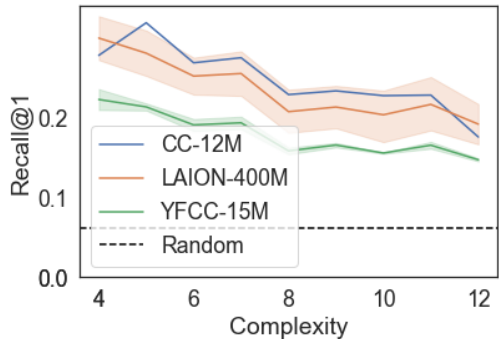


Figure 5. *Productivity Analysis on Hard Negatives of All Types.* We plot models’ Recall@1 on the overall hard negatives retrieval set against complexity, where each retrieval set contains hard negatives of all types. We find that models’ ability to correctly retrieve the ground-truth caption drops as complexity increases.

such false negatives is lower. This means there are more true negatives in the higher-complexity subsets, making retrieval easier for these models.

B.3. Retrieval results with all hard negatives at once

Productivity We present models’ retrieval performances over the whole productivity \mathcal{D}_{test}^{HN} dataset, where each retrieval set contains one ground truth caption and fifteen hard negatives, five for each of the three types HN-ATOM, HN-SWAP and HN-NEG. We find in Figure 5 that models’ Recall@1 performance decreases with complexity, which aligns with the findings on the separate retrieval sets for HN-ATOM, HN-SWAP and HN-NEG.

B.4. Qualitative analysis on systematicity evaluation

We perform a qualitative analysis to better understand why the LAION-400M trained models ViT-B-16 and ViT-L-14 show a large versus small performance drop from the Seen to Unseen Compounds split respectively. Table 13 presents examples where both ViT-B-16 and ViT-L-14 retrieve the correct caption successfully in the SC split and where ViT-B-16 fails in the UC split. Through this analysis, we find that the SC split for LAION-400M trained models is dominated by simple two-atom examples such as “purple couch”. The UC split, however, contains more complex examples that involve relationships such as “curtains on the window”. In particular, we find that the ViT-B-16 model struggles with the relationship “on” and often retrieves a wrong caption where “on” is replaced with “off” or where the object is replaced with an atomic foil. For example, ViT-B-16 retrieves “plants on bob and plants off building” incorrectly when the groundtruth caption is “plants on a building”. Nevertheless, the rank of the groundtruth caption is often still within the top three. This explains the narrower gap in ViT-B-16’s Re-

Table 8. *Systematicity Hard-Negative Dataset Analysis*. We report Recall@1,3,5 and Avg R@K results for all models on the D_{test}^{HN} hard-negative datasets. Model performance decreases from the Seen all compounds (SC) to the Unseen Compounds (UC) split, particularly for LAION-400M models.

Training dataset	Model	R@1		R@3		Avg R@K		
		SC	UC	SC	UC	SC	UC	
	Random	9.09	9.09	27.27	27.27	18.18	18.18	
	CC12M	31.25	26.60	70.70	68.58	50.97	47.59	
Image-to-text	YFCC15M	RN50	30.75	27.07	67.87	65.68	49.31	46.37
		RN101	30.32	28.50	67.39	67.00	48.85	47.75
	LAION400M	ViT-B-32	45.32	37.40	79.57	77.78	62.45	57.59
		ViT-B-16	49.69	43.12	83.20	82.51	66.45	62.81
		ViT-B-16+240	51.33	45.43	84.01	83.79	67.67	64.61
	ViT-L-14	52.16	47.92	83.52	82.53	67.84	65.22	

Table 9. *Systematicity HN-ATOM Dataset Analysis*. We report Recall@1,3 and Avg R@K results for all models on the D_{test}^{HN} subset with HN-ATOM. Model performance decreases from the Seen Compounds (SC) to the Unseen Compounds (UC) split, particularly for LAION-400M models.

Training dataset	Model	R@1		R@3		Avg R@K		
		SC	UC	SC	UC	SC	UC	
	Random	20.00	20.00	60.00	60.00	40.00	40.00	
	CC12M	56.19	53.21	93.60	91.84	74.90	72.52	
Image-to-text	YFCC15M	RN50	48.54	44.67	92.27	92.43	70.41	68.55
		RN101	48.19	44.35	91.82	91.88	70.00	68.12
	LAION400M	ViT-B-32	58.32	48.97	94.51	93.89	76.41	71.43
		ViT-B-16	61.98	53.79	95.44	95.26	78.71	74.53
		ViT-B-16+240	63.24	56.25	95.69	95.80	79.47	76.03
	ViT-L-14	64.91	58.89	95.82	95.95	80.37	77.42	

call@3 between Seen Compounds and Unseen Compounds. On the other hand, we see that ViT-L-14 continues to retrieve the correct caption even on the more challenging Unseen Compounds split, suggesting that a larger model size could improve compositional systematicity.

C. Additional Related Work

Evaluating learned representations By analyzing the properties of pretrained representations, our work continues a tradition of research in Computer Vision [3, 9–11, 14, 17] and Natural Language Processing [4, 6, 12, 13, 16, 18] that probes characteristics of representations themselves rather than their performance on downstream tasks. Instead of learning probes, we use retrieval for zero-shot evaluation in order to avoid scenarios where the learned probe compensates for the characteristics deficient in the original representations [1, 2, 5, 8, 20].

References

- [1] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, Mar. 2022.
- [2] Steven Cao, Victor Sanh, and Alexander M Rush. Low-complexity probing via finding subnetworks. *arXiv preprint arXiv:2104.03514*, 2021.
- [3] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
- [4] Jack Hessel and Alexandra Schofield. How effective is BERT without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online, Aug. 2021. Association for Computational Linguistics.
- [5] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

Table 10. *Systematicity HN-COMP Dataset Analysis*. We report Recall@1,3 and Avg R@K results for all models on the D_{test}^{HN} subset with HN-COMP. We observe a decrease in performance from the SC to UC split overall, except for the LAION-400M trained ViT-L-14 model.

Training dataset	Model	R@1		R@3		Avg R@K	
		SC	UC	SC	UC	SC	UC
	Random	14.29	14.29	42.86	42.86	28.57	28.57
	CC12M	57.45	55.42	84.09	83.83	70.77	69.63
	YFCC15M	51.32	49.02	80.39	79.55	65.85	64.28
		50.86	51.89	80.31	80.95	65.58	66.42
Image-to-text	ViT-B-32	66.91	66.57	89.73	89.07	78.32	77.82
	LAION400M	71.24	72.02	91.57	92.13	81.40	82.08
		72.32	72.75	91.89	91.82	82.10	82.28
		71.19	71.83	90.75	91.04	80.97	81.43

Table 11. *Systematicity Raw Dataset Analysis*. We report mean Recall@1 results for all models across k-fold evaluations. Model performance consistently decreases from Seen all Compounds (SC) to Unseen Compounds (UC) and from Unseen Compounds to Unseen Atoms (UA) splits, particularly for LAION-400M models.

Training dataset	Model	SC	UC	UA
	Random	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00
	CC-12M	19.92 ± 0.94	17.82 ± 0.99	15.02 ± 0.85
	YFCC-15M	16.30 ± 0.70	14.57 ± 0.69	12.80 ± 0.90
		17.10 ± 0.90	15.58 ± 1.04	13.62 ± 0.84
Image-to-text	ViT-B-16	35.61 ± 0.92	30.04 ± 1.42	25.88 ± 0.00
	LAION-400M	36.80 ± 0.90	31.10 ± 1.37	26.25 ± 0.00
		33.86 ± 0.97	29.00 ± 1.40	23.99 ± 0.00
		38.24 ± 0.70	32.70 ± 1.30	26.42 ± 0.00
	CC-12M	20.85 ± 0.98	18.15 ± 0.84	15.46 ± 1.10
	YFCC-15M	15.60 ± 0.79	14.05 ± 0.84	12.17 ± 0.64
		16.11 ± 0.84	14.47 ± 0.87	12.54 ± 0.66
Text-to-image	ViT-B-16	35.74 ± 0.76	29.58 ± 1.39	23.29 ± 0.00
	LAION-400M	37.25 ± 0.97	30.57 ± 1.33	24.26 ± 0.00
		33.66 ± 1.03	29.00 ± 1.40	22.10 ± 0.00
		38.69 ± 0.86	32.00 ± 0.90	25.61 ± 0.00

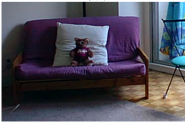

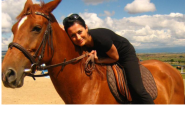



ing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

- [6] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.
- [7] 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.
- [8] Michael Lepori and R Thomas McCoy. Picking bert’s brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3637–3651, 2020.
- [9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online, July 2020. Association for Computational Linguistics.
- [10] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip, 2022.
- [11] Victor Milewski, Miryam de Lhoneux, and Marie-Francine Moens. Finding structural knowledge in multimodal-BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [12] Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online, Aug. 2021. Association for Computational Linguistics.

Table 12. *Productivity Raw Dataset Analysis*. We report mean Recall@1 results for all models across all complexities. We find that models' Recall@1 increases as the caption complexity increases.

Training dataset		Model	4	5	6	7	8	9	10	11	12	
		Random	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	
CC-12M		RN50	13.19	14.23	16.20	19.02	21.17	19.26	22.70	24.29	25.77	
Image-to-text		YFCC-15M	RN50	9.08	10.92	12.09	14.23	14.17	15.40	14.54	16.75	19.02
			RN101	11.10	11.47	11.90	14.60	15.34	16.50	16.99	18.90	20.67
		LAION-400M	ViT-B-16	20.80	20.00	25.89	27.30	29.63	29.02	30.31	34.11	37.36
			ViT-B-16+240	22.39	21.53	26.93	28.10	30.61	30.18	32.88	36.50	38.83
ViT-B-32	20.49		20.37	23.50	26.93	29.20	28.22	29.94	32.58	35.40		
		ViT-L-14	22.09	23.07	27.67	29.69	33.13	31.04	35.09	37.12	40.25	
CC-12M		RN50	12.52	15.03	15.52	17.85	17.30	19.75	21.66	22.52	25.58	
Text-to-image		YFCC-15M	RN50	8.04	9.82	9.82	12.15	12.94	13.62	13.37	14.97	15.15
			RN101	9.39	11.10	11.10	13.31	13.74	15.03	14.48	16.07	18.40
		LAION-400M	ViT-B-16	18.16	19.26	23.62	24.85	27.85	27.79	28.77	31.53	33.93
			ViT-B-16+240	18.96	20.67	25.46	26.13	28.83	29.02	31.41	32.88	37.24
ViT-B-32	17.55		18.65	22.21	23.50	26.20	26.13	27.36	28.83	32.21		
		ViT-L-14	19.88	20.43	24.97	26.63	30.37	30.00	32.76	34.42	36.99	

Table 13. *Systematicity Qualitative Analysis*. We present examples where LAION-400M trained ViT-B-16 and ViT-L-14 both perform well on the Seen Compounds (SC) split, and where ViT-B-16 performs poorly on the Unseen Compounds (UC) split.

Image	GT caption	ViT-B-16		ViT-L-14		
		R@1	Top 3 captions	R@1	Top 3 captions	
SC		purple couch	1	purple couch purple altar and brown couch purple commode and red couch	1	purple couch purple altar and brown couch purple desk and brown couch
		a white parked car	1	a white parked car a green parked car a white bike	1	a white parked car a green parked car a orange parked car
		a fully grown brown horse	1	a fully grown brown horse a fully grown brown mule and red horse a fully grown brown mule and yellow horse	1	a fully grown brown horse a fully grown brown mule and red horse a fully grown brown zebra and blue horse
UC		a cat on the sofa.	0	a cat on the console. a cat on the sofa. cat on counter and cat off sofa	1	a cat on the sofa. a cat on the console. badger on sofa and cat on console
		boat on the water	0	boat on the polish boat on the soda boat on the water	1	boat on the water boat on the lime ship on water and boat on rubber
		plants on a building	0	plants on bob and plants off building plants on a building court on building and plants off building	1	plants on a building park on a building billboard on building and plants off building

- [13] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.
- [14] Philipp J. Rösch and Jindřich Libovický. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1031–1041, Seattle, United States, July 2022. Association for Computational Linguistics.
- [15] 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.
- [16] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovered the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.
- [17] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality, 2022.
- [18] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics.
- [19] 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.
- [20] Yichu Zhou and Vivek Srikumar. DirectProbe: Studying representations without classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5070–5083, Online, June 2021. Association for Computational Linguistics.