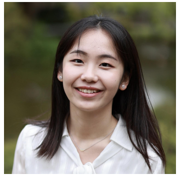
CREPE: Can Vision-Language Foundation Models Reason Compositionally?

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]

★ HIGHLIGHT @WED-AM-255

Compositionality

Compositionality enables human understanding of complex language

A yellow vase on top of a
black television

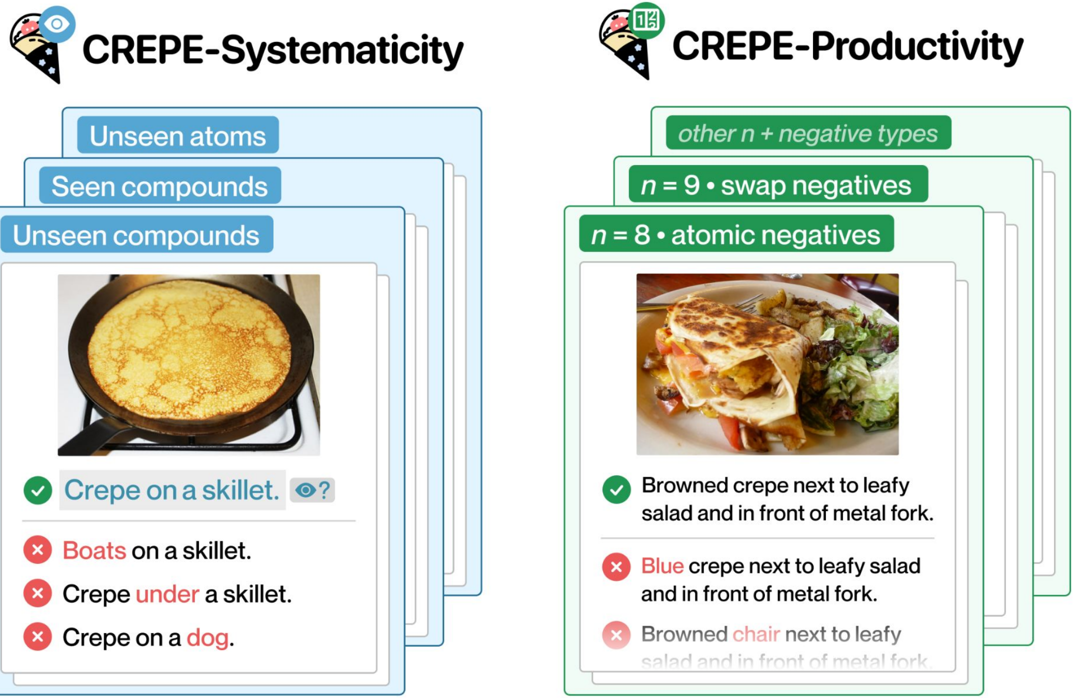
Compositionality enables human understanding of complex language and visual scenes.

A yellow vase on top of a black television

Compositionality enables human understanding of complex language and visual scenes.

A **yellow** $\boxed{\text{vase}}$ on top of a **black** $\boxed{\text{television}}$

CREPE measures two aspects of compositionality:



CREPE measures two aspects of compositionality: systematicity and productivity.

Seen compositions

A yellow **vase** on top of a black television

Tennis players wearing **pink** t-shirts

Unseen test composition

White and purple flowers in a **pink vase**

CREPE measures two aspects of compositionality: systematicity and productivity.

Complexity	Caption
n=4	A yellow vase on top of a television
n=5	A yellow vase on top of a black television
n=7	Plant inside a yellow vase on top of a black television
n=10	Plant inside a yellow vase on top of a black television in front of an old computer

We generate large-scale datasets of image-caption pairs and hard negative captions for image-to-text retrieval evaluation.

	Systematicity			Productivity
Training dataset	CC-12M	YFCC-15M	LAION-400M	Any
Test set Image-caption pairs	**385,777**	**385,777**	**373,703**	**17,553**
Hard negative captions	**325,523**	**316,668**	**309,342**	**183,855**

We evaluate vision-language models across 7 architectures trained with 4 algorithms on massive datasets.

	Algorithm	Training dataset size	Architectures
Systematicity, Productivity	CLIP	12M	Transformer + [RN50]
		15M	Transformer + [RN50, RN101]
		400M	Transformer + [ViT-B/32, ViT-B/16, ViT-B/16+240, ViT-L/14]
Productivity	CyCLIP	3M	Transformer + [RN50]
	FLAVA	14M	ViT-B/16 + ViT-B/16 + multimodal ViT
	ALBEF	70M	BERTbase + ViT-B/16 + multimodal BERTbase
	CLIP	400M	Transformer + [RN50, RN101, ViT-B/32, ViT-B/16, ViT-L/14]

We present 2 key takeaways from our experiments.

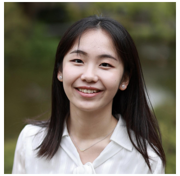
1. **State-of-the-art vision-language models do NOT exhibit systematicity or productivity;**
2. Neither emerges as we scale up the training dataset or model size.

We present 2 key takeaways from our experiments.

1. State-of-the-art vision-language models do NOT exhibit systematicity or productivity;
2. **Neither emerges as we scale up the training dataset or model size.**

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]

★ HIGHLIGHT @WED-AM-255

Humans Reason about Language and Images Compositionally

Humans Reason about Language and Images Compositionally

A yellow vase on top of a
black television

Humans Reason about Language and Images Compositionally

A **yellow** vase on top of a
black television

Humans Reason about Language and Images Compositionally

A yellow **vase** on top of a
black **television**

Humans Reason about Language and Images Compositionally

A yellow vase **on top of** a black television

Humans Reason about Language and Images Compositionally

A yellow vase on top of a
black television

Large-scale benchmarks for vision-language compositionality needed

VALSE

caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
image						

Winoground

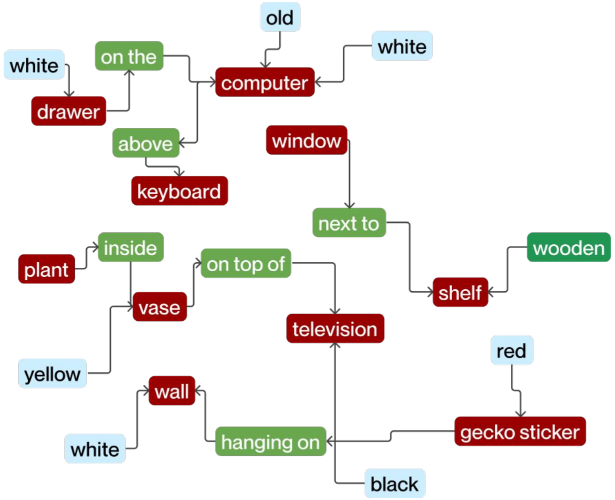
(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

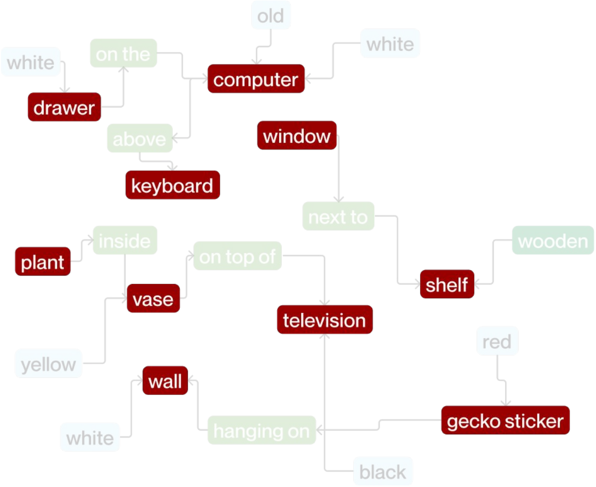
Images drawn from:
[1] Parcalabescu *et al. VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena*. 2021
[2] Thrush *et al. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality.* 2022.

We introduce CREPE: a benchmark to evaluate whether vision-language models exhibit compositionality

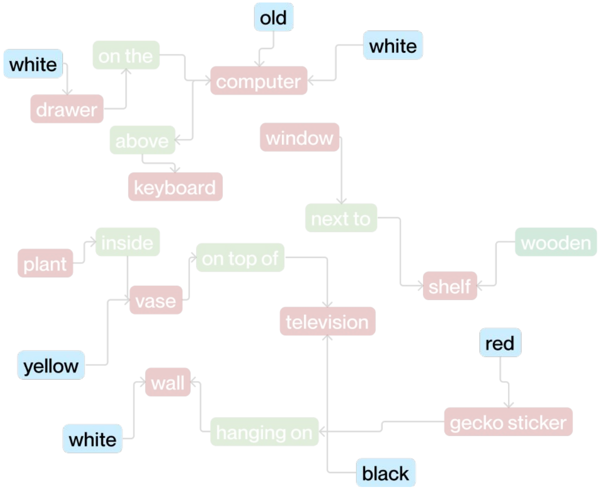
CREPE is constructed using scene graphs



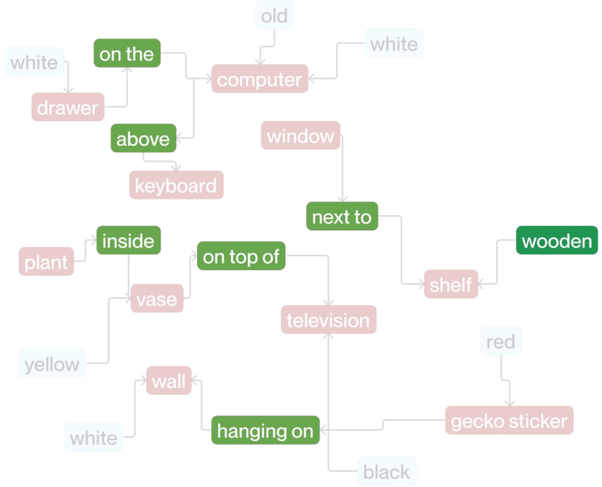
CREPE is constructed using scene graphs



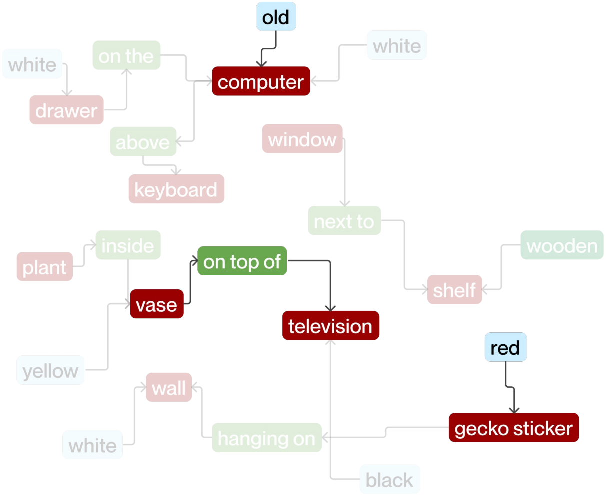
CREPE is constructed using scene graphs



CREPE is constructed using scene graphs



CREPE is constructed using scene graphs



Systematicity measures generalization to novel compositions

Seen compositions

A yellow **vase** on top of a black television

Tennis players wearing **pink** t-shirts

Systematicity measures generalization to novel compositions

Seen compositions

A yellow **vase** on top of a black television

Tennis players wearing **pink** t-shirts

Unseen test composition

White and purple flowers in a **pink** **vase**

Productivity measures understanding of increasingly complex captions

Complexity	Caption
n=4	A yellow vase on top of a television
n=5	A yellow vase on top of a black television
n=7	Plant inside a yellow vase on top of a black television
n=10	Plant inside a yellow vase on top of a black television in front of an old computer

CREPE evaluates models in both settings with image-to-text retrieval

- ✓ A yellow vase on top of a television
- ✗ Negative caption 1
- ✗ Negative caption 2
- ✗ Negative caption 3
- ✗ Negative caption 4

Randomly selecting negative captions introduces noise to evaluation

- ✓ A yellow vase on top of a television
- ✗ A black dog catching a frisbee
- ✗ Smiling man wearing sunglasses
- ✗ A tabby cat sleeping on a computer
- ✗ A red car next to a streetlight

Randomly selecting negative captions introduces noise to evaluation

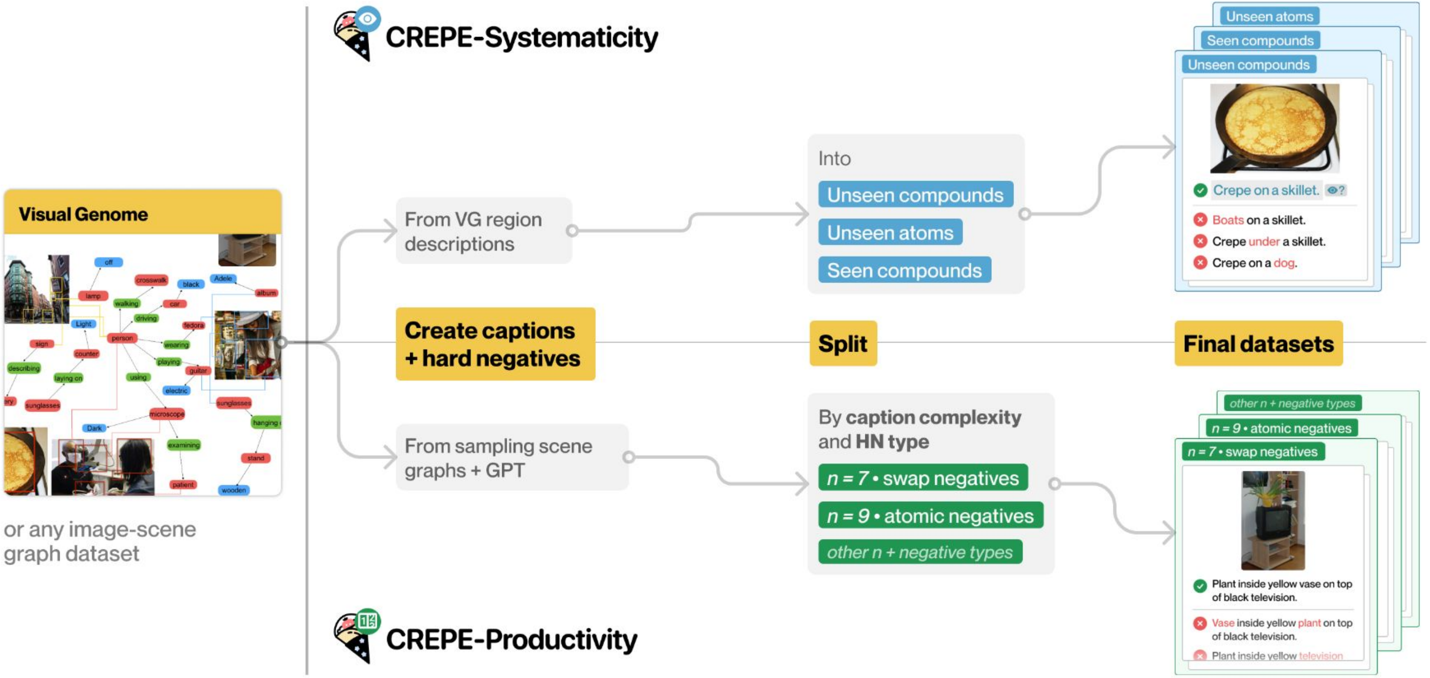
- ✓ **A yellow vase on top of a television**
- ✗ A black dog catching a frisbee
- ✗ Smiling man wearing sunglasses
- ✗ A tabby cat sleeping on a computer
- ✗ A red car next to a streetlight

CREPE uses hard-negatives to detect particular error modes

- ✓ A yellow vase on top of a television
- ✗ A **red** vase on top of a television
- ✗ A yellow vase on top of a **table**
- ✗ A **television** on top of a **yellow vase**
- ✗ A yellow vase **next to** a television

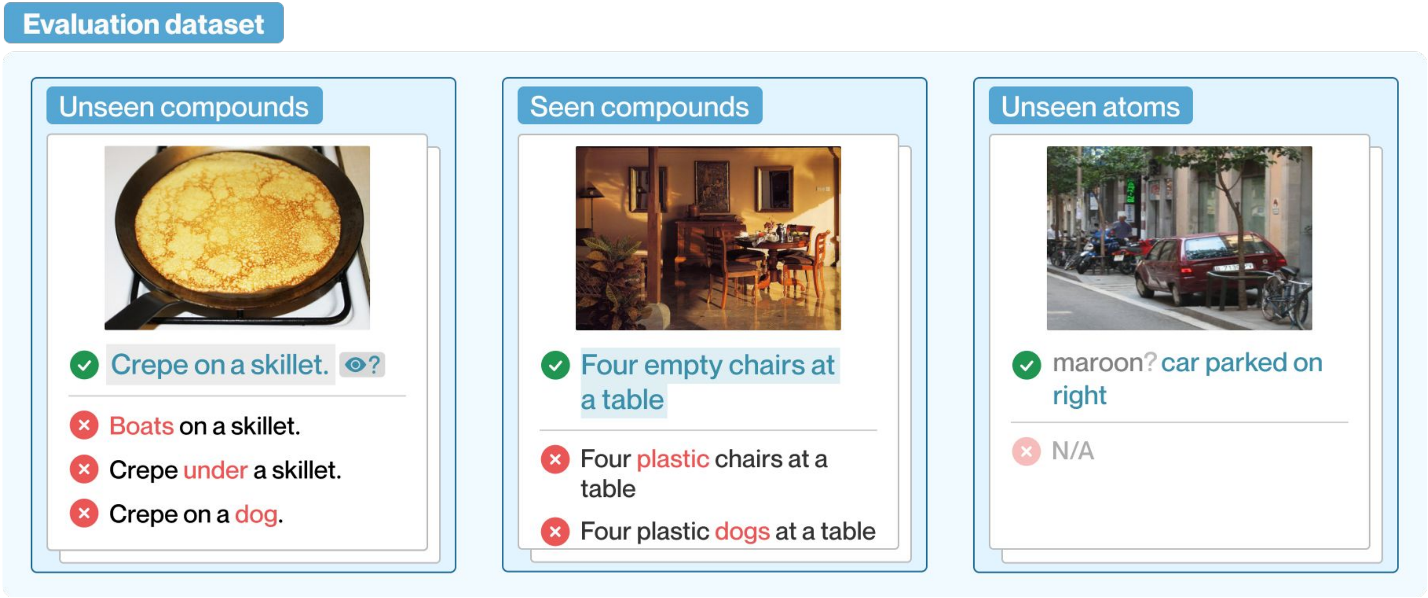
Making and Evaluating CREPE

At a glance

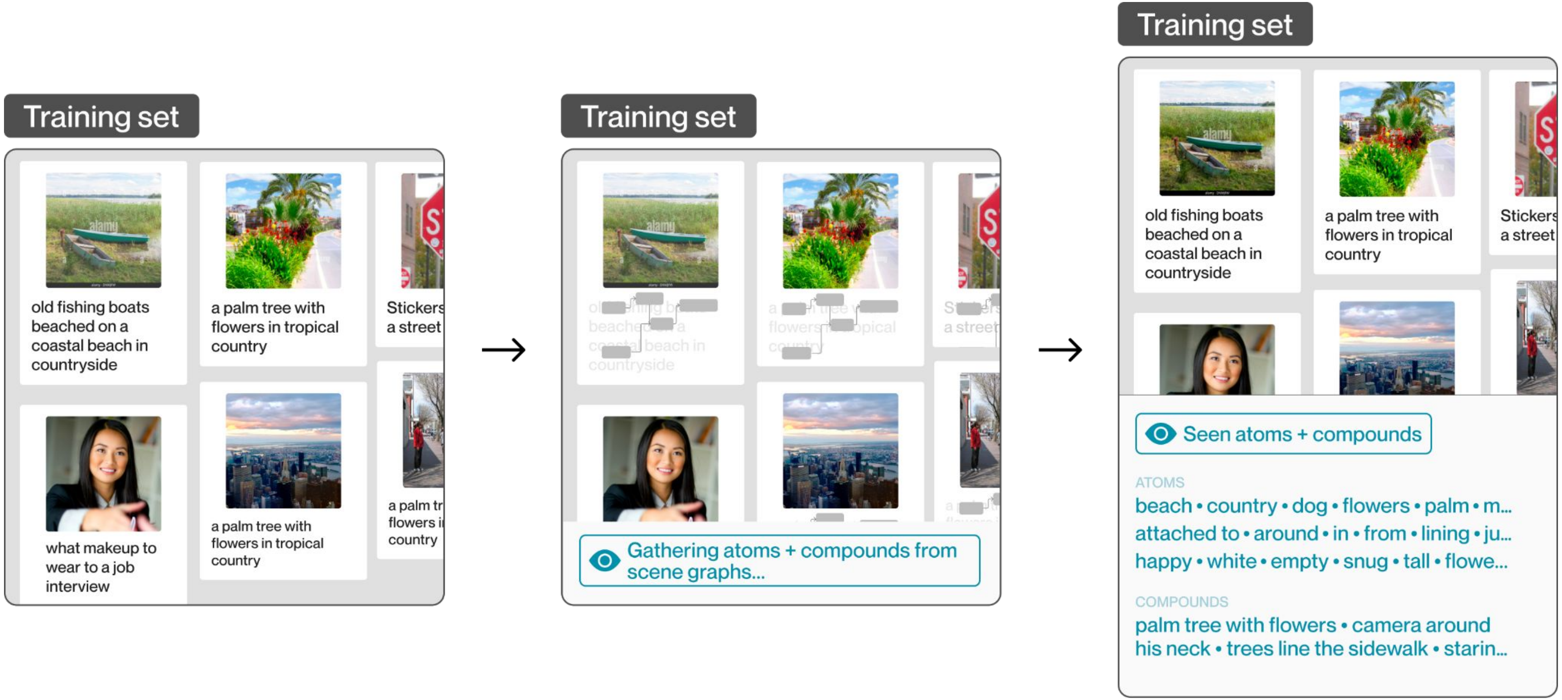


CREPE-Systematicity
Visual Genome
or any image-scene graph dataset
From VG region descriptions
Create captions + hard negatives
From sampling scene graphs + GPT
Into
Unseen compounds
Unseen atoms
Seen compounds
Split
By caption complexity and HN type
n = 7 • swap negatives
n = 9 • atomic negatives
other n + negative types
Final datasets
Unseen atoms
Seen compounds
Unseen compounds
Crepe on a skillet.
Boats on a skillet.
Crepe under a skillet.
Crepe on a dog.
other n + negative types
n = 9 • atomic negatives
n = 7 • swap negatives
Plant inside yellow vase on top of black television.
Vase inside yellow plant on top of black television.
CREPE-Productivity

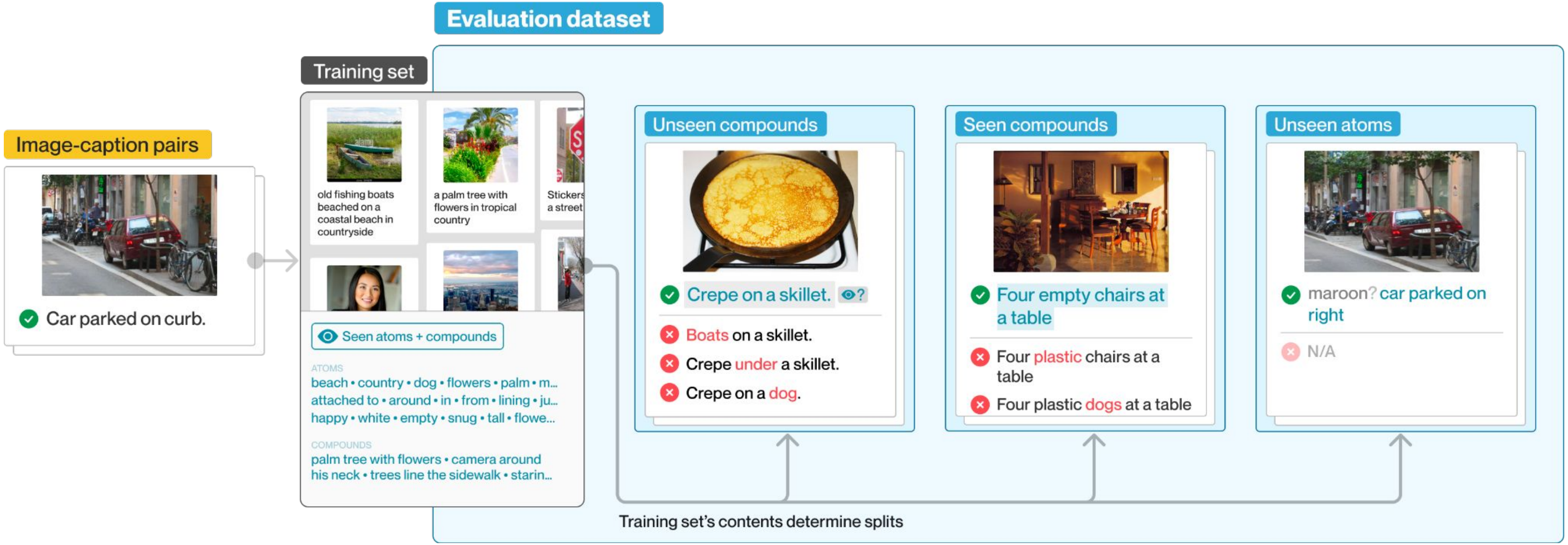
The Systematicity Dataset



The Systematicity Dataset



The Systematicity Dataset



The Systematicity Dataset

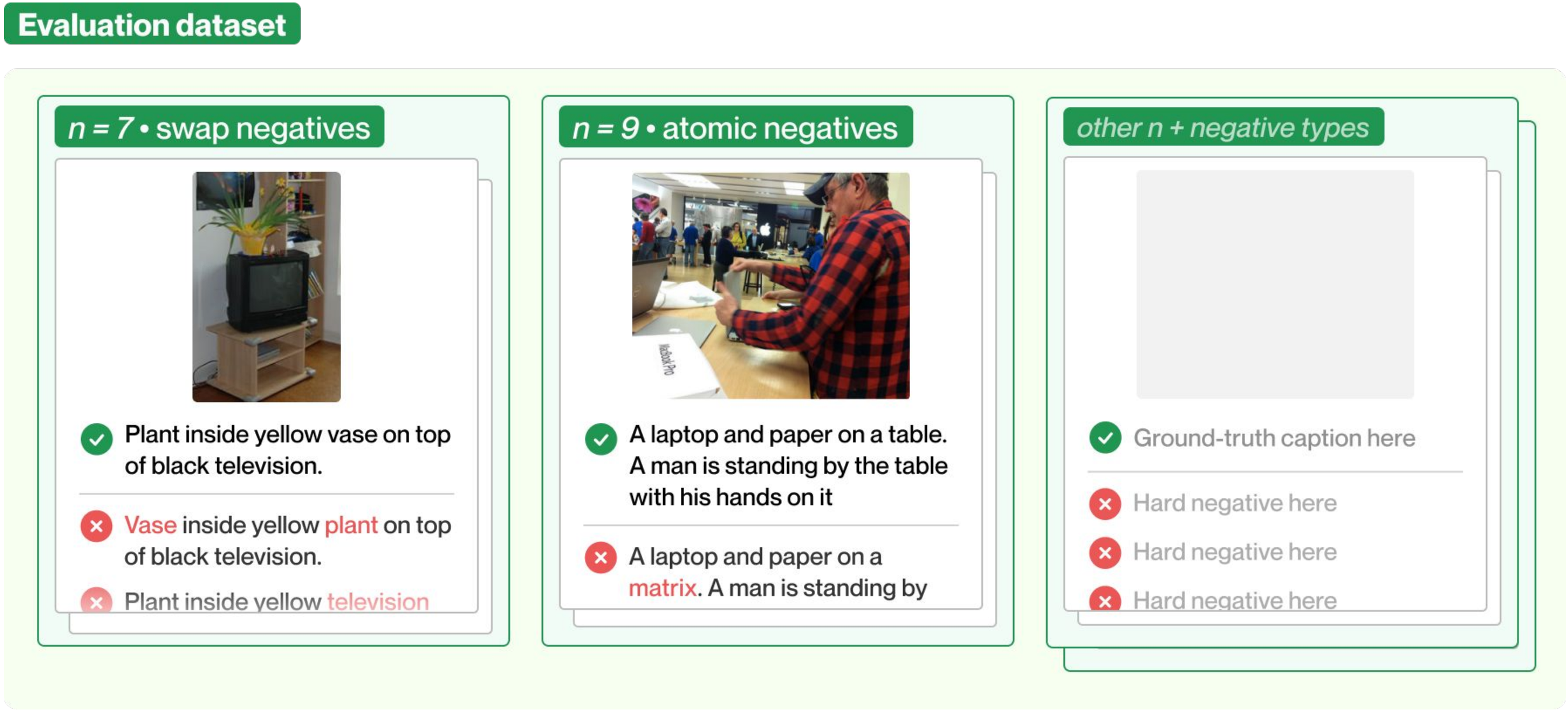
Boats on a skillet

HN-ATOM Caption

Crepe below a table and egg on a skillet

HN-COMP Caption

The Productivity Dataset

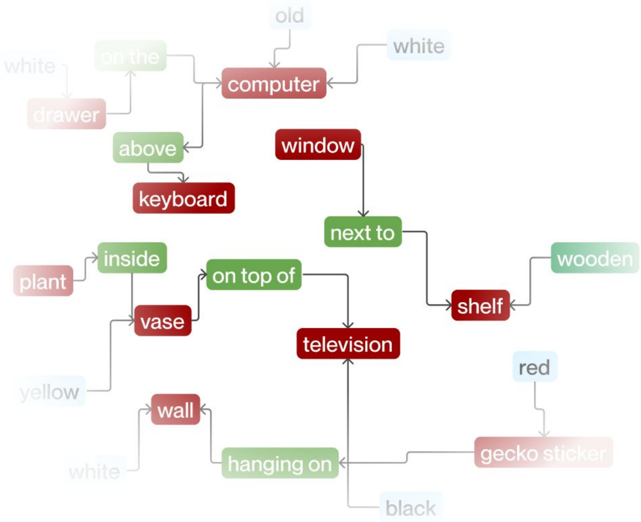


The Productivity Dataset

For a given complexity and hard negative type,

n = 7 • swap negatives

Image-scene graph pair



The Productivity Dataset

For a given complexity and hard negative type,

n = 7 • swap negatives

1: Perform a *n*-length random walk on the graph:

The Productivity Dataset

For a given complexity and hard negative type,

n = 7 • swap negatives

2: Generate a caption from subgraph. For n=8, we'll use a GPT prompt to do so..

Plant inside yellow vase on top of black television.

The Productivity Dataset

For a given complexity and hard negative type,

n = 7 • swap negatives

3: Generate all the hard negatives.

Plant inside yellow vase on top of black television.

Vase inside yellow plant on top of black television.

Plant inside yellow television on top of black vase.

Plant inside black vase on top of... yellow television.

Television inside of yellow vase on top of black plant.

...

The Productivity Dataset

Horse inside yellow vase on top of black television.

HN-Atom Caption

Plant inside black vase on top of yellow television.

HN-Swap Caption

Plant inside yellow vase not on top of black television.

HN-Neg Caption

Data Verification

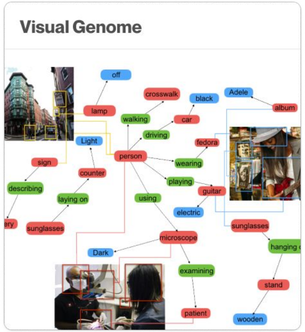
Two human annotators verified a subset of the ground truth and hard negative captions we generated for our evaluation datasets.

Ground truth captions	Productivity
Accuracy to image	87.9%
Pairwise annotator agreement	88.8%

Hard negative captions	Systematicity	Productivity
Genuine negative (incorrect statement about image)	86.0%	83.7%
Pairwise annotator agreement	83.7%	84.3%

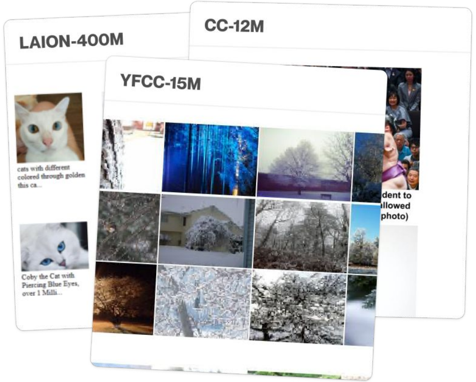
Experimental Setup

Datasets used



Visual Genome

The raw material; used to create the actual image-text pairs in the datasets.



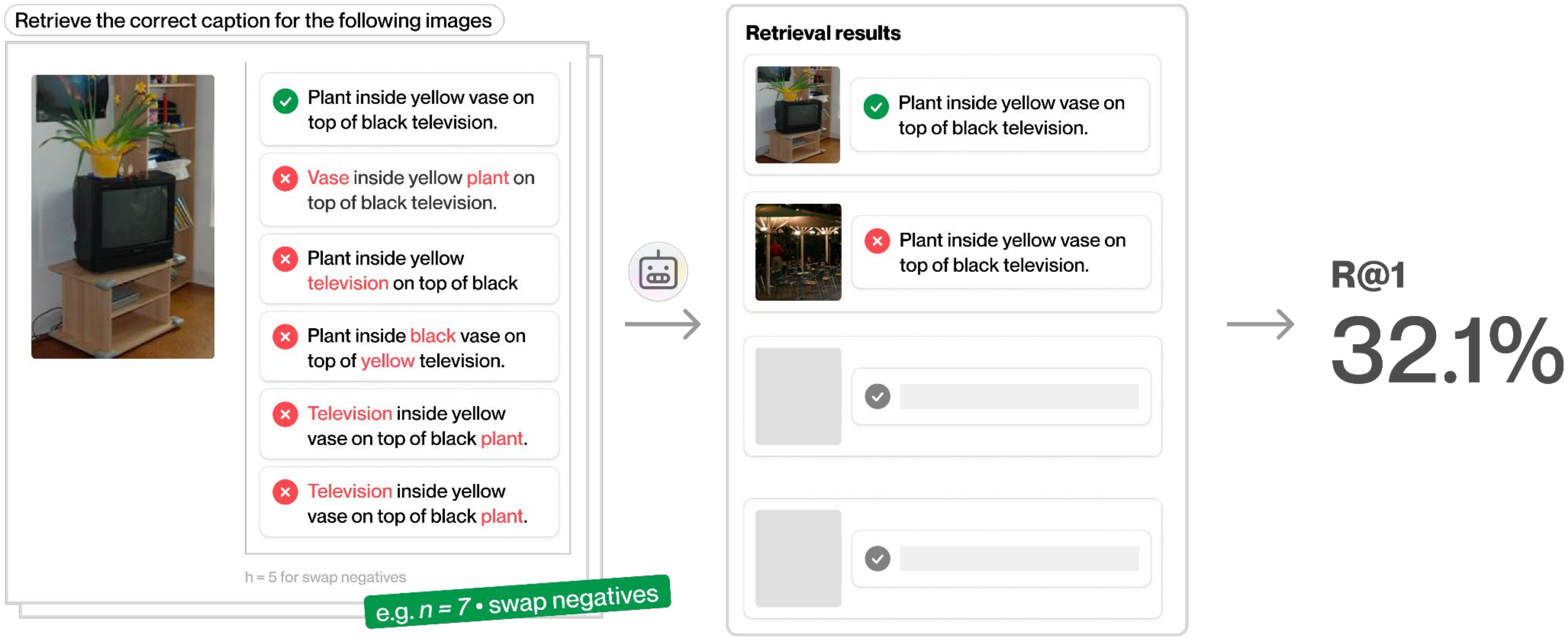
CC-12M, YFCC-15M, and LAION-400M

Training sets used to determine the splits of CREPE-Systematicity. Each training set above results in a different three-way split of the same captions.

We construct new **large-scale datasets for image-to-text retrieval** evaluation by leveraging Visual Genome's scene graphs.

	Systematicity			Productivity
Training dataset	CC-12M	YFCC-15M	LAION-400M	Any
Number of ground-truth image-text pairs in the test set	**385,777**	**385,777**	**373,703**	**17,553**
Number of hard negative texts in the test set	**325,523**	**316,668**	**309,342**	**183,855**

Retrieval metrics

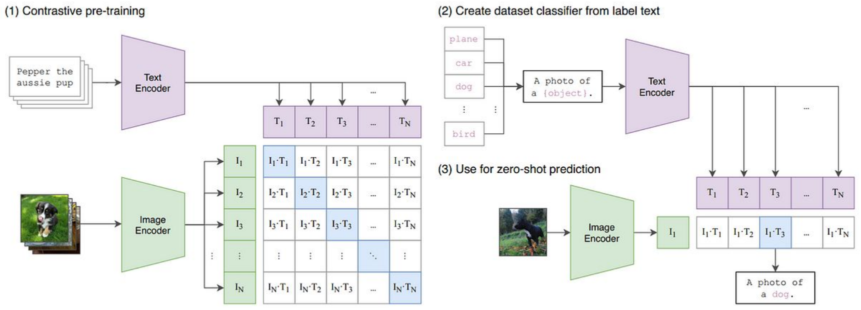


Models evaluated

Systematicity

A variety of **OpenCLIP's CLIP** models:

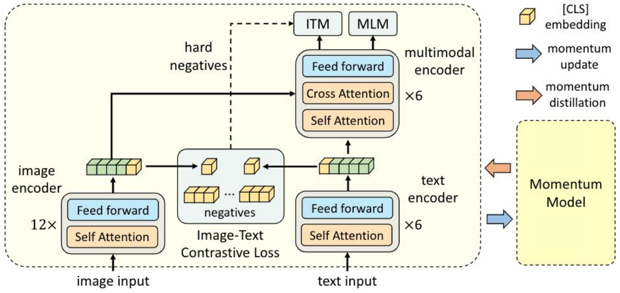
- trained on: CC12M, YFCC15M **or** LAION-400M,
- 6 different backbones total.



Productivity

We don't require knowledge of the models' training sets. Therefore:

- *OpenCLIP models used for systematicity, plus*
- **OpenAI's CLIP, CyCLIP, ALBEF, and FLAVA.**



Systematicity: Models' recall@1 decreases from the SC to UC split on the hard negatives test set with HN-ATOM, and HN-ATOM + HN-COMP.

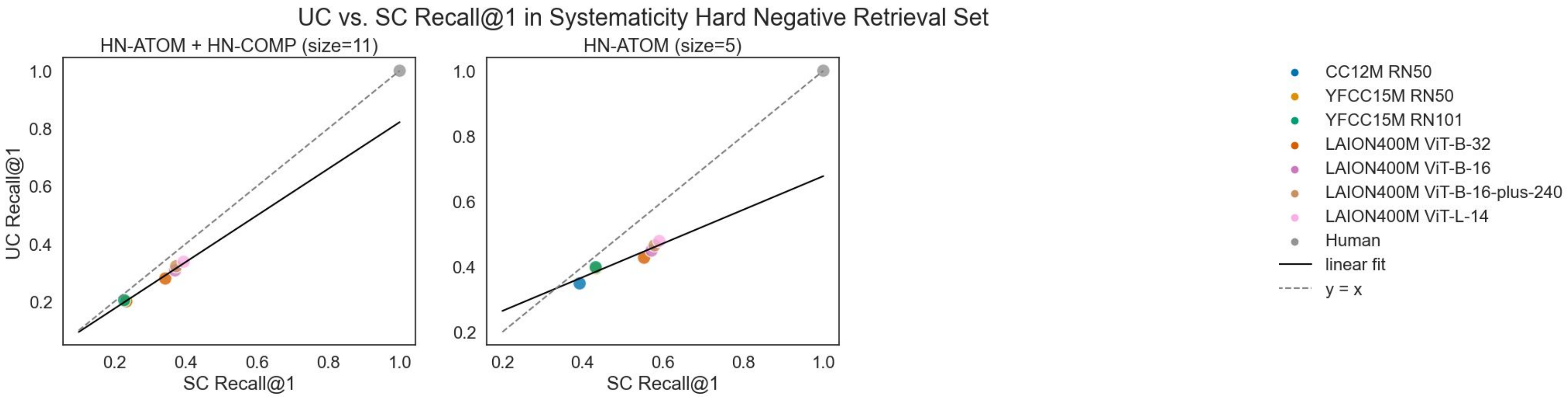


Figure 4. We plot models' recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types.

Systematicity: The drop is small for the CC-12M and YFCC-15M trained models and the most pronounced for LAION-400M-trained models.

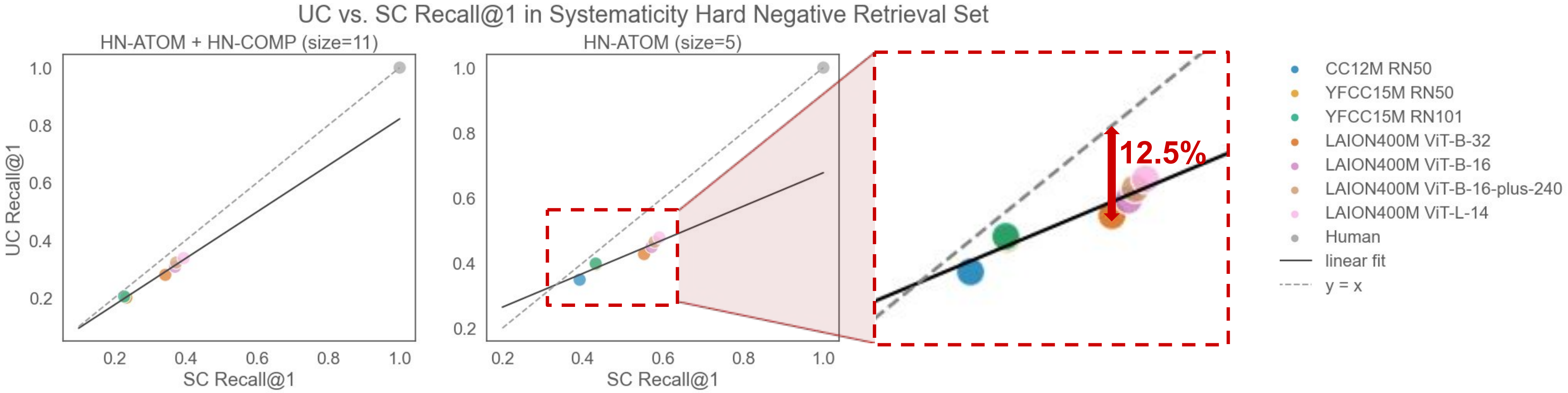


Figure 4. We plot models' recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types.

Systematicity: We find little to no difference in models' performance between the SC and UC split on the HN-COMP subset.

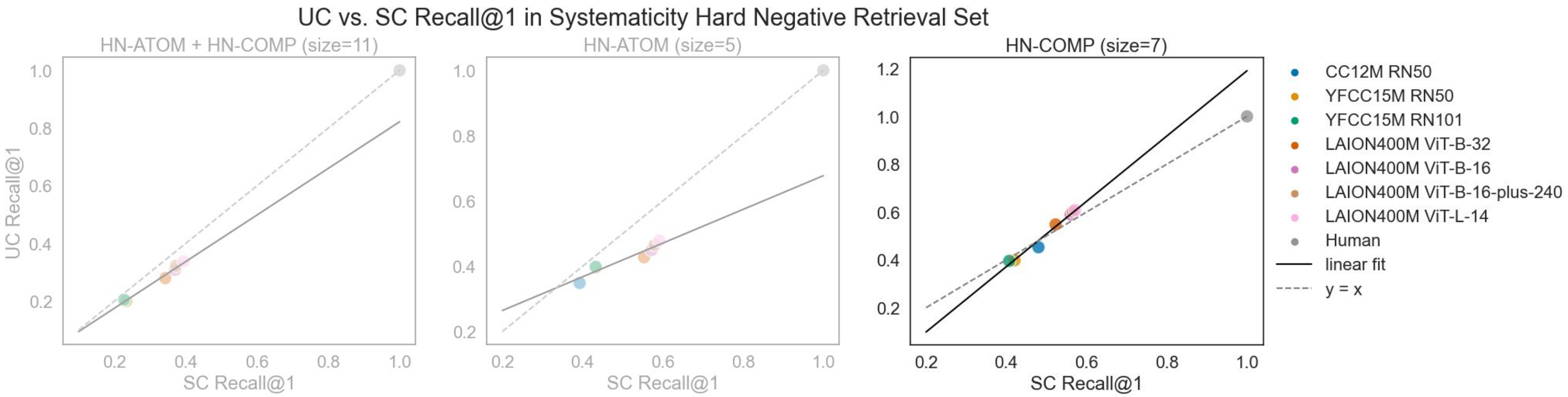


Figure 4. We plot models' recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types.

Productivity: OpenCLIP models' R@1 drops to random chance or below as complexity increases, particularly on the HN-ATOM and HN-SWAP sets.

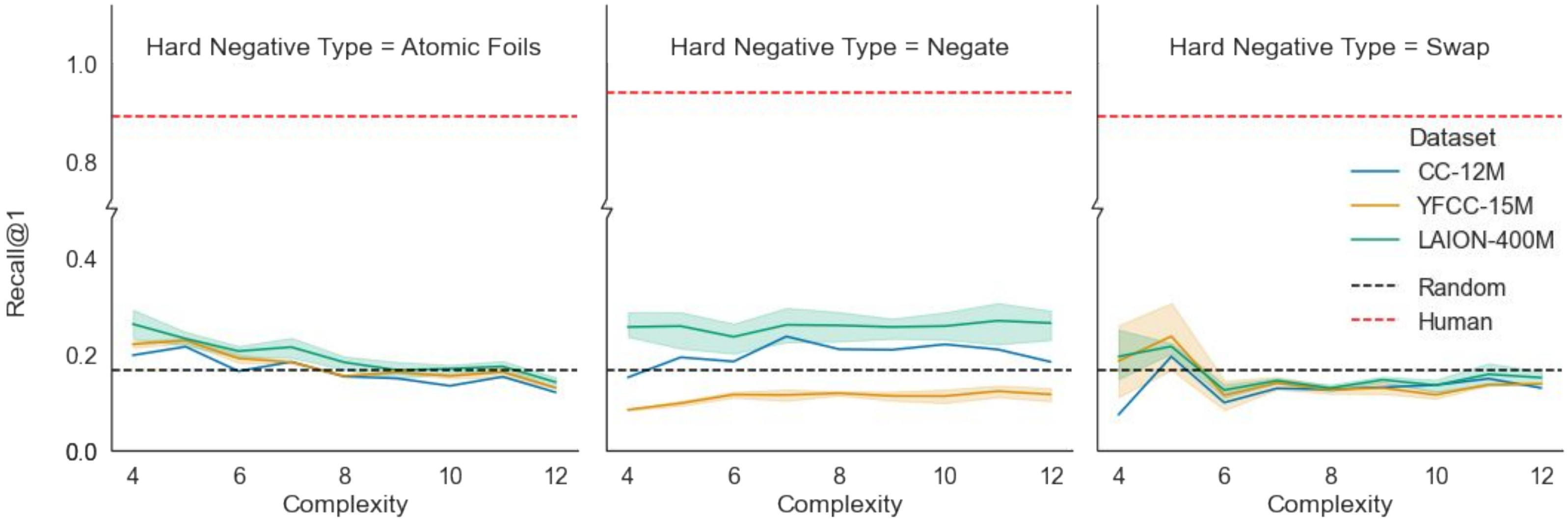


Figure 5. We plot models' recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets.

Productivity: OpenCLIP models' R@1 drops to random chance or below as complexity increases, particularly on the HN-ATOM and HN-SWAP sets.

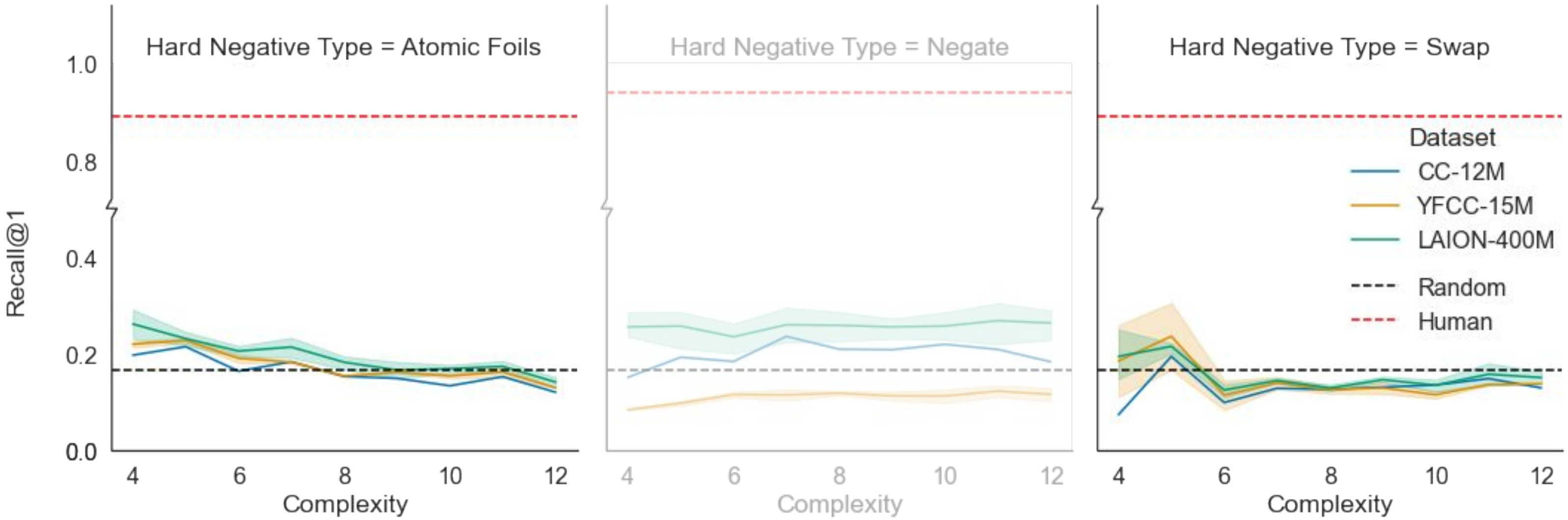


Figure 5. We plot models' recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets.

Productivity: Other VL models also demonstrate low performance and a downward trend as complexity increases, except for OpenAI CLIP models.

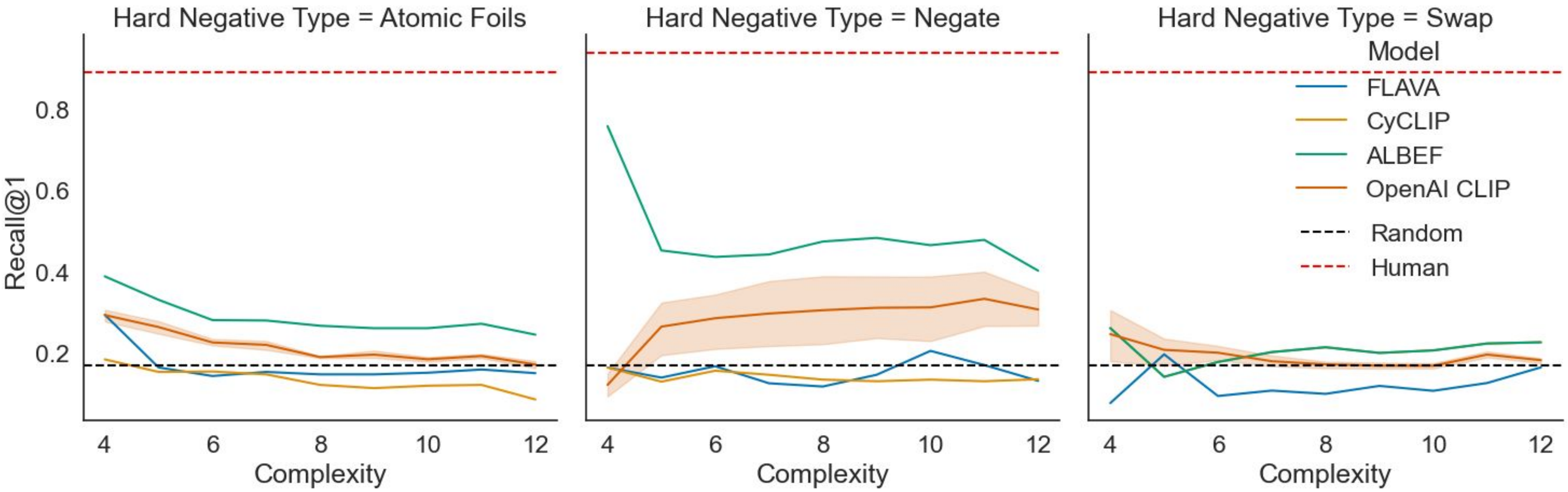


Figure 5. We plot models' recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets.

Productivity: Other VL models also demonstrate low performance and a downward trend as complexity increases, except for OpenAI CLIP models.

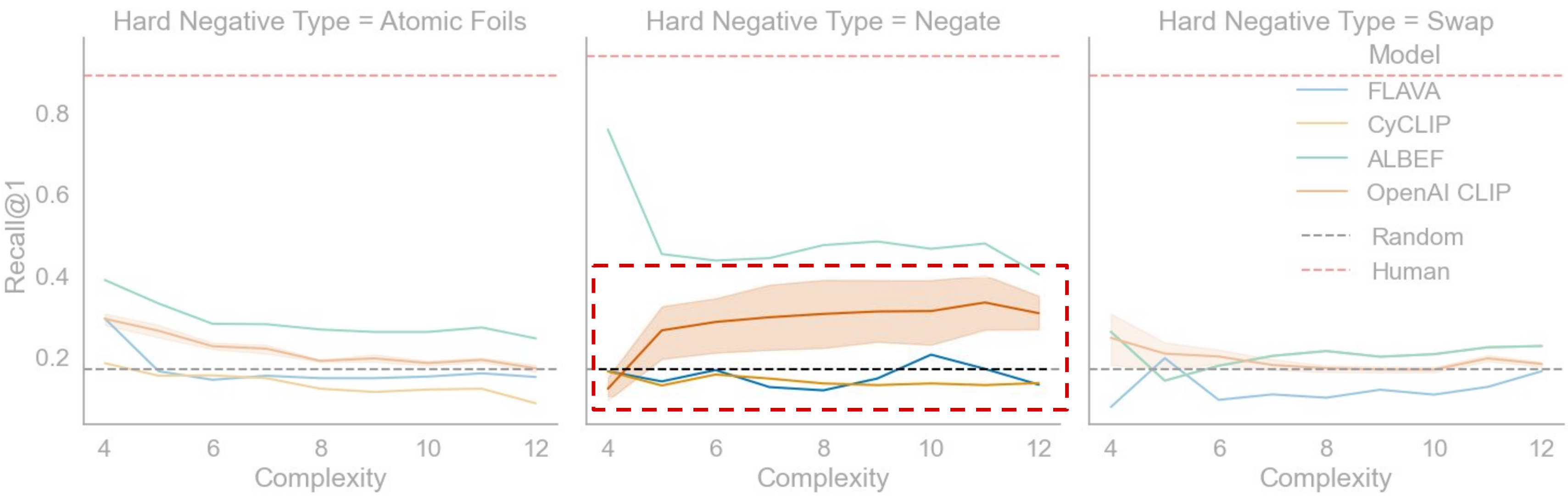


Figure 5. We plot models' recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets.

Systematicity and Productivity: We find no particular trends relating compositionality to training dataset size or model size.

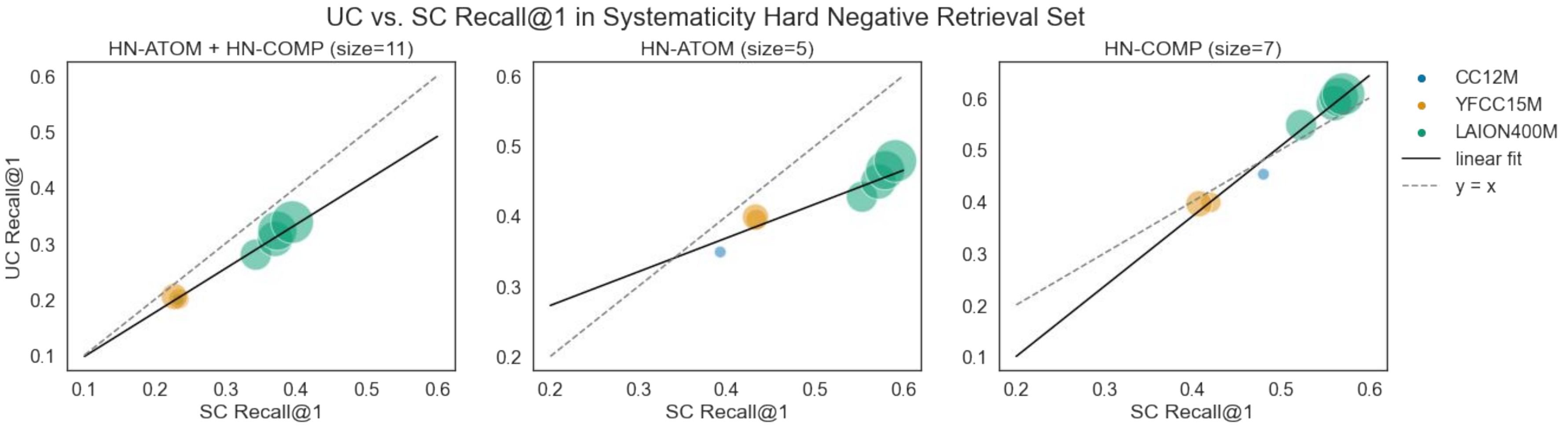


Figure 6. We plot models' recall@1 on the Seen Compounds vs. Unseen Compounds split of the systematicity retrieval set with hard negatives HN-ATOM, HN-COMP and both types, where the dot size represents model size.

Systematicity and Productivity: We find no particular trends relating compositionality to training dataset size or model size.

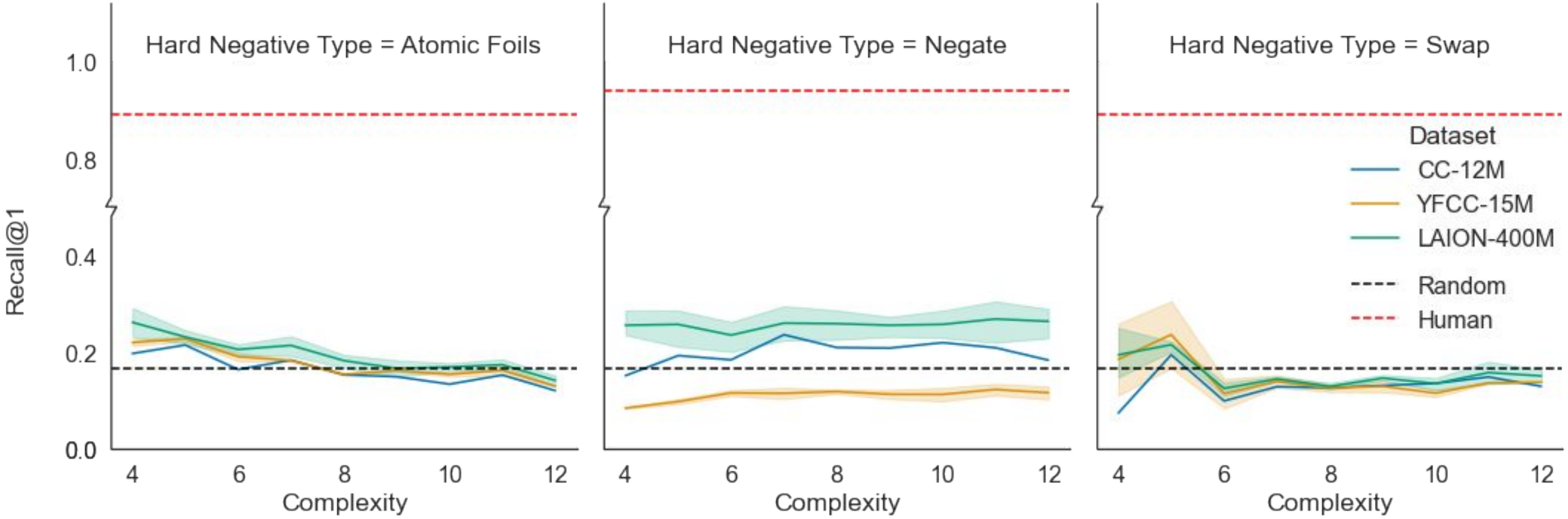


Figure 5. We plot models' recall@1 on the hard negatives retrieval set against complexity, averaged across all models pretrained on all three training datasets.

Systematicity and Productivity: We find no particular trends relating compositionality to training dataset size or model size.

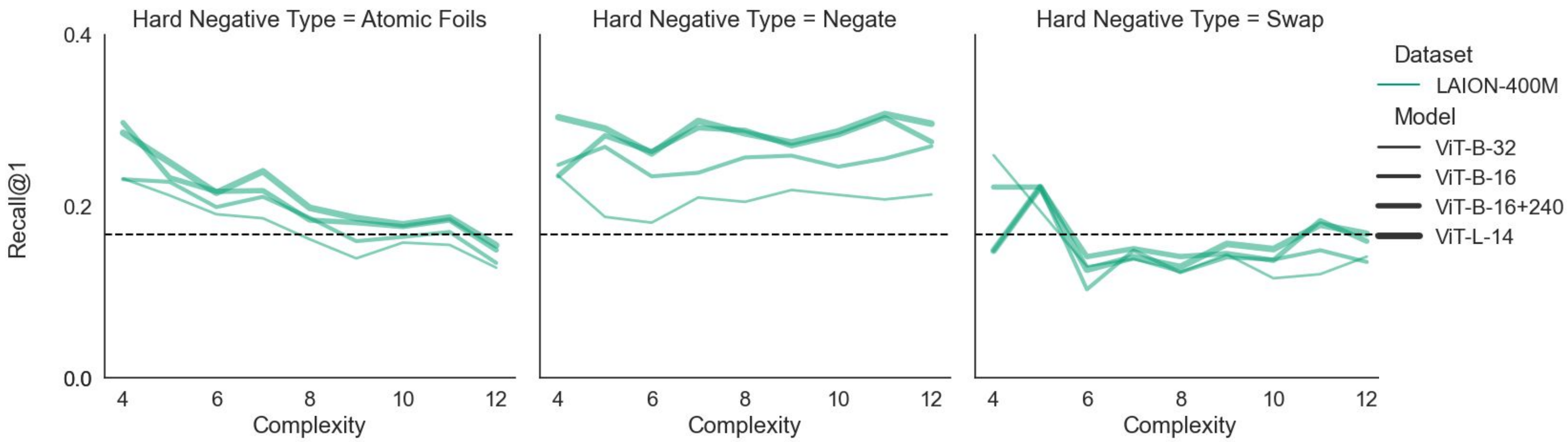


Figure 7. We plot the recall@1 of the four LAION-400M trained models of different sizes on the three hard negatives retrieval sets across complexities.

CREPE: Can Vision-Language Foundation Models Reason Compositionally?

No, state-of-the-art vision-language models do NOT exhibit systematicity or productivity, and compositionality is NOT likely to emerge as we scale up the training dataset or model size.

For more details, please refer to our paper from the QR code.

Code: https://github.com/RAIVNLab/CREPE.git

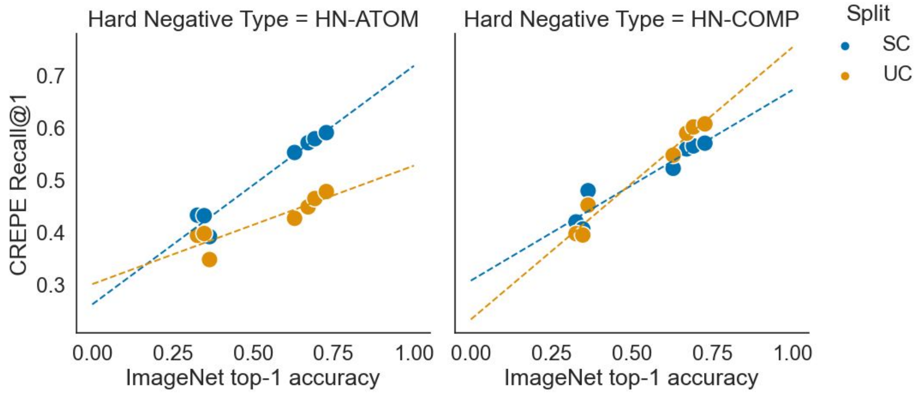
Contact: zixianma@cs.stanford.edu, jerryhong@cs.stanford.edu, mog29@cornell.edu

Backup slides

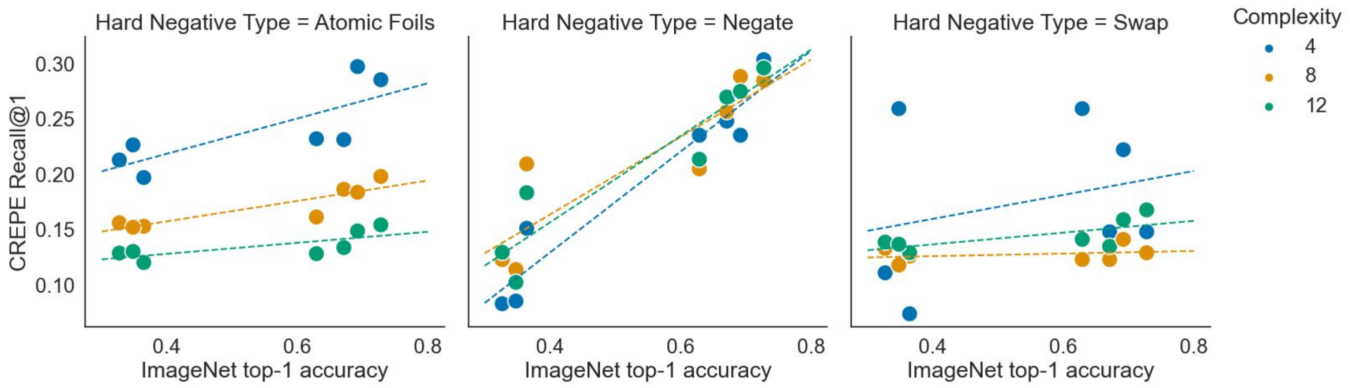
Results

Systematicity and Productivity: Zero-shot ImageNet accuracy strongly correlates with models' Recall@1 on the HN sets, except for HN-SWAP.

Systematicity



Productivity



We introduce CREPE: a benchmark to evaluate whether vision-language models exhibit compositionality

CREPE makes three major contributions:

2. We generate **a large quantity of hard negative captions** to support evaluation.

	Systematicity			Productivity
Training dataset	CC-12M	YFCC-15M	LAION-400M	Any
Number of hard negative texts in the test set	**325,523**	**316,668**	**309,342**	**183,855**

We present 4 key takeaways from our experiment.

1. **Systematicity**: Models' retrieval recall decreases when the compositions in the image are unseen.

We present 4 key takeaways from our experiment.

1. Systematicity: Models' retrieval recall decreases when the compositions in the image are unseen.
2. **Systematicity**: The decrease is largest — **12%** — for models trained on the largest dataset LAION-400M.

We present 4 key takeaways from our experiment.

1. Systematicity: Models' retrieval recall decreases when the compositions in the image are unseen.
2. Systematicity: The decrease is largest — 12% — for models trained on the largest dataset LAION-400M.
3. **Productivity**: Models' retrieval recall drops to **random chance or below** as caption complexity increases.

We present 4 key takeaways from our experiment.

1. Systematicity: Models' retrieval recall decreases when the compositions in the image are unseen.
2. Systematicity: The decrease is largest — 12% — for models trained on the largest dataset LAION-400M.
3. Productivity: Models' retrieval recall drops to random chance or below as caption complexity increases.
4. **Both**: We find no particular trends relating **training dataset or model size** to models' performance on our test sets.