

CLIP Under the Microscope: A Fine-Grained Analysis of Multi-Object Representation

Supplementary Material

7. Appendix

7.1. The SIMCO and ComCO Datasets

7.1.1. The SIMCO Dataset

The SIMCO dataset comprises 17 objects. These 17 objects are:

Cube	Sphere	Cylinder
Mug	Pentagon	Heart
Cone	Pyramid	Diamond
Moon	Cross	Snowflake
Leaf	Arrow	Star
Torus	Pot	

Using Blender software, a collection of images containing 2 to 5 objects has been created from these 17 objects. The total number of images in this dataset is approximately 85,000. Examples of these images can be seen in Figure 6.

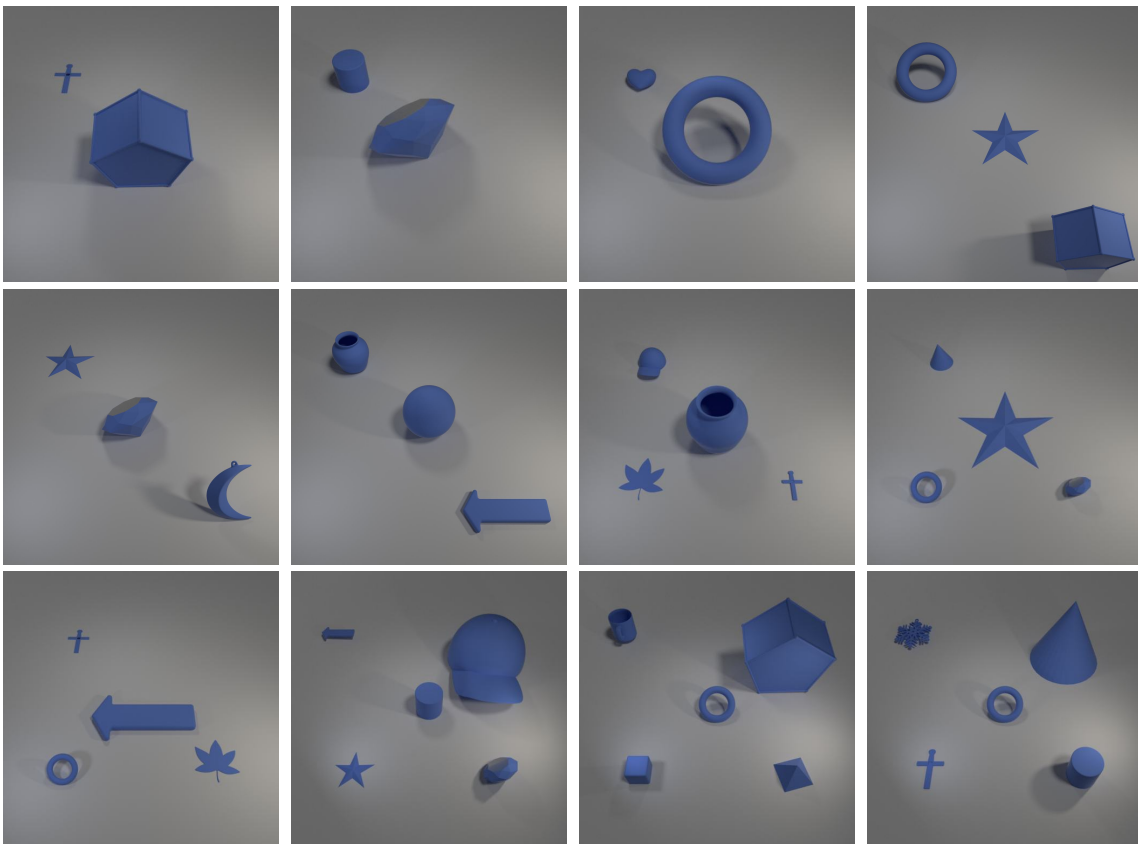


Figure 6. Examples of the SimCO dataset

7.1.2. The ComCO Dataset

The ComCO dataset contains 72 objects, as listed below:

The ComCO dataset contains 72 objects, as listed below:

person	bicycle	car	motorcycle	airplane	bus
train	truck	boat	traffic light	fire hydrant	street sign
stop sign	parking meter	bench	bird	cat	dog
horse	sheep	cow	dining table	cell phone	elephant
bear	zebra	giraffe	hat	backpack	umbrella
shoe	eye glasses	handbag	tie	suitcase	frisbee
skis	snowboard	kite	baseball bat	baseball glove	tennis racket
wine glass	hot dog	potted plant	teddy bear	hair drier	hair brush
skateboard	surfboard	bottle	plate	cup	fork
knife	spoon	bowl	banana	apple	sandwich
orange	broccoli	carrot	pizza	donut	cake
chair	couch	bed	mirror	window	desk
toilet	door	tv	laptop	mouse	remote
keyboard	microwave	oven	toaster	sink	refrigerator
blender	book	clock	vase	scissors	toothbrush

In this dataset, a collection of images containing 2 to 5 different objects has also been generated. The total number of images in this dataset is approximately 190,000. Various examples from this dataset can be seen in Figure 12.

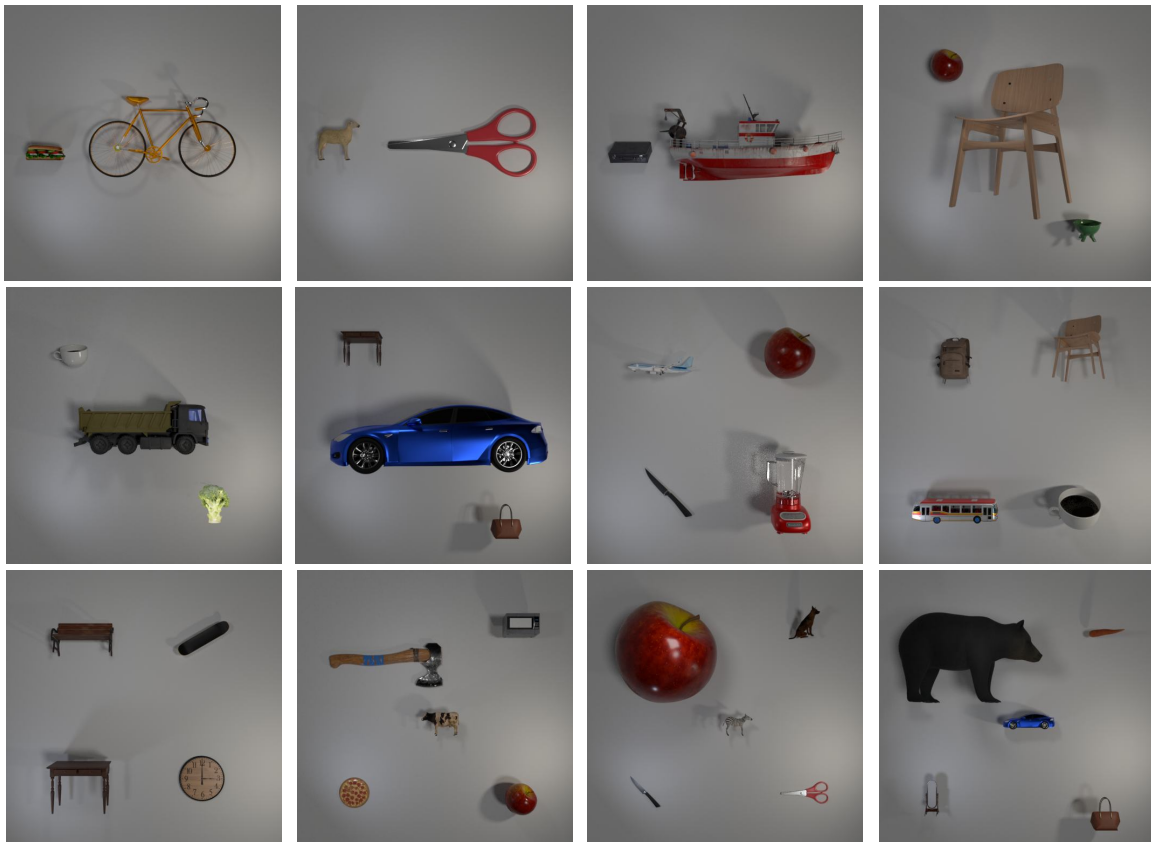


Figure 7. Examples of the ComCO dataset

7.2. Text-based Object Classification

7.2.1. Objective

The Text-based Object Classification experiment was designed to evaluate CLIP’s text encoder’s ability to represent individual objects within multi-object captions. Our goal was to quantify any potential bias in the representation of objects based on their position in the text.

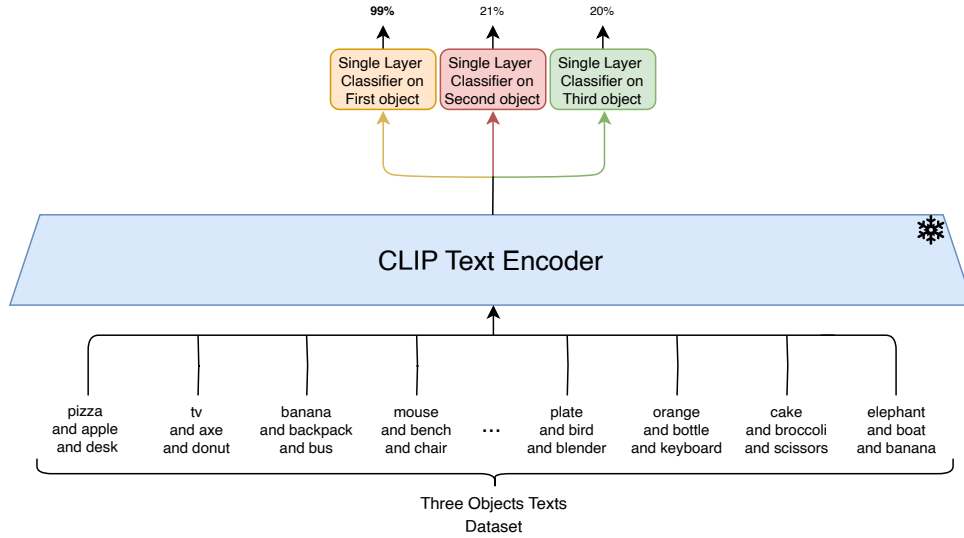


Figure 8. Illustration of the Text-based Object Classification experiment. The figure demonstrates how embeddings are calculated for multi-object captions using CLIP’s text encoder. A single-layer classifier is then trained on these embeddings to classify individual objects.

7.2.2. Methodology

1. Dataset Preparation:

- We used both the SimCO and ComCO datasets, which contain captions describing scenes with 2 to 5 objects.
- Each caption in the dataset follows a consistent format: “Object1 and Object2 and ... and ObjectN”.

2. Text Embedding Generation:

- For each multi-object caption, we used CLIP’s text encoder to generate a text embedding.
- This embedding is a high-dimensional vector representation of the entire caption.

3. Classifier Training:

- For each object position (1st, 2nd, 3rd, etc.), we trained a separate single-layer classifier.
- Input: The text embedding of the multi-object caption.
- Output: The predicted object class for that specific position.

4. Evaluation:

- We tested each classifier on a held-out portion of the dataset.
- For each caption, we recorded whether the classifier correctly identified the object at its respective position.
- We calculated the classification accuracy for each object position across all test captions.

We conducted the TOC experiment on various models under different scenarios, and the results are presented in Table 7. This experiment was repeated on both the SIMCO and ComCO datasets.

Table 7. Text-based Object Classification

Number of Objects	Dataset	Model	First Object	Second Object	Third Object	Fourth Object	Fifth Object
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	99.86	97.09	-	-	-
		<i>ViT-SO400M-SigLIP</i>	98.67	91.29	-	-	-
		<i>ViT-L-14 (datacomp)</i>	99.76	96.77	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.03	89.87	-	-	-
		<i>ViT-L-14 (laion2b)</i>	99.70	97.57	-	-	-
		<i>ViT-L-14 (openai)</i>	97.62	91.30	-	-	-
		<i>ViT-B-32 (openai)</i>	96.85	73.00	-	-	-
		<i>NegCLIP</i>	98.19	84.43	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	99.90	96.56	-	-	-
		<i>ViT-SO400M-SigLIP</i>	98.47	93.18	-	-	-
		<i>ViT-L-14 (datacomp)</i>	99.74	96.86	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.16	91.57	-	-	-
		<i>ViT-L-14 (laion2b)</i>	99.72	96.24	-	-	-
		<i>ViT-L-14 (openai)</i>	97.93	96.69	-	-	-
		<i>ViT-B-32 (openai)</i>	96.86	85.42	-	-	-
		<i>NegCLIP</i>	99.30	92.09	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	99.46	60.47	76.99	-	-
		<i>ViT-SO400M-SigLIP</i>	98.23	71.42	45.80	-	-
		<i>ViT-L-14 (datacomp)</i>	99.49	45.80	78.66	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.26	49.08	64.07	-	-
		<i>ViT-L-14 (laion2b)</i>	98.93	56.87	72.37	-	-
		<i>ViT-L-14 (openai)</i>	91.87	50.75	68.38	-	-
		<i>ViT-B-32 (openai)</i>	92.55	38.61	52.94	-	-
		<i>NegCLIP</i>	95.80	44.70	59.11	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	99.73	59.80	73.63	-	-
		<i>ViT-SO400M-SigLIP</i>	96.94	70.26	29.28	-	-
		<i>ViT-L-14 (datacomp)</i>	99.53	45.13	74.15	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.20	53.34	57.15	-	-
		<i>ViT-L-14 (laion2b)</i>	99.26	58.58	64.74	-	-
		<i>ViT-L-14 (openai)</i>	90.86	49.67	83.49	-	-
		<i>ViT-B-32 (openai)</i>	87.97	45.77	63.13	-	-
		<i>NegCLIP</i>	56.94	98.03	56.66	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	99.46	34.57	36.73	62.35	-
		<i>ViT-SO400M-SigLIP</i>	98.23	69.91	26.10	6.54	-
		<i>ViT-L-14 (datacomp)</i>	99.00	23.76	35.55	60.55	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.26	27.97	28.84	48.34	-
		<i>ViT-L-14 (laion2b)</i>	98.82	34.21	31.41	54.73	-
		<i>ViT-L-14 (openai)</i>	90.48	35.19	30.50	59.29	-
		<i>ViT-B-32 (openai)</i>	90.76	22.77	25.36	40.45	-
		<i>NegCLIP</i>	96.50	9.33	4.79	15.58	-
	ComCO	<i>ViT-H-14 (DFN)</i>	99.76	31.74	35.29	54.82	-
		<i>ViT-SO400M-SigLIP</i>	97.27	72.51	33.25	5.79	-
		<i>ViT-L-14 (datacomp)</i>	99.46	22.82	32.93	58.18	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.60	26.27	26.20	36.51	-
		<i>ViT-L-14 (laion2b)</i>	98.89	31.64	20.90	47.76	-
		<i>ViT-L-14 (openai)</i>	87.17	30.60	31.69	74.49	-
		<i>ViT-B-32 (openai)</i>	88.24	24.23	28.30	49.82	-
		<i>NegCLIP</i>	98.73	28.05	30.83	43.82	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	99.00	24.30	22.33	27.23	53.03
		<i>ViT-SO400M-SigLIP</i>	97.79	71.67	27.41	6.29	6.48
		<i>ViT-L-14 (datacomp)</i>	98.89	16.51	21.29	26.92	48.52
		<i>xlm-roberta-large-ViT-H-14</i>	99.46	17.15	16.63	20.18	35.64
		<i>ViT-L-14 (laion2b)</i>	98.43	25.51	19.81	23.15	41.07
		<i>ViT-L-14 (openai)</i>	89.79	26.33	20.74	24.69	50.29
		<i>ViT-B-32 (openai)</i>	92.73	15.67	17.03	19.58	33.62
		<i>NegCLIP</i>	96.83	15.50	17.54	22.58	36.40
	ComCO	<i>ViT-H-14 (DFN)</i>	99.80	19.44	20.79	24.86	42.38
		<i>ViT-SO400M-SigLIP</i>	97.63	70.57	32.34	5.42	5.72
		<i>ViT-L-14 (datacomp)</i>	99.13	14.75	19.89	25.72	47.11
		<i>xlm-roberta-large-ViT-H-14</i>	99.40	18.21	15.47	18.05	26.12
		<i>ViT-L-14 (laion2b)</i>	98.76	20.91	18.11	20.77	33.54
		<i>ViT-L-14 (openai)</i>	86.13	22.11	19.43	28.03	68.37
		<i>ViT-B-32 (openai)</i>	91.20	15.56	13.31	19.66	39.39
		<i>NegCLIP</i>	99.03	16.69	16.51	22.26	34.29

7.3. Text-based Object Retrieval

7.3.1. Objective

The Text-based Object Retrieval (TOR) experiment was designed to assess CLIP’s text encoder’s ability to retrieve individual objects from multi-object captions. This experiment aimed to investigate potential biases in object retrieval based on the object’s position within the caption.

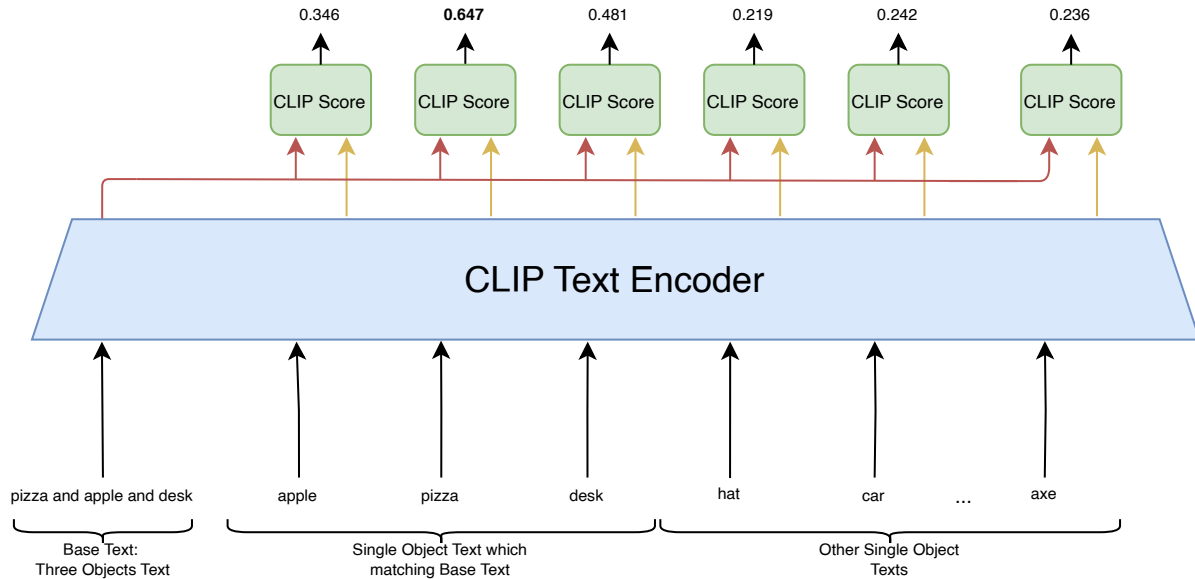


Figure 9. Visualization of the Text-based Object Retrieval experiment. This diagram illustrates the process of retrieving single-object texts based on multi-object captions using CLIP’s text encoder.

7.3.2. Methodology

1. Dataset Preparation:

- We utilized both the SimCO and ComCO datasets, containing captions describing scenes with 2 to 5 objects.
- Each multi-object caption followed the format: “Object1 and Object2 and ... and ObjectN”.
- We also prepared a set of single-object captions for each object class in our datasets.

2. Text Embedding Generation:

- We used CLIP’s text encoder to generate embeddings for all multi-object captions.
- Similarly, we generated embeddings for all single-object captions.

3. Similarity Computation:

- For each multi-object caption, we computed the cosine similarity between its embedding and the embeddings of all single-object captions.

4. Object Retrieval:

- For each multi-object caption, we identified the single-object caption with the highest similarity score.
- We recorded which object from the multi-object caption (1st, 2nd, 3rd, etc.) matched this retrieved single-object caption.

5. Evaluation:

- We calculated the percentage of times each object position (1st, 2nd, 3rd, etc.) was retrieved as the most similar.
- This percentage represents the retrieval accuracy for each object position.

We repeated the TOR experiment on various models across scenarios with captions containing 2 to 5 objects. This was done to confirm the presence of the discovered bias. The complete results of this experiment, which was conducted on both the SIMCO and ComCO datasets, can be observed in Table 8.

Table 8. Text-based Object Retrieval

Number of Objects	Dataset	Model	First Object	Second Object	Third Object	Fourth Object	Fifth Object
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	69.18	30.82	-	-	-
		<i>ViT-SO400M-SigLIP</i>	68.87	31.13	-	-	-
		<i>ViT-L-14 (datacomp)</i>	69.93	30.07	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	78.95	21.05	-	-	-
		<i>ViT-L-14 (laion2b)</i>	68.66	31.34	-	-	-
		<i>ViT-L-14 (openai)</i>	75.82	24.18	-	-	-
		<i>ViT-B-32 (openai)</i>	81.05	18.95	-	-	-
		<i>NegCLIP</i>	77.78	22.22	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	70.87	29.13	-	-	-
		<i>ViT-SO400M-SigLIP</i>	67.56	32.44	-	-	-
		<i>ViT-L-14 (datacomp)</i>	70.37	26.93	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	59.15	40.85	-	-	-
		<i>ViT-L-14 (laion2b)</i>	70.84	29.16	-	-	-
		<i>ViT-L-14 (openai)</i>	66.03	33.97	-	-	-
		<i>ViT-B-32 (openai)</i>	61.62	38.38	-	-	-
		<i>NegCLIP</i>	64.13	35.87	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	62.05	18.07	19.88	-	-
		<i>ViT-SO400M-SigLIP</i>	58.05	20.50	21.46	-	-
		<i>ViT-L-14 (datacomp)</i>	61.68	20.35	17.96	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	66.75	23.86	9.39	-	-
		<i>ViT-L-14 (laion2b)</i>	62.31	12.56	25.13	-	-
		<i>ViT-L-14 (openai)</i>	65.71	16.67	17.62	-	-
		<i>ViT-B-32 (openai)</i>	74.23	13.62	12.15	-	-
		<i>NegCLIP</i>	77.43	13.75	8.83	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	67.08	22.19	10.73	-	-
		<i>ViT-SO400M-SigLIP</i>	61.11	23.33	15.56	-	-
		<i>ViT-L-14 (datacomp)</i>	72.23	19.05	8.72	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	43.60	31.36	25.05	-	-
		<i>ViT-L-14 (laion2b)</i>	66.85	23.52	9.63	-	-
		<i>ViT-L-14 (openai)</i>	57.66	26.75	15.59	-	-
		<i>ViT-B-32 (openai)</i>	55.73	28.28	15.98	-	-
		<i>NegCLIP</i>	57.56	29.45	12.99	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	60.06	12.77	12.03	15.14	-
		<i>ViT-SO400M-SigLIP</i>	53.54	14.76	11.43	20.27	-
		<i>ViT-L-14 (datacomp)</i>	62.16	15.99	10.41	11.44	-
		<i>xlrm-roberta-large-ViT-H-14</i>	62.58	22.52	10.91	3.99	-
		<i>ViT-L-14 (laion2b)</i>	67.81	8.97	5.80	17.41	-
		<i>ViT-L-14 (openai)</i>	66.87	11.59	6.18	15.35	-
		<i>ViT-B-32 (openai)</i>	76.37	10.03	7.50	6.55	-
		<i>NegCLIP</i>	82.90	10.20	4.61	2.29	-
	ComCO	<i>ViT-H-14 (DFN)</i>	64.34	19.25	11.14	5.27	-
		<i>ViT-SO400M-SigLIP</i>	58.11	21.16	10.99	9.73	-
		<i>ViT-L-14 (datacomp)</i>	71.13	16.26	8.74	3.87	-
		<i>xlrm-roberta-large-ViT-H-14</i>	44.03	23.73	18.07	14.18	-
		<i>ViT-L-14 (laion2b)</i>	63.96	21.59	10.68	3.76	-
		<i>ViT-L-14 (openai)</i>	48.20	26.01	10.74	8.74	-
		<i>ViT-B-32 (openai)</i>	50.31	20.74	15.45	6.79	-
		<i>NegCLIP</i>	51.63	28.92	14.86	4.59	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	60.80	10.61	8.35	9.02	11.22
		<i>ViT-SO400M-SigLIP</i>	49.47	13.32	3.39	11.97	21.25
		<i>ViT-L-14 (datacomp)</i>	66.43	16.12	6.59	4.99	5.87
		<i>xlrm-roberta-large-ViT-H-14</i>	60.65	21.03	11.90	5.15	1.28
		<i>ViT-L-14 (laion2b)</i>	74.07	9.51	4.48	2.80	9.14
		<i>ViT-L-14 (openai)</i>	71.71	10.59	2.99	2.71	12.00
		<i>ViT-B-32 (openai)</i>	43.86	26.41	15.44	8.57	5.72
		<i>NegCLIP</i>	85.00	10.39	3.12	1.24	0.26
	ComCO	<i>ViT-H-14 (DFN)</i>	61.06	17.00	11.98	6.69	3.27
		<i>ViT-SO400M-SigLIP</i>	55.77	19.25	10.24	6.73	8.01
		<i>ViT-L-14 (datacomp)</i>	68.96	14.61	9.40	4.77	2.25
		<i>xlrm-roberta-large-ViT-H-14</i>	28.86	26.87	19.42	14.61	10.24
		<i>ViT-L-14 (laion2b)</i>	61.93	19.10	11.65	5.11	2.21
		<i>ViT-L-14 (openai)</i>	38.40	24.80	18.79	11.04	6.68
		<i>ViT-B-32 (openai)</i>	44.71	26.69	16.44	8.37	3.79
		<i>NegCLIP</i>	45.70	27.56	17.03	7.57	2.15

7.4. Image-based Object Classification

7.4.1. Objective

The Image-based Object Classification (IOC) experiment was designed to evaluate CLIP’s image encoder’s ability to represent individual objects within multi-object images. This experiment aimed to investigate potential biases in object classification based on the object’s size within the image.

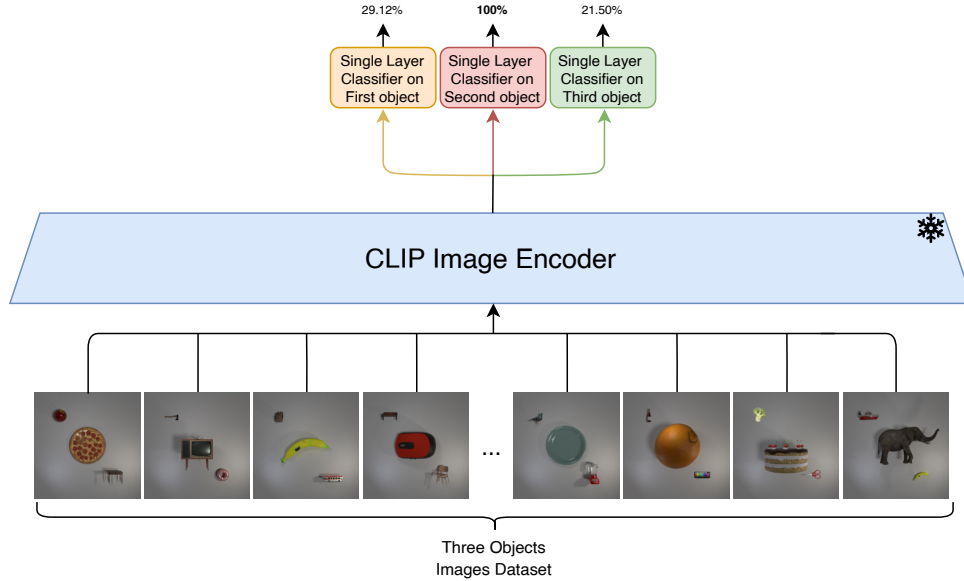


Figure 10. Illustration of the Image-based Object Classification experiment with the ComCO dataset. The diagram shows the process of classifying individual objects in K-object images using CLIP’s image encoder, with a single-layer classifier trained on the generated image embeddings

7.4.2. Methodology

1. Dataset Preparation:

- We utilized both the SimCO and ComCO datasets, containing images with 2 to 5 objects.
- In each image, one object was deliberately made larger than the others.
- The position of the larger object was varied across images to avoid position-based biases.

2. Image Embedding Generation:

- For each multi-object image, we used CLIP’s image encoder to generate an image embedding.
- This embedding is a high-dimensional vector representation of the entire image.

3. Classifier Training:

- We trained separate single-layer classifiers for each object position (large object, small object 1, small object 2, etc.).
- Input: The image embedding of the multi-object image.
- Output: The predicted object class for that specific position/size.

4. Evaluation:

- We tested each classifier on a held-out portion of the dataset.
- For each image, we recorded whether the classifier correctly identified the object at its respective position/size.
- We calculated the classification accuracy for each object position/size across all test images.

We conducted the IOC experiment on images from both datasets, focusing on scenarios with one significantly larger object in varying positions. The experiment was repeated across models, and the average results are shown in Table 9.

Table 9. Image-based Object Classification

Number of Objects	Dataset	Model	Large Object	Small Obj 1	Small Obj 2	Small Obj 3	Small Obj 4
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	88.1	14.29	-	-	-
		<i>ViT-SO400M-SigLIP</i>	97.62	16.67	-	-	-
		<i>ViT-L-14 (datacomp)</i>	83.33	11.9	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	78.57	21.43	-	-	-
		<i>ViT-L-14 (laion2b)</i>	66.67	11.9	-	-	-
		<i>ViT-L-14 (openai)</i>	64.29	0.00	-	-	-
		<i>ViT-B-32 (openai)</i>	61.9	0.00	-	-	-
		<i>NegCLIP</i>	40.48	7.14	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	100.0	26.36	-	-	-
		<i>ViT-SO400M-SigLIP</i>	100.0	33.9	-	-	-
		<i>ViT-L-14 (datacomp)</i>	100.0	42.35	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	40.85	-	-	-
		<i>ViT-L-14 (laion2b)</i>	100.0	31.29	-	-	-
		<i>ViT-L-14 (openai)</i>	99.8	41.29	-	-	-
		<i>ViT-B-32 (openai)</i>	99.8	35.81	-	-	-
		<i>NegCLIP</i>	99.6	41.95	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	100.0	35.65	41.57	-	-
		<i>ViT-SO400M-SigLIP</i>	99.8	42.8	49.03	-	-
		<i>ViT-L-14 (datacomp)</i>	100.0	39.94	51.28	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	99.9	48.42	56.28	-	-
		<i>ViT-L-14 (laion2b)</i>	99.8	45.56	56.08	-	-
		<i>ViT-L-14 (openai)</i>	98.98	39.73	50.46	-	-
		<i>ViT-B-32 (openai)</i>	96.12	38.1	51.58	-	-
		<i>NegCLIP</i>	97.04	42.59	59.35	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	100.0	29.12	21.5	-	-
		<i>ViT-SO400M-SigLIP</i>	100.0	30.94	29.94	-	-
		<i>ViT-L-14 (datacomp)</i>	100.0	36.56	33.5	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	33.69	32.31	-	-
		<i>ViT-L-14 (laion2b)</i>	100.0	35.44	30.31	-	-
		<i>ViT-L-14 (openai)</i>	99.94	33.31	34.31	-	-
		<i>ViT-B-32 (openai)</i>	99.94	29.0	32.94	-	-
		<i>NegCLIP</i>	99.81	33.88	43.0	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	100.0	40.06	34.06	41.31	-
		<i>ViT-SO400M-SigLIP</i>	100.0	47.0	38.5	41.06	-
		<i>ViT-L-14 (datacomp)</i>	100.0	48.94	38.38	45.06	-
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	48.19	35.81	46.38	-
		<i>ViT-L-14 (laion2b)</i>	100.0	50.5	41.81	43.94	-
		<i>ViT-L-14 (openai)</i>	100.0	45.19	38.38	39.0	-
		<i>ViT-B-32 (openai)</i>	100.0	38.06	31.5	37.25	-
		<i>NegCLIP</i>	100.0	42.0	37.25	46.94	-
	ComCO	<i>ViT-H-14 (DFN)</i>	100.0	16.64	14.13	12.38	-
		<i>ViT-SO400M-SigLIP</i>	100.0	18.95	15.57	17.57	-
		<i>ViT-L-14 (datacomp)</i>	100.0	20.64	21.01	19.01	-
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	20.45	18.45	16.51	-
		<i>ViT-L-14 (laion2b)</i>	100.0	19.76	17.57	18.89	-
		<i>ViT-L-14 (openai)</i>	99.94	19.32	21.89	22.39	-
		<i>ViT-B-32 (openai)</i>	100.0	21.58	21.83	22.26	-
		<i>NegCLIP</i>	100.0	21.89	23.64	31.33	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	100.0	34.0	30.0	30.38	21.62
		<i>ViT-SO400M-SigLIP</i>	100.0	38.5	34.7	27.38	25.62
		<i>ViT-L-14 (datacomp)</i>	100.0	40.38	36.12	32.0	24.75
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	41.56	39.56	36.69	32.81
		<i>ViT-L-14 (laion2b)</i>	100.0	43.88	39.5	34.0	28.94
		<i>ViT-L-14 (openai)</i>	100.0	42.19	36.38	32.81	31.94
		<i>ViT-B-32 (openai)</i>	98.81	36.25	35.38	33.88	26.06
		<i>NegCLIP</i>	99.19	40.88	37.94	37.56	28.94
	ComCO	<i>ViT-H-14 (DFN)</i>	100.0	13.88	9.38	9.32	11.94
		<i>ViT-SO400M-SigLIP</i>	100.0	15.51	13.88	14.57	14.76
		<i>ViT-L-14 (datacomp)</i>	100.0	18.2	15.07	16.07	18.32
		<i>xlm-roberta-large-ViT-H-14</i>	99.94	15.38	14.88	15.26	19.14
		<i>ViT-L-14 (laion2b)</i>	100.0	15.51	12.32	14.13	17.95
		<i>ViT-L-14 (openai)</i>	100.0	15.38	14.76	16.76	20.01
		<i>ViT-B-32 (openai)</i>	99.87	17.76	18.64	19.2	23.14
		<i>NegCLIP</i>	100	18.89	16.57	23.51	28.77

7.5. Image-based Object Retrieval

7.5.1. Objective

The Image-based Object Retrieval (IOR) experiment was designed to assess CLIP’s image encoder’s ability to retrieve individual objects from multi-object images. This experiment aimed to investigate potential biases in object retrieval based on the object’s size within the image.

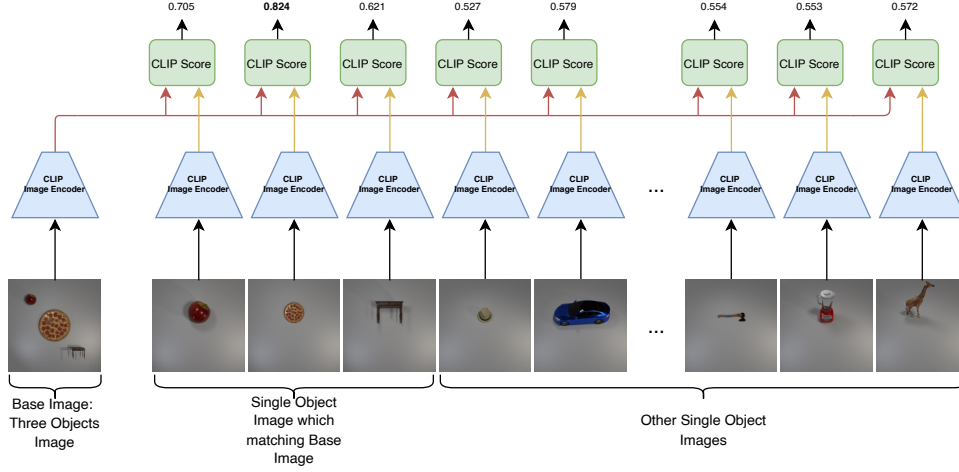


Figure 11. Visualization of the Image-based Object Retrieval experiment. This diagram illustrates the process of retrieving single-object images based on multi-object image inputs using CLIP’s image encoder. The experiment employs a base image containing three objects of varying sizes. CLIP scores are computed between the embedding of this multi-object image and embeddings of various single-object images.

7.5.2. Methodology

1. Dataset Preparation:

- We utilized both the SimCO and ComCO datasets, containing images with 2 to 5 objects.
- In each multi-object image, one object was deliberately made larger than the others.
- The position of the larger object was varied across images to avoid position-based biases.
- We also prepared a set of single-object images for each object class in our datasets.

2. Image Embedding Generation:

- We used CLIP’s image encoder to generate embeddings for all multi-object images.
- Similarly, we generated embeddings for all single-object images.

3. Similarity Computation:

- For each multi-object image, we computed the cosine similarity between its embedding and the embeddings of all single-object images.

4. Object Retrieval:

- For each multi-object image, we identified the single-object image with the highest similarity score.
- We recorded whether the retrieved single-object image corresponded to the large object or one of the small objects in the multi-object image.

5. Evaluation:

- We calculated the percentage of times the large object and each small object were retrieved as the most similar.
- This percentage represents the retrieval accuracy for each object size category (large object, small object 1, small object 2, etc.).

We conducted the IOR experiment on images from the SimCO and ComCO datasets with 2 to 5 objects, varying the position of the larger object to avoid location-based biases. The results are shown in Table 10.

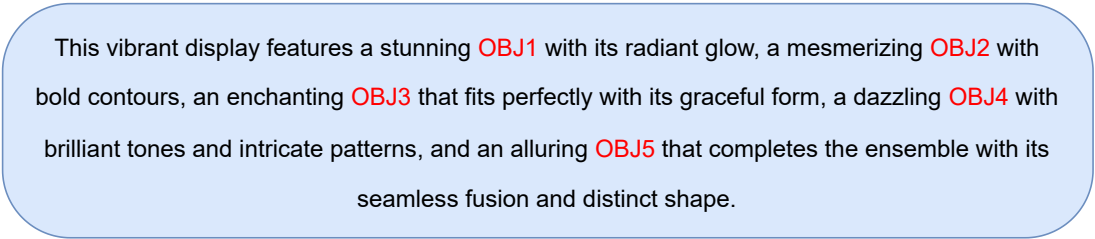
Table 10. Image-based Object Retrieval

Number of Objects	Dataset	Model	Large Object	Small Obj 1	Small Obj 2	Small Obj 3	Small Obj 4
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	99.11	0.89	-	-	-
		<i>ViT-SO400M-SigLIP</i>	91.67	8.33	-	-	-
		<i>ViT-L-14 (datacomp)</i>	91.96	8.04	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	94.92	5.08	-	-	-
		<i>ViT-L-14 (laion2b)</i>	92.86	7.14	-	-	-
		<i>ViT-L-14 (openai)</i>	87.88	12.12	-	-	-
		<i>ViT-B-32 (openai)</i>	90.24	9.76	-	-	-
		<i>NegCLIP</i>	94.64	5.36	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	97.35	2.65	-	-	-
		<i>ViT-SO400M-SigLIP</i>	95.13	4.87	-	-	-
		<i>ViT-L-14 (datacomp)</i>	89.85	10.15	-	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	93.89	6.11	-	-	-
		<i>ViT-L-14 (laion2b)</i>	94.84	5.16	-	-	-
		<i>ViT-L-14 (openai)</i>	83.7	16.30	-	-	-
		<i>ViT-B-32 (openai)</i>	86.86	13.14	-	-	-
		<i>NegCLIP</i>	83.3	16.7	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	93.80	0.65	5.55	-	-
		<i>ViT-SO400M-SigLIP</i>	83.27	5.61	11.12	-	-
		<i>ViT-L-14 (datacomp)</i>	77.16	5.81	17.04	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	80.21	5.12	14.66	-	-
		<i>ViT-L-14 (laion2b)</i>	76.57	9.57	13.86	-	-
		<i>ViT-L-14 (openai)</i>	72.07	8.66	19.27	-	-
		<i>ViT-B-32 (openai)</i>	61.14	14.69	24.17	-	-
		<i>NegCLIP</i>	59.13	14.91	25.96	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	96.52	1.71	17.8	-	-
		<i>ViT-SO400M-SigLIP</i>	90.5	5.47	4.03	-	-
		<i>ViT-L-14 (datacomp)</i>	89.65	6.09	4.26	-	-
		<i>xlm-roberta-large-ViT-H-14</i>	91.39	4.92	3.69	-	-
		<i>ViT-L-14 (laion2b)</i>	91.26	3.28	5.46	-	-
		<i>ViT-L-14 (openai)</i>	74.2	12.79	13.01	-	-
		<i>ViT-B-32 (openai)</i>	80.6	5.22	14.18	-	-
		<i>NegCLIP</i>	76.36	10.47	13.18	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	99.5	0.0	0.0	0.5	-
		<i>ViT-SO400M-SigLIP</i>	91.03	1.28	2.99	4.7	-
		<i>ViT-L-14 (datacomp)</i>	89.71	3.43	3.61	3.25	-
		<i>xlm-roberta-large-ViT-H-14</i>	92.47	2.08	2.60	2.86	-
		<i>ViT-L-14 (laion2b)</i>	86.92	4.67	3.74	4.67	-
		<i>ViT-L-14 (openai)</i>	70.55	13.01	7.53	8.9	-
		<i>ViT-B-32 (openai)</i>	52.17	18.84	13.04	15.94	-
		<i>NegCLIP</i>	74.4	10.4	7.2	8.0	-
	ComCO	<i>ViT-H-14 (DFN)</i>	95.86	2.55	1.27	0.32	-
		<i>ViT-SO400M-SigLIP</i>	94.03	2.24	1.49	2.24	-
		<i>ViT-L-14 (datacomp)</i>	93.3	3.91	1.12	16.8	-
		<i>xlm-roberta-large-ViT-H-14</i>	90.91	2.02	5.05	2.02	-
		<i>ViT-L-14 (laion2b)</i>	91.78	5.48	2.74	0.0	-
		<i>ViT-L-14 (openai)</i>	67.86	14.29	7.14	10.71	-
		<i>ViT-B-32 (openai)</i>	85.0	0.0	5.0	10.0	-
		<i>NegCLIP</i>	79.55	0.0	2.27	18.19	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	100.0	0.0	0.0	0.0	0.0
		<i>ViT-SO400M-SigLIP</i>	94.92	3.39	1.69	0.0	0.0
		<i>ViT-L-14 (datacomp)</i>	91.3	5.59	1.24	1.24	0.62
		<i>xlm-roberta-large-ViT-H-14</i>	77.42	11.83	5.38	3.23	2.15
		<i>ViT-L-14 (laion2b)</i>	81.01	8.86	5.06	1.27	0.38
		<i>ViT-L-14 (openai)</i>	77.14	8.57	5.71	5.71	2.86
		<i>ViT-B-32 (openai)</i>	68.75	25.0	6.25	0.0	0.0
		<i>NegCLIP</i>	58.62	17.24	15.52	5.17	3.45
	ComCO	<i>ViT-H-14 (DFN)</i>	95.16	1.61	1.61	0.0	1.61
		<i>ViT-SO400M-SigLIP</i>	80.0	0.0	0.0	0.0	20.0
		<i>ViT-L-14 (datacomp)</i>	90.91	4.55	0.0	0.0	4.55
		<i>xlm-roberta-large-ViT-H-14</i>	100.0	0.0	0.0	0.0	0.0
		<i>ViT-L-14 (laion2b)</i>	100.0	0.0	0.0	0.0	0.0
		<i>ViT-L-14 (openai)</i>	100.0	0.0	0.0	0.0	0.0
		<i>ViT-B-32 (openai)</i>	100.0	0.0	0.0	0.0	0.
		<i>NegCLIP</i>	50.0	0.0	0.0	50.0	0.0

7.6. Text-based Object Classification for Long Caption

In this section, we revisited the IOC experiment with a significant modification to the caption structure. Our objective was to investigate whether the previously observed bias persists in longer, more elaborate captions. We achieved this by expanding the caption template, incorporating additional descriptive phrases between object mentions.

The extended caption template used in this experiment was as follows:



This vibrant display features a stunning OBJ1 with its radiant glow, a mesmerizing OBJ2 with bold contours, an enchanting OBJ3 that fits perfectly with its graceful form, a dazzling OBJ4 with brilliant tones and intricate patterns, and an alluring OBJ5 that completes the ensemble with its seamless fusion and distinct shape.

Figure 12. Format for Extended Caption Template

This template allowed us to maintain a consistent structure while significantly increasing the caption length and complexity.

The results of this modified IOC experiment are presented in Table 11. Notably, the observed pattern closely resembles that of the standard IOC experiment. This similarity suggests that the bias identified in shorter captions persists even in more elaborate textual descriptions.

7.7. Text-based Object Retrieval for Long Caption

In this section, we aimed to examine the performance of various models in the IOR experiment when presented with longer caption formats. This approach mirrors our previous investigation, allowing us to draw comparisons between standard and extended caption scenarios.

We utilized the same extended caption template as in the previous section. The results of this experiment are presented in Table 12. Notably, the observed pattern closely aligns with that of the standard IOR experiment, suggesting a consistency in model behavior across different caption lengths.

Table 11. Text-based Object Classification on Long Captions

Number of Objects	Dataset	Model	First Object	Second Object	Third Object	Fourth Object	Fifth Object
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	100.0	89.01	-	-	-
		<i>ViT-SO400M-SigLIP</i>	100.0	93.83	-	-	-
		<i>ViT-L-14 (datacomp)</i>	100.0	63.22	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	99.82	51.83	-	-	-
		<i>ViT-L-14 (laion2b)</i>	100.0	85.88	-	-	-
		<i>ViT-L-14 (openai)</i>	99.65	98.26	-	-	-
		<i>ViT-B-32 (openai)</i>	100.0	72.69	-	-	-
		<i>NegCLIP</i>	100	89.59	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	99.99	99.86	-	-	-
		<i>ViT-SO400M-SigLIP</i>	100	99.48	-	-	-
		<i>ViT-L-14 (datacomp)</i>	100	98.89	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	99.95	92.84	-	-	-
		<i>ViT-L-14 (laion2b)</i>	100	99.03	-	-	-
		<i>ViT-L-14 (openai)</i>	99.99	99.99	-	-	-
		<i>ViT-B-32 (openai)</i>	99.59	99.45	-	-	-
		<i>NegCLIP</i>	99.94	98.99	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	99.34	43.49	89.66	-	-
		<i>ViT-SO400M-SigLIP</i>	100.0	65.26	49.76	-	-
		<i>ViT-L-14 (datacomp)</i>	100.0	30.47	37.20	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	97.78	22.96	27.23	-	-
		<i>ViT-L-14 (laion2b)</i>	99.65	57.67	35.51	-	-
		<i>ViT-L-14 (openai)</i>	99.13	86.67	58.22	-	-
		<i>ViT-B-32 (openai)</i>	96.26	54.19	44.88	-	-
		<i>NegCLIP</i>	98.30	67.60	65.90	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	99.31	78.44	84.15	-	-
		<i>ViT-SO400M-SigLIP</i>	99.93	67.22	76.89	-	-
		<i>ViT-L-14 (datacomp)</i>	98.98	85.77	65.64	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	99.21	38.60	60.10	-	-
		<i>ViT-L-14 (laion2b)</i>	98.81	82.72	74.31	-	-
		<i>ViT-L-14 (openai)</i>	99.41	96.44	82.18	-	-
		<i>ViT-B-32 (openai)</i>	95.59	81.91	76.09	-	-
		<i>NegCLIP</i>	98.62	74.29	81.70	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	99.17	24.74	67.00	41.46	-
		<i>ViT-SO400M-SigLIP</i>	100.0	46.75	24.40	20.93	-
		<i>ViT-L-14 (datacomp)</i>	100.0	15.27	17.79	43.03	-
		<i>xlrm-roberta-large-ViT-H-14</i>	98.87	13.34	12.67	15.85	-
		<i>ViT-L-14 (laion2b)</i>	99.56	36.03	19.23	34.51	-
		<i>ViT-L-14 (openai)</i>	98.22	70.29	40.54	50.71	-
		<i>ViT-B-32 (openai)</i>	97.47	41.20	25.18	24.31	-
		<i>NegCLIP</i>	98.93	49.58	35.89	35.40	-
	ComCO	<i>ViT-H-14 (DFN)</i>	98.34	62.49	70.25	42.34	-
		<i>ViT-SO400M-SigLIP</i>	99.90	39.28	58.01	32.51	-
		<i>ViT-L-14 (datacomp)</i>	97.95	71.61	37.24	48.50	-
		<i>xlrm-roberta-large-ViT-H-14</i>	99.34	20.38	21.45	25.08	-
		<i>ViT-L-14 (laion2b)</i>	98.41	66.90	51.43	38.87	-
		<i>ViT-L-14 (openai)</i>	96.39	88.74	62.87	75.1	-
		<i>ViT-B-32 (openai)</i>	96.81	62.50	59.19	22.93	-
		<i>NegCLIP</i>	98.50	45.93	40.11	68.58	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	97.44	18.82	53.68	26.08	47.45
		<i>ViT-SO400M-SigLIP</i>	100.0	20.35	19.30	12.57	18.40
		<i>ViT-L-14 (datacomp)</i>	99.74	17.57	19.29	41.34	23.67
		<i>xlrm-roberta-large-ViT-H-14</i>	99.09	12.51	8.49	8.63	30.25
		<i>ViT-L-14 (laion2b)</i>	99.69	60.13	28.18	49.20	54.92
		<i>ViT-L-14 (openai)</i>	96.26	70.36	44.68	36.7	48.1
		<i>ViT-B-32 (openai)</i>	96.79	30.71	15.25	12.58	41.30
		<i>NegCLIP</i>	99.35	32.26	22.22	16.39	62.63
	ComCO	<i>ViT-H-14 (DFN)</i>	97.45	43.49	29.20	17.91	1.13
		<i>ViT-SO400M-SigLIP</i>	98.46	45.21	32.54	26.64	1.18
		<i>ViT-L-14 (datacomp)</i>	92.76	40.83	17.56	9.8	1.05
		<i>xlrm-roberta-large-ViT-H-14</i>	99.84	13.18	11.02	8.26	45.38
		<i>ViT-L-14 (laion2b)</i>	97.39	41.48	19.5	9.4	1.26
		<i>ViT-L-14 (openai)</i>	92.81	68.46	31.85	9.8	1.24
		<i>ViT-B-32 (openai)</i>	95.85	42.62	22.24	9.18	0.9
		<i>NegCLIP</i>	99.16	27.60	19.78	21.80	69.08

Table 12. Text-based Object Retrieval For long template

Number of Objects	Dataset	Model	Accuracy	First Object	Second Object	Third Object	Fourth Object	Fifth Object
n = 2	SimCO	<i>ViT-H-14 (DFN)</i>	96.73	62.16	37.84	-	-	-
		<i>ViT-SO400M-SigLIP</i>	5.88	100.0	0.00	-	-	-
		<i>ViT-L-14 (datacomp)</i>	98.04	70.67	29.33	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	98.69	76.82	23.18	-	-	-
		<i>ViT-L-14 (laion2b)</i>	51.63	62.03	37.97	-	-	-
		<i>ViT-L-14 (openai)</i>	96.08	39.46	60.54	-	-	-
		<i>ViT-B-32 (openai)</i>	79.74	45.90	54.10	-	-	-
		<i>NegCLIP</i>	99.35	38.82	61.18	-	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	92.38	71.03	28.97	-	-	-
		<i>ViT-SO400M-SigLIP</i>	3.42	100.0	0.00	-	-	-
		<i>ViT-L-14 (datacomp)</i>	84.32	62.63	37.37	-	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	72.06	63.31	36.69	-	-	-
		<i>ViT-L-14 (laion2b)</i>	58.73	63.01	36.99	-	-	-
		<i>ViT-L-14 (openai)</i>	84.64	61.27	38.70	-	-	-
		<i>ViT-B-32 (openai)</i>	78.38	61.77	37.78	-	-	-
		<i>NegCLIP</i>	82.67	55.63	44.37	-	-	-
n = 3	SimCO	<i>ViT-H-14 (DFN)</i>	88.6	43.02	30.43	26.56	-	-
		<i>ViT-SO400M-SigLIP</i>	0.74	100.0	0.00	0.00	-	-
		<i>ViT-L-14 (datacomp)</i>	88.48	63.02	24.38	12.60	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	89.83	61.66	22.10	16.23	-	-
		<i>ViT-L-14 (laion2b)</i>	31.86	56.54	26.15	17.31	-	-
		<i>ViT-L-14 (openai)</i>	69.73	24.08	39.89	36.03	-	-
		<i>ViT-B-32 (openai)</i>	38.24	25.96	39.10	34.94	-	-
		<i>NegCLIP</i>	72.30	23.39	52.71	23.90	-	-
	ComCO	<i>ViT-H-14 (DFN)</i>	76.75	50.43	22.45	27.12	-	-
		<i>ViT-SO400M-SigLIP</i>	0.07	100.0	0.00	0.00	-	-
		<i>ViT-L-14 (datacomp)</i>	56.14	47.80	34.17	18.03	-	-
		<i>xlrm-roberta-large-ViT-H-14</i>	36.78	48.46	28.75	22.79	-	-
		<i>ViT-L-14 (laion2b)</i>	29.17	48.75	35.78	15.47	-	-
		<i>ViT-L-14 (openai)</i>	52.38	43.44	37.00	19.53	-	-
		<i>ViT-B-32 (openai)</i>	49.97	47.58	30.75	21.45	-	-
		<i>NegCLIP</i>	50.80	38.67	38.16	23.17	-	-
n = 4	SimCO	<i>ViT-H-14 (DFN)</i>	66.47	39.82	21.88	24.34	13.96	-
		<i>ViT-SO400M-SigLIP</i>	0.49	100.0	0.00	0.00	0.00	-
		<i>ViT-L-14 (datacomp)</i>	74.58	61.74	22.17	10.96	5.13	-
		<i>xlrm-roberta-large-ViT-H-14</i>	65.95	53.96	21.36	19.33	5.35	-
		<i>ViT-L-14 (laion2b)</i>	22.42	66.76	17.78	11.22	4.23	-
		<i>ViT-L-14 (openai)</i>	58.73	16.30	32.78	26.49	24.37	-
		<i>ViT-B-32 (openai)</i>	18.43	35.64	37.77	14.18	12.41	-
		<i>NegCLIP</i>	50.78	26.25	49.94	16.73	7.08	-
	ComCO	<i>ViT-H-14 (DFN)</i>	52.87	47.87	20.54	22.72	8.87	-
		<i>ViT-SO400M-SigLIP</i>	0.01	100.0	0.00	0.00	0.00	-
		<i>ViT-L-14 (datacomp)</i>	31.36	39.21	30.74	20.94	9.11	-
		<i>xlrm-roberta-large-ViT-H-14</i>	14.99	43.03	24.29	19.72	12.96	-
		<i>ViT-L-14 (laion2b)</i>	10.19	42.66	34.16	17.09	6.09	-
		<i>ViT-L-14 (openai)</i>	28.78	35.25	31.55	19.19	13.86	-
		<i>ViT-B-32 (openai)</i>	21.62	43.69	24.57	16.78	14.59	-
		<i>NegCLIP</i>	19.41	30.36	30.38	24.39	14.86	-
n = 5	SimCO	<i>ViT-H-14 (DFN)</i>	45.44	43.46	20.45	18.34	11.87	5.88
		<i>ViT-SO400M-SigLIP</i>	0.16	100.0	0.00	0.00	0.00	0.00
		<i>ViT-L-14 (datacomp)</i>	51.45	59.26	22.46	8.12	8.46	1.70
		<i>xlrm-roberta-large-ViT-H-14</i>	52.92	54.87	13.81	19.30	8.16	3.86
		<i>ViT-L-14 (laion2b)</i>	12.34	75.40	10.31	8.42	4.26	1.61
		<i>ViT-L-14 (openai)</i>	29.39	8.98	29.39	28.44	15.97	17.20
		<i>ViT-B-32 (openai)</i>	6.69	32.11	38.57	12.22	8.55	8.55
		<i>NegCLIP</i>	17.54	23.15	41.18	24.48	7.65	3.53
	ComCO	<i>ViT-H-14 (DFN)</i>	23.56	36.07	19.21	22.65	11.90	10.17
		<i>ViT-SO400M-SigLIP</i>	0.00	100.0	0.00	0.00	0.00	0.00
		<i>ViT-L-14 (datacomp)</i>	12.49	32.55	27.84	23.76	12.73	3.11
		<i>xlrm-roberta-large-ViT-H-14</i>	9.26	40.26	21.35	18.16	11.99	8.23
		<i>ViT-L-14 (laion2b)</i>	4.57	38.49	31.50	17.50	8.31	4.20
		<i>ViT-L-14 (openai)</i>	1.75	21.59	18.57	20.25	20.54	19.02
		<i>ViT-B-32 (openai)</i>	1.86	32.72	15.62	14.71	18.36	16.26
		<i>NegCLIP</i>	1.41	24.30	23.17	22.14	17.64	12.75

7.8. LAION Dataset Analysis

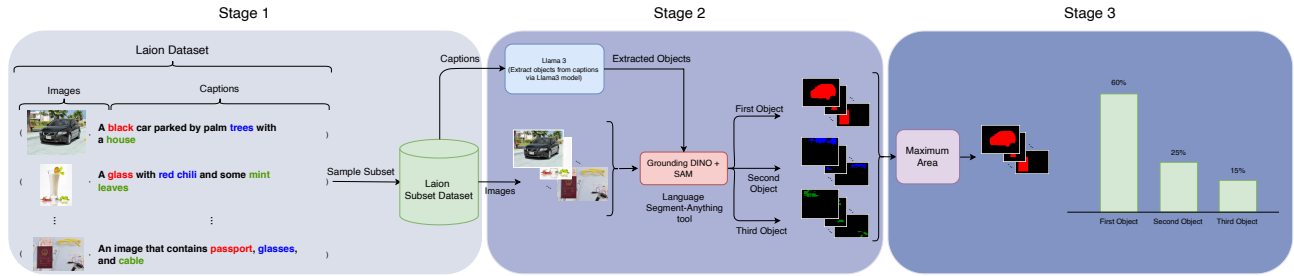


Figure 13. Process flow for LAION dataset analysis

To investigate the potential bias in CLIP’s training data, as discussed in Section 4.3, Claim 2, we conducted an analysis of the LAION dataset. This process, illustrated in Figure 13, consisted of three main stages:

7.8.1. Stage 1: Dataset Sampling

Due to the vast size of the LAION dataset (over 2 billion image-text pairs), we randomly selected a subset of 200,000 samples for our analysis. This subset maintained the diversity of the original dataset while making the analysis computationally feasible.

7.8.2. Stage 2: Object Extraction

For each image-caption pair in our subset:

1. We used the Llama 3 model to extract object mentions from the captions. This step allowed us to identify the objects described in each text without relying on manual annotation.
2. We applied the Grounding DINO + SAM (Segment Anything Model) tool to generate object masks for the corresponding images. This process enabled us to identify and segment individual objects within each image.

7.8.3. Stage 3: Analysis

With the extracted data, we performed the following analysis:

1. **Object Order:** We recorded the order in which objects were mentioned in each caption.
2. **Object Size:** Using the generated masks, we calculated the area of each object in the corresponding image.
3. **Correlation:** We examined the relationship between an object’s position in the caption and its size in the image.

As shown in Figure 14 This distribution strongly suggests a bias in the LAION dataset where larger objects tend to be mentioned earlier in image captions. This finding supports our hypothesis about the origin of CLIP’s text encoder bias, as discussed in Section 4.3 of the main paper.

7.9. COCO Dataset Analysis

In this section, we repeated the experiment conducted in Section 4.3 for different scenarios involving 2 to 5 objects. We divided the captions in the COCO dataset into four subsets: those mentioning 2 objects, 3 objects, 4 objects, and 5 objects. We then analyzed each subset to determine in what percentage of cases the largest object appeared in which position.

The results of this evaluation are presented in Figure 14. As can be observed, this trend is repeated across all scenarios: in most cases, the larger object appears earlier in the caption.

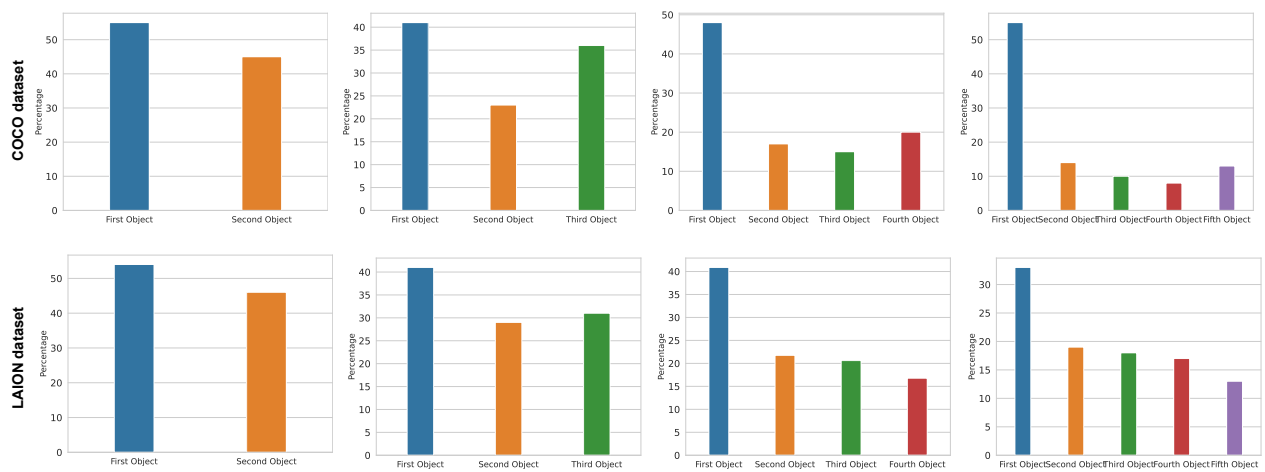


Figure 14. Distribution of larger object positions in captions for objects in COCO and LAION dataset

7.10. Object Categories from DomainNet

The DomainNet dataset objects were categorized into three groups based on their relative sizes: small, medium, and large. These categories were used to investigate potential bias in CLIP’s text embeddings, as discussed in Section 4.3, Claim 1. The full list of objects used in each category is presented below:

7.10.1. Small Objects

ant	anvil	apple	arm
asparagus	axe	banana	bandage
basket	bat	bee	belt
binoculars	bird	blackberry	blueberry
book	boomerang	bottlecap	bowtie
bracelet	brain	bread	broccoli
broom	bucket	butterfly	cactus
cake	calculator	calendar	camera
candle	carrot	cat	clarinet
clock	compass	cookie	crab
backpack	crown	cup	dog
donut	drill	duck	dumbbell
ear	envelope	eraser	eye
eyeglasses	feather	finger	fork
frog	hammer	hat	headphones
hedgehog	helmet	hourglass	jacket
keyboard	key	knife	lantern
laptop	leaf	lipstick	lobster
lollipop	mailbox	marker	megaphone
microphone	microwave	mosquito	mouse
mug	mushroom	necklace	onion
owl	paintbrush	parrot	peanut
pear	peas	pencil	pillow
pineapple	pizza	pliers	popsicle
postcard	potato	purse	rabbit
raccoon	radio	rake	rhinoceros
rifle	sandwich	saw	saxophone
scissors	scorpion	shoe	shovel
skateboard	skull	snail	snake
snorkel	spider	spoon	squirrel
stethoscope	strawberry	swan	sword
syringe	teapot	telephone	toaster
toothbrush	trombone	trumpet	umbrella
violin	watermelon	wheel	

7.10.2. Medium Objects

angel	bathtub	bear	bed
bench	bicycle	camel	cannon
canoe	cello	chair	chandelier
computer	cooler	couch	cow
crocodile	dishwasher	dolphin	door
dresser	drums	flamingo	guitar
horse	kangaroo	ladder	mermaid
motorbike	panda	penguin	piano
pig	sheep	stereo	stove
table	television	tiger	zebra

7.10.3. Large Objects

aircraft carrier	airplane	ambulance	barn
bridge	bulldozer	bus	car
castle	church	cloud	cruise ship
dragon	elephant	firetruck	flying saucer
giraffe	helicopter	hospital	hot air balloon
house	moon	mountain	palm tree
parachute	pickup truck	police car	sailboat
school bus	skyscraper	speedboat	submarine
sun	tent	The Eiffel Tower	Wall of China
tractor	train	tree	truck
van	whale	windmill	

7.11. Text to image generation

The biases observed in CLIP’s encoders have significant implications beyond image-text matching, particularly for text-to-image generation models that incorporate CLIP components. To investigate this impact, we focused on Stable Diffusion, a popular text-to-image generation model that utilizes CLIP’s text encoder in its pipeline. Stable Diffusion employs CLIP’s text encoder to process input prompts, creating text embeddings that guide the image generation process. Given our identification of biases in CLIP’s text encoder, especially the preference for objects mentioned earlier in text descriptions, we hypothesized

that these biases would manifest in the generated images. To test this hypothesis, we designed an experiment using prompts containing multiple objects from the COCO dataset. Our goal was to observe whether the order of objects in the text prompt influences their prominence or likelihood of appearance in the generated images.

Our experimental methodology consisted of three main steps. First, we created 1,000 multi-object prompts, each containing four distinct objects from the COCO dataset. Second, we used these prompts to generate images using three versions of Stable Diffusion: v1.4 [16], v2, and SD-XL [12]. Finally, to evaluate the presence of objects in the generated images, we employed YOLO v8 [15], a state-of-the-art object detection model. We configured YOLO v8 with a detection threshold of 0.25 and used it to validate which objects from the original prompt were present in the generated image.

This approach allowed us to quantitatively assess how CLIP’s text encoder biases propagate through the Stable Diffusion pipeline and manifest in the generated images. By comparing the frequency of object detection with their position in the input prompt, we could directly observe the impact of the text-side bias on the image generation process.

Table 13. Object presence in Stable Diffusion-generated images

Model	First Obj	Second Obj	Third Obj	Fourth Obj
<i>SD v1.4</i>	57.7	44.7	38.1	35.4
<i>SD V2</i>	62.5	49.7	47.5	42.2
<i>SD-XL</i>	79.2	69.3	59.4	64.0

Our findings, presented in Table 13, demonstrate a clear correlation between an object’s position in the text prompt and its likelihood of appearing in the generated image. This correlation aligns with our earlier observations of CLIP’s text encoder bias, suggesting that these biases significantly influence the output of text-to-image generation models.

7.12. Preliminary Method for Bias Mitigation

In our analysis, we observed a critical limitation in the text encoder of CLIP: it disproportionately prioritizes objects mentioned earlier in captions. This bias results in embeddings that heavily represent the first object while progressively diminishing the contribution of subsequent objects. To mitigate this, we explored a novel strategy to reduce positional dependence in object representations.

7.12.1. Proposed Solution

We propose splitting a given caption into multiple sub-captions, each focusing on a single object. By generating embeddings for each sub-caption and aggregating these embeddings, we aim to achieve a balanced representation that minimizes positional bias.

To evaluate this approach, we utilized the ComCO dataset, where objects in captions are separated by the conjunction ‘and’. This structure allowed straightforward decomposition of captions into sub-captions corresponding to individual objects. We conducted the image-text matching experiment (described in Section 5.1) under two conditions: (1) using original captions as-is and (2) using the aggregated embeddings from split captions. Results from this comparison are presented in Table 14.

7.12.2. Results and Observations

As shown in Table 14, the aggregated approach led to a substantial improvement in image-text matching accuracy. This outcome suggests that reducing the influence of positional bias can enhance the text encoder’s performance in multi-object scenarios. Our findings further underscore the potential of designing methods that neutralize word order effects, thereby enabling more robust and unbiased embeddings.

Table 14. Image-Text Matching Accuracy for ComCO Dataset with Original and Split Caption Aggregation Approaches. The first scenario represents results using original captions, while the second scenario reflects the aggregated embeddings of split captions.

Model	Original Captions (%)	Split Caption Aggregation (%)
<i>CLIP Datacomp</i> [6]	67.50	98.39
<i>CLIP Roberta</i>	64.75	97.35
<i>SIGLIP</i> [22]	72.36	99.05
<i>CLIP openAI</i>	52.23	88.56
<i>NegCLIP</i>	46.94	96.82

7.12.3. Limitations and Future Directions

We acknowledge that this solution, while effective for the ComCO dataset, is a heuristic and dataset-specific approach. Its generalizability remains limited. Nonetheless, this experiment demonstrates our commitment to exploring practical solutions and provides a foundation for future advancements.

Future work will focus on developing scalable methods to address positional bias. Possible directions include leveraging large language models (LLMs) to automate caption decomposition into sub-captions and modifying the positional embeddings in the text encoder to ensure equal representation of all objects. These efforts aim to provide a more comprehensive and generalizable solution, paving the way for improved robustness in vision-language models.