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

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade,
Mohammad Hossein Rohban, Mahdieh Soleymani Baghshah
Sharif University of Technology, Tehran, Iran
{reza.abbasi, ali.nazari02, aminreza.sefid, a.banayeean, rohban, soleymani}@sharif.edu

## Abstract

*Contrastive Language-Image Pre-training (CLIP) models excel in zero-shot classification, yet face challenges in complex multi-object scenarios. This study offers a comprehensive analysis of CLIP's limitations in these contexts using a specialized dataset, ComCO, designed to evaluate CLIP's encoders in diverse multi-object scenarios. Our findings reveal significant biases: the text encoder prioritizes first-mentioned objects, and the image encoder favors larger objects. Through retrieval and classification tasks, we quantify these biases across multiple CLIP variants and trace their origins to CLIP's training process, supported by analyses of the LAION dataset and training progression. Our image-text matching experiments show substantial performance drops when object size or token order changes, underscoring CLIP's instability with rephrased but semantically similar captions. Extending this to longer captions and text-to-image models like Stable Diffusion, we demonstrate how prompt order influences object prominence in generated images. For more details and access to our dataset and analysis code, visit our project repository: https://clip-oscope.github.io/.*

## 1. Introduction

The convergence of vision and language in artificial intelligence has led to the development of Vision-Language Models (VLMs) that can interpret and generate multimodal content. Among these, OpenAI's Contrastive Language-Image Pre-training (CLIP) model [13] has been particularly influential, demonstrating remarkable capabilities in zero-shot image classification and setting new standards for multimodal understanding [3, 5, 18, 20]. The success of CLIP has catalyzed a wide array of applications—from image retrieval and visual question answering to text-to-image generation—signifying a paradigm shift in how models perceive and relate visual and linguistic information.

Visual Language Models like CLIP face significant challenges in understanding and reasoning about complex scenes with multiple objects and intricate relationships. CLIP struggles to identify distinct objects and model their relationships accurately, especially when captions contain the same objects but differ in their relationships. This results in difficulty distinguishing between similar captions with different object relationships. Several benchmark datasets have been introduced to elucidate the limitations of existing models in capturing subtle relational nuances. Notably, Winoground [20], VL-CheckList [23], ARO [21], and CREPE [10] have been instrumental in evaluating models' capacities to accurately match images with semantically appropriate captions.

Numerous studies have addressed compositionality challenges in multi-object scenarios, often through end-to-end methods like fine-tuning with hard-negative samples [21] to improve model performance. However, these approaches have faced criticism and subsequent refinement, as seen in methods like SUGARCREPE [8] and [17], which generate negative captions with minor structural changes or LLMs to highlight semantic distinctions. While most focus on CLIP's ability to distinguish structurally similar yet conceptually different captions, few studies, such as Dumpala et al. [4], explore CLIP's performance on semantically equivalent but structurally distinct captions, revealing a gap in understanding CLIP's inconsistency with such prompts.

While previous studies have advanced our understanding of CLIP's limitations, our work uniquely focuses on CLIP's performance with semantically equivalent but structurally varied captions rather than simply distinguishing conceptually different captions. This shift enables a deeper examination of the model's grasp of language and visual content, where systematic errors reveal potential biases. Unlike prior works that primarily propose benchmarks or end-to-end solutions, we investigate the root causes of CLIP's behavior, delving into the mechanisms of both image and text encoders to uncover why the model displays biases and lacks robustness to certain linguistic and visual varia-

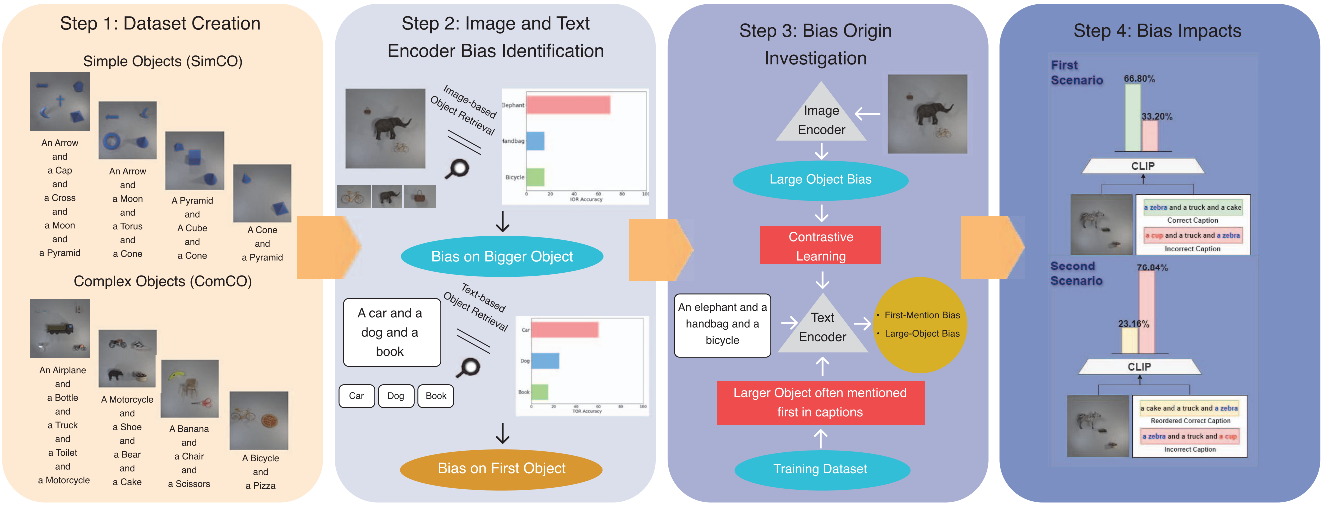

Step 1: Dataset Creation
Simple Objects (SimCO)
An Arrow and a Cap and a Cross and a Moon and a Pyramid
An Arrow and a Moon and a Torus and a Cone
A Pyramid and A Cube and a Cone
A Cone and a Pyramid
Complex Objects (ComCO)
An Airplane and a Bottle and a Truck and a Toilet and a Motorcycle
A Motorcycle and a Shoe and a Bear and a Cake
A Banana and a Chair and a Scissors
A Bicycle and a Pizza
Step 2: Image and Text Encoder Bias Identification
Image-based Object Retrieval
Elephant
Handbag
Bicycle
0
20
40
60
80
100
IOR Accuracy
Bias on Bigger Object
Text-based Object Retrieval
A car and a dog and a book
Car
Dog
Book
TOR Accuracy
Bias on First Object
Step 3: Bias Origin Investigation
Image Encoder
Large Object Bias
Contrastive Learning
An elephant and a handbag and a bicycle
Text Encoder
First-Mention Bias
Large-Object Bias
Larger Object often mentioned first in captions
Training Dataset
Step 4: Bias Impacts
First Scenario
66.80%
33.20%
CLIP
a zebra and a truck and a cake
Correct Caption
a cup and a truck and a zebra
Incorrect Caption
Second Scenario
76.84%
23.16%
CLIP
a cake and a truck and a zebra
Reordered Correct Caption
a zebra and a truck and a cup
Incorrect Caption

Figure 1. Overview of our key contributions. Step 1: We create ComCO dataset for controlled multi-object experiments. Step 2: We identify biases in CLIP's image encoder (favoring larger objects) and text encoder (prioritizing first-mentioned objects). Step 3: We investigate the origin of these biases, finding a connection to training data characteristics. Step 4: We demonstrate the practical impacts of these biases on image-text matching task, showing how they affect model performance in multi-object scenarios.

tions. To support this analysis, we introduce the **ComCO** dataset, purpose-built for examining CLIP's performance under *controlled* multi-object scenarios. Our study spans multiple versions of CLIP trained on diverse datasets and architectures, ensuring the broad applicability of our findings. This comprehensive approach aims to deepen our understanding of CLIP's limitations and pave the way for more adaptable vision-language models. Beyond CLIP, our insights have significant implications for text-to-image (T2I) generative models and multimodal large language models (MLLMs), where decoding CLIP's encoding intricacies can inform advancements in artificial intelligence across domains. As shown in Figure 1, our key contributions are as follows:

- **Development of Novel Dataset**: We introduce *ComCO*, a specialized dataset for creating *controlled* multi-object scenarios. Unlike previous benchmarks, ComCO allows control over object size and caption order, enabling precise analysis of model performance across compositional challenges and enhancing understanding of VLMs' strengths and weaknesses.
- **Encoder Analysis**: We conduct an in-depth examination of CLIP's image and text encoders in multi-object scenes, revealing weaknesses in preserving information for object distinction and identifying where compositional information is lost.
- **Bias Identification**: Our study reveals that CLIP's image encoder prefers larger objects, while the text encoder favors first-mentioned and visually larger objects, highlighting biases in CLIP's handling of visual and linguistic information.
- **Investigation of Bias Origins**: We explore the origins of these biases, showing that larger objects are often mentioned earlier in CLIP's training captions, and are favored in embeddings due to the abundance of their visual tokens. We substantiate this with analyses of the LAION dataset and CLIP's training progression.
- **Practical Impact**: We show how these biases affect performance in multi-object tasks, with significant drops in image-text matching accuracy in ComCO and COCO [9]. These biases also extend to text-to-image models, influencing object prominence based on prompt order.

These findings reveal how biases in CLIP's text and image encoders significantly reduce its performance in multi-object scenarios, emphasizing the need to address these biases to enhance vision-language models' robustness. Our work offers key insights into CLIP's behavior and lays groundwork for improving model performance in real-world applications.

## 2. Methodology

### 2.1. Dataset Design

To thoroughly evaluate the performance of CLIP models in multi-object scenarios under controlled conditions, we constructed the **ComCO** (Complex COCO Objects) dataset. Utilizing Blender software allowed us precise control over the number, location, and dimensions of objects in the images (see Appendix 7.1). The **ComCO** dataset comprises 72 objects derived from the COCO dataset. We generated images containing 2, 3, 4, and 5 objects. Each image is paired with a specific caption that accurately describes the objects present. This approach ensures high control over the dataset and minimizes confounding factors, providing a robust platform for evaluating the CLIP models.

We deliberately chose not to use text-to-image models for generating these datasets due to two main reasons. First, these models often lack the capability to produce high-quality, fully controlled multi-object images. Second, since CLIP is used in many of these models, utilizing them could introduce unwanted biases into our evaluations.

### 2.2. Experimental Framework for Encoder Analysis

The main goal of this study is to evaluate the performance of CLIP's text and image encoders separately in multi-object scenarios. We aim to analyze the impact and contribution of each object in the final output of the encoders. To achieve this, we conducted experiments using our designed ComCO dataset, with images and captions containing two to five objects. To ensure the generalizability of our findings, we also validated our results on the widely-used COCO dataset. We designed two sets of experiments: retrieval-based experiments and classification-based experiments. Given the consistency of the results in both types of experiments, we have included the classification results in the appendix 7.2 and 7.4 and explain the retrieval-based experiments bellow.

#### 2.2.1. TEXT-BASED OBJECT RETRIEVAL (TOR)

The Text-based Object Retrieval task evaluates how well CLIP's text encoder can identify individual objects within multi-object captions. As illustrated in Figure 2a, this experiment involves several steps: First, we use CLIP's text encoder to create embeddings for both multi-object captions and single-object captions. We then measure the similarity between each multi-object caption embedding and all single-object caption embeddings. The single-object caption with the highest similarity score is considered the ”retrieved” object. To assess performance, we calculate retrieval accuracy for each object position in the multi-object captions. This helps us identify any biases related to an object's position within a caption, such as favoring objects mentioned first or last.

#### 2.2.2. IMAGE-BASED OBJECT RETRIEVAL (IOR)

The Image-based Object Retrieval task is similar to TOR but focuses on CLIP's image encoder. As shown in Figure 2b, this experiment involves several steps: We begin by using CLIP's image encoder to generate embeddings for multi-object images and single-object images. We then compute similarity scores between each multi-object image embedding and all single-object image embeddings. The single-object image with the highest similarity score is considered the ”retrieved” object. To evaluate performance, we calculate retrieval accuracy for different object size categories (e.g., large, small) within the multi-object images. This allows us to determine if the image encoder shows any preference for objects of a particular size.

We also experimented with a variation of ComCO, called SimCO, where objects were replaced with simple geometric shapes from the CLEVR dataset. This was done to confirm that bias persists even with non-natural, geometric objects. Further details are provided in Appendix 7.1.

## 3. Results and Analysis

Our experiments revealed significant biases in both the text and image encoders of the CLIP model. This section presents our findings, organized by encoder type and focusing on retrieval tasks.

### 3.1. Text Encoder Biases

We observed a consistent bias in the text encoder towards the first object mentioned in descriptions. In the TOR experiment, the retrieval accuracy (as shown in Table 1) was highest for the first object, indicating its dominant influence on the overall text representation. This suggests that the text encoder prioritizes the initial object, leading to its more accurate retrieval compared to subsequent objects. The detailed results for the scenarios involving 2, 3, and 5 objects can be found in the appendix 7.3, and experiments on longer caption templates are in Appendix 7.6 and 7.7.

### 3.2. Image Encoder Biases

In multi-object images, the image encoder exhibited a strong bias towards larger objects. The Image-based Object Retrieval IOR experiment, detailed in Table 2, shows that larger objects were more frequently and accurately retrieved during single-object image searches. This finding highlights the image encoder's bias towards larger objects, which receive disproportionate emphasis in the final image representation. Further detailed results, specifically for scenarios with 2, 3, and 5 objects, are provided in the appendix 7.5.

### 3.3. COCO Dataset Experiments

To validate the generalizability of our findings from the synthetic dataset, we conducted similar experiments on the COCO dataset, which comprises real images with accompanying captions. This real-world dataset allowed us to investigate whether the previously observed biases persist in more naturalistic settings.

Due to the absence of single-object images for COCO objects, we approached the IOR experiment in two ways. First, we used single-object images from the DomainNet dataset [11] as retrieval targets. Second, we introduced an alternative approach called Image-to-Text Object Retrieval (I2TOR). In I2TOR, we used the textual names of COCO

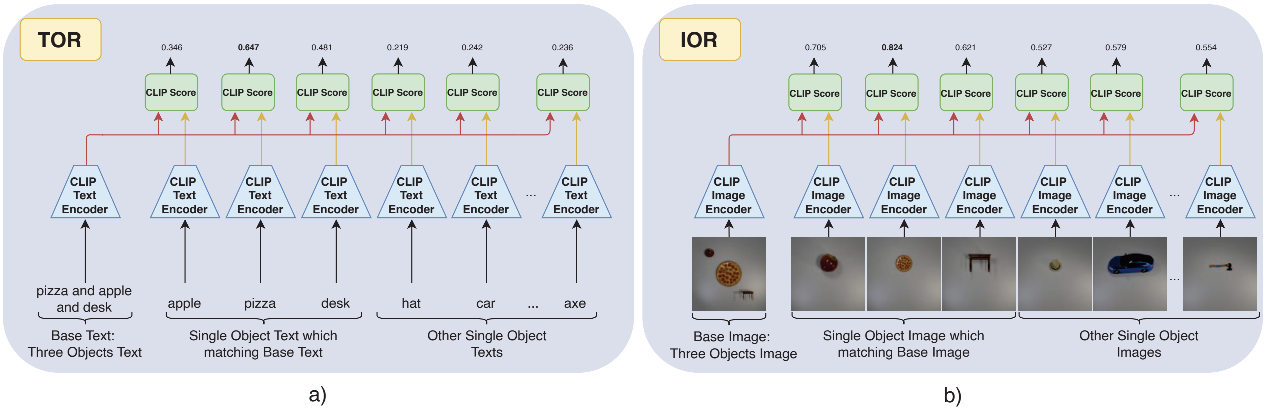

TOR
0.346
0.647
0.481
0.219
0.242
0.236
CLIP Score
CLIP Text Encoder
...
pizza and apple and desk
apple
pizza
desk
hat
car
...
axe
Base Text: Three Objects Text
Single Object Text which matching Base Text
Other Single Object Texts
a)
IOR
0.705
0.824
0.621
0.527
0.579
0.554
CLIP Score
CLIP Image Encoder
...
Base Image: Three Objects Image
Single Object Image which matching Base Image
Other Single Object Images
b)

Figure 2. Experimental setup for Text-based Object Retrieval (TOR) and Image-based Object Retrieval (IOR) tasks. a) TOR: The CLIP text encoder generates embeddings for multi-object and single-object texts. Cosine similarity scores are calculated between the base text embedding and single-object text embeddings to identify the most similar object. b) IOR: The CLIP image encoder generates embeddings for multi-object and single-object images. Cosine similarity scores are calculated between the base image embedding and single-object image embeddings to identify the most similar object.

Table 1. Performance on TOR for ComCO datasets

| Task | Model | First Obj | Second Obj | Third Obj | Fourth Obj |
|---|---|---|---|---|---|
| TOR | *CLIP LAION* | **63.96** | 21.59 | 10.68 | 3.76 |
| | *CLIP Datacomp* | **71.13** | 16.26 | 8.74 | 3.87 |
| | *CLIP Roberta* | **44.03** | 23.73 | 18.07 | 14.18 |
| | *SIGLIP* | **58.11** | 21.16 | 10.99 | 9.73 |
| | *CLIP openAI* | **50.31** | 20.74 | 14.45 | 6.79 |
| | *NegCLIP* | **51.63** | 28.92 | 14.86 | 4.59 |
| | *SugarCrepe* | **44.29** | 30.32 | 18.73 | 6.66 |

Table 2. Performance on IOR for ComCO datasets

| Task | Model | Large Object | Small Obj 1 | Small Obj 2 | Small Obj 3 |
|---|---|---|---|---|---|
| IOR | *CLIP LAION* | **85.45** | 6.36 | 5.45 | 2.73 |
| | *CLIP Datacomp* | **85.16** | 5.65 | 4.95 | 4.24 |
| | *CLIP Roberta* | **87.40** | 8.66 | 2.36 | 1.57 |
| | *SIGLIP* | **77.66** | 10.11 | 6.38 | 5.85 |
| | *CLIP openAI* | **65.22** | 17.39 | 8.70 | 8.70 |
| | *NegCLIP* | **61.67** | 15.00 | 13.33 | 10.00 |
| | *SugarCrepe* | **60.0** | 18.38 | 16.85 | 4.7 |

Table 3. Performance on TOR for coco dataset

| Task | Model | First Obj | Second Obj | Third Obj | Fourth Obj |
|---|---|---|---|---|---|
| TOR | *CLIP openAI* | **35.24** | 21.90 | 20.48 | 22.38 |
| | *CLIP LAION* | **67.89** | 13.76 | 8.26 | 10.09 |
| | *CLIP Datacomp* | **57.68** | 17.68 | 12.75 | 11.88 |
| | *CLIP Roberta* | **40.78** | 23.30 | 20.39 | 15.53 |
| | *SIGLIP* | **49.47** | 26.84 | 12.11 | 11.58 |
| | *NegCLIP* | **38.69** | 22.11 | 17.09 | 22.11 |

Table 4. Performance on IOR for coco dataset

| Task | Model | Large Object | Small Obj 1 | Small Obj 2 | Small Obj 3 |
|---|---|---|---|---|---|
| IOR | *CLIP openAI* | **43.02** | 28.82 | 17.13 | 11.03 |
| | *CLIP LAION* | **39.44** | 28.45 | 17.70 | 14.41 |
| | *CLIP Datacomp* | **36.71** | 29.55 | 19.13 | 14.61 |
| | *CLIP Roberta* | **36.71** | 28.61 | 19.82 | 14.86 |
| | *SIGLIP* | **36.63** | 28.29 | 20.02 | 15.06 |
| | *NegCLIP* | **44.04** | 28.86 | 16.48 | 10.62 |
| I2TOR | *CLIP openAI* | **51.49** | 24.87 | 13.68 | 9.97 |
| | *CLIP LAION* | **45.50** | 27.02 | 15.91 | 11.56 |
| | *CLIP Datacomp* | **46.64** | 26.82 | 14.53 | 12.01 |
| | *CLIP Roberta* | **44.69** | 26.98 | 16.04 | 12.29 |
| | *SIGLIP* | **47.09** | 27.07 | 15.10 | 10.74 |
| | *NegCLIP* | **49.04** | 27.07 | 14.08 | 9.81 |

objects instead of single-object images. These object names were embedded using CLIP's text encoder, allowing us to perform a retrieval task consistent with the IOR methodology while adapting to the constraints of the COCO dataset.

Tables 3 and 4 present the results of our COCO dataset experiments. In TOR, the first-mentioned object in COCO captions was retrieved with higher accuracy, which aligns with our earlier findings of bias in the text encoder. Similarly, in IOR, larger objects in COCO images were retrieved more accurately, consistent with the trends observed in our synthetic dataset experiments. The I2TOR results further confirmed this bias, demonstrating that even when using textual object representations, the bias towards larger objects persists.

Our experiments reveal two significant biases in the CLIP model: the text encoder shows a strong preference for the first mentioned object in textual descriptions, while the image encoder exhibits greater sensitivity to larger objects in images. These biases can significantly impact the overall system performance in various vision-language tasks, particularly in multi-object scenarios.

# 4. Origin of Bias in CLIP Models

In this section, we investigate the potential origins of the biases observed in CLIP models and provide evidence supporting our hypotheses.

## 4.1. Bias in the Image Encoder

The observed bias favoring larger objects within the image domain can be attributed to the architectural characteristics of Vision Transformers (ViT) [2] utilized in CLIP's image encoder. Our hypothesis is that larger objects, which occupy a greater number of patches in the ViT's patch-based image representation, exert a more significant influence on the final class (CLS) token representation. This bias is not exclusive to CLIP; it appears to be a consistent feature across ViT models, as demonstrated by our experiments detailed in the appendix.

To substantiate this hypothesis, we designed an experiment to quantify the attention allocated by the CLS token to each image patch. By calculating the cumulative attention received by each object from the CLS token, we could assess the influence of object size on attention allocation. We applied this analysis to our three-object ComCO dataset, and the results are illustrated in Figure 3. The findings confirm our hypothesis: larger objects indeed receive more attention from the CLS token.

## 4.2. Bias in the Text Encoder

We explore the bias present in the text encoder from two perspectives: the attention mechanism in the model structure and the model's training method.

### 4.2.1. Impact of Attention Mechanism

Text encoder models can be categorized based on their attention mechanisms: uni-directional (causal) attention and bi-directional attention. In models with causal attention, each token attends only to preceding tokens, whereas in bi-directional models, each token attends to all tokens in the sequence.

When OpenAI introduced the CLIP model, its text encoder employed causal attention, meaning each token could only attend to tokens before it and itself. This differs from typical self-attention mechanisms, where tokens attend to all other tokens. Most CLIP models use causal self-attention, with the exception of the variant using the XLM-Roberta text encoder, which also employs self-attention. However, as shown in Table 1, even this model exhibits the mentioned bias. This indicates that the bias does not originate from the attention mechanism itself.

### 4.2.2. Role of Training Method

To determine whether the observed bias is specific to CLIP models, we compared CLIP's text encoder with two other models designed to embed sentences into a meaningful semantic space: Sentence-BERT (SBERT) [14] and SimCSE [7]. The primary distinction is that CLIP's embedding space is shared between images and text, whereas SBERT and SimCSE operate solely in the text domain.

Table 5. Performance on TOC and TOR for ComCO datasets

| Task | Model | **First Obj** | **Second Obj** | **Third Obj** | **Fourth Obj** |
|---|---|---|---|---|---|
| TOR | *CLIP* | **56.28** | 22.71 | 13.17 | 7.48 |
| | *SBERT* | 29.02 | 19.80 | 17.50 | **33.57** |
| | *SimCSE* [7] | 27.59 | 19.07 | 17.76 | **34.83** |

We conducted the TOR experiment on our dataset using these models. As presented in Table 5, the bias observed in CLIP differs from that in the other models. This suggests that CLIP's unique training method, which aligns images and text in a shared embedding space through contrastive learning, contributes to the bias. Therefore, to uncover the root cause of the bias, we focus on the specifics of CLIP's training procedure.

## 4.3. Hypothesized Origin of Text-Side Bias in CLIP

We hypothesize that the text-side bias in CLIP, which favors objects mentioned earlier in text descriptions, originates from the image-side bias toward larger objects and is transferred to the text encoder during contrastive training. We present evidence supporting this hypothesis through two key claims and an analysis of the training progression.

**Claim 1: Larger Objects Have More Influence on Text Embeddings.** Building upon the established image-side bias discussed earlier, we posit that objects with larger physical sizes exert more influence on CLIP's text embeddings due to the alignment enforced during contrastive training. To test this, we categorized objects in the DomainNet dataset into large, medium, and small groups based on their relative physical sizes in real-world (with the full list of objects provided in the appendix 7.10). Specifically, objects smaller than a school bag were categorized as small, objects sized between a school bag and a medium-sized car were classified as medium, and objects larger than a car—up to significantly larger items—were considered large. We then constructed two sets of sentences, each containing four objects: one set with a large object mentioned first followed by three medium-sized objects, and another with a small object mentioned first followed by three medium-sized objects.

Figure 4.a compares the TOR accuracy for the first object in these two groups. The higher TOR accuracy for sentences beginning with large objects supports our hypothesis that larger objects, when mentioned first, have a more significant impact on the text embeddings due to the cross-modal alignment with their prominent representation in images.

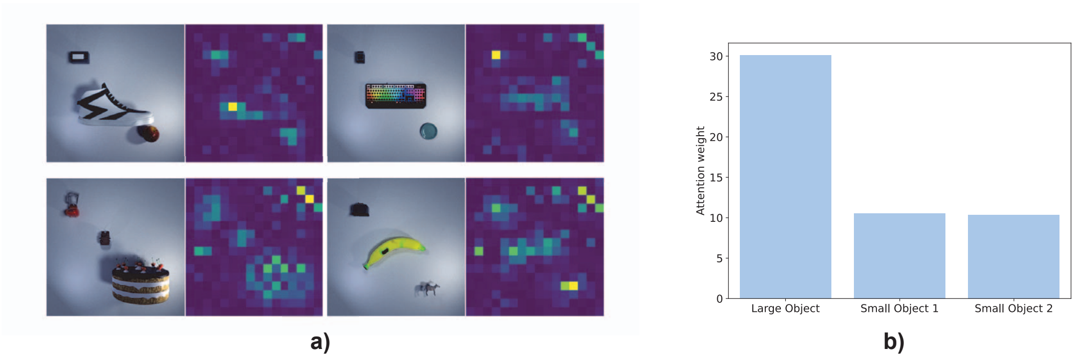


Figure 3. Attention allocation from the CLS token to objects of different sizes in the ComCO dataset. a) Qualitative results showing the CLS token's attention to each object. b) Quantitative analysis of attention distribution across 8,000 images, with each image containing one large and two small objects. The bar chart shows the average attention allocated to the large object versus the smaller ones, demonstrating a bias towards larger objects.

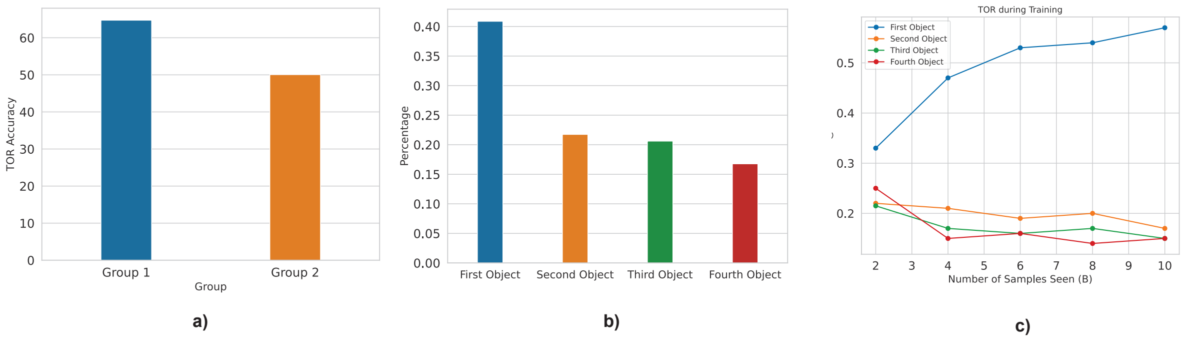


Figure 4. a) Top-1 Object Retrieval accuracy comparison for sentences where the first object is either large or small. The higher TOR accuracy for sentences beginning with large objects supports the hypothesis that larger objects, when mentioned first, exert a stronger influence on text embeddings due to cross-modal alignment with their prominent visual representation in images. b) Distribution of the position of the largest object within image captions from the LAION datasets. The results show a consistent bias where larger objects tend to be mentioned earlier in text descriptions. c) Progression of TOR rates across different training stages, indicating that text-side bias strengthens as the model is exposed to more data, suggesting the cumulative effect of image-side bias being transferred to the text encoder through contrastive learning.

**Claim 2: Caption Bias in Training Datasets.** To investigate potential biases in CLIP's training data, we analyzed both the LAION [19] and COCO datasets. Due to limited computational resources and the large size of the LAION dataset, which contains over 2 billion image-text pairs, we randomly selected a subset of 200,000 samples for our analysis. Using the Llama3 model, we extracted objects from the image captions and employed the Language Segment-Anything tool to generate object masks in the corresponding images, calculating their areas based on these masks. A detailed description of our LAION dataset analysis methodology can be found in Appendix 7.8.

Figure4.b shows the position of the largest object within each caption. The results indicate that, in the majority of cases, the largest object in an image is mentioned earlier in its caption. The same experiment was conducted on the COCO dataset, with detailed results and the distribution for two to five object scenarios provided in Appendix 7.9. This demonstrates a consistent bias in the training data, where larger objects are not only more visually prominent but are also described earlier in text annotations.

**Analysis of Bias Development During Training.** To further validate our hypothesis, we examined the progression of text-side bias during CLIP's training. We utilized model checkpoints from the LAION dataset at five training stages, corresponding to exposure to 2, 4, 6, 8, and 10 billion samples. We conducted TOR experiments at each stage, focusing on the retrieval accuracy for the first object mentioned in text descriptions.

Figure4.c depicts the evolution of the TOR rate across different training stages for scenarios with varying numbers of objects (from 3 to 8). The consistent upward trend in the TOR rate as the model is exposed to more training data suggests that the text-side bias strengthens over time, likely due to the cumulative effect of the image-side bias being transferred to the text encoder through contrastive learning.

**Incomplete Text Representation of CLIP** Here we want to theoretically highlight why the CLIP text encoder could learn an incomplete representation of the text. Let $\mathbf{z}$ and $\mathbf{w}$ represent a latent representation of an image content described in the caption, and such visual content not mentioned in the text, respectively. For example, $\mathbf{z}$ represents the fact that an image contains "a horse that is eating the grass." In this case, $\mathbf{w}$ might represent other details in the image, like the "horse color," "where the horse is located," etc. We assume a data generative process as follows:

$$\begin{aligned} I &:= g(\mathbf{z},\mathbf{w}) \\ T &:= h(\mathbf{z}), \end{aligned}$$

where $I$ is the image, and $T$ is its corresponding caption.

Now we want to learn a joint embedding of the image and text through the CLIP. Here, we assume that $f_\theta(.)$ and $i_\omega(.)$ as learnable functions that map the image and text into the joint embedding space, respectively.

**Theorem 1** *Let elements of $\mathbf{z}$ be independent, zero-mean, and unit-variance. The contrastive loss for the ideal text encoder, $i_\omega(T) = \mathbf{z}$ converges to that of a non-ideal incomplete one, i.e. $i_{\omega'}(T) = \mathbf{z}_s$, where $\mathbf{z}_s$ is the first $d-k$ dimensions of $\mathbf{z}$, with $k$ being a constant, and $d \to \infty$.*

Proof: The contrastive loss in making this learning happen can be written as:

$$\mathbb{E}_{\mathbf{z},\mathbf{z}',\mathbf{w}}\left\{\frac{\exp(sim(\mathbf{z},\mathbf{z}))}{\exp(sim(\mathbf{z},\mathbf{z}))+\sum_k \exp(sim(\mathbf{z},\mathbf{z}'_k))}\right\} \tag{1}$$

with

$$sim(\mathbf{z},\mathbf{z}') = S(f_\theta(g(\mathbf{z},\mathbf{w}), i_\omega(h(\mathbf{z}')))),$$

and $\mathbf{z}$ and $\{\mathbf{z}'_k | 1 \leq k \leq b\}$ are $b+1$ i.i.d. samples of the content in the representation space, and $S$ is some normalized similarity metric, e.g. cosine similarity, and $b+1$ is the batch size. We assume that elements of $\mathbf{z}$ are independent, unit-variance, and zero mean. We further assume that the dimensionality of $\mathbf{z}$, denoted as $d$, goes to infinity.

Under such conditions, and based on Law of Large Numbers, $\|\mathbf{z}\| \xrightarrow{p} \sqrt{d}$, when $d$ is large. Therefore, for any two independent copies of $\mathbf{z}$, $\mathbf{z}'_k$, we have $sim(\mathbf{z},\mathbf{z}'_k) = \mathbf{z}^\top \mathbf{z}'_k / (\|\mathbf{z}\| \|\mathbf{z}'_k\|) \xrightarrow{p} 0$.

It is evident that in the ideal case, $f_\theta(g(\mathbf{z},\mathbf{w})) = \mathbf{z}$ and also $i_\omega(h(\mathbf{z})) = \mathbf{z}$, so the contrastive loss would converge to $e/(e+b)$, as the numerator is $e$, and the second term in the denominator converges to $\exp(0) = 1$, according to the Mann-Wald's theorem.

However, we show that other learning of this representation could achieve the same amount of loss. For instance, let $\mathbf{z}_s$ be the first $d-k$ elements of $\mathbf{z}$, with $k$ being a *constant*. We show that if $f_{\theta'}(I) = \mathbf{z}_s$ and $i_{\omega'}(T) = \mathbf{z}_s$, the same loss would be achieved in the limit of large $d$. To see this, note that the numerator stays the same, i.e. $e$, while the second term in the denominator still converges to $b\exp(0) = b$.

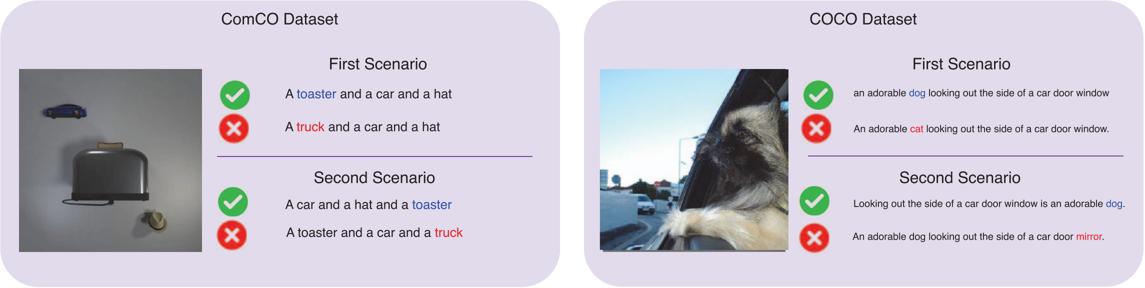
This means that even if the image and text encoder of the CLIP only partially recover the content embedding, they reach an excellent loss. But such possible incomplete representations of $\mathbf{z}$ are combinatorially large, making convergence of the CLIP to such local minima pretty likely. This makes the text encoding of CLIP be far from ideal. Furthermore, the text encoder would become *biased*, depending on which of such local minima it converges to. Based on this explanation, we would expect a text encoder that has learned a complete representation to exhibit such biases to a lesser degree. As mentioned earlier, the subject of learning text representations in VLMs that are discriminative of hard negatives (e.g. NegCLIP) has been around for few years. We tested one of strongest such models, [8], in our benchmark to validate the hypothesis that an incomplete text representation is one of the causes of the bias in the VLMs. We noticed that this model shows lower bias based on our benchmark (see the SugarCrepe model in tables 1 and 2).

We have developed an initial approach to address the identified bias in the CLIP model, which is presented in Appendix 7.12. While this method is specific to our current dataset, it represents a promising step toward addressing these challenges and can inspire further advancements. This work demonstrates our commitment to exploring practical solutions while maintaining the primary focus of this study on the analysis of bias and its implications.

## 5. Practical Impacts of Encoder Biases

The biases observed in CLIP's image and text encoders significantly impact model performance in real-world applications. This section explores how these biases manifest in image-text matching tasks, while further analyses of text-to-image generation impacts are presented in Appendix 7.11.

Our analysis in this section serves two primary purposes. First, it provides concrete evidence of how these theoretical biases can translate into practical limitations. Second, it offers insights into potential areas for improvement in vision-language models, particularly in handling complex, multi-



ComCO Dataset, First Scenario:

✓ A toaster and a car and a hat

✗ A truck and a car and a hat

ComCO Dataset, Second Scenario:

✓ A car and a hat and a toaster

✗ A toaster and a car and a truck

COCO Dataset, First Scenario:

✓ an adorable dog looking out the side of a car door window

✗ An adorable cat looking out the side of a car door window.

COCO Dataset, Second Scenario:

✓ Looking out the side of a car door window is an adorable dog.

✗ An adorable dog looking out the side of a car door mirror.

Figure 5. An example of the correct and incorrect caption structures in the first and second scenarios.

object scenarios. Through a series of carefully designed experiments, we illustrate how the biases in both text and image encoders can lead to unexpected or suboptimal results in tasks that are crucial for many downstream applications.

## 5.1. Image-Text Matching

Building upon our findings of biases in CLIP's image and text encoders, we now demonstrate how these biases tangibly affect the model's performance in image-caption matching tasks. We designed two experimental scenarios, conducted on both the ComCO and COCO datasets, to evaluate these biases. The results of these experiments are summarized in Table 6. To better illustrate the differences between these two scenarios, an example of the caption structures is shown in Figure 5. In each scenario, we created incorrect captions by switching one object in the caption with an object that is not present in the image. Additionally, GPT-4O [1] was used to rewrite the captions in the COCO dataset.

**First Scenario** In the first scenario, biases assist the model in distinguishing between the correct and incorrect captions. In the correct captions, the largest object in the image is placed at the beginning, aligning with the model's bias towards prioritizing first-mentioned objects and larger objects. For the incorrect captions, the non-existent object is deliberately placed at the beginning, which helps the model recognize the difference between the correct and incorrect captions more effectively. This positioning emphasizes the discrepancy early on, allowing the model to better detect the mismatch between the caption and the image. The performance of different models in this scenario can be seen in Table 6 under the "First Scenario" column.

**Second Scenario** In the second scenario, biases lead the model to make errors. The correct captions place the largest object at the end of the sentence, disrupting the model's bias towards objects mentioned earlier and its preference for larger objects. In the incorrect captions, the non-existent object is placed at the end, making it more difficult for the model to differentiate between correct and incorrect captions as its attention is drawn away from the critical discrepancies. The performance of different models in this scenario is shown in Table 6 under the "Second Scenario" column.

Table 6. Performance Comparison on Image-Text Matching for ComCO and COCO Datasets

| Dataset | Model | First Scenario | Second Scenario |
|---|---|---|---|
| ComCO | *CLIP Datacomp* [6] | **99.99** | 67.50 |
| | *CLIP Roberta* | **99.98** | 64.75 |
| | *SIGLIP* [22] | **99.49** | 72.36 |
| | *CLIP openAI* | **99.59** | 52.23 |
| | *NegCLIP* | **96.82** | 46.94 |
| | *SugarCrepe* | **98.55** | 60.43 |
| COCO | *CLIP Datacomp* [6] | **71.2** | 54.2 |
| | *CLIP Roberta* | **72.2** | 54.1 |
| | *SIGLIP* [22] | 64.8 | 39.5 |
| | *CLIP openAI* | **63.5** | 26.4 |
| | *NegCLIP* | **72** | 28.7 |
| | *SugarCrepe* | **80.0** | 40.9 |

By comparing these two scenarios, we demonstrate that biases in CLIP can either help or hinder the model's performance depending on how captions are structured. The experimental results, particularly with the use of GPT-4O for caption rephrasing in the COCO dataset, reveal how such biases can influence the accuracy of image-text matching tasks. These biases must be addressed to improve CLIP's robustness in real-world multi-object scenarios.

For further insights on how these biases affect text-to-image generation, refer to our extended experiments in Appendix 7.11.

## 6. Conclusion

This study uncovers biases in CLIP's encoders, with the text encoder favoring first-mentioned objects and the image encoder emphasizing larger ones, which impacts performance in multi-object tasks. Using the ComCO dataset, we highlighted these biases' effects on object representation and positioning, underscoring the need for balanced training. We attribute these biases to CLIP's contrastive framework, where alignment issues propagate across modalities. Addressing these biases is essential for vision-language advancements, as seen with models like Stable Diffusion.

# References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 5

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 1

[4] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *arXiv preprint arXiv:2406.11171*, 2024. 1

[5] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 1

[6] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 8, 17

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 5

[8] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024. 1, 7

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[10] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1

[11] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 3

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 17

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[14] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 5

[15] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 17

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 17

[17] Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5563–5573, 2024. 1

[18] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 1

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[20] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[21] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[22] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 8, 17

[23] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models

with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 1