# Vision-Language Models Do *Not* Understand Negation

Kumail Alhamoud[1] Shaden Alshammari[1] Yonglong Tian[*2] Guohao Li[3]
Philip H.S. Torr[3] Yoon Kim[1] Marzyeh Ghassemi[1]

[1] MIT [2]OpenAI [3] University of Oxford
https://NegBench.github.io

## Abstract

*Many practical vision-language applications require models that understand* negation*, e.g., when using natural language to retrieve images which contain certain objects but not others. Despite advancements in vision-language models (VLMs) through large-scale training, their ability to comprehend negation remains underexplored. This study addresses the question: how well do current VLMs understand negation? We introduce NegBench, a new benchmark designed to evaluate negation understanding across 18 task variations and 79k examples spanning image, video, and medical datasets. The benchmark consists of two core tasks designed to evaluate negation understanding in diverse multimodal settings: Retrieval with Negation and Multiple Choice Questions with Negated Captions. Our evaluation reveals that modern VLMs struggle significantly with negation, often performing at chance level. To address these shortcomings, we explore a data-centric approach wherein we finetune CLIP models on large-scale synthetic datasets containing millions of negated captions. We show that this approach can result in a 10% increase in recall on negated queries and a 28% boost in accuracy on multiple-choice questions with negated captions.*

## 1. Introduction

Joint embedding-based Vision-Language Models (VLMs), such as CLIP, have revolutionized how we approach multimodal tasks by learning a shared embedding space where both images and text are mapped together. This shared space enables a variety of applications, including cross-modal retrieval, video retrieval, text-to-image generation, image captioning, and even medical diagnosis [2, 20, 21, 23, 33, 35, 38, 40–42, 53]. By aligning visual and linguistic representations, these models achieve remarkable performance across domains and are able to model complex interactions between vision and language inputs.

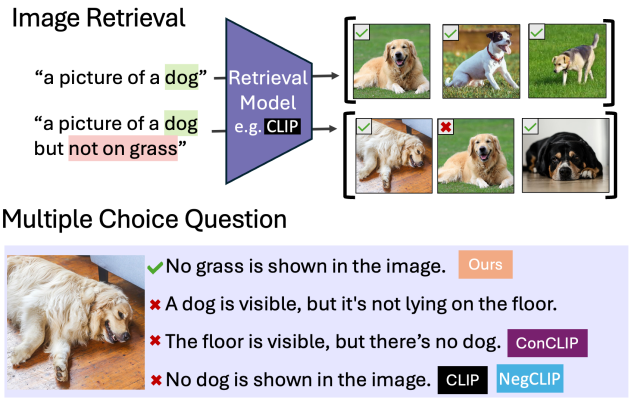



Figure 1. We present *NegBench* with image retrieval and multiple-choice tasks to evaluate negation understanding. CLIP-based models frequently misinterpret negation in both tasks, but we show how a synthetic data approach can improve performance.

Despite these advances, there is an emerging limitation: these models fail to handle *negation*, which is essential in many real-world scenarios. Negation enables precise communication by specifying what is false or absent [13, 17, 28, 29]. For example, a radiologist may search for images showing "bilateral consolidation with no evidence of pneumonia", or a safety inspector might query "construction sites with no barriers". Current benchmarks like CREPE and CC-Neg have introduced limited tests of negation, but they rely on rigid, templated examples that do not reflect the complexity of natural language queries [26, 43]. As a result, they fall short in evaluating how well VLMs understand negation in practical applications.

To comprehensively evaluate how well VLMs handle negation, we design a multi-level evaluation paradigm inspired by real-world information retrieval systems, where a coarse-grained retrieval step often precedes a fine-grained ranking or selection step [25, 31].

The first task, Retrieval-Neg, tests whether models can handle real-world queries that mix affirmative and negative statements, such as "a beach with no people" or "a building without windows." This task challenges the model to

* Yonglong Tian was at Google Deepmind during this work.

retrieve images from diverse datasets based on the presence of certain elements and the absence of others, simulating scenarios found in search engines, content moderation, and recommendation systems. By retrieving several potentially relevant matches (e.g., top-5 retrieval), Retrieval-Neg serves as the coarse-grained retrieval component of our evaluation.

The second task, MCQ-Neg, provides a fine-grained, structured evaluation that directly assesses specific failures in negation. In this task, the model must choose the correct description of an image from several closely related options, where the incorrect choices are hard negatives, differing only by what is affirmed or negated. For instance, in medical diagnostics, consider distinguishing between "The X-ray shows evidence of pneumonia but no evidence of pleural effusion" and "The X-ray shows evidence of pleural effusion but no evidence of pneumonia." These statements are linguistically similar but convey opposite diagnoses, requiring the model to parse subtle yet critical differences.

Through our evaluation pipeline, we uncover a surprising limitation: joint embedding-based VLMs frequently collapse affirmative and negated statements into similar embeddings, treating "a dog" and "no dog" as nearly indistinguishable. This affirmation bias reveals a significant shortcoming that was not sufficiently addressed in previous benchmarks like CREPE or CC-Neg.

Recognizing this critical gap, we then ask: If current models fail to understand negation, can we improve them? To tackle this, we propose a data-centric solution, introducing two large-scale synthetic datasets—CC12M-NegCap and CC12M-NegMCQ—designed to improve negation comprehension. Fine-tuning CLIP-based models on these datasets leads to substantial improvements, including a 10% increase in recall on negated queries and a 40% boost in accuracy on multiple-choice questions with negated captions.

The rest of the paper follows a challenge-diagnosis-solution structure. We introduce NegBench to evaluate negation comprehension, analyze VLMs' affirmation bias, and propose a data-driven solution using synthetic negation examples. We will open-source all models and data to foster research in negation understanding and its applications.

## 2. Related Work

Our work lies within the field of evaluating and advancing foundational vision-language models (VLMs). Joint-embedding models based on CLIP [34] show impressive generalization across visio-linguistic tasks like cross-modal retrieval, image captioning, and visual question answering [2, 20, 21, 33, 35, 38, 40–42] in diverse visual domains, extending beyond natural images to videos and medical images [3, 14, 23, 24, 30, 53]. We introduce a benchmark and data-centric approach to rigorously evaluate and improve negation understanding in these VLMs.

**Negation Understanding in Language and Vision.** Recent work showed that large language models perform sub-optimally when tasked with negation understanding [10, 47]. We go a step further by showing that vision-language models exhibit a more severe affirmation bias, completely failing to differentiate affirmative from negative captions.

Despite this critical limitation, existing benchmarks provide limited assessments of negation in VLMs. CREPE [26] and the concurrent work CC-Neg [43] are among the few vision-language benchmarks that include negation, but they focus on compositional understanding and rely on linguistic templates that fail to reflect the varied ways negation appears in real user queries. In contrast, our proposed benchmark, NegBench, leverages an LLM to generate natural-sounding negated captions, spanning a broader range of negation types and contexts across images, videos, and medical datasets. This systematic design enables a thorough evaluation of VLMs' ability to handle negation in multimodal settings, uncovering unique challenges and failure cases that have not been fully addressed in prior work.

**Improving CLIP for Compositionality and Negation.** Recent methods have explored improving the generalization abilities of CLIP-like VLMs for visio-linguistic compositionality and limited aspects of negation understanding. For instance, NegCLIP [50] employs composition-aware mining when finetuning CLIP to enhance compositional reasoning, while ConCLIP [43] modifies the CLIP loss to incorporate synthetic, template-based negation examples. In the medical domain, negation is a common feature in clinical text reports, often indicating the absence of specific pathologies [46]. Specialized models like BiomedCLIP [53] and CONCH [23] have been pretrained on millions of biomedical image-text pairs to address a variety of medical tasks, leveraging domain-specific knowledge from large-scale multimodal data. NegBench provides a systematic way to evaluate general-purpose and medical VLMs.

**Synthetic Data for Model Training.** It is common to use synthetic data to improve the performance of models in computer vision [1, 5, 16, 49]. Recent studies have shown that it is possible to use synthetic data to learn general vision-language representations, with some models trained entirely on synthetic images and captions achieving results comparable to real data [12, 44, 45]. Our approach is similar in spirit, but it constructs synthetic datasets to teach models a new, complex capability—*negation understanding*.

## 3. The Negation Benchmark (NegBench)

We design NegBench as a multi-level evaluation to assess the capacity of joint-based vision-language models to understand negation across different tasks: (1) coarse-grained retrieval, by accurately retrieving images that satisfy specified inclusions and exclusions, and (2) fine-

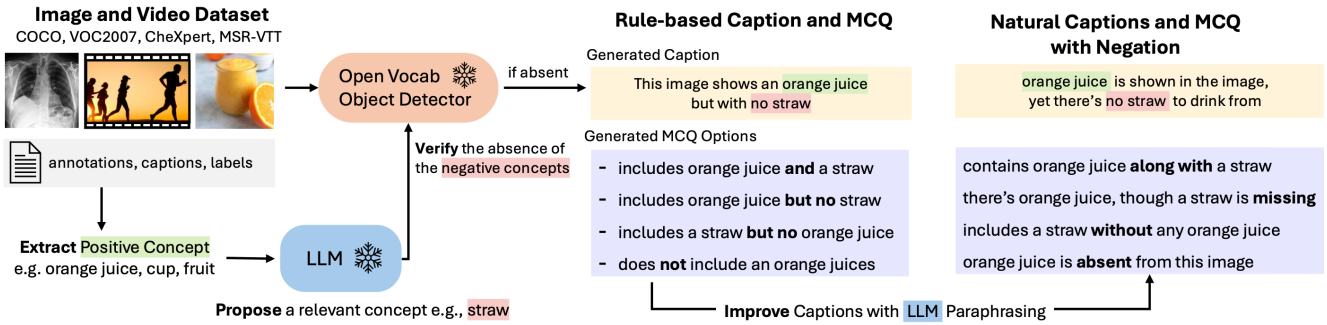



Figure 2. **General Pipeline for Constructing NegBench**. We start by extracting positive concepts from vision datasets. An LLM proposes negative concepts, which are verified with an object detector for datasets without explicit object annotations. We use templates to generate captions with negation, then paraphrase them by an LLM to ensure linguistic variety and robust evaluation of negation understanding.

grained question-answering, by selecting the correct description from closely related options, testing the model's detailed understanding of negation beyond simple retrieval.

In the Retrieval-Neg task, the model retrieves the top-5 images that match both affirmative and negative criteria within a query. In the MCQ-Neg task, the model selects the correct description of an image from options that differ only in the affirmation or negation of specific elements.

### 3.1. Transforming Datasets for Negation Evaluation

**General Dataset Transformation Overview.** To implement the two-stage evaluation pipeline of NegBench, we adapt several popular vision datasets, covering images (COCO [22], VOC2007 [7]), video (MSR-VTT [48]), and specialized medical imaging domains (CheXpert [15]). For each dataset, we identify positive elements $\{pos\}$, which represent objects or concepts present in the image, and negative elements $\{neg\}$, which are absent from the image but commonly associated with the present objects. When available, we use object-level annotations to identify these elements, as in COCO, VOC2007, and CheXpert; for other datasets, we derive positive and negative elements directly from the captions. This flexible approach allows NegBench to extend any vision dataset, whether it includes object-level annotations or captions, to evaluate negation comprehension across diverse tasks and data modalities.

In the Retrieval-Neg task, we modify standard captions by including negations, evaluating how models handle queries that specify both present and absent elements. For example, captions are modified as: "There is no $x$ in the image. [Original Caption]." or "[Original Caption]. There is no $x$ in the image." To introduce linguistic diversity, we use LLaMA 3.1 [6] to paraphrase these captions.

For the MCQ-Neg task, we generate multiple-choice questions (MCQs) for each image. The model must identify the correct description based on three linguistic templates: Affirmation, Negation, and Hybrid [18].

**1. Affirmation**: "This image includes **A** (and **C**)."
**2. Negation**: "This image does not include **B**."
**3. Hybrid**: "This image includes **A** but not **B**."

Each MCQ consists of one correct answer and three incorrect answers, which serve as hard negatives, misleading the model if it does not properly understand negation. A correct answer accurately describes the presence of $\{pos\}$ elements or negates $\{neg\}$ elements. A False Affirmation (e.g., "This image includes $x$" when $x \in \{neg\}$) or a False Negation (e.g., "This image does not include $x$" when $x \in \{pos\}$) highlights the model's failure to comprehend the image. The Hybrid template further evaluates the model's ability to combine affirmation and negation in the same caption. These MCQs are also paraphrased using LLaMA 3.1 to increase linguistic diversity.

### 3.2. Applicability Across Data Types and Domains

NegBench supports a wide range of data types and domains, enabling comprehensive negation evaluation.

**Video Understanding.** Video retrieval tasks introduce temporal complexity, where negation can involve both objects and actions that vary over time. Using MSR-VTT as an example, we prompt LLaMA 3.1 [6] to extract positive and negative elements from each video's caption. These elements may represent either objects present in the video or actions taking place. For Retrieval-Neg, we create captions specifying both the presence of some elements and the absence of others (e.g., "A person is cooking but not eating"). In MCQ-Neg, we generate multiple-choice questions where the model must select the description that most accurately represents a video segment, requiring it to reason about negation of objects and actions in dynamic scenes.

**Medical Image Interpretation with CheXpert.** Accurate negation understanding is critical in high-stakes domains like medical imaging. Using the CheXpert dataset [15], we focus on the most frequent condition *Lung Opacity* and design two binary classification tasks:

*Task 1: Affirmation Control Task.* This task evaluates the model's ability to associate images with specific medical conditions using affirmative statements.

> **Question**: Which option describes this image?
>
> A) This image shows Lung Opacity.
> B) This image shows Atelectasis.

*Task 2: Negation Understanding Task.* This task tests whether the model can correctly interpret negation, distinguishing the presence or absence of a medical condition.

> **Question**: Which option best describes the image?
>
> A) This image shows Lung Opacity.
> B) This image does *not* show Lung Opacity.

These extensions highlight the adaptability of NegBench to various data types and domains, from general images and videos to specialized medical imaging. This versatility ensures that NegBench provides rigorous, contextually relevant evaluations of negation understanding in VLMs.

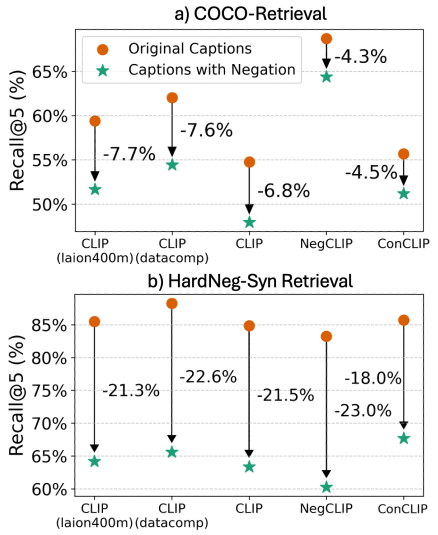



Figure 3. **Performance drop in recall@5 on (a) COCO and (b) HardNeg-Syn text-to-image retrieval with negated captions (green stars) compared to original captions (orange circles).** All models show substantial drops in performance, with NegCLIP experiencing the largest drop of 23.0% on HardNeg-Syn, which features hard negatives requiring stronger negation reasoning.

### 3.3. Synthetic Datasets for Controlled Evaluation

To rigorously test negation understanding, we construct *HardNeg-Syn*, a dataset that precisely controls object presence and absence by synthesizing hard negative images.

**Motivation and Benefits of Synthetic Data.** Synthetic data offers several advantages over traditional image datasets. First, by creating "hard negatives"—image pairs that differ only by a single object's presence or absence—we can evaluate the sensitivity of models to negation with minimal confounding variables. Additionally, image datasets like COCO and VOC2007 are limited in the range of visual concepts they cover; COCO has 80 objects while VOC2007 includes only 20. To expand this diversity, we prompt a large language model to propose a broader set of objects, which we use as targets in our synthetic dataset. This approach enables the generation of visually varied scenes that more comprehensively test negation comprehension across a wider array of objects and contexts.

**Construction Process for the HardNeg-Syn Evaluation Dataset.** We create 10,000 image pairs using Stable Diffusion [37], where each pair includes one image containing a target object and another where it is explicitly absent. To ensure accurate object presence or absence, we use the open-vocabulary object detector OWL-ViT [27].

## 4. NegBench Evaluations: Results and Insights

In this section, we benchmark the negation abilities of different VLMs using NegBench, comparing models based on their architecture, training data, and training objectives to reveal specific areas where negation understanding remains limited. Specifically, we evaluate five CLIP ViT-B/32 models on Retrieval-Neg and MCQ-Neg tasks. These include OpenAI CLIP [34], CLIP-laion400m [39], and CLIP-datacomp [9], which differ by pretraining dataset, as well as NegCLIP [50], trained to improve compositional language understanding, and ConCLIP [43], trained specifically to improve negation understanding. To handle the video dataset, MSR-VTT, we follow [3] and encode 4 uniformly sampled frames per video, averaging their features to obtain the video embedding. For medical tasks, we evaluate CONCH [23] and BioMedCLIP [53], two medical foundation VLMs. We also assess the impact of scaling up CLIP-laion400m (ViT-B, ViT-L, and ViT-H) to determine if larger embedding model sizes improve negation understanding. In addition, we investigate whether recent joint-embedding models trained with advanced objectives, such as SigLIP (ViT-L) [52], or AIMV2 [8] with Locked-image text Tuning [51], offer better performance on negation tasks.

**CLIP models struggle with negated queries in retrieval tasks.** We evaluate five CLIP-based models on the original COCO text-to-image retrieval task and its Retrieval-Neg version, where captions include negated statements. Across models, performance drops significantly on the negated task. In COCO retrieval (Figure 3a), CLIP-laion400m experiences a 7.7% drop in recall@5, with CLIP-datacomp and CLIP showing drops of 7.6% and 6.8%, respectively. In the more challenging HardNeg-Syn retrieval task (Figure 3b), the performance drops are even more pronounced due to the presence of hard negatives, *i.e.* images that closely resemble positive examples but differ by the exclusion of a single object. Here, NegCLIP, despite its promise for compositional

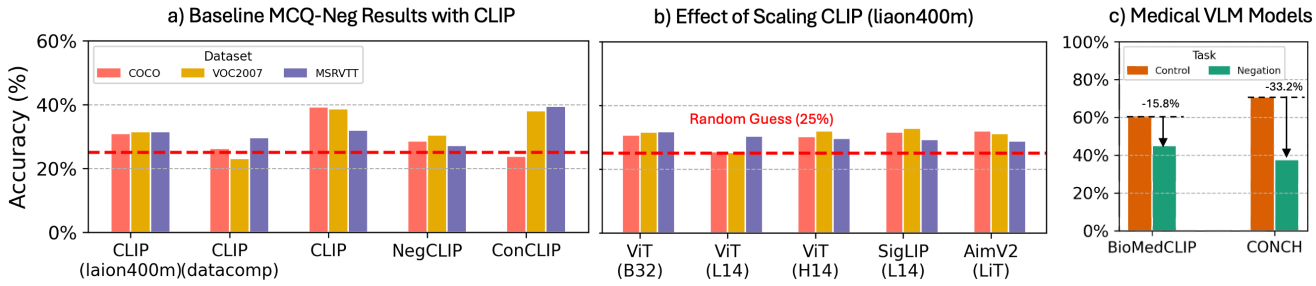



Figure 4. **MCQ-Neg performance across model families. (a)** CLIP-based models perform near random guessing (shown as a red dashed line), revealing their poor ability to handle negation. **(b)** Increasing model size (ViT-B→L→H) and using more advanced joint-embedding models (SigLIP, AIMV2) does not lead to better negation understanding, despite strong performance on other VLM tasks. **(c)** Medical VLMs experience large performance drops on negation MCQs, highlighting the risks of affirmation bias in high-stakes applications.

understanding, suffers a 23.0% drop, while ConCLIP, designed specifically for negation understanding, still declines by 18.0%. These results suggest that interpreting negation, particularly in the presence of hard negatives, remains a key challenge for retrieval tasks.

**MCQ-Neg reveals severe limitations in CLIP models.** Figure 4a shows that most models perform at or only slightly better than random guessing (indicated by the red dashed line at 25%) on the MCQ-Neg task. Interestingly, both NegCLIP and ConCLIP fall short from improving over the original OpenAI CLIP NegBench performance. Overall, these results reveal a fundamental limitation of CLIP-like pretraining objectives, which encourage strong associations between visual concepts and specific words, but struggle to interpret nuanced language like negation. Notably, the highest value is CLIP's accuracy on COCO, which is 39%. However, a score of sub 40% on a 4-way multiple-choice task is far below an acceptable level, demonstrating that models exhibit a serious lack of negation understanding.

**Bigger or newer is not (yet?) better at negation.** We show in Figure 4b that scaling up the model size from ViT-B/32 (86M parameters) to ViT-L/14 (307M parameters) and ViT-H/14 (632M parameters) does not improve negation understanding. We also evaluate the more recent joint-embedding models SigLIP (ViT-L/14) and AIMV2 (LiT), observing that they too fail to outperform baseline CLIP models on the MCQ-Neg task. Given that AIMV2 represents the state of the art on many vision-language tasks at the time of writing, this further highlights that negation remains a significantly under-addressed challenge in current VLMs.

**Critical failures in high-stakes medical tasks.** Figure 4c presents the results for the CheXpert MCQ-Neg task, where BioMedCLIP and CONCH exhibit substantial performance drops of 15.8% and 33.2%, respectively, when negation is introduced. This result is especially concerning in the context of medical diagnostics, where accurate interpretation of negation (e.g., the presence or absence of a condition such as Lung Opacity) is essential for correct diagnoses.

### 4.1. Why Do VLMs *Not* Understand Negation?

The results from NegBench reveal that CLIP VLMs struggle with different forms of negation understanding, motivating a deeper analysis into the underlying causes of these failures. In this section, we examine model performance across different MCQ types and analyze the embedding spaces of various models to uncover specific shortcut strategies that limit their negation comprehension.

**Model performance varies widely across MCQ types.** To understand why models struggle to perform better than random chance, we categorize the MCQs into three types based on the correct answer template: Affirmation, Negation, and Hybrid. Figure 5 compares model accuracy across these MCQ types. All models perform poorly on Negation MCQs, reflecting a general struggle with negation understanding (middle panel). In contrast, performance on Affirmation MCQs is substantially higher (left panel)—for instance, CLIP achieves 82% accuracy on Affirmation MCQs for VOC2007, but only 3% on Negation MCQs, revealing a severe affirmation bias in all models (except ConCLIP.)

To understand this behavior, we analyze the types of sentences models tend to select when making mistakes. Most models frequently choose Negation sentences that incorrectly negate existing objects (see template selection frequencies in the appendix). This likely stems from the task design: 67% of MCQs (Negation and Hybrid) do not contain a correct Affirmative option, which causes biased models to default to statements like "This image does not include {pos}." These results suggest that models trained with CLIP-like objectives often adopt shortcut strategies that ignore critical words such as "no." We refer to this tendency as the *affirmation bias* of CLIP-like models.

While ConCLIP appears less susceptible to affirmation bias, it does not outperform other models in NegBench, as its accuracy on Negation and Hybrid MCQs remains low. As we will show next, ConCLIP suffers from a different kind of bias that hinders its usability: it maps templated Hybrid captions to the same location in its embedding space.

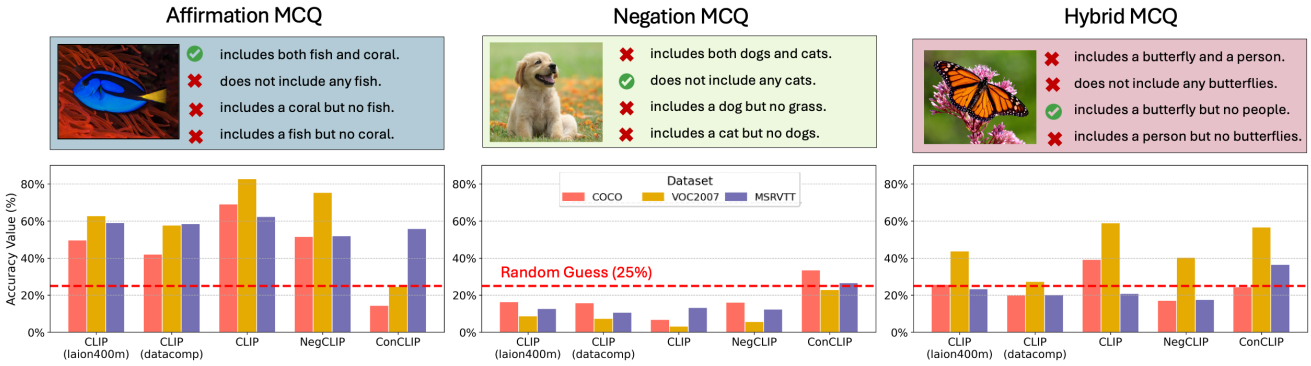
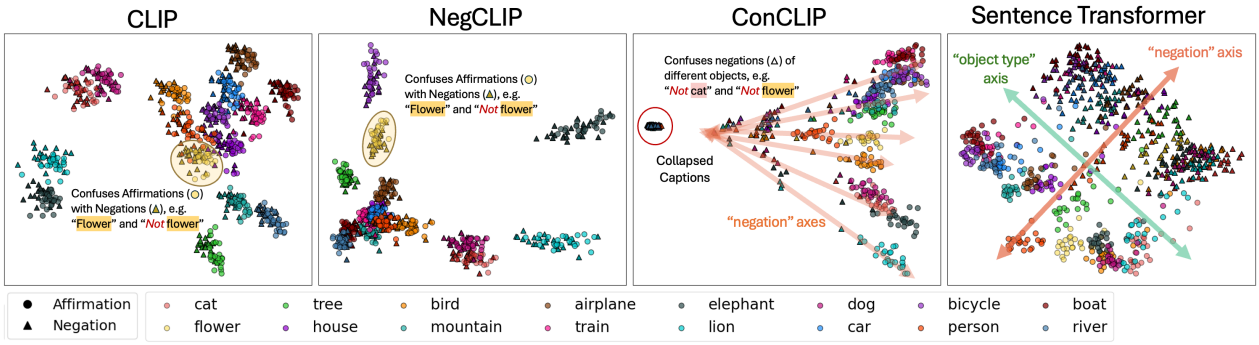
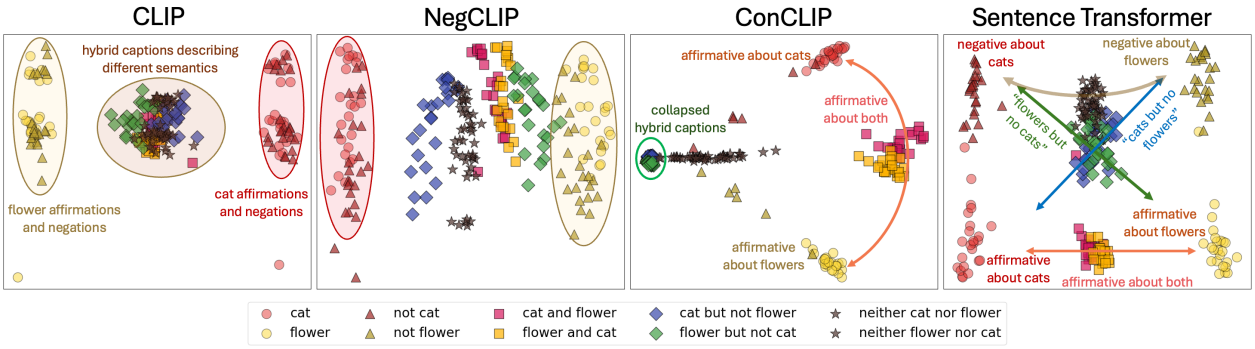



Figure 5. **Performance by MCQ type: Affirmation, Negation, and Hybrid.** CLIP-like models exhibit strong *affirmation bias*—they perform well on Affirmation MCQs (left panel), but fail on Negation MCQs (middle panel), often performing much below random chance.




(a) PCA embeddings for affirmative (dots) and negated (triangles) captions.




(b) PCA embeddings for hybrid captions (diamonds) and cases where two objects are negated (stars) or affirmed (squares).

Figure 6. **PCA Projections of Caption Embeddings Across Models.** CLIP and NegCLIP lack separation between affirmative and negated captions. ConCLIP treats all negated captions as identical, regardless of the object type, while the Sentence Transformer shows more ideal separability along both 'object type' and 'negation' dimensions.

**Embedding analysis reveals VLM shortcut strategies.** To investigate potential shortcut strategies, we analyze the embedding spaces of various models using 24 Affirmative ("X") and 24 Negated ("Not X") templates to create 48 captions per object. We apply PCA to the resulting embeddings (Figure 6a). The templates are detailed in the appendix.

We observe varying behaviors across models. The overlapping embeddings for affirmative and negated captions in *CLIP* and *NegCLIP* suggest that these models do not distinguish between positive and negative statements, possibly due to a "bag-of-words" shortcut strategy [11, 50] that overlooks negation words. This explains why both models incorrectly select the Negation template, which negates positive objects, in Figure 5. *CoNCLIP* separates positive and negative captions but fails to distinguish between negative captions of different objects, collapsing all negative caption embeddings toward a single point (red circle).

We include the embeddings of a text-only Sentence Transformer [36] as a reference that effectively differentiates affirmative and negated captions along distinct "object type" and "negation" axes, exemplifying ideal separation.

**Hybrid captions reveal more evidence of collapsed embeddings.** Figure 6b extends the previous analysis to hybrid captions that combine affirmations and negations. It provides further evidence that *ConCLIP* employs a shortcut strategy for embedding linguistic negation, with hybrid and negated captions collapsing towards a single point (green circle), indicating significant compression along the negation axis. While *CLIP* and *NegCLIP* struggle to distinguish affirmative from negative statements, *NegCLIP* shows better separation for hybrid captions, which appear collapsed in the CLIP embedding space. This suggests that NegCLIP's poor performance on Hybrid MCQs might be due to a misalignment between the text and image encoders, rather than an inability to understand hybrid sentence structure. In contrast, the *Sentence Transformer* effectively distinguishes between different caption types and provides semantically guided representations. For example, it aligns "flowers but not cats" along the line connecting "flowers" and "not cats."

# 5. A Data-Centric Approach for Improving Negation Understanding

We hypothesize that the tendency of CLIP-based models to rely on linguistic shortcuts, which hinders their negation understanding as explored in Section 4.1, stems from training data limitations. In CLIP, training data lacks examples with explicit negation, leaving it unable to distinguish negated and affirmed concepts. In contrast, ConCLIP's training data overfits to a single hybrid linguistic template, limiting its ability to generalize across varied negation structures. Next, we explore data-centric strategies to address these gaps, introducing a dataset that includes diverse negation examples spanning a range of linguistic styles.

## 5.1. Synthesizing a Fine-Tuning Negation Dataset

We augment the CC12M dataset [4], which contains approximately 10 million image-text pairs, to generate two synthetic datasets with negation: CC12M-NegCap and CC12M-NegMCQ. Our goal is to expose models to a wide variety of negation scenarios and improve their ability to encode negated statements. The process follows these steps:

1. **Object Extraction**: Using LLaMA 3.1 [6], we extract positive objects (those mentioned in the caption) and negative objects (contextually relevant but not present) from each image-caption pair in CC12M.
2. **Visual Verification**: An open-vocabulary object detector [27] verifies the presence of positive objects and ensures the absence of the negative objects in the image. This step is crucial to avoid introducing incorrect negations that could confuse the model.
3. **Caption Generation**: For each image, we generate multiple new captions that incorporate negated objects into the original captions. LLaMA 3.1 is used to ensure the generated captions are natural-sounding and reflect realistic negation scenarios found in retrieval queries.

We construct two variants of the synthetic dataset.**CC12M-NegCap** includes three captions per image with incorporated negated objects, totaling approximately 30 million captions. **CC12M-NegMCQ** includes four captions per image: one correct and three hard negatives based on object annotations, offering stronger training signals for fine-grained negation understanding and resulting in around 40 million captions. To balance broad retrieval with fine-grained negation capabilities, we introduce **CC12M-NegFull**, a comprehensive dataset that combines CC12M-NegCap and CC12M-NegMCQ. We will release the extracted object annotations for each image in CC12M, along with the corresponding URLs, and all the generated captions in CC12M-NegFull. This will help the community build on our dataset and advance research in negation understanding and multimodal retrieval.

## 5.2. Fine-Tuning with Negation-Enriched Data

**Standard CLIP Objective on CC12M-NegCap.** Let $\mathcal{B}_{\text{cap}} = \{(I_i, T_i)\}_{i=1}^{N}$ represent a batch of $N$ image-caption pairs from CC12M-NegCap, where each image $I_i$ is paired with a caption $T_i$ that describes present and absent objects in the image. For each batch $\mathcal{B}_{\text{cap}}$, we compute a similarity matrix $S \in \mathbb{R}^{N \times N}$, where each element $S_{j,k}$ represents the cosine similarity between the $j$-th image and the $k$-th caption. The CLIP objective applies a symmetric cross-entropy loss over this matrix, encouraging high similarity for correct image-caption pairs and low similarity for incorrect pairs. This loss is denoted as $\mathcal{L}_{\text{CLIP}}(\mathcal{B}_{\text{cap}})$ and provides the model with diverse negation examples in a contrastive learning setup.

**Multiple-Choice Objective on CC12M-NegMCQ.**

Let $\mathcal{B}_{\text{mcq}} = \{(I_i, \{T_{i,1}, \ldots, T_{i,C}\})\}_{i=1}^{M}$ be a batch of $M$ examples from CC12M-NegMCQ, where each image $I_i$ is paired with $C$ captions $\{T_{i,j}\}_{j=1}^{C}$. One caption correctly describes the image, while the others serve as hard negatives. For our experiments, we set $C = 4$. To fine-tune on CC12M-NegMCQ, we compute the cosine similarity between each image and its four caption options, generating a set of logits for each image-option pair.

The multiple-choice loss $\mathcal{L}_{\text{MCQ}}(\mathcal{B}_{\text{mcq}})$ is then computed by applying a cross-entropy loss over the logits, with the correct answer index as the target. This loss encourages the model to assign higher similarity to the correct caption and lower similarity to the hard negative captions:

$$\mathcal{L}_{\text{MCQ}}(\mathcal{B}_{\text{mcq}}) = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(\text{logits}_{i,c_i})}{\sum_{j=1}^{C}\exp(\text{logits}_{i,j})}, \quad (1)$$

where $c_i$ indicates the index of the correct caption describing the $i$-th image.

**Combined Training Objective.** The final objective combines the contrastive loss on CC12M-NegCap with the MCQ loss on CC12M-NegMCQ, weighted by $\alpha$ to balance their contributions. The total loss for one batch is:

$$\mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_{\text{CLIP}}(\mathcal{B}_{\text{cap}}) + (1-\alpha)\mathcal{L}_{\text{MCQ}}(\mathcal{B}_{\text{mcq}}). \quad (2)$$

**Evaluation Protocol.** To assess the impact of our data-centric approach, we fine-tune two pretrained models (OpenAI CLIP and NegCLIP) on CC12M-NegCap using the contrastive loss $\mathcal{L}_{\text{CLIP}}$. Additionally, we fine-tune both models on the combined CC12M-NegCap and CC12M-NegMCQ datasets using $\mathcal{L}_{\text{Total}}$ in Equation (2). For comparison, we fine-tune these models on the original CC12M dataset to isolate the effect of our negation-enriched datasets. Our goal is to demonstrate that CLIP models can significantly improve their understanding of negation with the right data.

We evaluate the models on two tasks: (i) text-to-image and text-to-video retrieval on COCO and MSR-VTT, both with and without negated queries, and (ii) image-to-text and video-to-text MCQ tasks, where models select the correct caption from four options. The results are shown in Table 1.

**Results.** Fine-tuning CLIP and NegCLIP on negation-enriched data leads to consistent and substantial improvements across both retrieval and MCQ tasks. On COCO, fine-tuning CLIP with CC12M-NegCap improves R-Neg@5 from 48.0% to 57.8%, while MCQ accuracy rises from 39.2% to 47.3% (**+8.1**). Similarly, NegCLIP's MCQ score improves from 28.6% to 40.4% (**+11.8**) with the same data. Larger MCQ gains are observed when training with CC12M-NegFull, which includes both contrastive and MCQ supervision: CLIP and NegCLIP achieve MCQ accuracies of 54.4% and 56.2%, respectively, corresponding to relative gains of **+15.2** and **+27.6** over their initial baselines. Similar trends also hold on the video dataset MSR-VTT. These results demonstrate that leveraging our high-quality synthetic dataset can effectively enhance VLM negation understanding.

**Ablation: Effect of varying $\alpha$.** The table below shows the impact of varying the weight factor $\alpha$ in the combined loss $\mathcal{L}_{\text{Total}} = \alpha\mathcal{L}_{\text{CLIP}} + (1-\alpha)\mathcal{L}_{\text{MCQ}}$ when fine-tuning CLIP on CC12M-NegFull. As $\alpha$ increases, more weight is placed on the original CLIP contrastive objective, while a lower $\alpha$ emphasizes the MCQ loss. Properly tuning $\alpha$ is important to balance between fine-grained MCQ and standard retrieval.

| $\alpha$ | 0 | 0.5 | 0.9 | 0.99 | 1 |
|---|---|---|---|---|---|
| COCO Recall@5 (%) | 33.9 | 37.3 | 47.6 | 54.2 | 58.5 |
| COCO MCQ Acc (%) | 59.4 | 53.7 | 54.6 | 54.4 | 47.3 |

## 6. Discussion and Conclusions

**Implications.** Our findings point to two broader implications for enhancing language understanding in VLMs.

| Model | Fine-tune data | R@5 (↑) | R-Neg@5 (↑) | MCQ (↑) |
|---|---|---|---|---|
| CLIP | None | 54.8 | 48.0 | 39.2 |
| | CC12M | 58.8 | 54.5 | 34.7 (↓4.5) |
| | **CC12M-NegCap** | **58.5** | **57.8** | **47.3** (↑8.1) |
| | **CC12M-NegFull** | **54.2** | **51.9** | **54.4** (↑15.2) |
| NegCLIP | None | 68.7 | 64.4 | 28.6 |
| | CC12M | 70.2 | 66.0 | 28.9 (↑0.3) |
| | **CC12M-NegCap** | **68.6** | **67.5** | **40.4** (↑11.8) |
| | **CC12M-NegFull** | **69.0** | **67.0** | **56.2** (↑27.6) |

(a) COCO Evaluation

| Model | Fine-tune data | R@5 (↑) | R-Neg@5 (↑) | MCQ (↑) |
|---|---|---|---|---|
| CLIP | None | 50.6 | 45.8 | 32.1 |
| | CC12M | 53.7 | 49.9 | 30.8 (↓1.3) |
| | **CC12M-NegCap** | **54.1** | **53.5** | **41.5** (↑9.4) |
| | **CC12M-NegFull** | **46.9** | **43.9** | **44.9** (↑12.8) |
| NegCLIP | None | 53.7 | 51.0 | 27.3 |
| | CC12M | 56.4 | 52.6 | 31.6 (↑4.3) |
| | **CC12M-NegCap** | **56.5** | **54.6** | **39.8** (↑12.5) |
| | **CC12M-NegFull** | **54** | **51.5** | **46.2** (↑18.9) |

(b) MSR-VTT Evaluation

Table 1. **Comparison of fine-tuning datasets** on performance metrics across COCO and MSR-VTT, fine-tuned on respective datasets and evaluated on retrieval and MCQs. Differences in MCQ accuracy from the baseline are shown, with increases of +8 or more highlighted. Fine-tuning on negation-enriched data significantly improves negation understanding (R-Neg and MCQ).

From a data perspective, pretraining datasets should include a diverse array of language constructs, especially those involving nuanced expressions like negation or complex syntactic structures, to help models capture the subtleties of human language. Currently, many VLMs are pretrained on datasets that primarily consist of straightforward, affirmative statements, which might limit the models' ability to understand more subtle language elements. From a learning perspective, our results suggest that a combination of contrastive learning and MCQ supervised training can improve coarse-grained retrieval and fine-grained negation understanding. We experimented with different values of $\alpha$ in Equation (2), which revealed a tradeoff in performance. This suggests that alternative or supplementary training objectives beyond contrastive learning could enhance models' sensitivity to nuanced language, enabling more robust applications in real-world settings where precise language interpretation is essential.

**Summary.** This paper introduces *NegBench* to systematically evaluate negation understanding in VLMs. Our findings reveal that CLIP-based models exhibit a strong affirmation bias, limiting their application in scenarios where negation is critical, such as medical diagnostics and safety monitoring. Through synthetic negation data, we offer a promising path toward more reliable models. While our synthetic data approach improves negation understanding, challenges remain, particularly with fine-grained negation differences.

**Acknowledgments.** This work was supported in part by a National Science Foundation (NSF) 22-586 Faculty Early Career Development Award (2339381), a Gordon & Betty Moore Foundation award, a Google Research Scholar award, and a UKRI grant Turing AI Fellowship (EP/W002981/1). The authors would like to thank Walter Gerych, Olawale Salaudeen, and Mark Hamilton for valuable discussions and feedback.

## References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126:961–972, 2018. 2

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022. 1, 2

[3] Santiago Castro and Fabian Caba. FitCLIP: Refining large-scale pretrained image-text models for zero-shot video understanding tasks. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 2, 4

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 7

[5] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019. 2

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 7

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 2010. 3

[8] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 4

[9] 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 M 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. In *NeurIPS Datasets and Benchmarks Track*, 2023. 4

[10] Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. This is not a dataset: A large negation benchmark to challenge large language models. In *EMNLP*. Association for Computational Linguistics, 2023. 2

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020. 6

[12] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 2

[13] Laurence R. Horn. *A Natural History of Negation*. University of Chicago Press, 1989. 1

[14] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024. 2

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 3

[16] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. In *International Conference on Learning Representations*, 2022. 2

[17] Michael P Jordan. The power of negation in english: Text, context and relevance. *Journal of pragmatics*, 29(6), 1998. 1

[18] Miren Itziar Laka Mugarza. *Negation in syntax–on the nature of functional categories and projections*. PhD thesis, Massachusetts Institute of Technology, 1990. 3

[19] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 13

[20] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17990–17999, 2022. 1, 2

[21] Zhengxin Li, Wenzhe Zhao, Xuanyi Du, Guangyao Zhou, and Songlin Zhang. Cross-modal retrieval and semantic refinement for remote sensing image captioning. *Remote Sensing*, 16(1):196, 2024. 1, 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[23] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 1, 2, 4

[24] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2

[25] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425, 2024. 1

[26] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *CVPR*, 2023. 1, 2

[27] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *ECCV*, 2022. 4, 7, 16

[28] Roser Morante and Eduardo Blanco. Recent advances in processing negation. *Natural Language Engineering*, 27(2): 121–130, 2021. 1

[29] Partha Mukherjee, Youakim Badr, Shreyesh Doppalapudi, Satish M Srinivasan, Raghvinder S Sangwan, and Rahul Sharma. Effect of negation in sentences on sentiment analysis and polarity detection. *Procedia Computer Science*, 185: 370–379, 2021. 1

[30] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021. 2

[31] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019. 1

[32] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288, 2024. 13

[33] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18092–18102, 2022. 1, 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 2, 4

[35] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 1, 2

[36] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*. Association for Computational Linguistics, 2019. 6

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4

[38] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023. 1, 2

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4

[40] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*. 1, 2

[41] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022.

[42] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. 1, 2

[43] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*, 2024. 1, 2, 4

[44] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2023. 2

[45] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024. 2

[46] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 2

[47] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 101–114, 2023. 2

[48] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[49] Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[50] 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 *ICLR*, 2023. 2, 4, 6, 12

[51] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022. 4

[52] 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. 4

[53] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 2, 4

# Appendix

## A. Evaluating LLaVA on NegBench MCQs

In the main paper, we proposed a novel evaluation paradigm for negation understanding, aimed at simulating real-world scenarios as closely as possible. We then proceeded to evaluate joint embedding-based VLMs, particularly CLIP models, which are the dominant models for multimodal retrieval tasks, in addition to being popular for text-to-image generation, image captioning, and medical multimodal tasks. However, we recognize that there are other VLMs that can be useful in certain settings. In particular, instruction-tuned VLMs like LLaVA open up the path for conversational VLM chatbots. In this section, we evaluate LLaVA on the three natural image MCQ tasks in NegBench (COCO, VOC2007, and HardNeg-Syn). The results are in Figure 7.

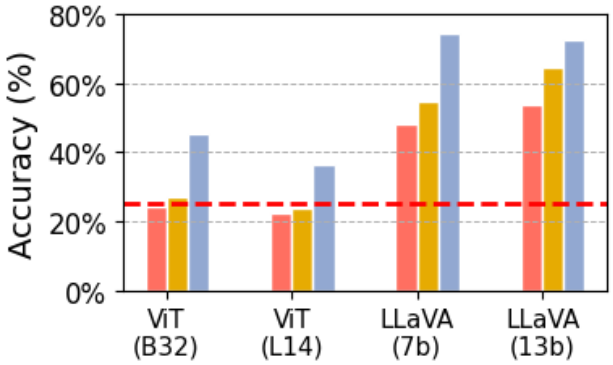



Figure 7. **Caption.**

**LLaVA, an instruction-tuned VLM, demonstrates improvement.** Figure 7 shows that LLaVA significantly outperforms CLIP models on the MCQ-Neg tasks. This is particularly notable because LLaVA uses a CLIP ViT-L/14 vision encoder, which we have shown in Figure 4 to struggle with negation. The key advantage of LLaVA might be in its use of the Vicuna LLM for text encoding. Unlike CLIP, which is pretrained on vision-language pairs that predominantly contain affirmative image captions, LLMs like Vicuna are trained on diverse textual corpora that include both affirmations and negations. This broader exposure allows LLaVA to better interpret negated statements. Additionally, LLaVA uses a learned projection layer to align vision and language representations, in contrast to CLIP's contrastive learning objective, which tends to ignore word order and subtle linguistic cues like negation [50]. We further explore these differences in Figure 8.

**Limitations of LLaVA as a retrieval system.** While LLaVA demonstrates improved negation understanding, it has significant limitations as a retrieval model compared to CLIP. CLIP learns a joint image-text embedding space, making it highly efficient for retrieval tasks by simply embedding both images and texts, and then computing cosine similarities. In contrast, LLaVA processes a single image-text pair at a time and generates text output, which makes image-to-text retrieval feasible only if all possible captions can fit into the model's context window. For MCQ-Neg, we applied this method by presenting the image alongside all possible captions and prompting LLaVA to select the correct one. However, this approach does not scale well with a large number of candidates and is not applicable for text-to-image retrieval, where fitting all dataset images into the context window is impractical. Therefore, advancing models like CLIP is crucial for real-world multimodal retrieval with negation. In the paper, we explored the data-centric reasons behind CLIP's failures in negation understanding and proposed synthetic data strategies to address them.

## B. A Closer Look at VLM Negation Failures

To better understand the negation failures of VLMs, we further analyze the models' tendency to select specific template types when answering multiple-choice questions (MCQs) and provide further analysis into the embedding space of these models.

### B.1. Template Selection Frequency

Figure 8 analyzes the frequency with which different models select specific template types (Affirmation, Negation, Hybrid) when answering multiple-choice questions, regardless of the correct answer. This analysis helps to reveal potential biases in model behavior and understand why models may struggle with negation. As shown in Figure 5 from the paper, most models perform poorly on Negation MCQs, reflecting a general struggle with negation understanding.

### B.2. Template Selection Frequency

Figure 8 analyzes how often different models select templates of type Affirmation, Negation, or Hybrid when answering multiple-choice questions—regardless of whether the selected answer is correct. This helps reveal systematic biases in model decision-making.

We observe that most CLIP-based models strongly over-select Negation templates, even when the correct answer is an Affirmation or Hybrid statement. This aligns with the results in Figure 5, where models struggle with Negation MCQs and tend to default to negated statements. This behavior supports our earlier claim of an *affirmation bias*: models trained with CLIP-like objectives tend to ignore function words like "not" and collapse positive and negative statements in their embedding space.

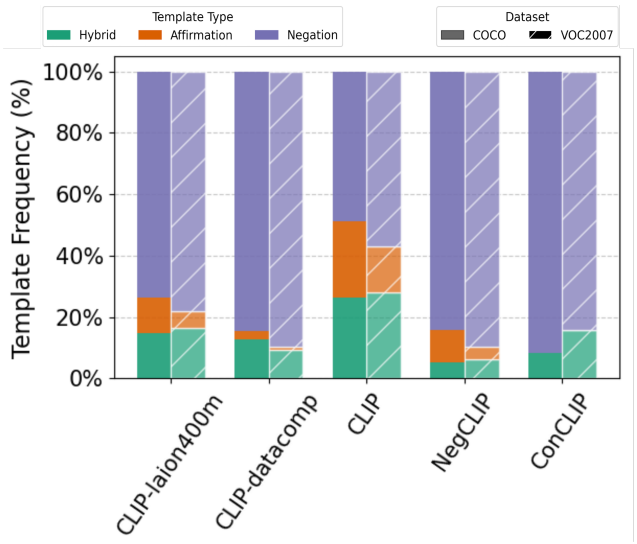



Figure 8. Template selection frequency for various models on COCO and VOC2007 datasets, broken down by template type (Affirmation, Negation, Hybrid).

Table 2. MCQ Total Accuracy (%) across different datasets for various models

| Model | COCO | VOC2007 | HardNeg-Syn |
|---|---|---|---|
| CLIP-OpenAI | 16.27% | 14.47% | 18.24% |
| CLIP-Laion400M | 24.26% | 27.01% | 44.60% |
| CLIP-datacomp | 19.73% | 19.72% | 34.10% |
| NegCLIP | 10.21% | 8.51% | 17.03% |
| ConCLIP | 15.20% | 20.43% | 11.10% |
| CLIP-L14 | 22.44% | 23.69% | 36.51% |
| CLIP-H14 | 32.14% | 38.26% | 36.98% |

### B.3. Template Embedding Analysis

This subsection provides further details about the embedding analysis presented in Figure 6 of the main paper. We achieve this by:

1. Specifying templates used to generate the embeddings.
2. Expanding the embedding analysis to more models.

To generate the embeddings for the PCA projections, we used five categories of templates: Affirmation (single object), Negation (single object), Affirmation (two objects), Hybrid (one object affirmed, one negated), and Double Negation (two objects negated). Each category contains 24 templates, except for Affirmation (two objects) which has 23. The templates vary sentence structure and wording while maintaining the same core meaning.

- **Affirmation (single object):** 24 templates. Examples: "This image includes A", "A is present in this image", "This image shows A", "A is depicted in this image", "A appears in this image".
- **Negation (single object):** 24 templates. Examples: "This image does not include A", "A is not present in this image", "This image lacks A", "A is not depicted in this image", "A does not appear in this image".
- **Affirmation (two objects):** 23 templates. Examples: "This image includes A and B", "A and B are present in this image", "This image shows A and B", "A and B are depicted in this image", "A and B appear in this image".
- **Hybrid (one object affirmed, one negated):** 24 templates. Examples: "This image includes A but not B", "A is present in this image but not B", "This image shows A but not B", "This image features A but not B", "A appears in this image but not B".
- **Double Negation (two objects negated):** 24 templates. Examples: "This image includes neither A nor B", "Neither A nor B are present in this image", "This image shows neither A nor B", "Neither A nor B are depicted in this image", "Neither A nor B appear in this image".

While Figure 6 focused on CLIP, NegCLIP, and ConCLIP, Figure 9 presents an additional visualization with PCA projections for other CLIP models (varying in size and pretraining datasets). This broader analysis will provide a more comprehensive view of how different CLIP models handle negation in the embedding space.

## C. Additional Insights and Context

**D.1 How does this work fit into the broader landscape of negation and compositionality research?**

Prior benchmarks such as CREPE and CC-Neg introduced limited forms of negation in vision-language tasks, focusing on compositionality or constrained template-based generation. More recently, SPEC [32] proposed fine-grained VQA tasks with a subset evaluating negation understanding. NaturalBench [19] presents a vision-centric QA protocol that reveals large performance gaps between humans and top-tier VLMs (e.g., GPT-4o, Qwen2-VL), often caused by answer biases such as a tendancy to say "Yes." over "No."

Our work complements and extends these efforts with several contributions:

- We introduce **NegBench**, a large-scale benchmark with 79K examples across retrieval and MCQ tasks, spanning images, video, and medical domains.
- We design **naturalistic negation prompts** using LLMs, covering a broad range of negation types and avoiding rigid linguistic templates.
- We generate **70M+ synthetic negation-enriched training samples**, supporting both contrastive and multiple-choice learning objectives.
- We conduct extensive experiments showing that our models **outperform prior negation-specific models (e.g., ConCLIP)** as well as SOTA VLMs (e.g., AIMv2) on negation tasks.

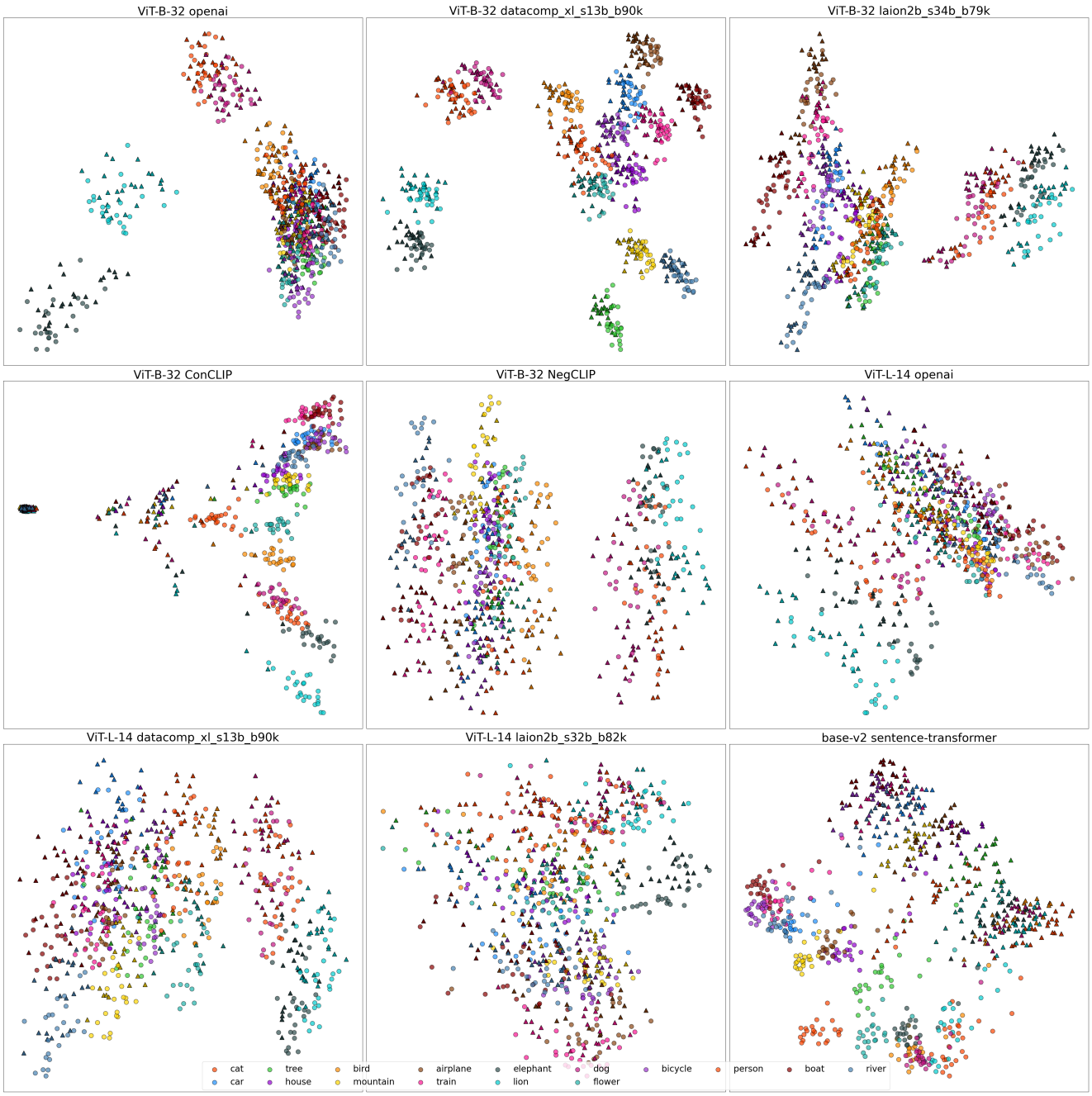



Figure 9. PCA projections of caption embeddings for various CLIP models and the Sentence Transformer. Each point represents a caption embedding. This figure complements Figure 6 by providing a broader view of embedding separation across different VLMs.

**D.2 What is the significance of model scaling experiments and comparisons to recent architectures like AIMv2?**

A common intuition is that larger models may better capture fine-grained distinctions such as negation. To evaluate this, we scale CLIP across ViT-B, L, and H variants, and additionally assess newer joint-embedding models such as SigLIP and AIMv2. Despite stronger performance on standard retrieval tasks, these models still struggle on MCQ-Neg and do not meaningfully close the gap—indicating that increased capacity alone does not resolve negation failures.

**D.3 How are negative object queries constructed in retrieval and MCQ settings?**

For datasets with dense annotations (COCO, VOC2007),

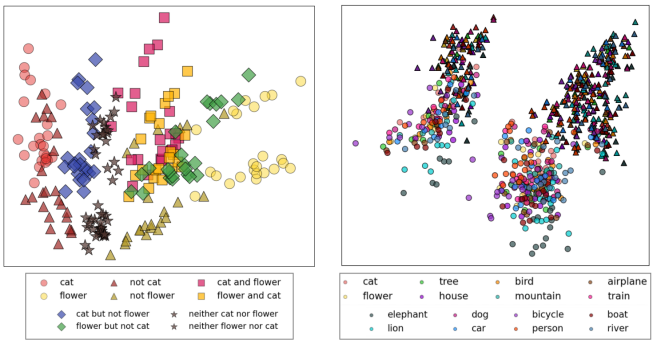



Figure 10. PCA projections of caption embeddings for finetuned CLIP model on CC12M-NegCap. Each point represents a caption embedding.

we construct a co-occurrence matrix to identify object pairs that frequently appear together. We then generate negated prompts by selecting a plausible object that is *absent* from the current image but typically co-occurs with present objects. This ensures that the negation is realistic and visually grounded, rather than relying on unlikely or artificially constructed distractors.

**D.4 What is the significance of the medical experiment, despite its simplicity?**

The medical retrieval experiment uses a simple binary decision setup, which offers a clean, interpretable upper bound on model capability. Models are tasked with distinguishing statements like "has pneumonia" versus "does not have pneumonia." Despite the simplicity, we observe large performance drops under negation (up to 33%) for domain-specialized VLMs such as BioMedCLIP and CONCH. This reveals a persistent failure mode with real-world clinical implications, where affirming or negating a condition must be handled with precision to avoid dire consequences.

## D. Dataset and Task Summary for NegBench

We provide a summary of the datasets and tasks used in NegBench, a framework designed to evaluate Visual Language Models (VLMs) on their understanding of negation across different modalities, including images, videos, and medical imaging. The benchmark includes both retrieval and multiple-choice question (MCQ) tasks, with two variations: templated and LLM-paraphrased. For synthetic data, we generate 10,000 images using Stable Diffusion, which serve as hard negatives for one another, enabling a more focused evaluation of negation comprehension in text-to-image retrieval tasks.

Each dataset contributes to either Retrieval-Neg or MCQ-Neg tasks, except for CheXpert, which has two distinct tasks (Affirmation Control and Negation Understanding) in both MCQ and binary classification formats. Additionally, we utilize original retrieval captions for COCO (5,000) and MSR-VTT (1,000), expanding the overall dataset size. VOC2007 does not include a Retrieval-Neg task as it lacks retrieval-style captions.

The total number of task variations across all datasets in NegBench is 18, and the total number of samples across all tasks and variations is 79,239. Table 3 summarizes the datasets, tasks, task versions, and sizes.

- **COCO**: 5,000 retrieval captions and 5,914 MCQ questions, resulting in 10,000 retrieval problems and 11,828 MCQ problems with templated and LLM-paraphrased variations.
- **VOC2007**: 5,032 MCQ questions, leading to 10,064 total samples. No retrieval task is provided due to the absence of retrieval-style captions.
- **MSR-VTT**: 1,000 retrieval captions and 1,000 MCQ questions, resulting in 2,000 samples per task, including both variations.
- **CheXpert**: Two MCQ tasks (4-choice) and two binary classification tasks. The 4-choice MCQ covers 690 samples for affirmation and 1,587 for negation, while the binary tasks each include 690 samples.
- **HardNeg-Syn**: 10,000 synthetic images, used to create 20,000 retrieval and 20,000 MCQ problems across templated and LLM-paraphrased versions.

Table 3. **Summary of datasets and tasks in NegBench.** Each task includes both templated and LLM-paraphrased versions, except for CheXpert tasks, which are templated only due to their straightforwardness (they directly evaluate diagnostic capabilities in the presence of negation words). The HardNeg-Syn dataset contains 10,000 synthetic images as hard negatives, offering a more targeted evaluation of negation understanding. The total number of task variations is 18, with a total of 79,239 samples across all tasks and variations.

| Dataset | Task | Templated | LLM-Paraphrased | Task Size | Notes |
|---|---|---|---|---|---|
| **COCO** | Retrieval-Neg | ✓ | ✓ | 10,000 | Image retrieval with negated captions. |
| | MCQ-Neg | ✓ | ✓ | 11,828 | MCQ task with affirmative, negated, and hybrid options. |
| **VOC2007** | MCQ-Neg | ✓ | ✓ | 10,064 | MCQ task. No Retrieval-Neg for VOC2007. |
| **MSR-VTT** | Retrieval-Neg | ✓ | ✓ | 2,000 | Video retrieval task with negated captions. |
| | MCQ-Neg | ✓ | ✓ | 2,000 | Video-based MCQ task with temporal context. |
| **CheXpert** (4-choice) | Affirmation Control MCQ | ✓ | – | 690 | Medical image MCQ with 4 choices. |
| | Negation Understanding MCQ | ✓ | – | 1,587 | MCQ task with negation. |
| **CheXpert** (binary) | Affirmation Control | ✓ | – | 690 | Binary classification of medical images. |
| | Negation Understanding | ✓ | – | 690 | Binary classification, negated statements. |
| **HardNeg-Syn** | Retrieval-Neg | ✓ | ✓ | 20,000 | Synthetic image retrieval task. |
| | MCQ-Neg | ✓ | ✓ | 20,000 | MCQ task for synthetic images with 4 answer choices. |

### D.1. Details of HardNeg-Syn Construction

**Object Label Selection**

We gather a wide range of object text labels from existing datasets like ImageNet.

**Scene Description**

For each selected object label (`A`), LLaMA 3.1 generates:
A `{background_description}` and a related object `{B}`, crafting realistic scene contexts.

**Image Generation**

Using Stable Diffusion, we generate pairs of images:

> **Positive Image:** `{background_description}` with `{A}` next to `{B}`.
>
> **Negative Image:** `{background_description}` with `{A}`, excluding `{B}` in the negative prompt to ensure its absence.

**Verification**

We use OWL-ViT [27] to verify the presence and absence of `A` and `B`.

**Caption Generation**

Captions are generated using templates and paraphrased with LLaMA 3.1 for naturalness.

## E. Visualizing the NegBench Evaluation Tasks

In Figures Figures 11 to 14, we visualize a few samples from the NegBench retrieval and MCQ tasks we introduced in the paper. We note that the datasets are diverse in terms of the nature of visual domain and real-world applicability.

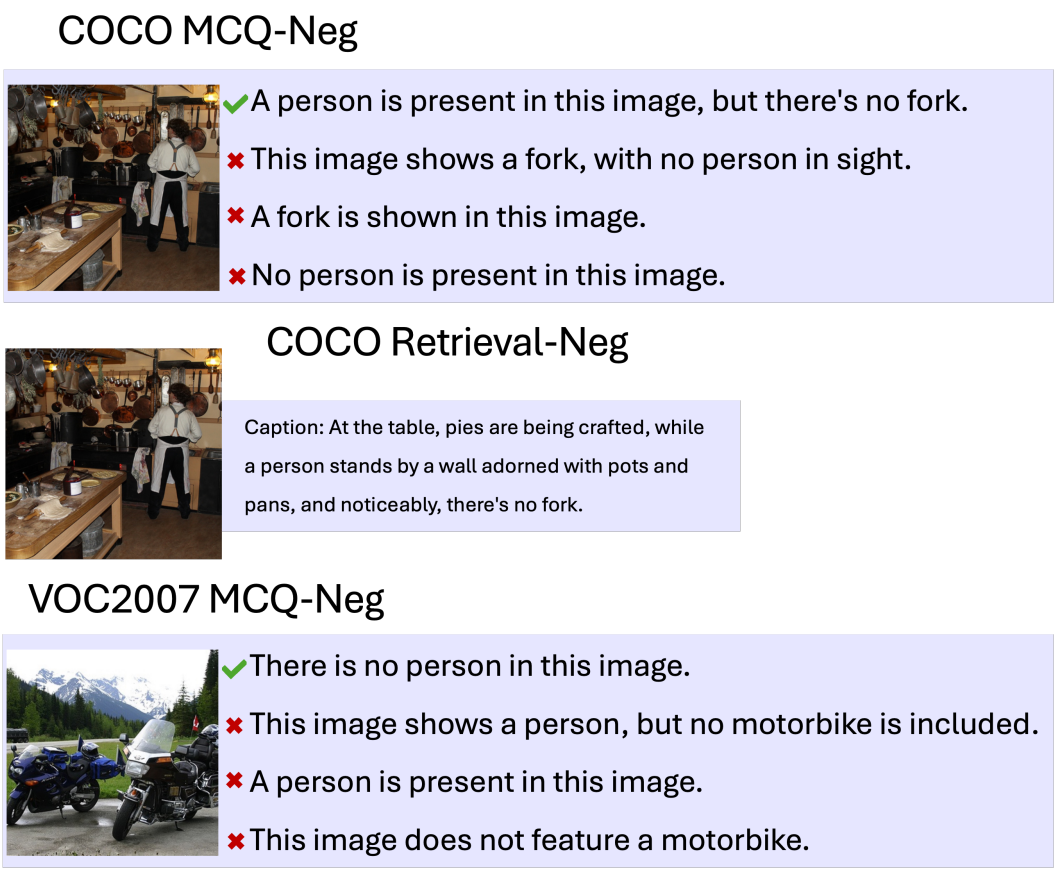



Figure 11. Examples of COCO and VOC2007 tasks, including Retrieval with negated captions and MCQ with negation.

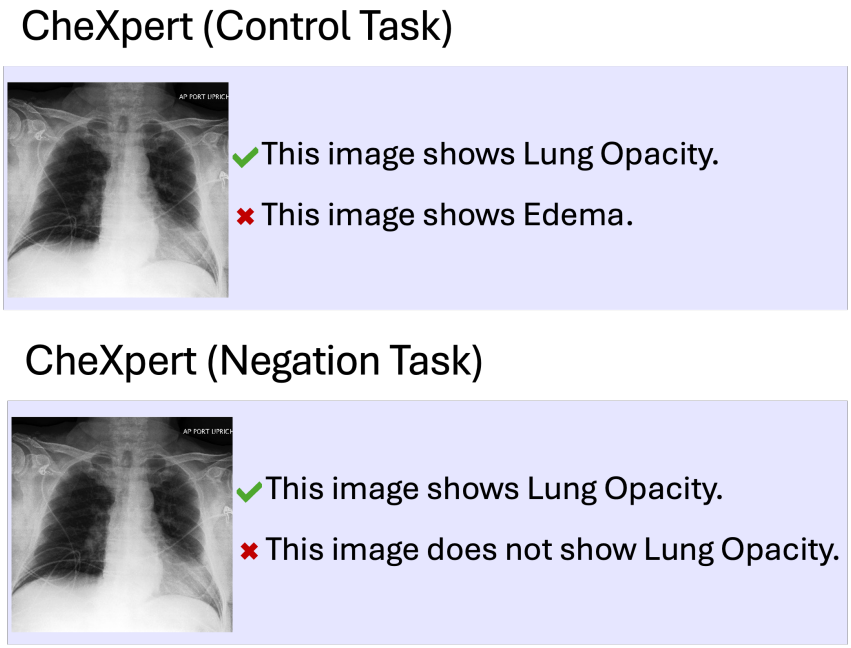



Figure 12. Examples of CheXpert MCQ tasks, including the Affirmation Control task and the Negation task.

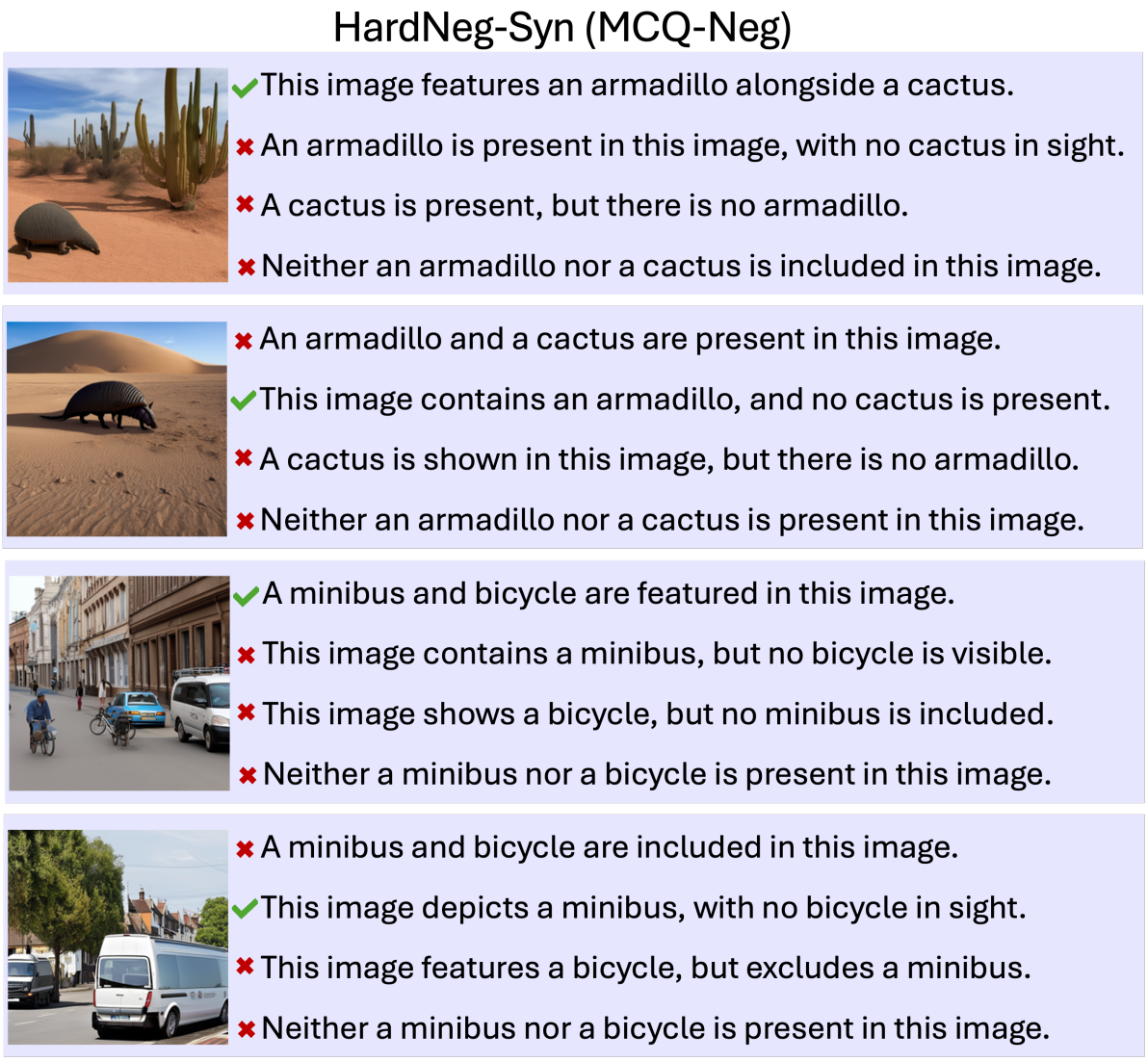



Figure 13. Examples of HardNeg-Syn (MCQ-Neg) tasks. Images in this dataset are constructed in pairs, with each pair differing by a single object (the cactus in the first pair), making the dataset particularly suitable for studying negation understanding.

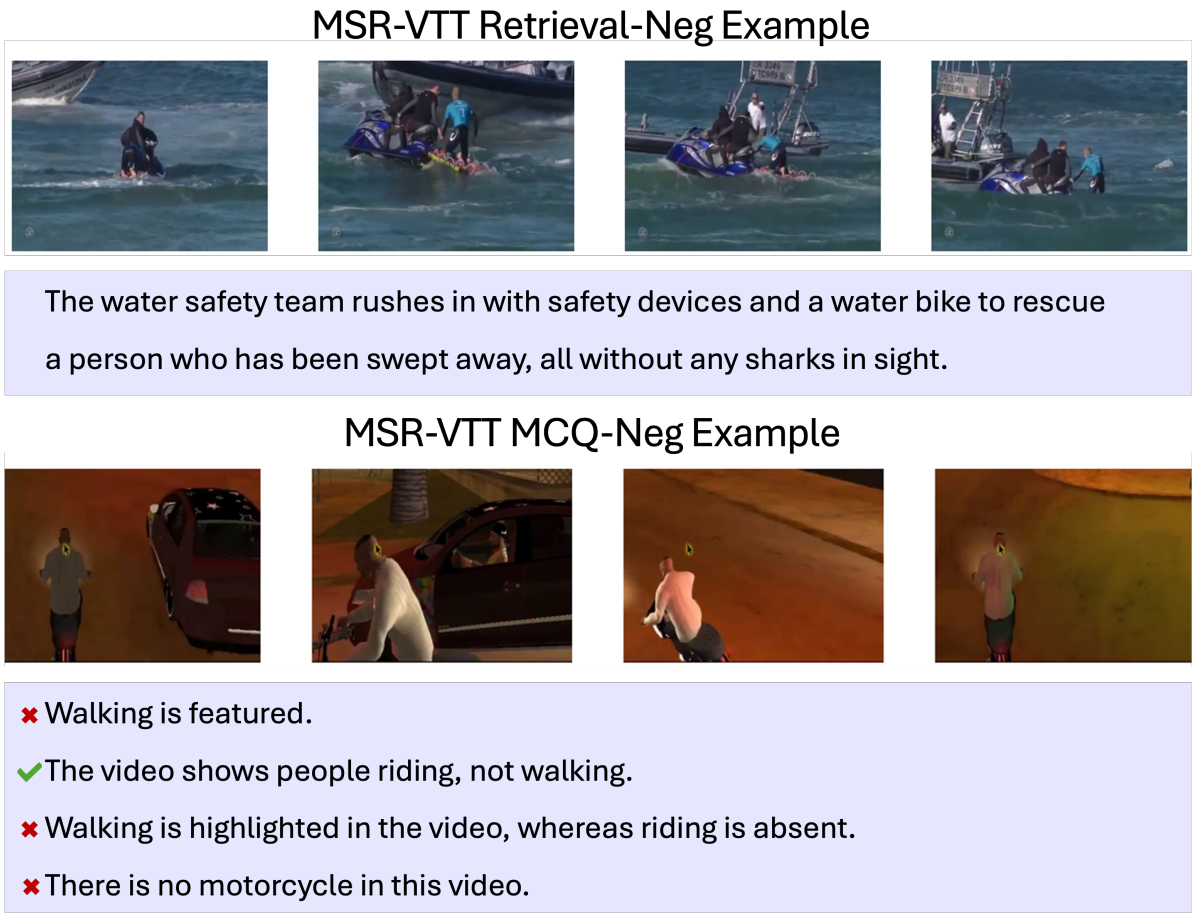



Figure 14. Examples of MSR-VTT tasks, including Retrieval-Neg (with negated captions about a complex water rescue scene) and MCQ-Neg (with answer choices about the presence or absence of actions like walking).