CVPR
#2817

Vision-Language Models Do *Not* Understand Negation

Anonymous CVPR submission

Paper ID 2817

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* 79*k 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 40% 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, 18, 19, 21, 30, 32, 35, 38–40, 49]. 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.

Despite these advances, there is an emerging limitation: these models fail to handle *negation*, which is essential in many real-world scenarios. Negation enables

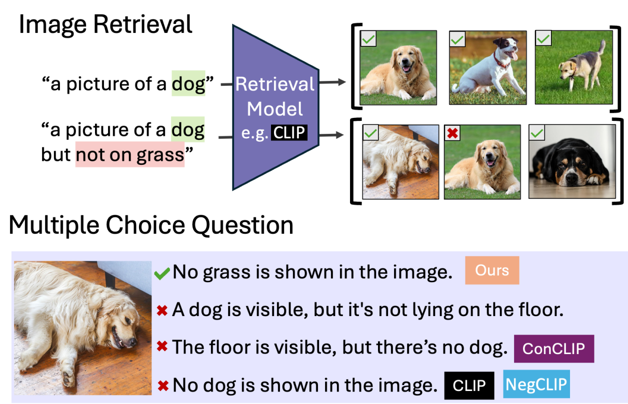


Figure 1. We present *NegBench* with image retrieval and multiple-choice tasks to evaluate negation understanding. CLIP-based models frequently misinterpret negation in both tasks, but we show how a synthetic data approach can improve performance.

precise communication by specifying what is false or absent [12, 16, 26, 27]. 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 [24, 41]. 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 [23, 29].

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 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—Syn-Neg-Cap and Syn-Neg-MCQ—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 [31] show impressive generalization across visio-linguistic tasks like cross-modal retrieval, image captioning, and visual question answering [2, 18, 19, 30, 32, 35, 38–40] in diverse visual domains, extending beyond natural images to videos and medical images [3, 13, 21, 22, 28, 49]. 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 [9, 45]. 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 [24] and the concurrent work CC-Neg [41] 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 [48] employs composition-aware mining when finetuning CLIP to enhance compositional reasoning, while ConCLIP [41] 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 [44]. Specialized models like Biomed-CLIP [49] and CONCH [21] 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, 15, 47]. 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 [11, 42, 43]. 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-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

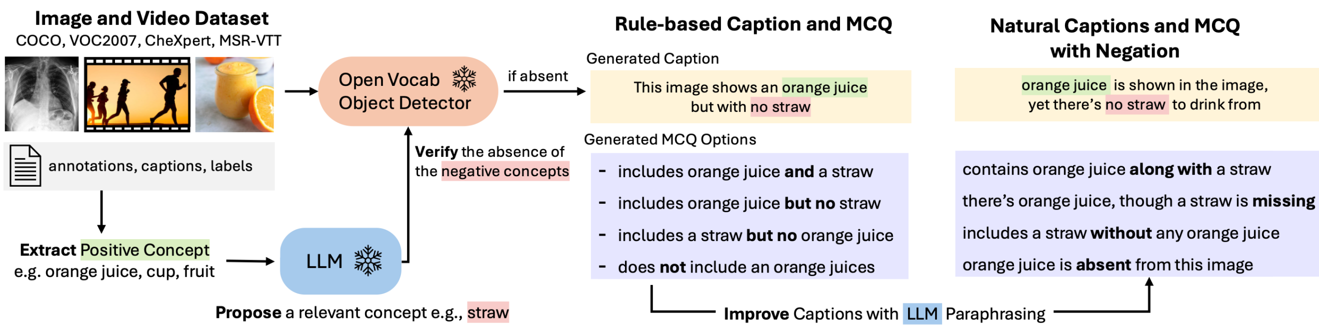


Figure 2. **General Pipeline for Constructing NegBench**. We start by extracting positive concepts from vision datasets. An LLM proposes negative concepts, which are verified with an object detector for datasets without explicit object annotations. We use templates to generate captions with negation,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

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 [20], VOC2007 [7]), video (MSR-VTT [46]), and specialized medical imaging domains (CheXpert [14]). 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 [17].

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 [14], 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.

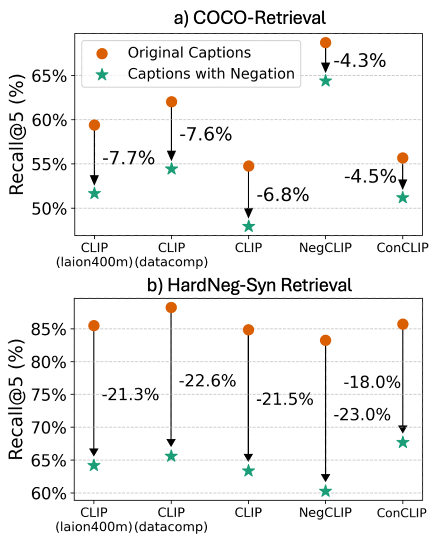


Figure 3. **Performance drop in recall@5 on (a) COCO and (b) HardNeg-Syn text-to-image retrieval with negated captions (green stars) compared to original captions (orange circles).** All models show substantial drops in performance, with NegCLIP experiencing the largest drop of 23.0% on HardNeg-Syn, 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 [34], 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 [25].

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 [31], CLIP-laion400m [37], and CLIP-datacomp [8], which differ by pretraining dataset, as well as NegCLIP [48], trained to improve compositional language understanding, and ConCLIP [41], 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 CLIP video embedding. For medical tasks, we evaluate CONCH [21] and BioMedCLIP [49], two medical foundation VLMs. We also assess the impact of scaling up CLIP-laion400m (ViT-B, ViT-L, and ViT-H) to determine if model size improves negation understanding.

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 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 worse than random guessing (indicated by the red dashed line at 25%) on the MCQ-Neg task, with CLIP-base achieving only 15% on COCO and 8% on VOC2007. These results reveal a fun-

CVPR #2817

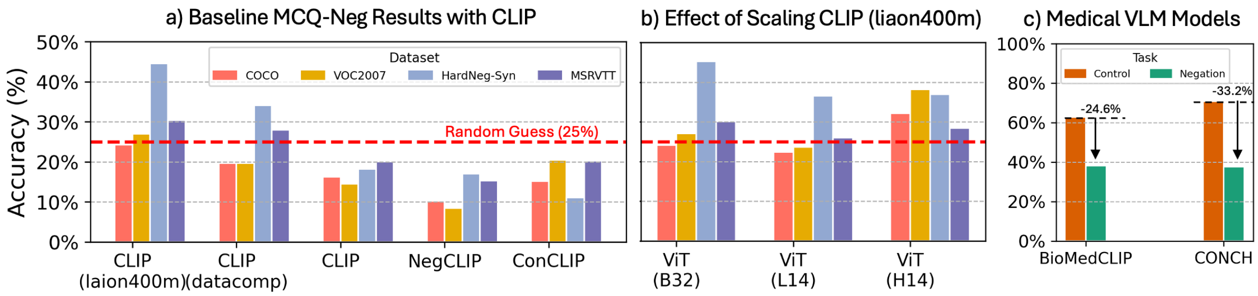


Figure 4. **MCQ-Neg performance for (a) baseline CLIP models, (b) larger model sizes, and (c) medical VLMs.** (a) CLIP-based models mostly perform worse than random guessing (shown as a red dashed line) on most datasets. (b) Scaling up CLIP models does not significantly improve negation understanding. (c) Medical VLMs experience a significant drop in performance on negation MCQs.

damental limitation of CLIP's pretraining objective, which encourages strong associations between visual concepts and specific words, but struggles to interpret negation. Notably, CLIP-laion400m performs better, reaching over 40% accuracy on the HardNeg-Syn dataset. This improvement likely stems from the fact that both CLIP-laion400m and Stable Diffusion (used to generate the HardNeg-Syn dataset) were trained on the LAION dataset [36]. However, a score of 40% on a 4-way multiple-choice task is still far below an acceptable level, demonstrating that even under this setup, models exhibit a serious lack of negation understanding.

Scaling CLIP does not address the negation problem. As shown in Figure 4b, 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 qualitatively improve negation understanding. While ViT-H/14 performs slightly better on COCO and VOC2007, it underperforms on HardNeg-Syn and MSR-VTT compared to ViT-B/32. These results suggest that increasing model size alone is not an effective strategy for addressing the fundamental issues with negation understanding.

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 24.6% 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 and favorable patient outcomes.

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 perform below 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, with evaluations conducted in two settings: one using LLaMA 3.1 to paraphrase answer choices into natural-sounding sentences, and another using rigid linguistic templates. All models perform poorly on Negation MCQs, reflecting a general struggle with negation understanding.

Most models tend to select Negation sentences regardless of whether answers are templated or LLM-paraphrased, as seen in the selection frequencies visualized in the appendix. This behavior likely arises from task design, where 67% of MCQs (Negation and Hybrid) lack a correct affirmative option, leading models to default to "This image does not include {pos}." These results suggest that models trained with CLIP-like objectives often adopt shortcut strategies that ignore specific words like "no."

The template-based results reveal more biases in model behavior. For instance, ConCLIP outperforms on Hybrid MCQs, achieving the highest accuracy, but fails entirely on Affirmation MCQs, scoring 0% on both image datasets. This bias is particularly prominent in the rigid template structure, where ConCLIP is skewed towards constructs like "This image includes X but not Y." In fact, as we will show next, ConCLIP maps all templated Hybrid captions to the same location in its embedding space.

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 [10, 48] that

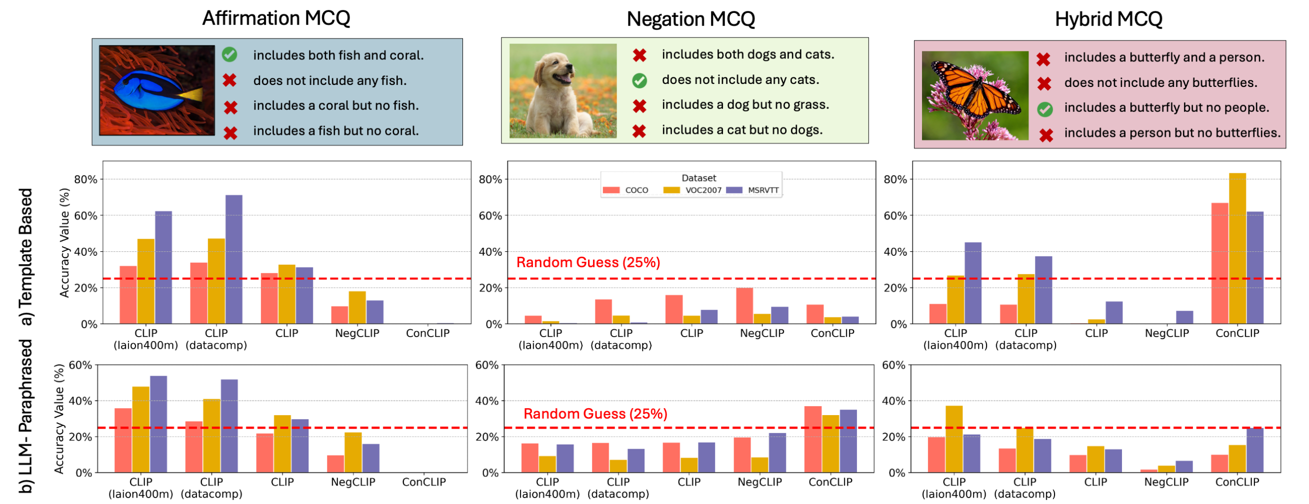
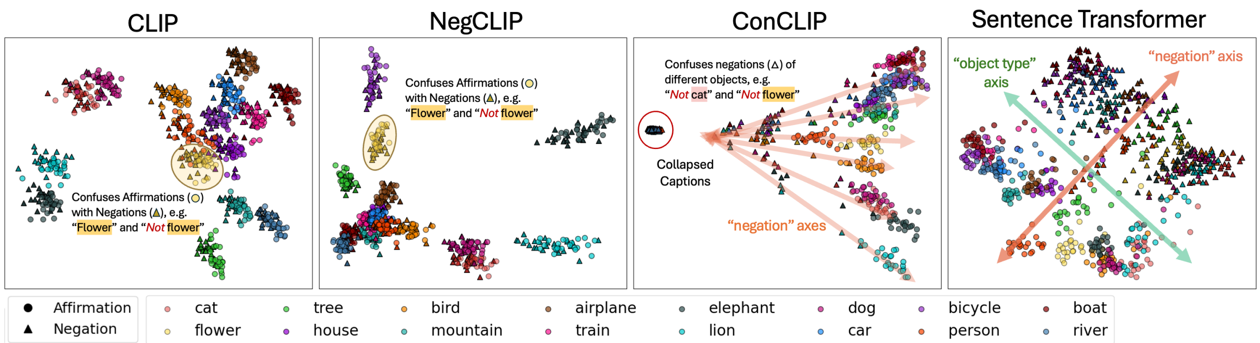
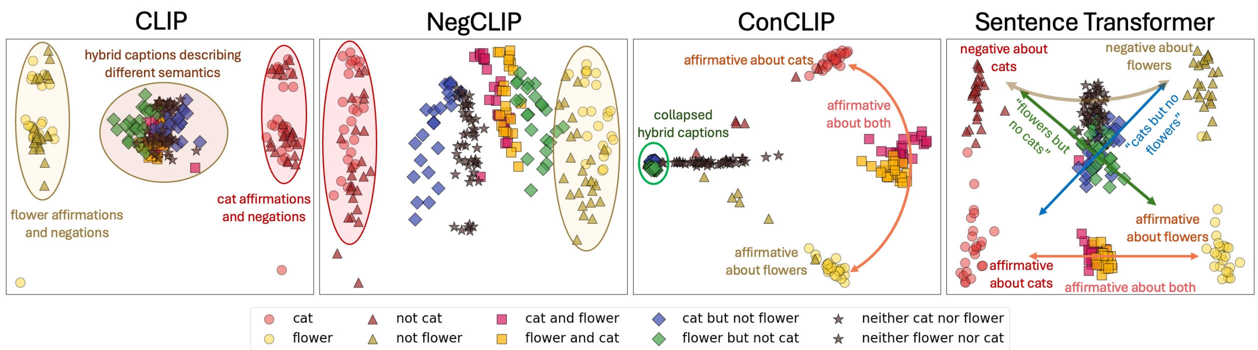


Figure 5. **Performance by MCQ type (Affirmation, Negation, Hybrid) across (a) template-based and (b) LLM-paraphrased answer choices.** VLMs show significant biases towards specific templates (*e.g*. ConCLIP with Hybrid). Template selection frequency (analyzed in the appendix) confirms that CLIP defaults to Negation answers, especially when a positive object is incorrectly negated.



(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.

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 [33] 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 [25] 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:

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

CVPR #2817

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 α 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 CC12M-NegCap leads to significant improvements in handling negated queries in retrieval. On COCO, CLIP's R-Neg@5 score increases by 10%, while the gap between R@5 and R-Neg@5 narrows from 6.8% to 0.7%, indicating that the finetuned model performs nearly as well on negated queries as on standard ones. A similar pattern is seen in MSR-VTT.

However, fine-tuning on CC12M-NegCap alone does not improve performance on the MCQ task, suggesting that the contrastive objective is insufficient for learning fine-grained negation understanding. To address this, we fine-tune CLIP and NegCLIP on the combined CC12M-NegFull dataset using Equation (2), yielding substantial improvements on MCQ tasks. On COCO-MCQ, for instance, NegCLIP's accuracy rises from 10.2% to 51.0%, a 40.8% increase.

Ablation: Effect of varying α. The table below shows the impact of varying the weight factor α 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 α increases, more weight is placed on the original CLIP contrastive objective, while a lower α emphasizes the MCQ loss. Properly tuning α is important to balance between fine-grained MCQ and standard retrieval.

α	0	0.5	0.9	0.99	1
COCO Recall@5 (%)	33.9	37.3	47.6	54.2	58.5
COCO MCQ Acc (%)	61.0	54.7	50.5	46.9	14.7

Model	Fine-tune data	R@5 (↑)	R-Neg@5 (↑)	MCQ (↑)
CLIP	None	54.8	48.0	16.3
	CC12M	58.8	54.5	11.2 (↓5.1)
	CC12M-NegCap	**58.5**	**57.8**	**14.7** (↓1.6)
	CC12M-NegFull	**54.2**	**51.9**	**46.9** (↑30.6)
NegCLIP	None	68.7	64.4	10.2
	CC12M	70.2	66.0	10.6 (↑0.4)
	CC12M-NegCap	**68.6**	**67.5**	**12.5** (↑2.3)
	CC12M-NegFull	**69.0**	**67.0**	**51.0** (↑40.8)

(a) COCO Evaluation

Model	Fine-tune data	R@5 (↑)	R-Neg@5 (↑)	MCQ (↑)
CLIP	None	50.6	45.8	20.1
	CC12M	53.7	49.9	16.9 (↓3.2)
	CC12M-NegCap	**54.1**	**53.5**	**20.1** (0.0)
	CC12M-NegFull	**46.9**	**43.9**	**35.6** (↑15.5)
NegCLIP	None	53.7	51.0	15.3
	CC12M	56.4	52.6	16.8 (↑1.5)
	CC12M-NegCap	**56.5**	**54.6**	**18.9** (↑3.6)
	CC12M-NegFull	**54**	**51.5**	**36.6** (↑21.3)

(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 +1 or more highlighted. Fine-tuning on negation-enriched data significantly improves negation understanding (R-Neg and MCQ).

6. Discussion and Conclusions

Implications. Our findings point to two broader implications for enhancing language understanding in VLMs. 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 contrastive learning alone may not be sufficient for fine-grained language distinctions. We experimented with different values of α in Equation (2), which revealed a tradeoff in performance: higher values improved coarse-grained retrieval but diminished performance on fine-grained multiple-choice questions. 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.

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] 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

[9] 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

[10] 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. 5

[11] 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

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

[13] 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

[14] 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

[15] 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

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

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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

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

[27] 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

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

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

[30] 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

[31] 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

[32] 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

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

[34] 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

[35] 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

[36] 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. 5

[37] 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

[38] 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

[39] 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.

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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, 5

[49] 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