

# Vision-Language Models Do Not Understand Negation

Anonymous CVPR submission

Paper ID 2817

## Abstract

001 *Many practical vision-language applications require models*  
 002 *that understand negation, e.g., when using natural lan-*  
 003 *guage to retrieve images which contain certain objects but*  
 004 *not others. Despite advancements in vision-language mod-*  
 005 *els (VLMs) through large-scale training, their ability to*  
 006 *comprehend negation remains underexplored. This study*  
 007 *addresses the question: how well do current VLMs under-*  
 008 *stand negation? We introduce NegBench, a new bench-*  
 009 *mark designed to evaluate negation understanding across*  
 010 *18 task variations and 79k examples spanning image, video,*  
 011 *and medical datasets. The benchmark consists of two core*  
 012 *tasks designed to evaluate negation understanding in di-*  
 013 *verse multimodal settings: Retrieval with Negation and*  
 014 *Multiple Choice Questions with Negated Captions. Our*  
 015 *evaluation reveals that modern VLMs struggle significantly*  
 016 *with negation, often performing at chance level. To address*  
 017 *these shortcomings, we explore a data-centric approach*  
 018 *wherein we finetune CLIP models on large-scale synthetic*  
 019 *datasets containing millions of negated captions. We show*  
 020 *that this approach can result in a 10% increase in recall on*  
 021 *negated queries and a 40% boost in accuracy on multiple-*  
 022 *choice questions with negated captions.*

## 023 1. Introduction

024 Joint embedding-based Vision-Language Models (VLMs),  
 025 such as CLIP, have revolutionized how we approach multi-  
 026 modal tasks by learning a shared embedding space where  
 027 both images and text are mapped together. This shared  
 028 space enables a variety of applications, including cross-  
 029 modal retrieval, video retrieval, text-to-image generation,  
 030 image captioning, and even medical diagnosis [2, 18, 19,  
 031 21, 30, 32, 35, 38–40, 49]. By aligning visual and linguistic  
 032 representations, these models achieve remarkable per-  
 033 formance across domains and are able to model complex  
 034 interactions between vision and language inputs.

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

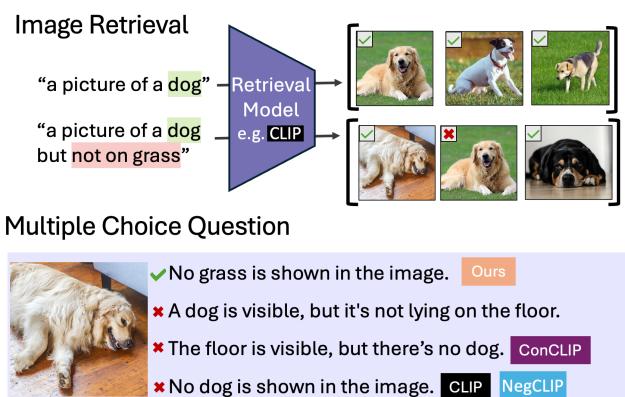


Figure 1. We present *NegBench* with image retrieval and multiple-choice tasks to evaluate negation understanding. CLIP-based models frequently misinterpret negation in both tasks, but we show how a synthetic data approach can improve performance.

038 precise communication by specifying what is false or ab-  
 039 sent [12, 16, 26, 27]. For example, a radiologist may search  
 040 for images showing “bilateral consolidation with no evi-  
 041 dence of pneumonia”, or a safety inspector might query  
 042 “construction sites with no barriers”. Current benchmarks  
 043 like CREPE and CC-Neg have introduced limited tests of  
 044 negation, but they rely on rigid, templated examples that  
 045 do not reflect the complexity of natural language queries  
 046 [24, 41]. As a result, they fall short in evaluating how well  
 047 VLMs understand negation in practical applications.

048 To comprehensively evaluate how well VLMs handle  
 049 negation, we design a multi-level evaluation paradigm in-  
 050 spired by real-world information retrieval systems, where a  
 051 coarse-grained retrieval step often precedes a fine-grained  
 052 ranking or selection step [23, 29].

053 The first task, Retrieval-Neg, tests whether models can  
 054 handle real-world queries that mix affirmative and negative  
 055 statements, such as “a beach with no people” or “a build-  
 056 ing without windows.” This task challenges the model to  
 057 retrieve images from diverse datasets based on the presence  
 058 of certain elements and the absence of others, simulating  
 059 scenarios found in search engines, content moderation, and  
 060 recommendation systems. By retrieving several potentially

061 relevant matches (e.g., top-5 retrieval), Retrieval-Neg serves  
062 as the coarse-grained retrieval component of our evaluation.

063 The second task, MCQ-Neg, provides a fine-grained,  
064 structured evaluation that directly assesses specific failures  
065 in negation. In this task, the model must choose the cor-  
066 rect description of an image from several closely related  
067 options, where the incorrect choices are hard negatives, dif-  
068 fering only by what is affirmed or negated. For instance, in  
069 medical diagnostics, consider distinguishing between “The  
070 X-ray shows evidence of pneumonia but no evidence of  
071 pleural effusion” and “The X-ray shows evidence of pleural  
072 effusion but no evidence of pneumonia.” These statements  
073 are linguistically similar but convey opposite diagnoses, re-  
074quiring the model to parse subtle yet critical differences.

075 Through our evaluation pipeline, we uncover a surprising  
076 limitation: joint embedding-based VLMs frequently col-  
077 lapse affirmative and negated statements into similar em-  
078 beddings, treating “a dog” and “no dog” as nearly indis-  
079 tinguishable. This affirmation bias reveals a significant  
080 shortcoming that was not sufficiently addressed in previous  
081 benchmarks like CREPE or CC-Neg.

082 Recognizing this critical gap, we then ask: If cur-  
083 rent models fail to understand negation, can we improve  
084 them? To tackle this, we propose a data-centric solution,  
085 introducing two large-scale synthetic datasets—Syn-Neg-  
086 Cap and Syn-Neg-MCQ—designed to improve negation  
087 comprehension. Fine-tuning CLIP-based models on these  
088 datasets leads to substantial improvements, including a 10%  
089 increase in recall on negated queries and a 40% boost in ac-  
090 curacy on multiple-choice questions with negated captions.

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

## 097 2. Related Work

098 Our work lies within the field of evaluating and advanc-  
099 ing foundational vision-language models (VLMs). Joint-  
100 embedding models based on CLIP [31] show impressive  
101 generalization across visio-linguistic tasks like cross-modal  
102 retrieval, image captioning, and visual question answering  
103 [2, 18, 19, 30, 32, 35, 38–40] in diverse visual domains,  
104 extending beyond natural images to videos and medical im-  
105 ages [3, 13, 21, 22, 28, 49]. We introduce a benchmark and  
106 data-centric approach to rigorously evaluate and improve  
107 negation understanding in these VLMs.

108 **Negation Understanding in Language and Vision.** Re-  
109 cent work showed that large language models perform sub-  
110 optimally when tasked with negation understanding [9, 45].  
111 We go a step further by showing that vision-language mod-  
112 els exhibit a more severe affirmation bias, completely fail-

113 113 ing to differentiate affirmative from negative captions.

114 Despite this critical limitation, existing benchmarks pro-  
115 vide limited assessments of negation in VLMs. CREPE [24]  
116 and the concurrent work CC-Neg [41] are among the few  
117 vision-language benchmarks that include negation, but they  
118 focus on compositional understanding and rely on linguis-  
119 tic templates that fail to reflect the varied ways negation ap-  
120 pears in real user queries. In contrast, our proposed bench-  
121 mark, NegBench, leverages an LLM to generate natural-  
122 sounding negated captions, spanning a broader range of  
123 negation types and contexts across images, videos, and  
124 medical datasets. This systematic design enables a thor-  
125ough evaluation of VLMs’ ability to handle negation in mul-  
126 timodal settings, uncovering unique challenges and failure  
127 cases that have not been fully addressed in prior work.

128 **Improving CLIP for Compositionality and Negation.**  
129 Recent methods have explored improving the generaliza-  
130 tion abilities of CLIP-like VLMs for visio-linguistic com-  
131 positionality and limited aspects of negation under-  
132 standing. For instance, NegCLIP [48] employs composi-  
133 tion-aware mining when finetuning CLIP to enhance compo-  
134 sitional reasoning, while ConCLIP [41] modifies the CLIP  
135 loss to incorporate synthetic, template-based negation ex-  
136 amples. In the medical domain, negation is a common fea-  
137 ture in clinical text reports, often indicating the absence of  
138 specific pathologies [44]. Specialized models like Biomed-  
139 CLIP [49] and CONCH [21] have been pretrained on mil-  
140 lions of biomedical image-text pairs to address a variety of  
141 medical tasks, leveraging domain-specific knowledge from  
142 large-scale multimodal data. NegBench provides a system-  
143 atic way to evaluate general-purpose and medical VLMs.

144 **Synthetic Data for Model Training.** It is common to use  
145 synthetic data to improve the performance of models in  
146 computer vision [1, 5, 15, 47]. Recent studies have shown  
147 that it is possible to use synthetic data to learn general  
148 vision-language representations, with some models trained  
149 entirely on synthetic images and captions achieving results  
150 comparable to real data [11, 42, 43]. Our approach is similar  
151 in spirit, but it constructs synthetic datasets to teach models  
152 a new, complex capability—*negation understanding*.

## 153 3. The Negation Benchmark (NegBench)

154 We design NegBench as a multi-level evaluation to as-  
155 sess the capacity of joint-based vision-language models  
156 to understand negation across different tasks: (1) coarse-  
157 grained retrieval, by accurately retrieving images that sat-  
158 isfy specified inclusions and exclusions, and (2) fine-  
159 grained question-answering, by selecting the correct de-  
160 scription from closely related options, testing the model’s  
161 detailed understanding of negation beyond simple retrieval.

162 In the Retrieval-Neg task, the model retrieves the top-  
163 5 images that match both affirmative and negative criteria  
164 within a query. In the MCQ-Neg task, the model selects the

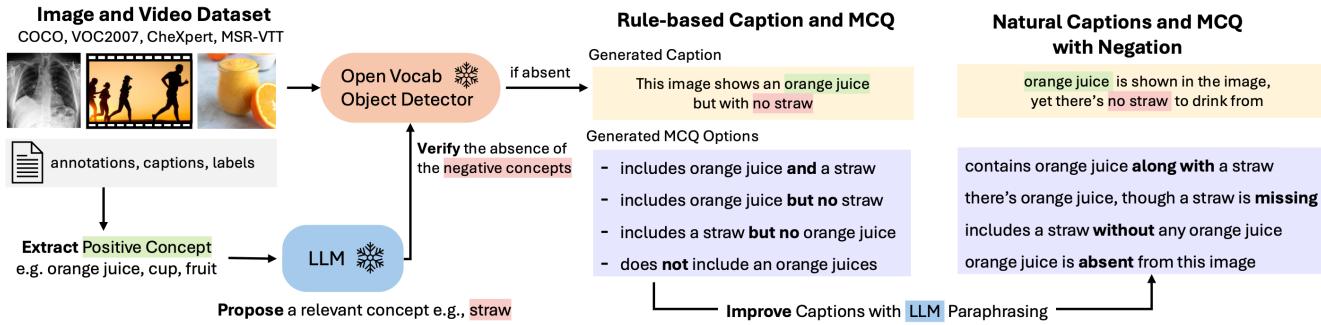


Figure 2. **General Pipeline for Constructing NegBench.** We start by extracting positive concepts from vision datasets. An LLM proposes negative concepts, which are verified with an object detector for datasets without explicit object annotations. We use templates to generate captions with negation,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

165 correct description of an image from options that differ only  
166 in the affirmation or negation of specific elements.

### 167 3.1. Transforming Datasets for Negation Evaluation

168 **General Dataset Transformation Overview.** To implement  
169 the two-stage evaluation pipeline of NegBench, we  
170 adapt several popular vision datasets, covering images  
171 (COCO [20], VOC2007 [7]), video (MSR-VTT [46]), and  
172 specialized medical imaging domains (CheXpert [14]). For  
173 each dataset, we identify positive elements  $\{pos\}$ , which  
174 represent objects or concepts present in the image, and neg-  
175 ative elements  $\{neg\}$ , which are absent from the image but  
176 commonly associated with the present objects. When avail-  
177 able, we use object-level annotations to identify these el-  
178 ements, as in COCO, VOC2007, and CheXpert; for other  
179 datasets, we derive positive and negative elements directly  
180 from the captions. This flexible approach allows NegBench  
181 to extend any vision dataset, whether it includes object-level  
182 annotations or captions, to evaluate negation comprehen-  
183 sion across diverse tasks and data modalities.

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

191 For the MCQ-Neg task, we generate multiple-choice  
192 questions (MCQs) for each image. The model must identify  
193 the correct description based on three linguistic templates:  
194 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**.”

195 Each MCQ consists of one correct answer and three in-  
196 correct answers, which serve as hard negatives, misleading  
197 the model if it does not properly understand negation. A  
198

199 correct answer accurately describes the presence of  $\{pos\}$   
200 elements or negates  $\{neg\}$  elements. A False Affirma-  
201 tion (e.g., “This image includes  $x$ ” when  $x \in \{neg\}$ ) or  
202 a False Negation (e.g., “This image does not include  $x$ ”  
203 when  $x \in \{pos\}$ ) highlights the model’s failure to com-  
204 prehend the image. The Hybrid template further evaluates  
205 the model’s ability to combine affirmation and negation in  
206 the same caption. These MCQs are also paraphrased using  
207 LLaMA 3.1 to increase linguistic diversity.

### 208 3.2. Applicability Across Data Types and Domains

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

211 **Video Understanding.** Video retrieval tasks introduce tem-  
212 poral complexity, where negation can involve both objects  
213 and actions that vary over time. Using MSR-VTT as an ex-  
214 ample, we prompt LLaMA 3.1 [6] to extract positive and  
215 negative elements from each video’s caption. These el-  
216 ements may represent either objects present in the video or  
217 actions taking place. For Retrieval-Neg, we create cap-  
218 tions specifying both the presence of some elements and  
219 the absence of others (e.g., “A person is cooking but not  
220 eating”). In MCQ-Neg, we generate multiple-choice ques-  
221 tions where the model must select the description that most  
222 accurately represents a video segment, requiring it to reason  
223 about negation of objects and actions in dynamic scenes.

224 **Medical Image Interpretation with CheXpert.** Accurate  
225 negation understanding is critical in high-stakes domains  
226 like medical imaging. Using the CheXpert dataset [14], we  
227 focus on the most frequent condition *Lung Opacity* and de-  
228 sign two binary classification tasks:

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

232 **Question:** Which option describes this image?

- A) This image shows Lung Opacity.
- B) This image shows Atelectasis.

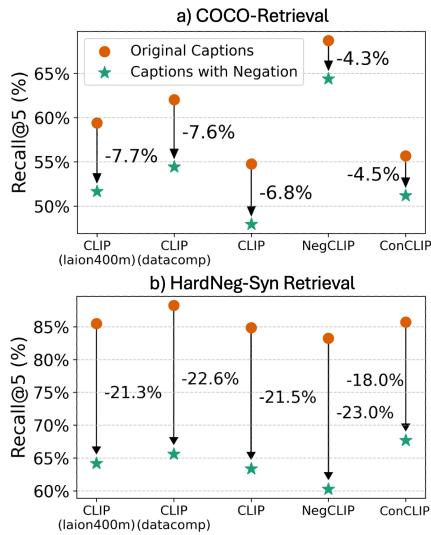
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231

233 *Task 2: Negation Understanding Task.* This task tests  
 234 whether the model can correctly interpret negation, distin-  
 235 guishing the presence or absence of a medical condition.

236 **Question:** Which option best describes the image?

- 237 A) This image shows Lung Opacity.  
 238 B) This image does *not* show Lung Opacity.

239 These extensions highlight the adaptability of NegBench  
 240 to various data types and domains, from general images and  
 241 videos to specialized medical imaging. This versatility en-  
 sures that NegBench provides rigorous, contextually rele-  
 242 vant evaluations of negation understanding in VLMs.



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

### 3.3. Synthetic Datasets for Controlled Evaluation

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

251 **Motivation and Benefits of Synthetic Data.** Synthetic  
 252 data offers several advantages over traditional image  
 253 datasets. First, by creating “hard negatives”—image  
 254 pairs that differ only by a single object’s presence or ab-  
 255 sence—we can evaluate the sensitivity of models to nega-  
 256 tion with minimal confounding variables. Additionally, im-  
 257 age datasets like COCO and VOC2007 are limited in the  
 258 range of visual concepts they cover; COCO has 80 objects  
 259 while VOC2007 includes only 20. To expand this diversity,  
 260 we prompt a large language model to propose a broader  
 261 set of objects, which we use as targets in our synthetic

262 dataset. This approach enables the generation of visually  
 263 varied scenes that more comprehensively test negation com-  
 264 prehension across a wider array of objects and contexts.

### Construction Process for the HardNeg-Syn Evaluation

265 **Dataset.** We create 10,000 image pairs using Stable Diffusion [34], where each pair includes one image contain-  
 266 ing a target object and another where it is explicitly absent.  
 To ensure accurate object presence or absence, we use the  
 267 open-vocabulary object detector OWL-ViT [25].

## 4. NegBench Evaluations: Results and Insights

268 In this section, we benchmark the negation abilities of dif-  
 269 ferent VLMs using NegBench, comparing models based  
 270 on their architecture, training data, and training objectives  
 271 to reveal specific areas where negation understanding re-  
 272 mains limited. Specifically, we evaluate five CLIP ViT-B/32  
 273 models on Retrieval-Neg and MCQ-Neg tasks. These in-  
 274 clude OpenAI CLIP [31], CLIP-laion400m [37], and CLIP-  
 275 datacomp [8], which differ by pretraining dataset, as well  
 276 as NegCLIP [48], trained to improve compositional lan-  
 277 guage understanding, and ConCLIP [41], trained specif-  
 278 ically to improve negation understanding. To handle the  
 279 video dataset, MSR-VTT, we follow [3] and encode 4 uni-  
 280 formly sampled frames per video, averaging their features  
 281 to obtain the CLIP video embedding. For medical tasks, we  
 282 evaluate CONCH [21] and BioMedCLIP [49], two medical  
 283 foundation VLMs. We also assess the impact of scaling up  
 284 CLIP-laion400m (ViT-B, ViT-L, and ViT-H) to determine if  
 model size improves negation understanding.

285 **CLIP models struggle with negated queries in retrieval**  
 286 **tasks.** We evaluate five CLIP-based models on the origi-  
 287 nal COCO text-to-image retrieval task and its Retrieval-Neg  
 288 version, where captions include negated statements. Across  
 289 models, performance drops significantly on the negated  
 290 task. In COCO retrieval (Figure 3a), CLIP-laion400m expe-  
 291 riences a 7.7% drop in recall@5, with CLIP-datacomp and  
 292 CLIP showing drops of 7.6% and 6.8%, respectively. In the  
 293 more challenging HardNeg-Syn retrieval task (Figure 3b),  
 294 the performance drops are even more pronounced due to the  
 295 presence of hard negatives, *i.e.* images that closely resemble  
 296 positive examples but differ by the exclusion of a single ob-  
 297 ject. Here, NegCLIP, despite its promise for compositional  
 298 understanding, suffers a 23.0% drop, while ConCLIP, de-  
 299 signed specifically for negation understanding, still declines  
 300 by 18.0%. These results suggest that interpreting negation,  
 301 particularly in the presence of hard negatives, remains a key  
 302 challenge for retrieval tasks.

303 **MCQ-Neg reveals severe limitations in CLIP models.**  
 304 Figure 4a shows that most models perform worse than ran-  
 305 dom guessing (indicated by the red dashed line at 25%) on  
 306 the MCQ-Neg task, with CLIP-base achieving only 15% on  
 307 COCO and 8% on VOC2007. These results reveal a fun-

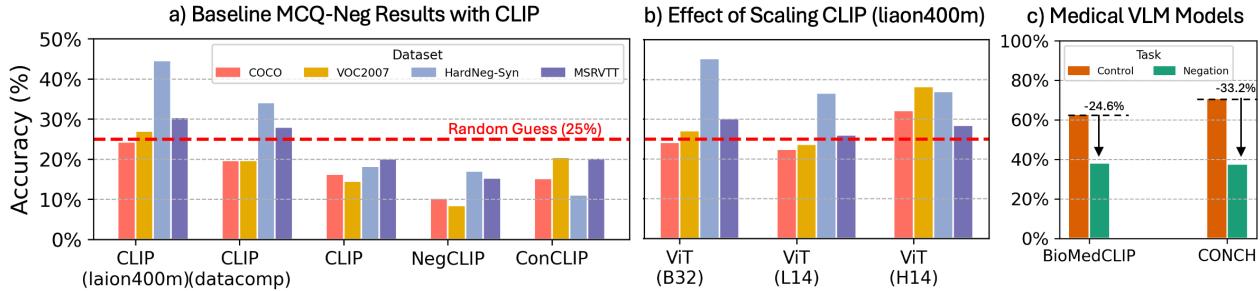


Figure 4. **MCQ-Neg performance for (a) baseline CLIP models, (b) larger model sizes, and (c) medical VLMs.** (a) CLIP-based models mostly perform worse than random guessing (shown as a red dashed line) on most datasets. (b) Scaling up CLIP models does not significantly improve negation understanding.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

308 fundamental limitation of CLIP’s pretraining objective, which  
309 encourages strong associations between visual concepts and  
310 specific words, but struggles to interpret negation. Notably,  
311 CLIP-laion400m performs better, reaching over 40% accuracy  
312 on the HardNeg-Syn dataset. This improvement likely  
313 stems from the fact that both CLIP-laion400m and Stable  
314 Diffusion (used to generate the HardNeg-Syn dataset) were  
315 trained on the LAION dataset [36]. However, a score of  
316 40% on a 4-way multiple-choice task is still far below an  
317 acceptable level, demonstrating that even under this setup,  
318 models exhibit a serious lack of negation understanding.

319 **Scaling CLIP does not address the negation problem.**  
320 As shown in Figure 4b, scaling up the model size from  
321 ViT-B/32 (86M parameters) to ViT-L/14 (307M parameters)  
322 and ViT-H/14 (632M parameters) does not qualita-  
323 tively improve negation understanding. While ViT-H/14  
324 performs slightly better on COCO and VOC2007, it un-  
325 derperforms on HardNeg-Syn and MSR-VTT compared to  
326 ViT-B/32. These results suggest that increasing model size  
327 alone is not an effective strategy for addressing the funda-  
328 mental issues with negation understanding.

329 **Critical failures in high-stakes medical tasks.** Figure 4c  
330 presents the results for the CheXpert MCQ-Neg task, where  
331 BioMedCLIP and CONCH exhibit substantial performance  
332 drops of 24.6% and 33.2%, respectively, when negation is  
333 introduced. This result is especially concerning in the  
334 context of medical diagnostics, where accurate interpretation of  
335 negation (e.g., the presence or absence of a condition such  
336 as Lung Opacity) is essential for correct diagnoses and fa-  
337 vorable patient outcomes.

#### 338 4.1. Why Do VLMs Not Understand Negation?

339 The results from NegBench reveal that CLIP VLMs strug-  
340 gle with different forms of negation understanding, moti-  
341 vating a deeper analysis into the underlying causes of these  
342 failures. In this section, we examine model performance  
343 across different MCQ types and analyze the embedding  
344 spaces of various models to uncover specific shortcut strate-  
345 gies that limit their negation comprehension.

**Model performance varies widely across MCQ types.**  
346 To understand why models perform below random chance,  
347 we categorize the MCQs into three types based on the  
348 correct answer template: Affirmation, Negation, and Hy-  
349 brid. Figure 5 compares model accuracy across these MCQ  
350 types, with evaluations conducted in two settings: one us-  
351 ing LLaMA 3.1 to paraphrase answer choices into natural-  
352 sounding sentences, and another using rigid linguistic tem-  
353 plates. All models perform poorly on Negation MCQs, re-  
354 flecting a general struggle with negation understanding.  
355

356 Most models tend to select Negation sentences regard-  
357 less of whether answers are templated or LLM-paraphrased,  
358 as seen in the selection frequencies shown to the right of the  
359 figure. This behavior likely arises from task design, where  
360 67% of MCQs (Negation and Hybrid) lack a correct affir-  
361 mative option, leading models to default to “This image  
362 does not include {pos}.” These results suggest that mod-  
363 els trained with CLIP-like objectives often adopt shortcut  
364 strategies that ignore specific words like “no.”

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

374 **Embedding analysis reveals VLM shortcut strategies.**  
375 To investigate potential shortcut strategies, we analyze the  
376 embedding spaces of various models using 24 Affirmative  
377 (“X”) and 24 Negated (“Not X”) templates to create 48 cap-  
378 tions per object. We apply PCA to the resulting embeddings  
379 (Figure 6a). The templates are detailed in the appendix.

380 We observe varying behaviors across models. The over-  
381 lapping embeddings for affirmative and negated captions in  
382 CLIP and NegCLIP suggest that these models do not dis-  
383tinguish between positive and negative statements, possi-  
384 bly due to a “bag-of-words” shortcut strategy [10, 48] that

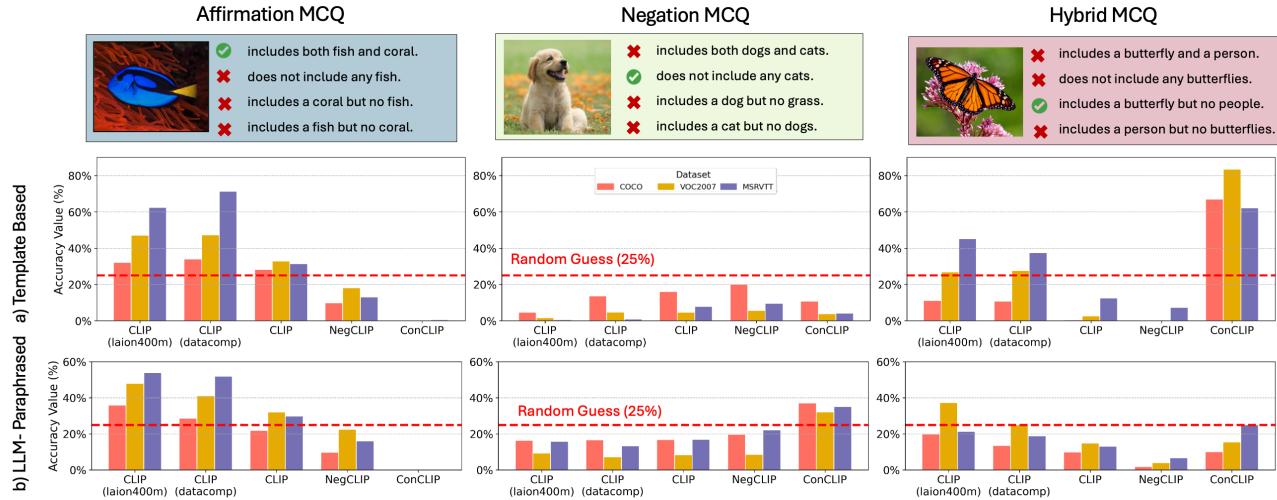
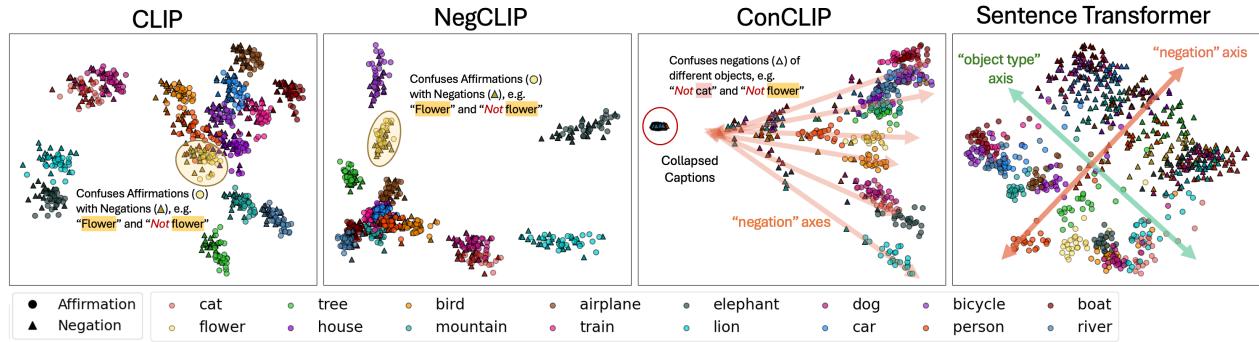
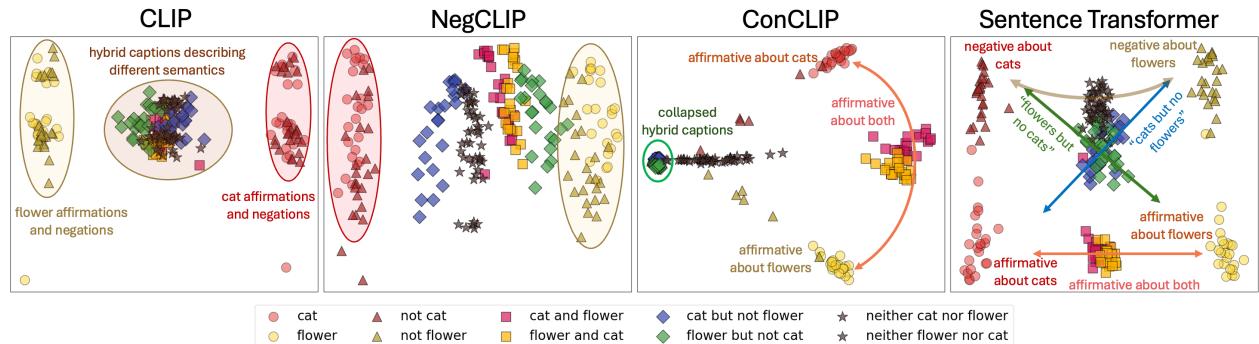


Figure 5. **Performance by MCQ type (Affirmation, Negation, Hybrid) across (a) template-based and (b) LLM-paraphrased answer choices.** VLMs show significant biases towards specific templates (e.g. ConCLIP with Hybrid). Template selection frequency (analyzed in the appendix) confirms that CLIP defaults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

385 overlooks negation words. This explains why both models  
386 incorrectly select the Negation template, which negates positive  
387 objects, in Figure 5. CoNCLIP separates positive and  
388 negative captions but fails to distinguish between negative

captions of different objects, collapsing all negative caption  
389 embeddings toward a single point (red circle).

390 We include the embeddings of a text-only Sentence  
391 Transformer [33] as a reference that effectively differenti-  
392

393     ates affirmative and negated captions along distinct “object  
394     type” and “negation” axes, exemplifying ideal separation.

395     **Hybrid captions reveal more evidence of collapsed em-  
396     beddings.** Figure 6b extends the previous analysis to hy-  
397     brid captions that combine affirmations and negations. It  
398     provides further evidence that *ConCLIP* employs a shortcut  
399     strategy for embedding linguistic negation, with hybrid and  
400     negated captions collapsing towards a single point (green  
401     circle), indicating significant compression along the nega-  
402     tion axis. While *CLIP* and *NegCLIP* struggle to distinguish  
403     affirmative from negative statements, *NegCLIP* shows bet-  
404     ter separation for hybrid captions, which appear collapsed  
405     in the *CLIP* embedding space. This suggests that *Neg-  
406     CLIP*’s poor performance on Hybrid MCQs might be due to  
407     a misalignment between the text and image encoders, rather  
408     than an inability to understand hybrid sentence structure. In  
409     contrast, the *Sentence Transformer* effectively distinguishes  
410     between different caption types and provides semantically  
411     guided representations. For example, it aligns “flowers but  
412     not cats” along the line connecting “flowers” and “not cats.”

## 413     5. A Data-Centric Approach for Improving 414     Negation Understanding

415     We hypothesize that the tendency of *CLIP*-based models to  
416     rely on linguistic shortcuts, which hinders their negation un-  
417     derstanding as explored in Section 4.1, stems from training  
418     data limitations. In *CLIP*, training data lacks examples with  
419     explicit negation, leaving it unable to distinguish negated  
420     and affirmed concepts. In contrast, *ConCLIP*’s training data  
421     overfits to a single hybrid linguistic template, limiting its  
422     ability to generalize across varied negation structures. Next,  
423     we explore data-centric strategies to address these gaps, in-  
424     troducing a dataset that includes diverse negation examples  
425     spanning a range of linguistic styles.

### 426     5.1. Synthesizing a Fine-Tuning Negation Dataset

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

- 433     1. **Object Extraction:** Using LLaMA 3.1 [6], we extract  
434         positive objects (those mentioned in the caption) and  
435         negative objects (contextually relevant but not present)  
436         from each image-caption pair in CC12M.
- 437     2. **Visual Verification:** An open-vocabulary object detec-  
438         tor [25] verifies the presence of positive objects and en-  
439         sures the absence of the negative objects in the image.  
440         This step is crucial to avoid introducing incorrect nega-  
441         tions that could confuse the model.
- 442     3. **Caption Generation:** For each image, we generate mul-

443     tiple new captions that incorporate negated objects into  
444     the original captions. LLaMA 3.1 is used to ensure the  
445     generated captions are natural-sounding and reflect real-  
446     istic negation scenarios found in retrieval queries.

447     We construct two variants of the synthetic dataset.  
448     **CC12M-NegCap** includes three captions per image with  
449     incorporated negated objects, totaling approximately 30  
450     million captions. **CC12M-NegMCQ** includes four cap-  
451     tions per image: one correct and three hard negatives based  
452     on object annotations, offering stronger training signals  
453     for fine-grained negation understanding and resulting in  
454     around 40 million captions. To balance broad retrieval with  
455     fine-grained negation capabilities, we introduce **CC12M-  
456     NegFull**, a comprehensive dataset that combines CC12M-  
457     NegCap and CC12M-NegMCQ. We will release the ex-  
458     tracted object annotations for each image in CC12M, along  
459     with the corresponding URLs, and all the generated cap-  
460     tions in CC12M-NegFull. This will help the community  
461     build on our dataset and advance research in negation un-  
462     derstanding and multimodal retrieval.

### 463     5.2. Fine-Tuning with Negation-Enriched Data

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

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

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

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

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

492 where  $c_i$  indicates the index of the correct caption de-  
493 scribing the  $i$ -th image.

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

$$498 \quad \mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_{\text{CLIP}}(\mathcal{B}_{\text{cap}}) + (1 - \alpha) \mathcal{L}_{\text{MCQ}}(\mathcal{B}_{\text{mcq}}). \quad (2)$$

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

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

515 **Results.** Fine-tuning CLIP and NegCLIP on CC12M-  
516 NegCap leads to significant improvements in handling  
517 negated queries in retrieval. On COCO, CLIP’s R-Neg@5  
518 score increases by 10%, while the gap between R@5 and  
519 R-Neg@5 narrows from 6.8% to 0.7%, indicating that the  
520 finetuned model performs nearly as well on negated queries  
521 as on standard ones. A similar pattern is seen in MSR-VTT.

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

530 **Ablation: Effect of varying  $\alpha$ .** The table below shows the  
531 impact of varying the weight factor  $\alpha$  in the combined loss  
532  $\mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_{\text{CLIP}} + (1 - \alpha) \mathcal{L}_{\text{MCQ}}$  when fine-tuning CLIP  
533 on CC12M-NegFull. As  $\alpha$  increases, more weight is placed  
534 on the original CLIP contrastive objective, while a lower  $\alpha$   
535 emphasizes the MCQ loss. Properly tuning  $\alpha$  is important to  
536 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 (%)	61.0	54.7	50.5	46.9	14.7

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

(a) COCO Evaluation

Model	Fine-tune data	R@5 ( $\uparrow$ )	R-Neg@5 ( $\uparrow$ )	MCQ ( $\uparrow$ )
CLIP	None	50.6	45.8	20.1
	CC12M	53.7	49.9	16.9 ( $\downarrow$ 3.2)
	CC12M-NegCap	<b>54.1</b>	<b>53.5</b>	<b>20.1</b> (0.0)
	CC12M-NegFull	<b>46.9</b>	<b>43.9</b>	<b>35.6</b> ( $\uparrow$ 15.5)
NegCLIP	None	53.7	51.0	15.3
	CC12M	56.4	52.6	16.8 ( $\uparrow$ 1.5)
	CC12M-NegCap	<b>56.5</b>	<b>54.6</b>	<b>18.9</b> ( $\uparrow$ 3.6)
	CC12M-NegFull	<b>54</b>	<b>51.5</b>	<b>36.6</b> ( $\uparrow$ 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  $\uparrow$ 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  $\alpha$  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.

570

571

## 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 Ilama 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 Haghgoor, 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

- 683 [25] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim  
684 Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh  
685 Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran  
686 Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil  
687 Houlsby. Simple open-vocabulary object detection with vi-  
688 sion transformers. *ECCV*, 2022. 4, 7, 15
- 689 [26] Roser Morante and Eduardo Blanco. Recent advances in  
690 processing negation. *Natural Language Engineering*, 27(2):  
691 121–130, 2021. 1
- 692 [27] Partha Mukherjee, Youakim Badr, Shreyesh Doppalapudi,  
693 Satish M Srinivasan, Raghvinder S Sangwan, and Rahul  
694 Sharma. Effect of negation in sentences on sentiment analy-  
695 sis and polarity detection. *Procedia Computer Science*, 185:  
696 370–379, 2021. 1
- 697 [28] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell.  
698 Clip-it! language-guided video summarization. *Advances  
699 in neural information processing systems*, 34:13988–14000,  
700 2021. 2
- 701 [29] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy  
702 Lin. Multi-stage document ranking with bert. *arXiv preprint  
703 arXiv:1910.14424*, 2019. 1
- 704 [30] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gon-  
705 zalez, Trevor Darrell, and Anna Rohrbach. On guiding vi-  
706 sual attention with language specification. In *Proceedings of  
707 the IEEE/CVF Conference on Computer Vision and Pattern  
708 Recognition*, pages 18092–18102, 2022. 1, 2
- 709 [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya  
710 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,  
711 Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning  
712 transferable visual models from natural language super-  
713 vision. In *ICML*. PMLR, 2021. 2, 4
- 714 [32] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong  
715 Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu.  
716 Denseclip: Language-guided dense prediction with context-  
717 aware prompting. In *Proceedings of the IEEE/CVF con-  
718 ference on computer vision and pattern recognition*, pages  
719 18082–18091, 2022. 1, 2
- 720 [33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence  
721 embeddings using siamese bert-networks. In *EMNLP*. Asso-  
722 ciation for Computational Linguistics, 2019. 6
- 723 [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz,  
724 Patrick Esser, and Björn Ommer. High-resolution image syn-  
725 thesis with latent diffusion models. In *Conference on Com-  
726 puter Vision and Pattern Recognition (CVPR)*, pages 10684–  
727 10695, 2022. 4
- 728 [35] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowd-  
729 hury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip  
730 for all things zero-shot sketch-based image retrieval, fine-  
731 grained or not. In *Proceedings of the IEEE/CVF Conference  
732 on Computer Vision and Pattern Recognition*, pages 2765–  
733 2775, 2023. 1, 2
- 734 [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu,  
735 Cade Gordon, Ross Wightman, Mehdi Cherti, Theo  
736 Coombes, Arush Katta, Clayton Mullis, Mitchell Worts-  
737 man, et al. Laion-5b: An open large-scale dataset for training  
738 next generation image-text models. *Advances in Neural In-  
739 formation Processing Systems*, 35:25278–25294, 2022. 5
- 740 [37] Christoph Schuhmann, Romain Beaumont, Richard Vencu,  
741 Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo  
742 Coombes, Arush Katta, Clayton Mullis, Mitchell Worts-  
743 man, Patrick Schramowski, Srivatsa R Kundurthy, Katherine  
744 Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia  
745 Jitsev. LAION-5b: An open large-scale dataset for train-  
746 ing next generation image-text models. In *Thirty-sixth Con-  
747 ference on Neural Information Processing Systems Datasets  
748 and Benchmarks Track*, 2022. 4
- 749 [38] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal,  
750 Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt  
751 Keutzer. How much can clip benefit vision-and-language  
752 tasks? In *International Conference on Learning Representa-  
753 tions*. 1, 2
- 754 [39] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei  
755 Cai. Proposalclip: Unsupervised open-category object pro-  
756 posal generation via exploiting clip cues. In *Proceedings of  
757 the IEEE/CVF Conference on Computer Vision and Pattern  
758 Recognition*, pages 9611–9620, 2022.
- 759 [40] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport:  
760 What and where pathways for robotic manipulation. In *Con-  
761 ference on robot learning*, pages 894–906. PMLR, 2022. 1,  
762 2
- 763 [41] Jaisidh Singh, Ishaaan Shrivastava, Mayank Vatsa, Richa  
764 Singh, and Aparna Bharati. Learn “no” to say “yes” bet-  
765 ter: Improving vision-language models via negations. *arXiv  
766 preprint arXiv:2403.20312*, 2024. 1, 2, 4
- 767 [42] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and  
768 Dilip Krishnan. Stablerep: Synthetic images from text-to-  
769 image models make strong visual representation learners. In  
770 *NeurIPS*, 2023. 2
- 771 [43] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip  
772 Krishnan, and Phillip Isola. Learning vision from mod-  
773 els rivals learning vision from data. In *Proceedings of  
774 the IEEE/CVF Conference on Computer Vision and Pattern  
775 Recognition*, pages 15887–15898, 2024. 2
- 776 [44] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew  
777 Y Ng, and Pranav Rajpurkar. Expert-level detection  
778 of pathologies from unannotated chest x-ray images via self-  
779 supervised learning. *Nature Biomedical Engineering*, 6(12):  
780 1399–1406, 2022. 2
- 781 [45] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and  
782 Trevor Cohn. Language models are not naysayers: an anal-  
783 ysis of language models on negation benchmarks. In *Pro-  
784 ceedings of the 12th Joint Conference on Lexical and Com-  
785 putational Semantics (\* SEM 2023)*, pages 101–114, 2023.  
786 2
- 787 [46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large  
788 video description dataset for bridging video and language. In  
789 *Proceedings of the IEEE conference on computer vision and  
790 pattern recognition*, pages 5288–5296, 2016. 3
- 791 [47] Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo  
792 Zhao. Real-fake: Effective training data synthesis through  
793 distribution matching. In *The Twelfth International Confer-  
794 ence on Learning Representations*, 2024. 2
- 795 [48] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri,  
796 Dan Jurafsky, and James Zou. When and why vision-

797 language models behave like bags-of-words, and what to do  
798 about it? In *ICLR*, 2023. [2](#), [4](#), [5](#), [12](#)

799 [49] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu,  
800 Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao,  
801 Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal  
802 biomedical foundation model pretrained from fifteen million  
803 scientific image-text pairs. *arXiv preprint arXiv:2303.00915*,  
804 2023. [1](#), [2](#), [4](#)

805 **Appendix**806 **A. Evaluating LLaVA on NegBench MCQs**

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

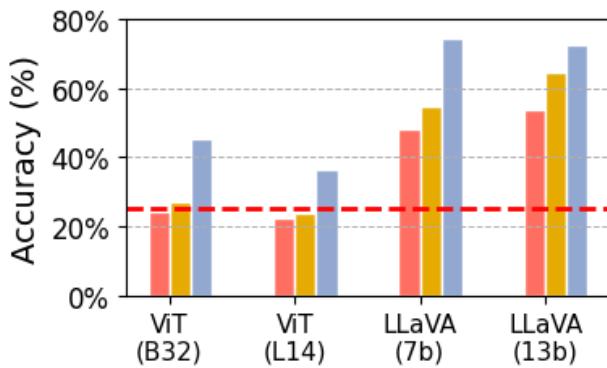


Figure 7. Caption.

820 **LLaVA, an instruction-tuned VLM, demonstrates im-  
 821 provement.** Figure 7 shows that LLaVA significantly out-  
 822 performs CLIP models on the MCQ-Neg tasks. This is par-  
 823 ticularly notable because LLaVA uses a CLIP ViT-L/14 vi-  
 824 sion encoder, which we have shown in Figure 4 to struggle  
 825 with negation. The key advantage of LLaVA might be in  
 826 its use of the Vicuna LLM for text encoding. Unlike CLIP,  
 827 which is pretrained on vision-language pairs that predom-  
 828 inantly contain affirmative image captions, LLMs like Vi-  
 829 cuna are trained on diverse textual corpora that include both  
 830 affirmations and negations. This broader exposure allows  
 831 LLaVA to better interpret negated statements. Additionally,  
 832 LLaVA uses a learned projection layer to align vision and  
 833 language representations, in contrast to CLIP’s contrastive  
 834 learning objective, which tends to ignore word order and  
 835 subtle linguistic cues like negation [48]. We further explore  
 836 these differences in Figure 8.

837 **Limitations of LLaVA as a retrieval system.** While  
 838 LLaVA demonstrates improved negation understanding, it

839 has significant limitations as a retrieval model compared  
 840 to CLIP. CLIP learns a joint image-text embedding space,  
 841 making it highly efficient for retrieval tasks by simply em-  
 842 bedding both images and texts, and then computing cosine  
 843 similarities. In contrast, LLaVA processes a single image-  
 844 text pair at a time and generates text output, which makes  
 845 image-to-text retrieval feasible only if all possible captions  
 846 can fit into the model’s context window. For MCQ-Neg, we  
 847 applied this method by presenting the image alongside all  
 848 possible captions and prompting LLaVA to select the cor-  
 849 rect one. However, this approach does not scale well with  
 850 a large number of candidates and is not applicable for text-  
 851 to-image retrieval, where fitting all dataset images into the  
 852 context window is impractical. Therefore, advancing mod-  
 853 els like CLIP is crucial for real-world multimodal retrieval  
 854 with negation. In the paper, we explored the data-centric  
 855 reasons behind CLIP’s failures in negation understanding  
 856 and proposed synthetic data strategies to address them.

857 **B. A Closer Look at VLM Negation Failures**

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

863 **B.1. Template Selection Frequency**

864 Figure 8 analyzes the frequency with which different mod-  
 865 els select specific template types (Affirmation, Negation,  
 866 Hybrid) when answering multiple-choice questions, regard-  
 867 less of the correct answer. This analysis helps to reveal po-  
 868 tential biases in model behavior and understand why mod-  
 869 els may struggle with negation. As shown in Figure 5 from  
 870 the paper, most models perform poorly on Negation MCQs,  
 871 reflecting a general struggle with negation understanding.

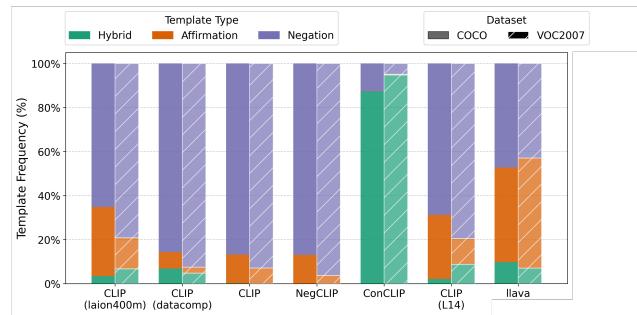


Figure 8. Template selection frequency for various models on COCO and VOC2007 datasets, broken down by template type (Affirmation, Negation, Hybrid).

872 Figure 8 shows that most CLIP-based models tend to  
 873 overselect the Negation template, even when the correct an-  
 874 swer is an affirmation or a hybrid statement. This behavior

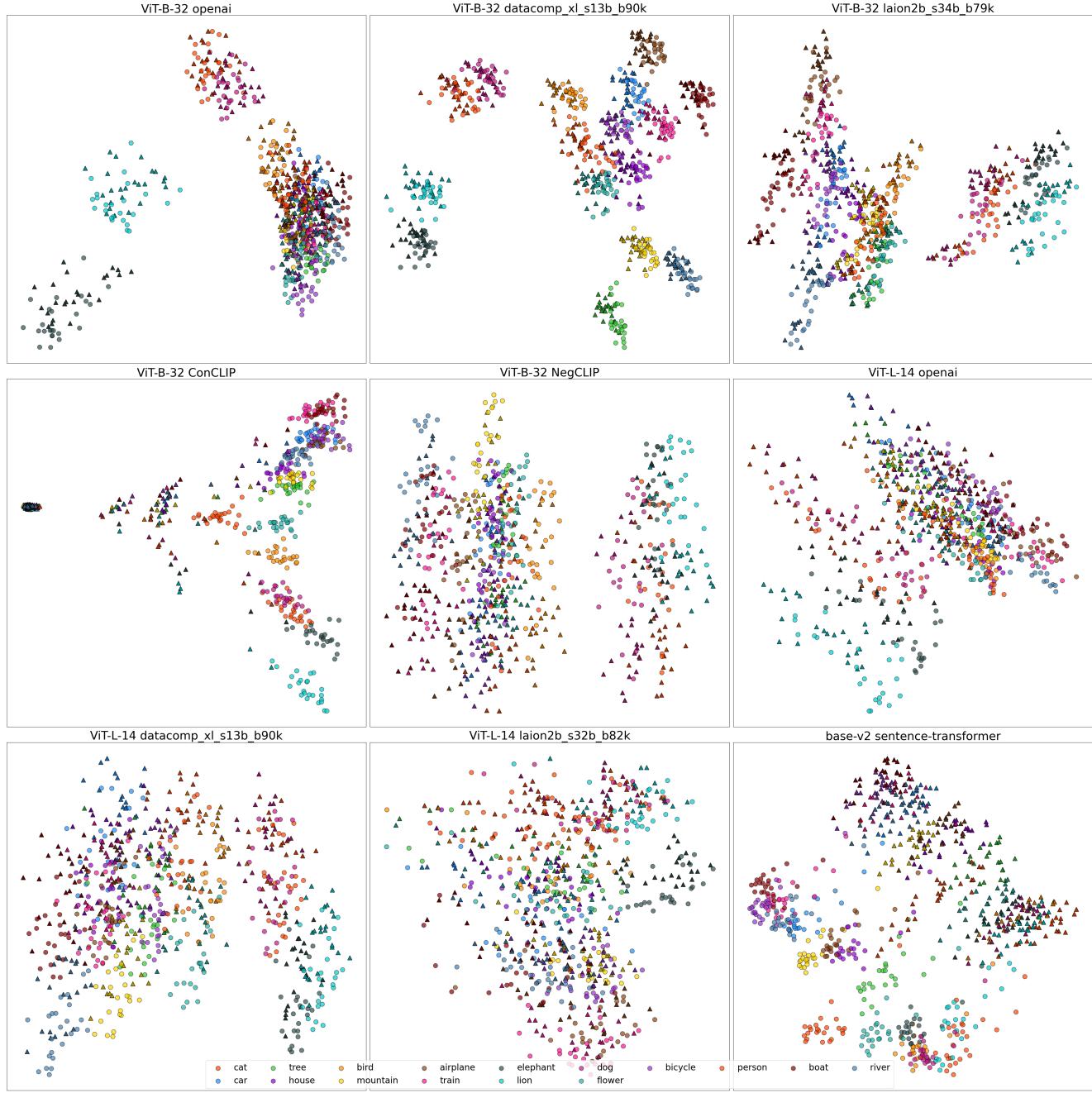


Figure 9. PCA projections of caption embeddings for various CLIP models and the Sentence Transformer. Each point represents a caption embedding. This figure complements Figure 6 by providing a broader view of embedding separation across different VLMs.

875 is consistent with the observation in Figure 5 that most mod- 882  
876 els perform poorly on Negation MCQs, where they must se- 883  
877 lect a caption that does not reference the positive object. In 884  
878 contrast, ConCLIP overselects the Hybrid template, which 885  
879 combines affirmation and negation in the same statement. 886  
880 This bias likely stems from ConCLIP’s training data, which 887  
881 overfits to a single hybrid linguistic template.

882 On the other hand, the LLaVA model exhibits a more bal- 883  
883 anced template selection frequency, indicating that it is less 884  
884 susceptible to the biases observed in CLIP-based models.

## B.2. Template Embedding Analysis

885 This subsection provides further details about the embed- 886  
886 ding analysis presented in Figure 6 of the main paper. We 887  
887

888 achieve this by:

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

891 To generate the embeddings for the PCA projections, we  
892 used five categories of templates: Affirmation (single ob-  
893 ject), Negation (single object), Affirmation (two objects),  
894 Hybrid (one object affirmed, one negated), and Double  
895 Negation (two objects negated). Each category contains  
896 24 templates, except for Affirmation (two objects) which  
897 has 23. The templates vary sentence structure and wording  
898 while maintaining the same core meaning.

- 899 • **Affirmation (single object):** 24 templates. Examples:  
900 "This image includes A", "A is present in this image",  
901 "This image shows A", "A is depicted in this image", "A  
902 appears in this image".  
903 • **Negation (single object):** 24 templates. Examples: "This  
904 image does not include A", "A is not present in this im-  
905 age", "This image lacks A", "A is not depicted in this  
906 image", "A does not appear in this image".  
907 • **Affirmation (two objects):** 23 templates. Examples:  
908 "This image includes A and B", "A and B are present in  
909 this image", "This image shows A and B", "A and B are  
910 depicted in this image", "A and B appear in this image".  
911 • **Hybrid (one object affirmed, one negated):** 24 tem-  
912 plates. Examples: "This image includes A but not B",  
913 "A is present in this image but not B", "This image shows  
914 A but not B", "This image features A but not B", "A ap-  
915 pears in this image but not B".  
916 • **Double Negation (two objects negated):** 24 templates.  
917 Examples: "This image includes neither A nor B", "Nei-  
918 ther A nor B are present in this image", "This image  
919 shows neither A nor B", "Neither A nor B are depicted  
920 in this image", "Neither A nor B appear in this image".

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

## 927 C. Dataset and Task Summary for NegBench

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

939 Each dataset contributes to either Retrieval-Neg or  
940 MCQ-Neg tasks, except for CheXpert, which has two dis-  
941 tinct tasks (Affirmation Control and Negation Under-  
942 standing) in both MCQ and binary classification formats. Ad-  
943 ditionally, we utilize original retrieval captions for COCO  
944 (5,000) and MSR-VTT (1,000), expanding the overall  
945 dataset size. VOC2007 does not include a Retrieval-Neg  
946 task as it lacks retrieval-style captions.

947 The total number of task variations across all datasets  
948 in NegBench is 18, and the total number of samples across  
949 all tasks and variations is 79,239. Table 2 summarizes the  
950 datasets, tasks, task versions, and sizes.

- 951 • **COCO:** 5,000 retrieval captions and 5,914 MCQ ques-  
952 tions, resulting in 10,000 retrieval problems and 11,828  
953 MCQ problems with templated and LLM-paraphrased  
954 variations.  
955 • **VOC2007:** 5,032 MCQ questions, leading to 10,064 total  
956 samples. No retrieval task is provided due to the absence  
957 of retrieval-style captions.  
958 • **MSR-VTT:** 1,000 retrieval captions and 1,000 MCQ  
959 questions, resulting in 2,000 samples per task, including  
960 both variations.  
961 • **CheXpert:** Two MCQ tasks (4-choice) and two binary  
962 classification tasks. The 4-choice MCQ covers 690 sam-  
963 ples for affirmation and 1,587 for negation, while the bi-  
964 nary tasks each include 690 samples.  
965 • **HardNeg-Syn:** 10,000 synthetic images, used to create  
966 20,000 retrieval and 20,000 MCQ problems across tem-  
967 plated and LLM-paraphrased versions.